



**INSTITUTO
FEDERAL**

Alagoas

INSTITUTO FEDERAL DE ALAGOAS

CAMPUS MACEIÓ

CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

IRIS MAYARA VIEIRA NUNES

**ANÁLISE DA EVASÃO E PERMANÊNCIA EM CURSOS SUPERIORES DE
TECNOLOGIA DA INFORMAÇÃO NO BRASIL**

MACEIÓ, AL

2026

IRIS MAYARA VIEIRA NUNES

ANÁLISE DA EVASÃO E PERMANÊNCIA EM CURSOS SUPERIORES DE
TECNOLOGIA DA INFORMAÇÃO NO BRASIL

Trabalho de Conclusão de Curso apresentado
ao Curso de Bacharelado em Sistemas de
Informação no Instituto Federal de Alagoas,
campus Maceió, como requisito parcial para a
obtenção de grau de Bacharel(a) em Sistemas
de Informação.

Orientador(a): Prof. Dr. Edison Camilo de
Moraes Júnior

MACEIÓ, AL

2026



Dados Internacionais de Catalogação na Publicação
Instituto Federal de Alagoas
Campus Maceió
Biblioteca Benevides Monte

004

N972a

Nunes, Iris Mayara Vieira.

Análise da evasão e permanência em cursos superiores de Tecnologia de Informação no Brasil [recurso eletrônico] / Iris Mayara Vieira Nunes. – Dados eletrônicos (1 arquivo : 7,71 MB). – 2026.

Sistema requerido: Adobe Acrobat Reader.

Modo de acesso: Internet.

Orientação: Prof. Dr. Edison Camilo de Moraes Júnior.

Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Instituto Federal de Alagoas, *Campus Maceió*, Maceió, 2026.

1. Ensino superior – Evasão e permanência. 2. Tecnologia da Informação – Brasil. 3. Business Intelligence (BI). 4. Dados educacionais. I. Título.

IRIS MAYARA VIEIRA NUNES

ANÁLISE DA EVASÃO E PERMANÊNCIA EM CURSOS SUPERIORES DE
TECNOLOGIA DA INFORMAÇÃO NO BRASIL

Trabalho de Conclusão de Curso apresentado
ao Curso de Bacharelado em Sistemas de
Informação no Instituto Federal de Alagoas,
campus Maceió, como requisito parcial para a
obtenção de grau de Bacharel(a) em Sistemas
de Informação.

Aprovado em: 26 de junho de 2026

BANCA EXAMINADORA

Documento assinado digitalmente
 EDISON CAMILO DE MORAES JÚNIOR
Data: 03/07/2026 14:26:14-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Edison Camilo de Moraes Júnior

Instituto Federal de Alagoas – IFAL

Documento assinado digitalmente
 JAILTON CARDOSO DA CRUZ
Data: 03/07/2026 11:11:30-0300
Verifique em <https://validar.iti.gov.br>

Jailton Cardoso da Cruz

Instituto Federal de Alagoas – IFAL

Documento assinado digitalmente
 MARCIO ROBERIO DA COSTA FERRO
Data: 03/07/2026 11:46:08-0300
Verifique em <https://validar.iti.gov.br>

Márcio Robério da Costa Ferro

Instituto Federal de Alagoas – IFAL

AGRADECIMENTOS

A Deus, por me sustentar em todos os momentos, me dando forças quando pensei em desistir e serenidade para seguir em frente, mesmo diante das dificuldades.

À minha mãe, por todo amor e apoio ao longo dessa caminhada. Cada palavra de incentivo e cada gesto de cuidado foram essenciais para que eu chegasse até aqui.

À minha filha, que chegou há pouco tempo e mudou completamente a minha vida, dando um novo sentido à minha trajetória. Nesta reta final, foi nela que encontrei coragem para continuar. Sua chegada ressignificou todo o meu esforço, e tudo o que faço será sempre por nós.

Aos meus colegas de curso, que tornaram as noites mais leves e fáceis, compartilhando comigo essa jornada de aprendizado e superação.

A todos os professores do curso, que contribuíram para a minha formação acadêmica e profissional, compartilhando conhecimentos e experiências que levarei para toda a vida, em especial ao professor orientador, que tive o prazer de conhecer ainda no início do curso e de contar com seu apoio até aqui, contribuindo com paciência, dedicação e orientação ao longo de todo o percurso.

E a todos que, de alguma forma, contribuíram para a realização deste sonho, o meu mais sincero e profundo agradecimento.

RESUMO

Este trabalho realiza uma pesquisa exploratória sobre a evasão e a permanência em cursos da área de Tecnologia da Informação no Brasil, abrangendo um total de 109 cursos, no período de 2014 a 2024, sob as perspectivas temporal, regional e sociodemográfica. A pesquisa caracteriza-se como quantitativa, de natureza aplicada e descritiva, com abordagem longitudinal, fundamentada em dados oficiais provenientes do Censo da Educação Superior, disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). A análise foi realizada por meio da organização e exploração dos dados utilizando técnicas de *Business Intelligence* (BI) e modelagem dimensional, com a construção de um painel analítico (*dashboard*) que possibilitou a visualização e interpretação de indicadores relacionados às matrículas, evasão e conclusão dos cursos. Os resultados evidenciam um cenário crítico, com taxa de evasão superior a 60% no contexto nacional, indicando que a expansão das matrículas em cursos de Tecnologia da Informação não tem sido acompanhada pela efetiva formação de profissionais na área. Como contribuição, o estudo evidencia o potencial do uso de dados educacionais abertos associados a ferramentas analíticas para apoiar a compreensão de fenômenos educacionais e subsidiar processos de tomada de decisão no contexto da educação superior.

Palavras-chave: ensino superior; tecnologia da informação; business intelligence; dados educacionais; análise de dados.

ABSTRACT

This study analyzes patterns of dropout and persistence in Information Technology undergraduate programs in Brazilian higher education between 2014 and 2024, considering temporal, regional, and sociodemographic perspectives. The research is characterized as quantitative, applied, and descriptive, with a longitudinal approach based on official secondary data from the Higher Education Census provided by the National Institute for Educational Studies and Research Anísio Teixeira (Inep). The analysis was conducted through the organization and exploration of data using Business Intelligence techniques and dimensional modeling, including the development of an analytical dashboard that enabled the visualization and interpretation of indicators related to enrollment, dropout, and program completion. The results reveal a critical scenario, with a dropout rate exceeding 60% at the national level, indicating that the expansion of enrollments in Information Technology programs has not been accompanied by a proportional increase in the effective training of professionals in this field. As a contribution, this study highlights the potential of open educational data combined with analytical tools to support the understanding of educational phenomena and to assist evidence-based decision-making in higher education.

Keywords: higher education; information technology education; business intelligence; educational data; data analysis.

SUMÁRIO

1 INTRODUÇÃO.....	9
1.1 OBJETIVOS.....	10
1.1.1 Objetivo Geral.....	10
1.1.2 Objetivos Específicos.....	10
1.2 QUESTÕES DE PESQUISA.....	11
1.3 DELIMITAÇÃO DO TRABALHO.....	11
1.4 ESTRUTURA DO TRABALHO.....	12
2 FUNDAMENTAÇÃO TEÓRICA.....	13
2.1 EVASÃO E PERMANÊNCIA NO ENSINO SUPERIOR DE TECNOLOGIA DA INFORMAÇÃO.....	13
2.2 GESTÃO DE DADOS PÚBLICOS.....	14
2.3 BUSINESS INTELLIGENCE (BI).....	15
2.4 MODELAGEM DIMENSIONAL.....	17
2.4.1 Star Schema.....	17
2.4.2 Snowflake.....	18
2.5 TOMADA DE DECISÃO BASEADA EM DADOS.....	19
3 METODOLOGIA.....	21
3.1 COLETA DOS DADOS.....	21
3.2 PROCESSAMENTO E ETL.....	22
3.3 MODELAGEM DOS DADOS.....	31
3.4 VISUALIZAÇÃO DOS DADOS.....	35
3.5 DEFINIÇÃO DAS MÉTRICAS.....	38
4 RESULTADOS E DISCUSSÃO.....	41
4.1 VISÃO GERAL E EVOLUÇÃO TEMPORAL DO FLUXO DISCENTE (2014–2024)..	41
4.2 ANÁLISE POR CURSOS E TENDÊNCIAS DE FORMAÇÃO TECNOLÓGICA.....	42
4.3 CONTRASTES REGIONAIS E DESAFIOS DE RETENÇÃO TERRITORIAL.....	44
4.4 PERFIL SOCIODEMOGRÁFICO E DIVERSIDADE NO CAMPO DE TI.....	45
4.5 IMPACTO DAS MODALIDADES E REDES DE ENSINO NA EFETIVIDADE FORMATIVA.....	46
5 CONSIDERAÇÕES FINAIS.....	49
6 DISPONIBILIDADE DE DADOS E CÓDIGO.....	52
REFERÊNCIAS.....	53

1 INTRODUÇÃO

A evasão no ensino superior brasileiro configura-se como um dos principais desafios da gestão educacional contemporânea, revelando-se um fenômeno complexo e multifatorial. Embora influenciada por uma combinação de fatores acadêmicos, econômicos e pessoais, essa problemática assume contornos ainda mais preocupantes nos cursos de Tecnologia da Informação e Comunicação (TIC) (SBC, 2018). Nesse cenário, a interrupção da trajetória acadêmica não representa apenas uma perda institucional, mas também um entrave ao desenvolvimento econômico e social do país.

O setor de TIC consolidou-se como um dos principais influenciadores da economia brasileira, sendo responsável por aproximadamente 6,5% do Produto Interno Bruto (PIB) nacional (Brasscom, 2024). Dados da Associação Brasileira das Empresas de TIC indicam que esse protagonismo se reflete em um cenário de alta empregabilidade, com uma demanda anual estimada em cerca de 159 mil novos profissionais e níveis de remuneração que podem variar entre 2,9 e 3,2 vezes a média salarial nacional (Brasscom, 2021).

Essa expressividade econômica tem impactado diretamente as IES, nas quais se observa um crescimento significativo no interesse por carreiras tecnológicas. Segundo dados do Censo da Educação Superior, os cursos vinculados à Computação e às áreas de TIC figuram entre os mais procurados do país. Contudo, apesar desse crescimento da demanda, os cursos da área de têm apresentado índices de evasão superiores à média observada em outras áreas do ensino superior, revelando um cenário que demanda análises baseadas em dados para melhor compreensão. (Semesp, 2023).

Nesse contexto, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) disponibiliza, anualmente, os microdados do Censo da Educação Superior, os quais constituem uma fonte relevante para a análise da evasão e permanência acadêmica. Entretanto, o elevado volume e a complexidade dessas bases dificultam sua exploração por gestores e pesquisadores, sobretudo quando se busca compreender o fenômeno de forma integrada, considerando variáveis temporais, regionais, institucionais e sociodemográficas. Assim, torna-se imprescindível a adoção de ferramentas capazes de transformar grandes conjuntos de dados em informações estratégicas para a tomada de decisão.

As técnicas de *Business Intelligence* (BI) destacam-se como um recurso relevante para o apoio à gestão educacional, ao possibilitarem a integração, a análise e a visualização interativa de dados. A aplicação dessas técnicas permite a identificação de padrões de evasão

e permanência, subsidiando a formulação de estratégias institucionais e políticas públicas orientadas por evidências.

Diante desse contexto, este Trabalho de Conclusão de Curso tem como objetivo realizar uma análise exploratória sobre a evasão e a permanência em cursos superiores de Tecnologia da Informação no Brasil, com base nos dados do Censo da Educação Superior referentes ao período de 2014 a 2024. Para tanto, propõe-se o desenvolvimento de uma solução de apoio à decisão fundamentada em técnicas de BI, que possibilite a exploração dos dados sob as perspectivas temporal, regional, institucional e sociodemográfica. Espera-se que os resultados obtidos contribuam para uma compreensão mais aprofundada do fenômeno da evasão acadêmica em cursos de TIC e auxiliem gestores educacionais na definição de ações voltadas à redução do abandono.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Analisar a evasão e a permanência em cursos superiores de Tecnologia da Informação no Brasil, a partir de dados do Censo da Educação Superior referentes ao período de 2014 a 2024, por meio do desenvolvimento de uma solução de apoio à decisão baseada em técnicas de BI.

1.1.2 Objetivos Específicos

- Realizar a coleta dos microdados do Censo da Educação Superior referentes ao período de 2014 a 2024, disponibilizados pelo Inep.
- Desenvolver um processo de Extração, Transformação e Carga (ETL) para tratamento, limpeza, padronização e consolidação dos dados.
- Projetar e implementar um modelo dimensional no formato *Star Schema*, definindo tabelas de fato e dimensões adequadas à análise de evasão e permanência.
- Carregar e estruturar o modelo de dados em uma plataforma de BI, garantindo integridade, desempenho e consistência analítica.
- Desenvolver um dashboard interativo para visualização das taxas de evasão e conclusão, permitindo análises por ano, região, categoria administrativa da instituição, modalidade de ensino e demais variáveis selecionadas.

1.2 QUESTÕES DE PESQUISA

Sendo assim, o problema de pesquisa que guia este trabalho consiste na seguinte questão: Quais os padrões de evasão e permanência em cursos de Tecnologia da Informação no ensino superior brasileiro?

Para responder adequadamente à questão principal, a pesquisa investigará os seguintes tópicos:

Q1: Qual é a taxa geral de evasão e conclusão nos cursos de Tecnologia da Informação no período de 2014 a 2024?

Q2: Como evoluiu o número de ingressantes e concluintes nos cursos de Tecnologia da Informação nos últimos 10 anos?

Q3: Qual é a proporção de alunos concluintes em relação aos ingressantes ao longo dos anos?

Q4: Quais cursos da área de Tecnologia da Informação apresentam os maiores volumes de ingressantes e concluintes?

Q5: Qual é a diferença na taxa de conclusão entre cursos tecnólogos e de bacharelado em TI?

Q6: Quais estados concentram o maior número de ingressantes e concluintes nos cursos de Tecnologia da Informação?

Q7: Como a taxa de evasão e permanência se distribui entre as diferentes regiões do Brasil?

Q8: Qual é a distribuição de alunos por gênero nos cursos de Tecnologia da Informação?

Q9: Existe diferença significativa nas taxas de evasão entre instituições públicas e privadas?

Q10: Como as taxas de evasão e permanência variam entre as modalidades de ensino presencial e a distância?

1.3 DELIMITAÇÃO DO TRABALHO

A presente pesquisa delimita-se à análise da evasão e permanência em cursos de graduação enquadrados na área de TIC, conforme adaptação de área específica da CINE/Unesco, utilizada pelo Censo da Educação Superior do Inep. O recorte temporal compreende o período de 2014 a 2024, permitindo uma análise histórica que abrange tanto o cenário pré-pandemia quanto as transformações educacionais recentes.

Geograficamente, o estudo abrange todo o território brasileiro, com segmentação por regiões e estados. No que tange à dimensão técnica, o trabalho foca no desenvolvimento de

uma solução de BI para visualização descritiva e diagnóstica. Portanto, não compõe o escopo desta pesquisa a aplicação de modelos preditivos de inteligência artificial ou a realização de entrevistas qualitativas com ex-alunos, restringindo-se ao processamento e interpretação dos microdados governamentais disponíveis.

1.4 ESTRUTURA DO TRABALHO

Este trabalho está organizado em cinco capítulos que detalham o percurso da pesquisa. Após esta introdução, o segundo capítulo apresenta a fundamentação teórica, abordando as causas da evasão e permanência em TI e os conceitos fundamentais de BI. O terceiro capítulo descreve a metodologia, caracterizada como quantitativa e documental, detalhando as etapas de tratamento dos microdados do Inep. No quarto capítulo, são apresentados e discutidos os resultados obtidos por meio da solução de BI desenvolvida. Por fim, o quinto capítulo expõe as considerações finais, sintetizando os principais achados e sugerindo trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 EVASÃO E PERMANÊNCIA NO ENSINO SUPERIOR DE TECNOLOGIA DA INFORMAÇÃO

A evasão no ensino superior é definida pela literatura acadêmica como um fenômeno multifatorial e dinâmico, que transcende o ato isolado do abandono escolar. Embora o modelo seminal de Tinto (1975) já enfatizasse a importância da integração acadêmica e social, perspectivas atualizadas do próprio autor (Tinto, 2017) reforçam que a persistência discente exige um alinhamento contínuo entre a autoeficácia do aluno, o suporte institucional proativo e o senso de pertencimento. Sob essa ótica, relatórios recentes da OCDE (2021) reiteram que a evasão gera impactos tripartites: prejuízos institucionais pela ociosidade de recursos, frustração das expectativas de ascensão socioeconômica do indivíduo e perda de capital humano qualificado para a sociedade. No cenário brasileiro, o Inep (2022) e a Semesp (2023) demonstram que essa problemática é agravada por determinantes macroestruturais, evidenciando que as deficiências na formação básica e as desigualdades financeiras são preditores significativos do desengajamento discente.

Na área de Tecnologia da Informação (TI), a literatura identifica uma dicotomia de forças causais: vetores de "expulsão", vinculados a entraves pedagógicos, e vetores de "atração", relacionados à dinâmica do setor produtivo (Correia; Mendonça; Silva, 2024). As barreiras pedagógicas são frequentemente associadas ao rigor técnico e à abstração lógica exigidos pelos currículos de Computação. Conforme aponta a Sociedade Brasileira de Computação (SBC, 2022), a carência de competências matemáticas e analíticas oriundas do ensino básico provoca uma descontinuidade cognitiva nos semestres iniciais. Esse "choque de realidade" manifesta-se em taxas elevadas de reprovação e retenção em disciplinas de Cálculo e Algoritmos, consolidando-se como um dos principais catalisadores do desânimo acadêmico e da consequente evasão precoce.

Simultaneamente, o setor de Tecnologia da Informação apresenta a singularidade da evasão motivada pela empregabilidade precoce. Diferentemente de profissões tradicionalmente regulamentadas, que exigem o diploma para o exercício legal da função, o mercado tecnológico tende a priorizar o domínio prático e o desenvolvimento de competências específicas (Silveira; Tonini, 2023). Essa característica, aliada à expansão acelerada de tecnologias como a Inteligência Artificial e à crescente demanda por capital humano, leva as empresas a recrutarem talentos ainda em fases iniciais da formação

universitária (Brasscom, 2021), forçando o estudante a realizar uma escolha pragmática pela inserção profissional imediata em detrimento da conclusão do curso (Oliveira Junior, 2022).

Por fim, a permanência estudantil é modulada por variáveis sociodemográficas que afetam o engajamento identitário. Cheryan *et al.* (2017) destacam que o ambiente predominantemente masculinizado das ciências exatas pode comprometer o sentimento de pertencimento de estudantes do sexo feminino. A literatura sugere que essa falta de identificação cultural e institucional atua como um fator de risco silencioso, perpetuando disparidades de gênero e padrões assimétricos de abandono na área (Cheryan *et al.*, 2017). A compreensão dessa complexidade sistêmica exige, portanto, metodologias analíticas que permitam o tratamento de grandes massas de dados para a identificação de padrões de vulnerabilidade.

2.2 GESTÃO DE DADOS PÚBLICOS

A gestão de dados públicos no cenário brasileiro é fundamentada pelo paradigma do Governo Aberto, que busca promover a transparência, a prestação de contas e a participação social. Segundo Zancanaro, Santos e Baudu (2017), os Dados Governamentais Abertos são ativos essenciais para que a sociedade possa monitorar políticas públicas e colaborar com o Estado na resolução de problemas complexos. No entanto, a literatura aponta que a simples disponibilização do dado não garante a sua utilidade prática. Para Stair e Reynolds (2018), existe uma distinção fundamental entre dado e informação: enquanto o dado é apenas um registro bruto e isolado, a informação é o resultado do dado processado e organizado de forma que possua valor e significado para o tomador de decisão.

No âmbito da educação superior, o Inep atua como o principal gestor de dados, sendo responsável pelo Censo da Educação Superior. Embora o acesso aos microdados seja garantido, a arquitetura apresenta desafios operacionais significativos, caracterizando-se por uma volumetria massiva, multiplicidade de variáveis e relacionamentos relacionais complexos que impõem barreiras técnicas à extração de evidências fidedignas (Inep, 2022). Esse cenário evidencia o paradoxo das organizações contemporâneas, definidas como "ricas em dados e pobres em informação", onde a saturação de registros sem o devido tratamento obscurece a percepção dos fenômenos e retarda a resposta institucional (Turban; Pollard; Wood, 2021).

Nesse contexto, a distinção entre dado e informação é central para a viabilização de políticas públicas baseadas em evidências. Enquanto o dado representa um registro discreto e descontextualizado, a informação emerge apenas após um processamento sistemático que

confere relevância e propósito ao conjunto (Stair; Reynolds, 2018). Para mitigar inconsistências e garantir a integridade analítica, Davenport e Harris (2017) preconizam a aplicação de fluxos de ETL. Esse processo técnico de depuração e padronização é o que permite converter grandes massas de dados governamentais em ativos informacionais confiáveis, assegurando que a análise final seja resiliente a ruídos estatísticos e erros de interpretação.

Por fim, a eficácia da gestão de dados públicos reside na convergência entre a qualidade da base e as ferramentas de apoio à decisão. A literatura sobre qualidade de dados enfatiza que dimensões como completude, precisão e atualidade são os pilares que sustentam a confiabilidade dos indicadores gerados (Batini; Scannapieco, 2016). Assim, a transformação dos microdados do Inep em ferramentas de monitoramento estratégico depende de uma infraestrutura técnica que sistematize o ciclo de vida do dado, permitindo que a complexidade do cenário educacional brasileiro seja traduzida em informações acionáveis para o enfrentamento de problemas estruturais, como a evasão e a permanência estudantil.

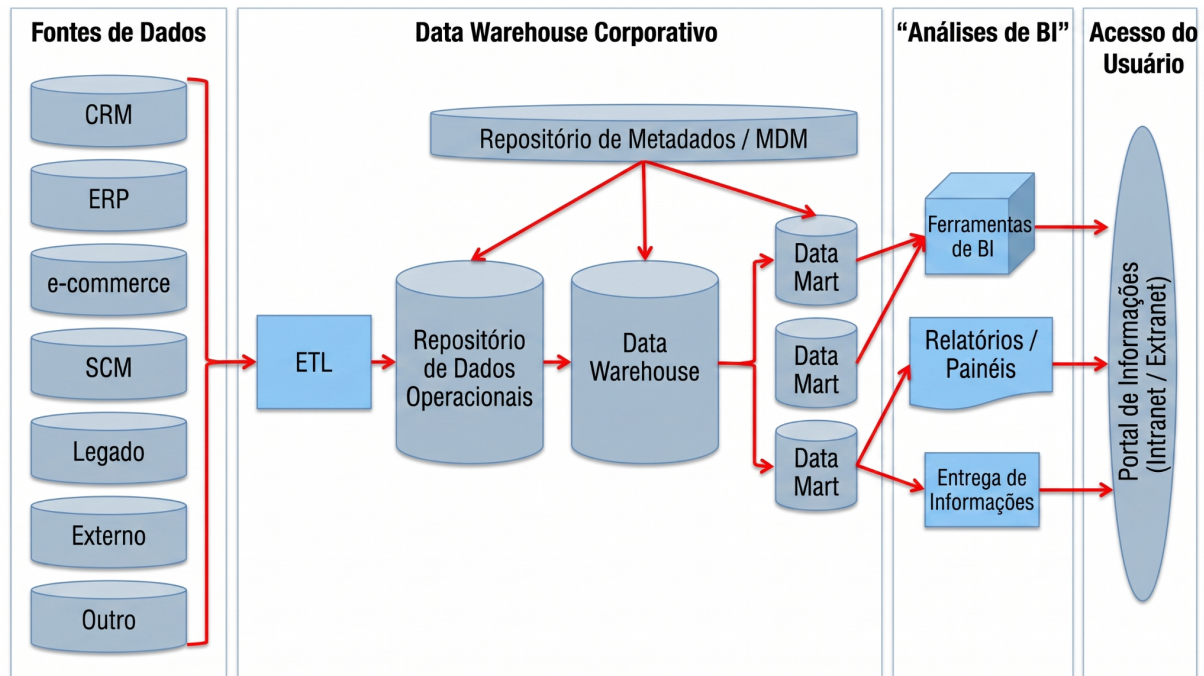
2.3 BUSINESS INTELLIGENCE (BI)

O termo *Business Intelligence* consolidou-se na literatura de Sistemas de Informação para descrever uma arquitetura e um conjunto de ferramentas que visam a conversão de dados brutos em informações acionáveis para a tomada de decisão. De acordo com Turban, Pollard e Wood (2021), o conceito moderno de BI evoluiu dos antigos Sistemas de Apoio à Decisão (SAD), passando a englobar não apenas ferramentas de análise retrospectiva, mas um processo holístico que abrange a coleta, o armazenamento e a gestão preditiva do conhecimento organizacional. Aprofundando essa perspectiva no contexto da Indústria 4.0, Tavera Romero *et al.* (2021) definem o BI como um ecossistema tecnológico crítico que integra bases de dados e algoritmos avançados para apresentar informações complexas de forma ágil aos tomadores de decisão, garantindo vantagem competitiva em mercados altamente dinâmicos.

A fundamentação tecnológica e estrutural do BI, contudo, repousa sobre a existência de um repositório centralizado e confiável. Segundo o modelo seminal de Inmon (2005), a base de um ambiente analítico eficaz é o Data Warehouse (DW), caracterizado classicamente como um banco de dados orientado por assunto, integrado, não volátil e histórico. A literatura clássica enfatiza que, sem a consolidação provida pelo DW, as iniciativas de BI tornam-se fragmentadas, uma vez que a integração de diferentes fontes de dados é o que garante a

padronização e a consistência das informações utilizadas no processo analítico, assegurando maior confiabilidade na tomada de decisão (Inmon, 2005; Stair; Reynolds, 2018). Esse fluxo de integração e a estrutura de camadas do sistema podem ser observados na Figura 1.

Figura 1 – Arquitetura de Business Intelligence.



Fonte: Adaptado de Turban, Pollard e Wood (2021).

O BI atua na lacuna entre os dados operacionais, gerados pelos sistemas de processamento de transações, e a necessidade de visão estratégica da alta gestão. Davenport e Harris (2017) argumentam que o BI permite que as organizações deixem de agir baseadas em intuições e empirismo para adotarem uma cultura estritamente orientada a dados (*data-driven*). Para os autores, a maturidade analítica de um sistema é medida pela sua capacidade de evoluir de análises puramente descritivas para diagnósticas e preditivas, fornecendo subsídios que expliquem as causas e os padrões subjacentes aos fenômenos observados.

Além da infraestrutura de armazenamento, a literatura contemporânea destaca o papel crítico das ferramentas de visualização de dados e exploração analítica. Refletindo a transição do mercado para o chamado *Self-Service BI*, Sharda, Delen e Turban (2020) reiteram que o grande objetivo atual do BI é a democratização dos dados. Essa abordagem permite que usuários de negócios explorem grandes massas de informações de forma autônoma e

interativa, sem a necessidade de conhecimentos técnicos profundos em programação ou linguagem de banco de dados.

2.4 MODELAGEM DIMENSIONAL

A modelagem dimensional é uma técnica de design de banco de dados especificamente voltada para ambientes analíticos e sistemas de apoio à decisão. Seu objetivo primordial é maximizar a performance das consultas e facilitar a compreensão intuitiva dos dados pelos usuários finais (Kimball; Ross, 2013). Diferente do design de bancos de dados transacionais, que busca a normalização para evitar redundâncias e garantir a integridade dos dados, a modelagem dimensional recorre à desnormalização estratégica, focando na eficiência da recuperação de grandes volumes de dados. (Sharda; Delen; Turban, 2021).

Segundo Silberschatz, Korth e Sudarshan (2020), essa arquitetura é composta por duas estruturas relacionais fundamentais: as tabelas de fatos e as tabelas de dimensões, que frequentemente assumem a topologia de um *Star Schema*. As tabelas de fatos compõem o núcleo do modelo, armazenando as métricas quantitativas que representam os eventos do processo de negócio. Em contrapartida, Coronel e Morris (2020) detalham que as tabelas de dimensões contêm os atributos descritivos que fornecem o contexto analítico para os fatos, permitindo aos gestores responderem a perguntas vitais como "quem", "onde", "quando" e "como" determinado evento ocorreu. Para Sharda, Delen e Turban (2021), a sinergia entre fatos e dimensões é o mecanismo essencial que converte dados brutos em um formato navegável, suportando a exploração interativa de tendências e padrões ocultos.

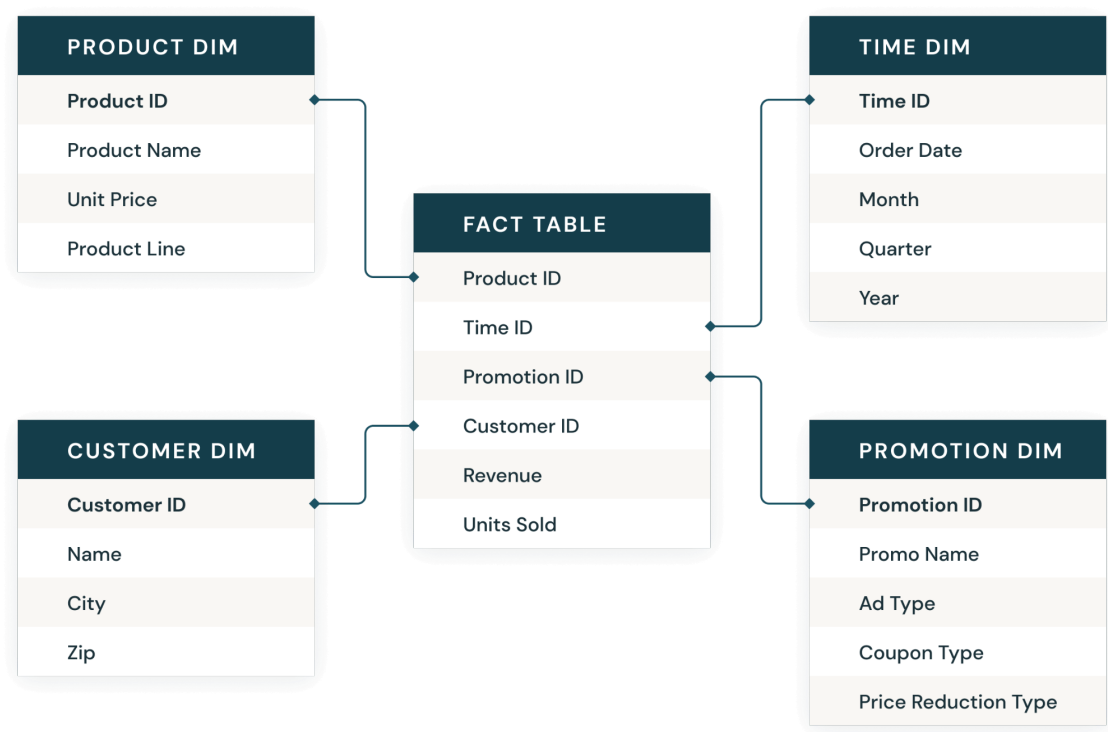
2.4.1 Star Schema

O *Star Schema* é a arquitetura dimensional mais difundida na literatura de BI. Ele é caracterizado por uma tabela de fato central cercada por várias tabelas de dimensões desnormalizadas, conectadas através de chaves estrangeiras (Kimball; Ross, 2013). De acordo com Silberschatz, Korth e Sudarshan (2020), a principal vantagem dessa topologia é a simplicidade das consultas, uma vez que requer apenas um nível de junção entre o fato e qualquer dimensão para extrair a informação.

A configuração visual e lógica dessa estrutura é apresentada na Figura 2, evidenciando a centralidade do evento em relação aos seus descritores. Sharda, Delen e Turban (2021) reiteram que o uso do *Star Schema* é recomendado para a maioria dos ambientes analíticos

devido à facilidade com que ferramentas de visualização de dados interpretam seu relacionamento, além de oferecer o melhor desempenho em operações de filtragem e agregação.

Figura 2 – Representação Gráfica do Modelo Star Schema



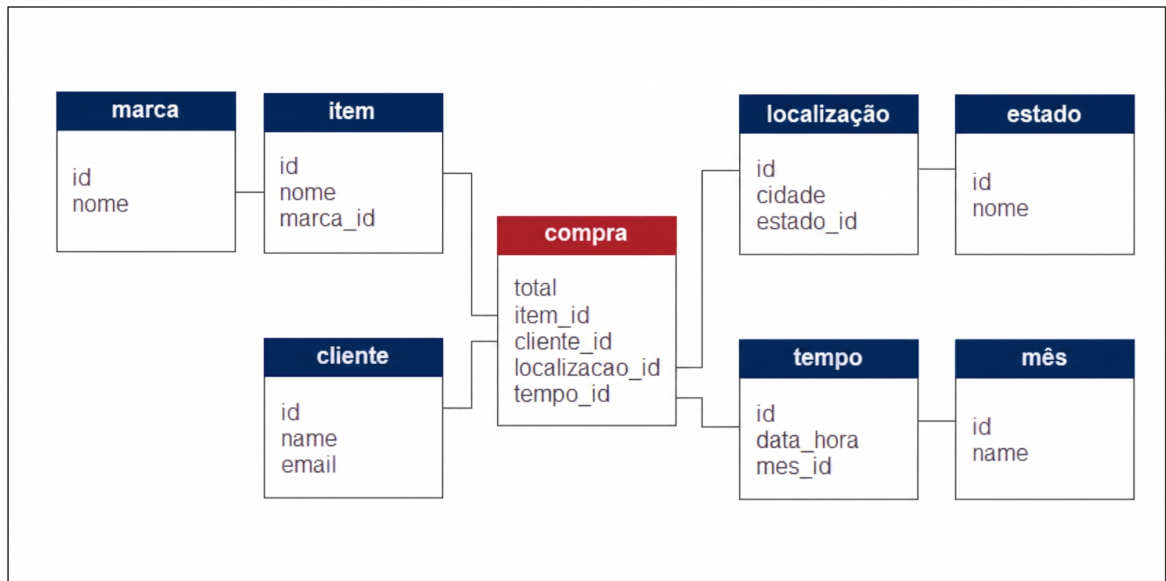
Fonte: Databricks (2024).

2.4.2 Snowflake

O esquema floco de neve (*Snowflake Schema*) é uma variação do modelo em estrela na qual as tabelas de dimensões são normalizadas, fragmentando hierarquias extensas para eliminar redundâncias e otimizar o armazenamento (Kimball; Ross, 2013; Coronel; Morris, 2020).

Embora o modelo otimize o espaço de armazenamento e facilite a manutenção da integridade dos dados mestre, Silberschatz, Korth e Sudarshan (2020) advertem que essa arquitetura pode degradar a performance das consultas, visto que exige múltiplos *joins* para reconstruir o contexto da dimensão. A Figura 3 ilustra como as dimensões se expandem para níveis secundários de detalhamento. Para Sharda, Delen e Turban (2021), a escolha por esse esquema deve ser justificada por necessidades técnicas específicas de armazenamento ou quando os dados de origem já possuem uma estrutura altamente hierarquizada.

Figura 3 – Representação Gráfica do Modelo Snowflake.



Fonte: Adaptado de PhoenixNAP (2024).

2.5 TOMADA DE DECISÃO BASEADA EM DADOS

A tomada de decisão baseada em dados é o processo de fundamentar escolhas estratégicas em evidências quantitativas e análises objetivas, minimizando a dependência de heurísticas cognitivas e intuições subjetivas. Segundo Marr (2017), a capacidade de tomar decisões inteligentes não reside no volume de dados coletados, mas na habilidade de formular as perguntas corretas que orientem a extração de valor desses ativos. Para o autor, os dados devem constituir a base estruturante da inteligência estratégica, permitindo que a organização se antecipe a crises e identifique oportunidades de melhoria de forma preditiva.

Dentro da perspectiva da Gestão do Conhecimento, Nonaka e Takeuchi (2019) argumentam que a decisão eficaz surge da conversão do conhecimento explícito, de dados e fatos registrados, em conhecimento tácito. Nesse contexto, os sistemas de informação atuam como mediadores que organizam o fluxo de dados para que o tomador de decisão possa interpretar o contexto e agir de forma assertiva. A qualidade desse processo é diretamente proporcional à qualidade dos dados utilizados. Redman (2016) adverte que decisões baseadas em dados imprecisos ou mal geridos podem ser mais prejudiciais do que aquelas baseadas puramente na intuição, cunhando o conceito de que a integridade do dado é o alicerce de qualquer governança analítica.

A natureza da decisão no serviço público e em instituições complexas também é discutida sob a ótica da Teoria de Sistemas. Checkland e Poulter (2020) sugerem que a tomada de decisão deve considerar a multidimensionalidade dos sistemas sociais. Ao utilizar bases de dados para informar políticas, o gestor deve ser capaz de realizar o que os autores chamam de "pensamento sistêmico", enxergando como uma variável é influenciada por um conjunto de outros fatores externos e internos capturados pelos sistemas de informação.

Complementarmente, Kitchin (2022) destaca que a revolução dos dados transformou a tomada de decisão pública. Para o autor, vivemos na era da "governança algorítmica", onde a infraestrutura técnica que processa o dado define os limites do que pode ser decidido. Assim, a eficácia de uma decisão baseada em dados depende de uma infraestrutura que permita a transparência dos processos analíticos, garantindo que os resultados gerados por ferramentas de BI sejam auditáveis e socialmente responsáveis (Kitchin, 2022).

3 METODOLOGIA

A presente pesquisa caracteriza-se como um estudo de natureza aplicada, com abordagem quantitativa e objetivos descritivos. O delineamento deste trabalho segue o modelo de pesquisa aplicada, e o procedimento técnico será estruturado em fases fundamentadas no ciclo de vida de projetos de BI e Ciência de Dados, compreendendo a coleta, o tratamento e a visualização dos dados, conforme detalhado nas seções seguintes.

3.1 COLETA DOS DADOS

A base de dados utilizada nesta pesquisa é composta pelos microdados do Censo da Educação Superior, disponibilizados anualmente pelo Inep, conforme ilustrado na Figura 4.

Para a constituição da série histórica analisada, foram selecionados 11 conjuntos de microdados, correspondentes ao período de 2014 a 2024, contemplando um conjunto para cada ano de referência. A definição desse intervalo temporal teve como objetivo garantir uma série histórica suficientemente extensa para a análise de tendências estruturais no ensino superior, bem como possibilitar a comparação entre o período anterior e posterior ao contexto pandêmico da Covid-19.

Figura 4 – Página de acesso aos Microdados do Censo da Educação Superior.



Fonte: Inep (2026).

Cada conjunto anual é disponibilizado pelo Inep em arquivo compactado no formato .zip, contendo os arquivos de dados em formato CSV, além da documentação técnica correspondente, incluindo dicionário de dados, leiautes e instrumentos de coleta.

Dentre os arquivos disponibilizados, esta pesquisa concentrou-se especificamente naqueles referentes ao cadastro de cursos e ao cadastro de IES, identificados pelos padrões MICRODADOS_CADASTRO_CURSOS e MICRODADOS_CADASTRO_IES.

O arquivo de cadastro de cursos reúne informações acadêmicas relacionadas à oferta e ao funcionamento dos cursos, sendo utilizado para a extração de variáveis referentes ao fluxo discente, tais como quantitativos de ingressantes, matriculados e concluintes, além de dados sobre modalidade de ensino e turno de funcionamento. Por sua vez, o arquivo de cadastro de IES foi empregado para obtenção de informações institucionais, incluindo categoria administrativa e localização geográfica das instituições.

Os arquivos foram obtidos por meio de download direto no portal oficial do Inep e posteriormente organizados em estrutura hierárquica no ambiente em nuvem Google Drive, com o objetivo de viabilizar a integração com as ferramentas utilizadas posteriormente.

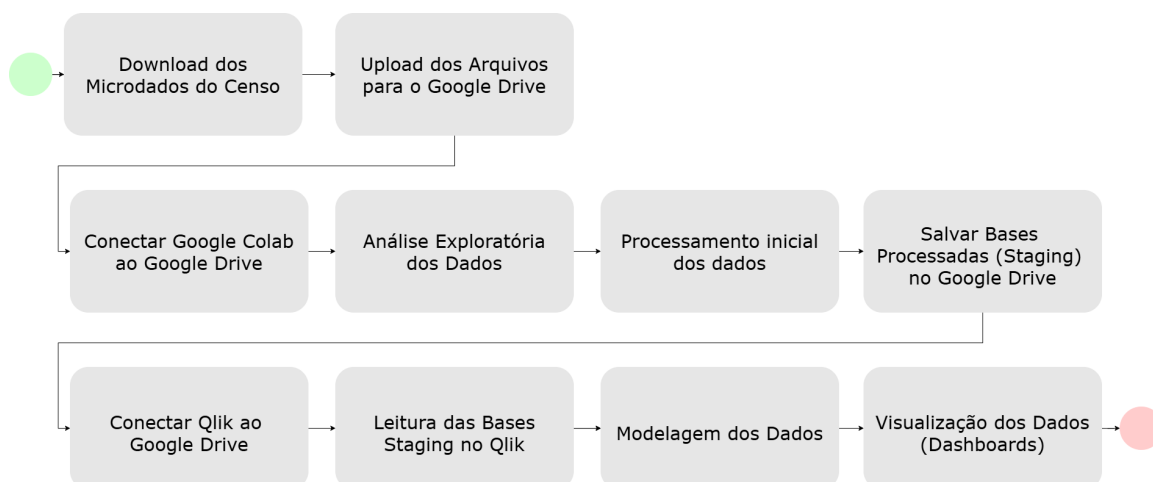
Considerando os 11 anos analisados, a base original totalizou aproximadamente 3,6 milhões de registros acumulados de cadastro de cursos e 27 mil registros do cadastro de IES ao longo da série histórica, distribuídos em mais de 300 variáveis por ano. O volume acumulado de dados brutos ultrapassou aproximadamente 2 GB.

Essa volumetria evidencia a dimensão nacional e longitudinal da base analisada, bem como sua complexidade estrutural, justificando a adoção de procedimentos sistematizados de tratamento, integração e modelagem de dados, descritos na seção subsequente.

3.2 PROCESSAMENTO E ETL

Após a etapa de coleta e organização dos microdados brutos, procedeu-se à fase de tratamento e processamento dos dados, com o objetivo de garantir consistência, padronização e adequação da base ao escopo analítico da pesquisa. O detalhamento completo do fluxo de dados, contemplando as etapas de ETL, encontra-se apresentado no diagrama de fluxo, conforme ilustrado na Figura 5.

Figura 5 – Diagrama de fluxo de ETL.



Fonte: Elaborado pelo autor (2026).

Inicialmente, foi realizada uma análise exploratória das bases brutas com o objetivo de compreender a estrutura, a tipologia das variáveis e a qualidade das informações disponibilizadas. Essa etapa foi conduzida na plataforma Google Colaboratory, utilizando a linguagem Python 3.1.13 e a biblioteca Pandas, amplamente empregada para manipulação e análise de dados estruturados.

A escolha do Google Colaboratory se justifica por se tratar de um ambiente de desenvolvimento baseado em nuvem que permite a execução de notebooks interativos sem necessidade de configuração local de infraestrutura computacional, além de facilitar o compartilhamento e a reprodutibilidade das análises. Enquanto a utilização da linguagem Python deve-se à sua ampla adoção na área de ciência de dados e à disponibilidade de bibliotecas especializadas para tratamento e exploração de grandes volumes de dados. Em particular, a biblioteca Pandas, que oferece estruturas eficientes para manipulação de dados tabulares, possibilitando operações de filtragem, agregação e verificação de consistência de forma ágil, que são características essenciais para o processamento dos microdados educacionais utilizados neste estudo.

A análise exploratória permitiu a inspeção preliminar das bases, conforme apresentado na Figura 6, por meio de comandos como `head()`, para visualização das primeiras linhas; `shape`, para identificação da dimensionalidade dos conjuntos de dados; e `info()`, para verificação dos tipos de dados e da presença de valores nulos.

Figura 6 – Trecho do código de análise exploratória no Google Colab.

```
[7] # Ver as primeiras linhas
df.head()

  NU_ANO_CENSO  NO_REGIAO  CO_REGIAO  NO_UF  SG_UF  CO_UF  NO_MUNICIPIO  CO_MUNICIPIO  IN_CAPITAL  TP_DIMENSAO  ...  QT_MAT_APOIO_SOCIAL  QT_CONC_APOIO_SOCIAL  QT_ATIV_EXTRACURRICULAR  QT_ING_A
0      2024      NaN      NaN      NaN      NaN      NaN      NaN      NaN      NaN      2  ...              0              0              0
1      2024      NaN      NaN      NaN      NaN      NaN      NaN      NaN      NaN      2  ...              0              0              0
2      2024      NaN      NaN      NaN      NaN      NaN      NaN      NaN      NaN      2  ...              0              0              0
3      2024      NaN      NaN      NaN      NaN      NaN      NaN      NaN      NaN      2  ...              0              0              0
4      2024      NaN      NaN      NaN      NaN      NaN      NaN      NaN      NaN      2  ...              0              0              4
5 rows x 223 columns

[8] # Ver o tamanho da base (linhas, colunas)
df.shape

(720349, 223)

[9] # Tipos de dados e valores nulos
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 720349 entries, 0 to 720348
Columns: 223 entries, NU_ANO_CENSO to QT_CONC_MOB_ACADEMICA
dtypes: float64(5), int64(200), object(18)
memory usage: 1.2+ GB
```

Fonte: Elaborado pelo autor (2026).

Nessa fase, observou-se um aumento gradual e esperado no volume de registros ao longo dos anos analisados. Igualmente, verificou-se que, entre os anos de 2022 e 2024, houve a inclusão de novas colunas nos conjuntos de dados disponibilizados pelo Inep. Contudo, alterações estruturais, como a adição de variáveis, modificações na nomenclatura ou mudanças no escopo de determinadas colunas, encontram-se devidamente documentadas nos dicionários de dados dos microdados. Dessa forma, foi possível identificar e compreender tais modificações durante o processo de tratamento das bases. As novas variáveis introduzidas nesses anos não impactaram o escopo analítico definido para este estudo, uma vez que não estavam relacionadas ao conjunto de atributos selecionados para a análise.

Adicionalmente, realizou-se a verificação sistemática de valores ausentes por meio do comando `isnull().sum()`, conforme ilustrado na Figura 7. A análise evidenciou quantidade expressiva de valores nulos em colunas relacionadas à localização geográfica.

Figura 7 – Resultado da verificação de valores nulos na base de cursos.

```
[6] ✓ 1s # Contar valores nulos por coluna
df.isnull().sum()[df.isnull().sum() > 0]
```

	0
NO_REGIAO	11778
CO_REGIAO	11778
NO_UF	11778
SG_UF	11778
CO_UF	11778
NO_MUNICIPIO	11778
CO_MUNICIPIO	11778
IN_CAPITAL	11778
TP_GRAU_ACADEMICO	2046

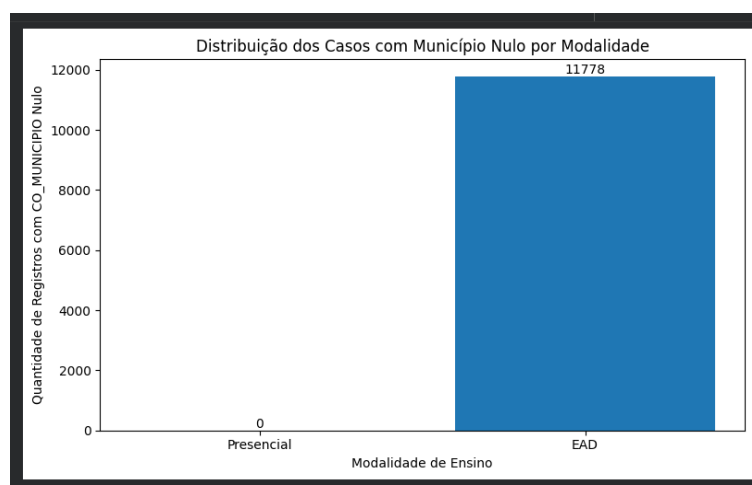
dtype: int64

Fonte: Elaborado pelo autor (2026).

A investigação detalhada desses registros permitiu identificar uma característica estrutural da base: nos microdados de cursos ofertados na modalidade a distância (EaD), determinadas variáveis geográficas apresentam valores nulos, em razão da lógica de registro adotada pelo sistema de origem. Nesses casos, a ausência de dados não configura inconsistência, mas sim uma particularidade da modelagem original dos microdados.

Essa relação entre modalidade de ensino e ausência de informações geográficas foi analisada por meio de visualização gráfica utilizando a biblioteca Matplotlib, conforme apresentado na Figura 8, permitindo evidenciar a correlação entre a variável TP_MODALIDADE_ENSINO e a incidência de valores nulos nos atributos de localização.

Figura 8 – Relação entre modalidade de ensino e incidência de valores nulos em variáveis geográficas.



Fonte: Elaborado pelo autor (2026).

Após a análise exploratória, implementou-se um procedimento automatizado por meio de estrutura de repetição, permitindo o processamento padronizado dos 11 arquivos anuais da base de cursos e dos 11 arquivos correspondentes à base de IES. Essa abordagem garantiu uniformidade no tratamento dos dados e reprodutibilidade do processo.

O processamento incluiu a filtragem dos registros referentes exclusivamente a cursos de graduação, excluindo-se aqueles relativos a cursos sequenciais de formação específica e demais níveis de ensino não contemplados no escopo da pesquisa. Adicionalmente, a base foi restrita aos cursos pertencentes à área de TIC, conforme a adaptação da CINE/Unesco adotada pelo Inep. Essa delimitação assegura que a análise reflita as particularidades do fluxo discente no setor tecnológico, objeto central deste estudo.

A Figura 9 apresenta um trecho do script utilizado para a realização da filtragem dos cursos de graduação e da seleção da área específica considerada.

Figura 9 – Trecho do script de filtragem dos cursos de graduação e da área TIC.

```
AREA_TIC = 'Computação e Tecnologias da Informação e Comunicação (TIC)'  #(TIC)
NIVEL_GRADUACAO = 1  #Graduação: 1

df_computacao_tic_grad = df.loc[
    (df['NO_CINE_AREA_ESPECIFICA'] == AREA_TIC) &
    (df['TP_NIVEL_ACADEMICO'] == NIVEL_GRADUACAO)
].copy()
```

Fonte: Elaborado pelo autor (2026).

Posteriormente, procedeu-se à seleção de atributos, com vistas à redução de dimensionalidade e à retenção das variáveis pertinentes ao escopo analítico do estudo. Na base de cursos, foram selecionadas 63 variáveis dentre as 223 originalmente disponíveis. Na base de IES, o conjunto foi restrito a 13 colunas consideradas essenciais para a análise.

A seleção das variáveis foi orientada por critérios de aderência aos objetivos da pesquisa. A descrição detalhada das variáveis selecionadas encontra-se apresentada no Quadro 1 e no Quadro 2, elaborados com base no dicionário de dados disponibilizado pelo Inep.

Quadro 1 – Variáveis selecionadas da base de cursos.

Nº	Nome da Variável	Descrição	Tipo	Categoria
1	NUM_ANO_CENSO	Ano de referência do Censo da Educação Superior	Numérico	
2	CO_IES	Código da Instituição	Numérico	
3	CO_CURSO	Nome do Curso	Numérico	
4	NO_CURSO	Código do Curso	Texto	
5	TP_GRAU_ACADEMICO	Tipo do grau acadêmico conferido ao aluno pela conclusão dos requisitos exigidos pelo curso	Numérico	1. Bacharelado 2. Licenciatura 3. Tecnológico 4. Bacharelado e Licenciatura (.) Não aplicável
6	TP_MODALIDADE_ENSINO	Tipo de modalidade de ensino do curso	Numérico	1. Presencial 2. Curso a distância
7	QT_ING	Quantidade de ingressantes	Numérico	
8	QT_ING_FEM	Quantidade de ingressantes do sexo feminino	Numérico	
9	QT_ING_MAS	Quantidade de ingressantes do sexo masculino	Numérico	
10	QT_ING_DIURNO	Quantidade de ingressantes no turno diurno	Numérico	
11	QT_ING_NOTURNO	Quantidade de ingressantes no turno noturno	Numérico	
12	QT_ING_0_17	Quantidade de ingressantes - Até 17 anos de idade	Numérico	
13	QT_ING_18_24	Quantidade de ingressantes - De 18 a 24 anos de idade	Numérico	
14	QT_ING_25_29	Quantidade de ingressantes - De 25 a 29 anos de idade	Numérico	
15	QT_ING_30_34	Quantidade de ingressantes - De 30 a 34 anos de idade	Numérico	
16	QT_ING_35_39	Quantidade de ingressantes - De 35 a 39 anos de idade	Numérico	
17	QT_ING_40_49	Quantidade de ingressantes - De 40 a 49 anos de idade	Numérico	
18	QT_ING_50_59	Quantidade de ingressantes - De 50 a 59 anos de idade	Numérico	
19	QT_ING_60 MAIS	Quantidade de ingressantes - De 60 ou mais anos de idade	Numérico	

20	QT_ING_BRANCA	Quantidade de ingressantes - Cor/Raça Branca	Numérico	
21	QT_ING_PRETA	Quantidade de ingressantes - Cor/Raça Preta	Numérico	
22	QT_ING_PARDA	Quantidade de ingressantes - Cor/Raça Parda	Numérico	
23	QT_ING_AMARELA	Quantidade de ingressantes - Cor/Raça Amarela	Numérico	
24	QT_ING_INDIGENA	Quantidade de ingressantes - Cor/Raça Indígena	Numérico	
25	QT_ING_CORND	Quantidade de ingressantes - Cor/Raça não dispõe da informação ou não declarada	Numérico	
26	QT_MAT	Quantidade de matrículas	Numérico	
27	QT_MAT_FEM	Quantidade de matrículas do sexo feminino	Numérico	
28	QT_MAT_MAS	Quantidade de matrículas do sexo masculino	Numérico	
29	QT_MAT_DIURNO	Quantidade de matrículas no turno diurno	Numérico	
30	QT_MAT_NOTURNO	Quantidade de matrículas no turno noturno	Numérico	
31	QT_MAT_0_17	Quantidade de matrículas - Até 17 anos de idade	Numérico	
32	QT_MAT_18_24	Quantidade de matrículas - De 18 a 24 anos de idade	Numérico	
33	QT_MAT_25_29	Quantidade de matrículas - De 25 a 29 anos de idade	Numérico	
34	QT_MAT_30_34	Quantidade de matrículas - De 30 a 34 anos de idade	Numérico	
35	QT_MAT_35_39	Quantidade de matrículas - De 35 a 39 anos de idade	Numérico	
36	QT_MAT_40_49	Quantidade de matrículas - De 40 a 49 anos de idade	Numérico	
37	QT_MAT_50_59	Quantidade de matrículas - De 50 a 59 anos de idade	Numérico	
38	QT_MAT_60 MAIS	Quantidade de matrículas - De 60 ou mais anos de idade	Numérico	
39	QT_MAT_BRANCA	Quantidade de matrículas - Cor/Raça Branca	Numérico	
40	QT_MAT_PRETA	Quantidade de matrículas - Cor/Raça Preta	Numérico	
41	QT_MAT_PARDA	Quantidade de matrículas - Cor/Raça Parda	Numérico	

42	QT_MAT_AMARELA	Quantidade de matrículas - Cor/Raça Amarela	Numérico	
43	QT_MAT_INDIGENA	Quantidade de matrículas - Cor/Raça Indígena	Numérico	
44	QT_MAT_CORND	Quantidade de matrículas - Cor/Raça não dispõe da informação ou não declarada	Numérico	
45	QT_CONC	Quantidade de concluintes	Numérico	
46	QT_CONC_FEM	Quantidade de concluintes do sexo feminino	Numérico	
47	QT_CONC_MAS	Quantidade de concluintes do sexo masculino	Numérico	
48	QT_CONC_DIURNO	Quantidade de concluintes no turno diurno	Numérico	
49	QT_CONC_NOTURNO	Quantidade de concluintes no turno noturno	Numérico	
50	QT_CONC_0_17	Quantidade de concluintes - Até 17 anos de idade	Numérico	
51	QT_CONC_18_24	Quantidade de concluintes - De 18 a 24 anos de idade	Numérico	
52	QT_CONC_25_29	Quantidade de concluintes - De 25 a 29 anos de idade	Numérico	
53	QT_CONC_30_34	Quantidade de concluintes - De 30 a 34 anos de idade	Numérico	
54	QT_CONC_35_39	Quantidade de concluintes - De 35 a 39 anos de idade	Numérico	
55	QT_CONC_40_49	Quantidade de concluintes - De 40 a 49 anos de idade	Numérico	
56	QT_CONC_50_59	Quantidade de concluintes - De 50 a 59 anos de idade	Numérico	
57	QT_CONC_60 MAIS	Quantidade de concluintes - De 60 ou mais anos de idade	Numérico	
58	QT_CONC_BRANCA	Quantidade de concluintes - Cor/Raça Branca	Numérico	
59	QT_CONC_PRETA	Quantidade de concluintes - Cor/Raça Preta	Numérico	
60	QT_CONC_PARDA	Quantidade de concluintes - Cor/Raça Parda	Numérico	
61	QT_CONC_AMARELA	Quantidade de concluintes - Cor/Raça Amarela	Numérico	
62	QT_CONC_INDIGENA	Quantidade de concluintes - Cor/Raça Indígena	Numérico	
63	QT_CONC_CORND	Quantidade de concluintes - Cor/Raça não	Numérico	

		dispõe da informação ou não declarada		
--	--	---------------------------------------	--	--

Fonte: Elaborado pelo autor com base no dicionário de dados do Inep (2024).

Quadro 2 – Variáveis selecionadas da base de instituições.

Nº	Nome da Variável	Descrição	Tipo	Categoria
1	CO_IES	Código único de identificação da IES	Numérico	
2	NO_IES	Nome da IES	Texto	
3	SG_IES	Sigla da IES	Texto	
4	TP_ORGANIZACAO_ACADEMICA	Tipo de Organização Acadêmica da IES	Numérico	1. Universidade 2. Centro Universitário 3. Faculdade 4. Instituto Federal de Educação, Ciência e Tecnologia 5. Centro Federal de Educação Tecnológica
5	TP_REDE	Rede de Ensino	Numérico	1. Pública 2. Privada
6	CO_UF_IES	Código da Unidade da Federação da sede administrativa ou reitoria da IES	Numérico	
7	NO_UF_IES	Nome da Unidade da Federação da sede administrativa ou reitoria da IES	Texto	
8	SG_UF_IES	Sigla da Unidade da Federação da sede administrativa ou reitoria da IES	Texto	
9	CO_MUNICIPIO_IES	Código do Município da sede administrativa ou reitoria da IES	Numérico	
10	NO_MUNICIPIO_IES	Nome do Município da sede administrativa ou reitoria da IES	Texto	
11	NO_REGIAO_IES	Nome da região geográfica da sede administrativa ou reitoria da IES	Texto	
12	NO_MESORREGIAO_IES	Nome da mesorregião geográfica da sede administrativa ou reitoria da IES	Texto	
13	NO_MICRORREGIAO_IES	Nome da microrregião geográfica da sede administrativa ou reitoria da IES	Texto	

Fonte: Elaborado pelo autor com base no dicionário de dados do Inep (2024).

Os dados filtrados e tratados nessa etapa foram então concatenados, originando duas bases consolidadas em camada de *staging*, uma referente aos cursos e outra às IES, ambas estruturadas para subsidiar as etapas subsequentes de análise e modelagem. As bases

resultantes foram armazenadas em ambiente Google Drive, garantindo persistência dos arquivos, organização do fluxo de trabalho e facilidade de acesso nas fases posteriores do processo analítico.

3.3 MODELAGEM DOS DADOS

Após a etapa de transformação dos dados no ambiente Python, iniciou-se a fase de Load do processo ETL, na qual os *datasets* já tratados e padronizados foram carregados na plataforma Qlik. Nessa etapa, os dados tabulares resultantes foram organizados e convertidos em um modelo multidimensional do tipo *Star Schema*, explorando a tecnologia associativa da ferramenta para otimizar consultas analíticas e proporcionar maior desempenho nas análises.

A escolha da plataforma para o desenvolvimento deste projeto fundamenta-se em suas características técnicas voltadas à integração, à modelagem e à exploração de dados de maneira eficiente. Um dos principais fatores que motivaram sua adoção é a presença de uma linguagem de script própria para extração e transformação de dados, cuja sintaxe apresenta semelhanças com a linguagem SQL, amplamente utilizada em bancos de dados, além da familiaridade da pesquisadora com a ferramenta. Essa proximidade facilita o aprendizado e a utilização da ferramenta por usuários que já possuem familiaridade com consultas estruturadas, permitindo realizar operações de seleção, transformação e carga de dados de forma relativamente intuitiva.

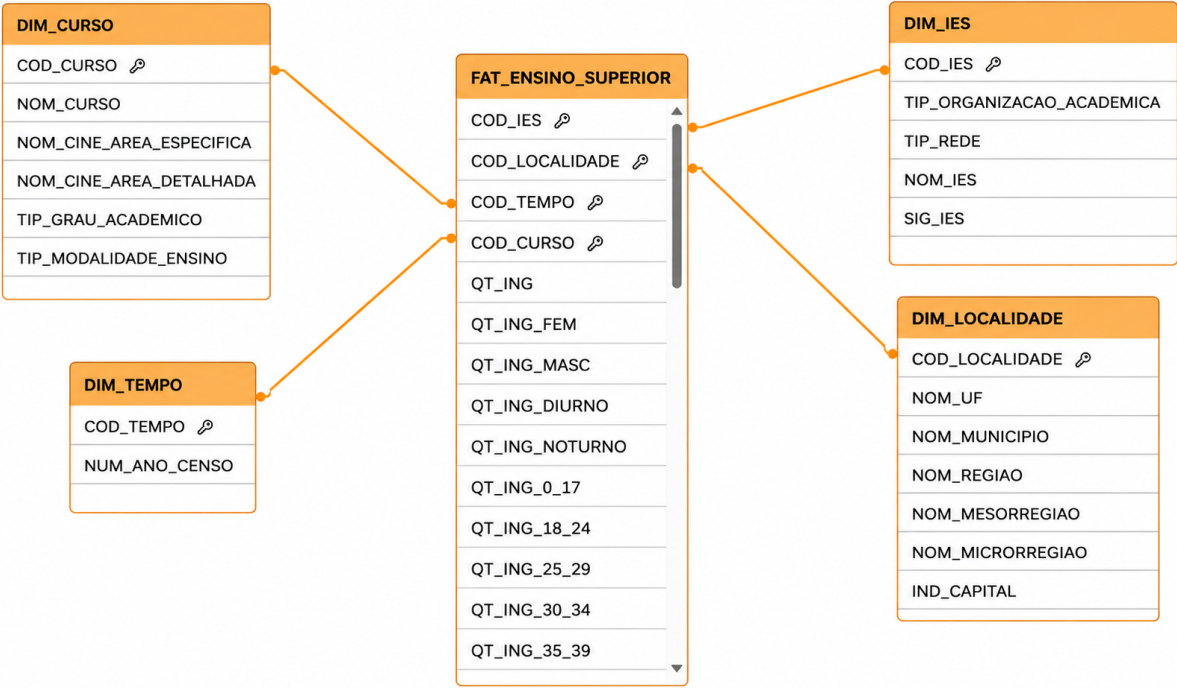
Outro aspecto metodológico relevante é a facilidade de conexão com diferentes fontes de dados externas. A plataforma permite estabelecer conexão direta com as planilhas armazenadas no Google Drive, possibilitando a atualização e o consumo dos dados sem a necessidade de exportação manual dos arquivos para outros formatos. Essa funcionalidade contribui para a automatização do processo de atualização das bases e reduz a possibilidade de inconsistências decorrentes da manipulação manual.

Os conjuntos de dados foram carregados a partir do Google Drive e inicialmente armazenados em duas tabelas de *staging* na memória do Qlik. Posteriormente, essas tabelas intermediárias foram utilizadas como fonte para a modelagem analítica, permitindo maior eficiência no processamento e organização das estruturas dimensionais.

Conforme apresentado na Figura 10, a arquitetura do modelo foi centralizada na tabela fato FAT_ENSINO_SUPERIOR, relacionada a quatro dimensões analíticas: DIM_TEMPO, responsável pela organização cronológica do período de 2014 a 2024;

DIM_LOCALIDADE, contendo atributos geográficos; DIM_CURSO, com características específicas das graduações; e DIM_IES, com informações institucionais das IES.

Figura 10 – Modelo dimensional da base de dados



Fonte: Elaborado pelo autor (2026).

Para o carregamento de cada dimensão, foram selecionadas as variáveis pertinentes a cada entidade analítica. De igual modo, realizou-se a padronização de variáveis textuais por meio da função Upper(), garantindo uniformidade na capitalização dos registros. Essa prática reduz inconsistências causadas por variações entre letras maiúsculas e minúsculas, assegurando maior precisão nos filtros e seleções realizadas nos painéis analíticos.

Foi adotada ainda uma convenção de nomenclatura para padronização dos campos. As variáveis foram renomeadas utilizando prefixos de três letras que indicam o tipo de atributo representado, como COD para identificadores numéricos, SIG para siglas e NOM para atributos descritivos textuais. Essa estratégia facilita a leitura do modelo de dados, tornando mais clara a função de cada variável dentro das dimensões e da tabela fato. De forma análoga, as tabelas também foram padronizadas segundo a lógica, adotando os prefixos FAT para a tabela fato e DIM para as tabelas dimensionais. Essa convenção contribui para a

organização do modelo analítico, melhora a manutenção do script de carregamento e favorece a compreensão da estrutura de dados durante o processo de desenvolvimento e análise.

Para a definição das chaves primárias das dimensões, utilizou-se a função `AutoNumber()` do Qlik, responsável por gerar identificadores inteiros únicos a partir de combinações de campos. Esse procedimento contribui para a otimização da estrutura de armazenamento e melhora o desempenho das associações entre tabelas.

Como exemplo, no carregamento da dimensão `DIM_LOCALIDADE`, foi utilizado como parâmetro para a função `AutoNumber()` a concatenação do código da unidade federativa com o código do município, gerando um identificador único para a dimensão. Essa estratégia permitiu criar uma chave substituta eficiente para a modelagem dimensional, sem perder a referência aos códigos oficiais presentes no conjunto de dados original. O trecho correspondente do script é apresentado na Figura 11.

Figura 11 – Trecho do script Qlik para o carregamento da dimensão de localidade.

```
DIM_LOCALIDADE:
LOAD DISTINCT
    AutoNumber(CO_UF_IES & CO_MUNICIPIO_IES) AS COD_LOCALIDADE,
    UPPER(SG_UF_IES) AS SIG_UF,
    UPPER(NO_UF_IES) AS NOM_UF,
    UPPER(NO_MUNICIPIO_IES) AS NOM_MUNICIPIO,
    UPPER(NO_REGIAO_IES) AS NOM_REGIAO,
    UPPER(NO_MESORREGIAO_IES) AS NOM_MESORREGIAO,
    UPPER(NO_MICRORREGIAO_IES) AS NOM_MICRORREGIAO
RESIDENT STG_IES;
```

Fonte: Elaborado pelo autor (2026).

De forma análoga, procedeu-se ao carregamento da tabela fato, na qual foram incorporadas as chaves associativas e as medidas quantitativas provenientes dos microdados. O carregamento considerou as particularidades estruturais dos *datasets* disponibilizados pelo Inep, que apresentam um elevado número de indicadores agregados relacionados ao fluxo discente.

Essas medidas contemplam diferentes dimensões de análise, como ingressantes, matriculados e concluintes, além de recortes por sexo, turno, faixa etária e raça/cor. Dessa forma, a tabela fato `FAT_ENSINO_SUPERIOR` passou a concentrar os indicadores quantitativos do fenômeno analisado, enquanto as dimensões armazenam os atributos descritivos utilizados para segmentação e análise dos dados.

Durante o carregamento, foram utilizadas as chaves geradas anteriormente nas dimensões, COD_TEMPO, COD_CURSO, COD_IES e COD_LOCALIDADE, permitindo estabelecer as associações necessárias para a estrutura do modelo dimensional. O trecho do script responsável pelo carregamento da tabela fato é apresentado na Figura 12.

Figura 12 – Trecho do script Qlik para o carregamento da tabela fato.

```

1  FAT_ENSINO_SUPERIOR:
2  LOAD
3      AutoNumber(NU_ANO_CENSO)           AS COD_TEMPO,
4      AutoNumber(CO_CURSO)              AS COD_CURSO,
5      AutoNumber(CO_IES)                AS COD_IES,
6      AutoNumber(CO_UF_IES & CO_MUNICIPIO_IES) AS COD_LOCALIDADE,
7      QT_ING,
8      QT_ING_FEM,
9      QT_ING_MASC,
10     QT_ING_DIURNO,
11     QT_ING_NOTURNO,
12     QT_ING_0_17,
13     QT_ING_18_24,
14     QT_ING_25_29,
15     QT_ING_30_34,
16     QT_ING_35_39,
17     QT_ING_40_49,
18     QT_ING_50_59,
19     QT_ING_60 MAIS,
20     QT_ING_BRANCA,
21     QT_ING_PRETA,
22     QT_ING_PARDA,
23     QT_ING_AMARELA,
24     QT_ING_INDIGENA,
25     QT_ING_CORND,
26     QT_MAT,
27     QT_MAT_FEM,
28     QT_MAT_MASC,
29     QT_MAT_DIURNO,
30     QT_MAT_NOTURNO,
31     QT_MAT_0_17,
32     QT_MAT_18_24,
33     QT_MAT_25_29,
34     QT_MAT_30_34,
35     QT_MAT_35_39,
36     QT_MAT_40_49,
37     QT_MAT_50_59,
38     QT_MAT_60 MAIS,
39     QT_MAT_BRANCA,
40     QT_MAT_PRETA,
41     QT_MAT_PARDA,
42     QT_MAT_AMARELA,
43     QT_MAT_INDIGENA,
44     QT_MAT_CORND,
45     QT_CONC,
46     QT_CONC_FEM,
47     QT_CONC_MASC,
48     QT_CONC_DIURNO,
49     QT_CONC_NOTURNO,
50     QT_CONC_0_17,
51     QT_CONC_18_24,
52     QT_CONC_25_29,
53     QT_CONC_30_34,
54     QT_CONC_35_39,
55     QT_CONC_40_49,
56     QT_CONC_50_59,
57     QT_CONC_60 MAIS,
58     QT_CONC_BRANCA,
59     QT_CONC_PRETA,
60     QT_CONC_PARDA,
61     QT_CONC_AMARELA,
62     QT_CONC_INDIGENA,
63     QT_CONC_CORND
64  RESIDENT STG_CURSOS;
65

```

Fonte: Elaborado pelo autor (2026).

Após a finalização do carregamento, as tabelas intermediárias de *staging* foram removidas da memória, garantindo maior eficiência no consumo de recursos computacionais e concluindo, assim, a construção do modelo analítico de dados.

3.4 VISUALIZAÇÃO DOS DADOS

A etapa de visualização de dados consistiu no desenvolvimento de um painel interativo na plataforma Qlik, atuando como a camada de apresentação (*front-end*) do modelo analítico previamente construído. O objetivo desta fase foi apresentar, de forma integrada, os principais indicadores de fluxo discente relacionados à evasão e à permanência em cursos de Tecnologia da Informação, traduzindo o grande volume de métricas quantitativas da tabela fato em representações gráficas claras e objetivas.

Além dessa robustez no processamento, a ferramenta oferece recursos avançados para o design da informação. A organização visual do painel foi estruturada seguindo o princípio de hierarquia da informação (Knaflitz, 2015). Na seção superior, foram disponibilizados os indicadores-chave de desempenho (*KPIs*) em caixas de texto destacadas, incluindo a taxa de evasão, a taxa de conclusão e o quantitativo total de ingressantes e concluintes no período analisado. Posicionar essas métricas no topo garante que o usuário obtenha uma visão macro do cenário logo no primeiro contato com a tela.

Para apoiar a interpretação analítica e detalhar as métricas e os *KPIs*, foram escolhidas visualizações específicas, justificadas pela natureza de cada dado:

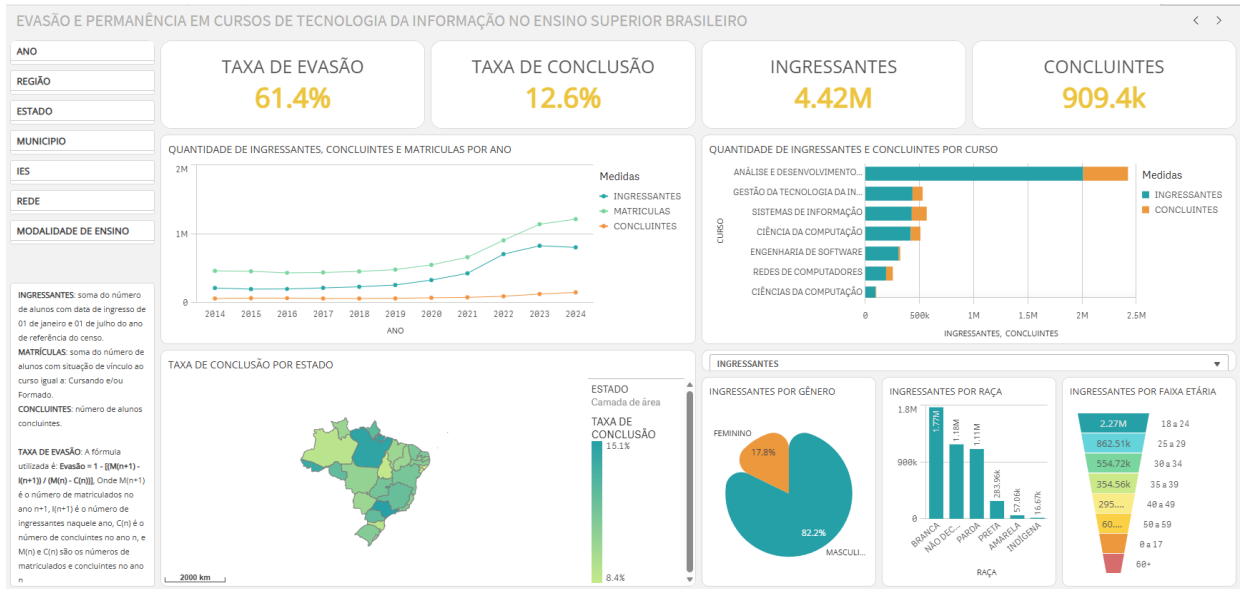
(i) Gráfico de linhas: adotado para representar a evolução temporal de ingressantes, matrículas e concluintes entre 2014 e 2024, por ser a estrutura visual mais adequada para evidenciar tendências, quedas estruturais e sazonalidades ao longo do tempo;

(ii) Gráfico de barras horizontais: utilizado para o ranqueamento dos cursos conforme o volume de ingressantes e concluintes. A orientação horizontal foi preferida para acomodar rótulos textuais longos, como nomes dos cursos, facilitando a leitura sem a necessidade de inclinar o texto;

(iii) Mapa coroplético: aplicado para a análise por unidade federativa, utilizando um gradiente de cores que permite identificar, de forma intuitiva e espacial, as diferenças regionais e as discrepâncias nas taxas de conclusão pelo país.

A disposição inicial desses elementos pode ser observada na Figura 13, que ilustra a interface do painel.

Figura 13 – Tela do painel interativo desenvolvido no Qlik.



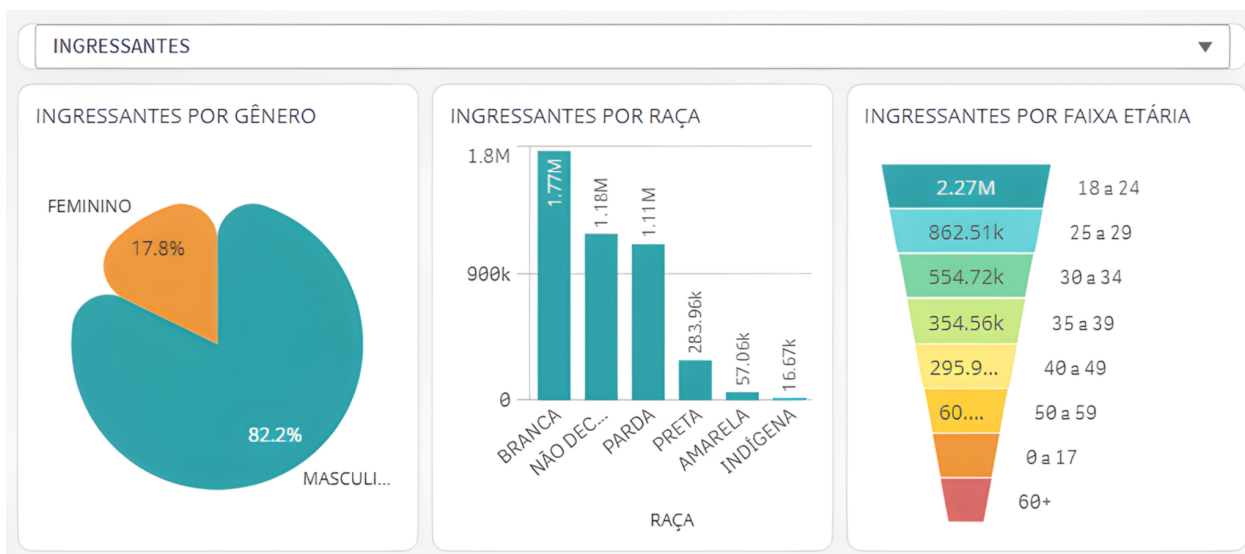
Fonte: Elaborado pelo autor (2026).

Adicionalmente, o painel contemplou a caracterização do perfil discente, com a segmentação de ingressantes e concluintes por gênero, raça/cor e faixa etária. A utilização de gráficos de distribuição para essas variáveis sociodemográficas viabiliza a identificação de disparidades e padrões de permanência.

O desenvolvimento desta seção específica do painel demandou uma abordagem técnica diferenciada, descrita a seguir, para atender a uma peculiaridade estrutural dos microdados do Inep. Nos *datasets* originais, as informações sociodemográficas não são apresentadas em formato granular, mas sim por meio de colunas com quantitativos previamente agrupados para cada curso e instituição, como, por exemplo, os campos QT_CONC_50_59 para concluintes entre 50 e 59 anos, ou QT_CONC_BRANCA para concluintes autodeclarados brancos.

Para otimizar o espaço na interface e evitar a duplicação excessiva de elementos visuais, que ocorreria caso fossem criados painéis separados para ingressantes, matriculados e concluintes, foi implementada uma solução interativa utilizando uma variável de controle. Conforme ilustrado na Figura 14, foi inserido um menu suspenso na parte superior da seção, permitindo ao usuário definir qual métrica deseja analisar no momento. Ao selecionar, por exemplo, "INGRESSANTES", a plataforma Qlik altera condicionalmente as expressões subjacentes de todos os gráficos dessa aba, passando a somar as medidas correspondentes à entrada de alunos.

Figura 14 – Painel de perfil discente com seletor dinâmico de métricas.

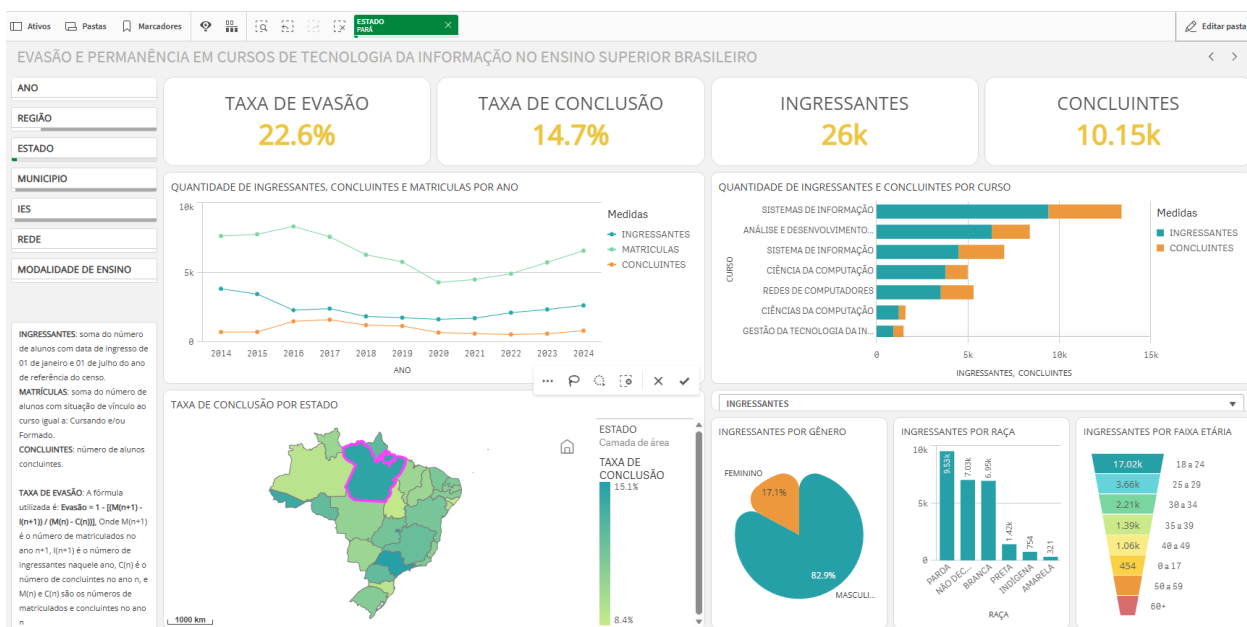


Fonte: Elaborado pelo autor (2026).

Um dos diferenciais da interface construída é a exploração da tecnologia associativa do Qlik. A interação do usuário ocorre por meio de filtros globais dispostos em um menu lateral e pela própria interação direta com os gráficos. Os filtros abrangem ano, região, estado, município, instituição de ensino, rede e modalidade. Diferentemente de consultas tradicionais, a tecnologia associativa permite que a seleção de qualquer elemento, como um clique em um estado específico no mapa ou em um curso na barra de ranqueamento, filtre instantaneamente e de forma dinâmica todas as demais visualizações da tela.

A Figura 15 demonstra a aplicação de filtros cruzados, evidenciando como a interface responde ativamente às interações do usuário. Ao selecionar um valor de dimensão diretamente em um dos gráficos, a plataforma Qlik adiciona automaticamente esse critério à barra de filtros. Em consequência dessa ação, todo o painel é reestruturado e as taxas e os quantitativos são recalculados instantaneamente, adaptando todas as visualizações da tela ao novo contexto de análise definido pelo usuário.

Figura 15 – Demonstração da aplicação de filtros dinâmicos no painel.



Fonte: Elaborado pelo autor (2026).

Por fim, com vistas à transparência metodológica e à usabilidade, foi incorporado um glossário técnico na própria interface. Esse recurso contém a descrição exata dos indicadores e das fórmulas utilizadas nas expressões da ferramenta, incluindo o cálculo da taxa de evasão com base no fluxo de matriculados, ingressantes e concluintes, assegurando que o usuário final compreenda as premissas matemáticas que embasam as visualizações apresentadas.

3.5 DEFINIÇÃO DAS MÉTRICAS

Para a quantificação dos fenômenos de retenção e abandono nos cursos da área de Tecnologia da Informação, foram estabelecidas métricas fundamentadas no fluxo anual de estudantes reportado pelo Censo da Educação Superior do Inep. A transparência dessas métricas é garantida pela inclusão de um glossário técnico na interface da solução desenvolvida, permitindo que o usuário compreenda a origem e o significado dos indicadores apresentados no painel analítico.

A construção dos indicadores baseou-se nas variáveis agregadas disponibilizadas nos microdados do censo, que reportam anualmente o volume de ingressantes, matriculados e concluintes por curso e instituição. A partir dessas informações foram definidos os seguintes indicadores base utilizados nos cálculos:

(i) Ingressantes (I): soma dos estudantes que iniciaram vínculo acadêmico no curso no ano de referência.

(ii) Matriculados (M): total de estudantes com vínculo ativo no curso no respectivo ano.

(iii) Concluintes (C): estudantes que integralizaram a carga horária do curso e obtiveram diploma no período analisado.

Esses três indicadores representam as principais variáveis de fluxo educacional presentes na base do Inep e constituem o ponto de partida para a construção das métricas de desempenho acadêmico utilizadas neste estudo.

A principal métrica adotada foi a Taxa de Evasão, calculada de forma longitudinal para estimar a perda de estudantes entre dois períodos consecutivos. A lógica aplicada no processamento dos dados segue a metodologia proposta pelo Instituto Lobo para o Desenvolvimento da Educação, Ciência e Tecnologia, amplamente utilizada em estudos sobre evasão no ensino superior brasileiro.

A equação utilizada é apresentada na Equação 1.

$$Evasão_n = 1 - (M(n + 1) - I(n + 1)) / (M_n - C_n) \quad (1)$$

Onde:

- n representa o ano de referência da análise;
- $M(n + 1)$ corresponde ao total de alunos matriculados no ano subsequente;
- $I(n + 1)$ representa o número de ingressantes no ano subsequente;
- M_n indica o total de alunos matriculados no ano base;
- C_n representa o número de concluintes no mesmo período.

Essa equação permite estimar o volume de estudantes que não permaneceram no sistema entre dois anos consecutivos, descontando tanto os novos ingressantes quanto os estudantes que concluíram o curso no período. Em termos conceituais, o cálculo busca identificar a proporção de alunos veteranos que deixam de renovar sua matrícula no período subsequente, caracterizando uma saída do sistema no período analisado, interpretada como evasão no contexto de dados agregados.

A metodologia deriva do cálculo da taxa de permanência, sendo a evasão definida como o complemento dessa taxa em relação a 100%. Essa abordagem é frequentemente denominada na literatura como evasão anual, evasão média anual ou taxa longitudinal de evasão, sendo amplamente empregada em análises baseadas em dados agregados do Censo da Educação Superior.

Complementarmente, foi definida a Taxa de Conclusão, apresentada na Equação 2, que representa uma métrica de sucesso acadêmico ao indicar a proporção de estudantes que efetivamente finalizaram o curso em relação ao total de matrículas ativas.

$$Conclusão_n = C_n / M_n \quad (2)$$

Essa métrica permite avaliar o desempenho do curso sob a perspectiva da diplomação, funcionando como um indicador complementar à evasão.

A adoção dessa metodologia se justifica pela natureza agregada dos microdados do Inep, que não permitem o acompanhamento individual da trajetória acadêmica dos estudantes ao longo do tempo. Diferentemente de sistemas acadêmicos institucionais, os microdados do censo não contêm identificadores únicos de alunos que possibilitem análises de coorte ou rastreamento longitudinal individual. Dessa forma, métodos baseados em fluxo agregado entre anos consecutivos representam a alternativa mais adequada para estimar indicadores de evasão em nível nacional.

Entretanto, é importante destacar algumas limitações inerentes a essa abordagem metodológica. Primeiramente, a métrica não distingue diferentes tipos de desligamento acadêmico, como transferência para outro curso ou instituição, trancamento de matrícula ou mudança de modalidade de ensino. Assim, parte dos estudantes classificados como evasores pode, na realidade, ter migrado para outra trajetória acadêmica dentro do próprio sistema educacional. Além disso, variações administrativas na forma de registro das matrículas ao longo dos anos podem influenciar pequenas oscilações nos indicadores.

Apesar dessas limitações, o método adotado é amplamente reconhecido na literatura especializada e tem sido utilizado em diversos estudos sobre evasão no ensino superior brasileiro justamente por sua compatibilidade com os dados públicos disponibilizados pelo Inep.

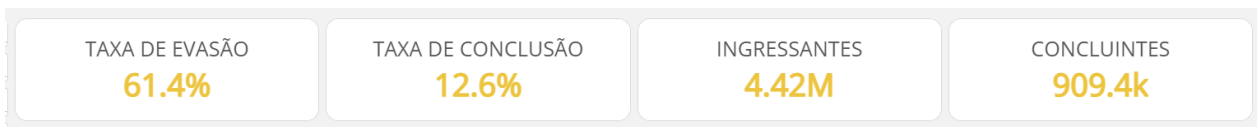
Por fim, as métricas foram implementadas diretamente no ambiente analítico da plataforma Qlik, permitindo que os cálculos sejam reprocessados dinamicamente conforme os filtros aplicados pelo usuário, como região geográfica, modalidade de ensino ou período analisado. Dessa forma, garante-se consistência estatística na visualização dos indicadores em diferentes níveis de agregação dos dados.

4 RESULTADOS E DISCUSSÃO

4.1 VISÃO GERAL E EVOLUÇÃO TEMPORAL DO FLUXO DISCENTE (2014–2024)

A análise dos dados consolidados entre 2014 e 2024, abrangendo um total de 109 cursos de 1841 instituições de ensino, revela um cenário desafiador para o ensino superior na área de Tecnologia da Informação no Brasil sob as perspectivas temporal, regional e sociodemográfica. Como resposta à Q1, o painel analítico evidencia uma taxa de evasão de 61,4%, representando um percentual elevado no contexto educacional aqui delimitado, o indicador mais expressivo identificado na análise. Esse resultado indica que, embora uma parcela dos estudantes permaneça vinculada ao sistema educacional, a efetividade formativa apresenta limitações, refletidas na taxa de conclusão de 12,6%, como mostrado na Figura 16. Tal cenário sugere a existência de dificuldades no fluxo acadêmico, evidenciando trajetórias prolongadas, retenção discente e obstáculos à diplomação.

Figura 16 – Indicadores do panorama nacional geral (2014–2024).



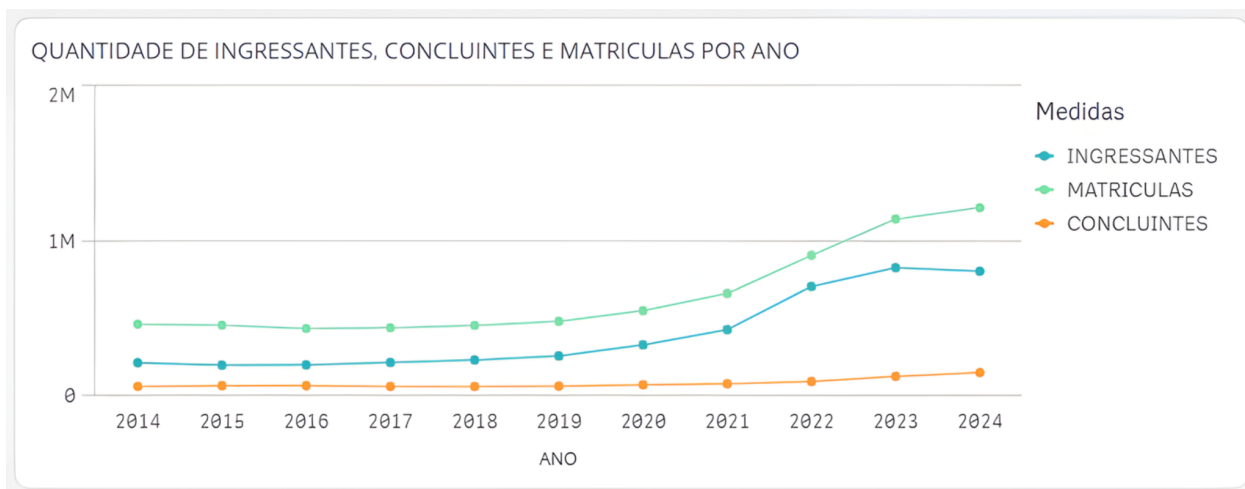
Fonte: Elaborado pelo autor (2026).

Em relação à Q2, a evolução temporal dos indicadores apresentada na Figura 17 evidencia um crescimento progressivo do volume de matrículas ao longo do período analisado, com aumento mais acentuado a partir de 2021. Esse movimento pode estar associado ao contexto pandêmico e pós-pandêmico, que acelerou processos de transformação digital na sociedade e ampliou a demanda por formação em áreas tecnológicas. Além disso, a expansão da modalidade de ensino a distância contribuiu para ampliar o acesso ao ensino superior, permitindo a entrada de novos perfis de estudantes.

No que se refere à Q3 observa-se que, apesar do crescimento expressivo das matrículas e dos ingressos, a série de concluintes permaneceu relativamente estável ao longo do mesmo período. Esse comportamento evidencia que a ampliação do acesso não foi acompanhada, no curto prazo, por mecanismos igualmente eficazes de permanência e progressão acadêmica. Tal cenário reforça o apresentado nos relatórios da OCDE (2021) de que a evasão no ensino superior reflete, além de perdas individuais para os estudantes,

ineficiências institucionais e sociais, associadas à interrupção do processo formativo e à subutilização de recursos educacionais.

Figura 17 – Gráfico de quantidade de ingressantes, concluintes e matrículas por ano.



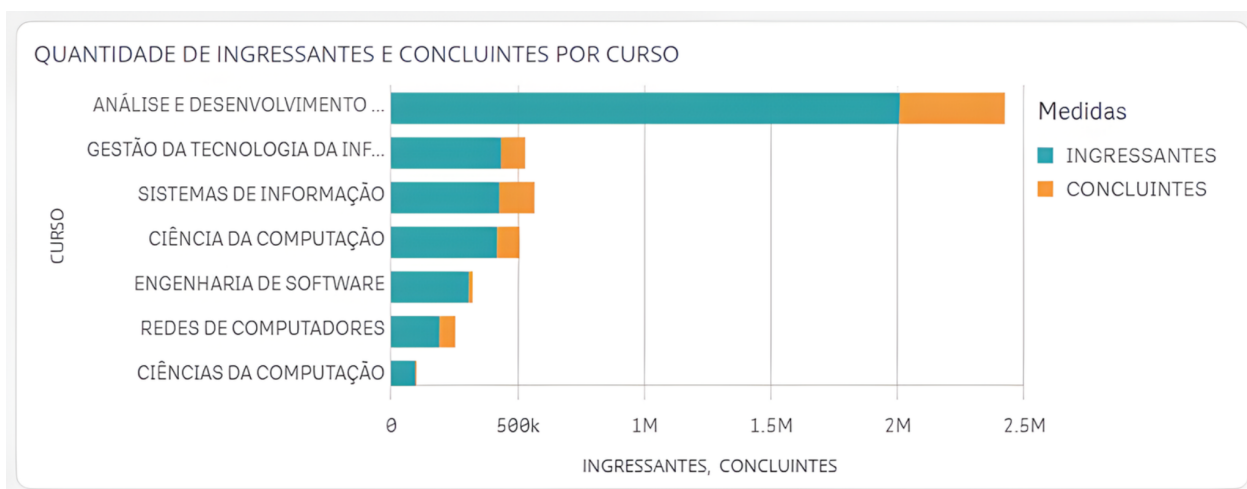
Fonte: Elaborado pelo autor (2026).

Ressalta-se, contudo, que parte dos estudantes ingressantes entre 2021 e 2024 ainda se encontra em processo de integralização curricular, considerando a duração média dos cursos da área. Dessa forma, os próximos censos educacionais serão fundamentais para verificar se a expansão recente se traduzirá em aumento efetivo de concluintes ou em novos ciclos de evasão.

4.2 ANÁLISE POR CURSOS E TENDÊNCIAS DE FORMAÇÃO TECNOLÓGICA

A partir dessa visão geral, a segmentação dos dados por curso evidencia padrões relevantes na escolha das formações dentro da área de Tecnologia da Informação. No que se refere à Q4, os cursos de Análise e Desenvolvimento de Sistemas e Gestão de Tecnologia da Informação lideram o ranking, concentrando os maiores volumes de ingressantes e concluintes ao longo do período analisado, conforme apresentado na Figura 18.

Figura 18 – Gráfico de quantidade de ingressantes e concluintes por curso.



Fonte: Elaborado pelo autor (2026).

Em um setor marcado pela rápida evolução tecnológica e pela elevada demanda por profissionais qualificados, formações de menor duração tendem a ser percebidas como alternativas mais rápidas de inserção no mercado de trabalho.

Além disso, a ampliação da oferta desses cursos, principalmente na modalidade a distância e em instituições privadas, contribuiu significativamente para sua disseminação no sistema educacional brasileiro. Esse movimento acompanha tendências já identificadas na literatura sobre educação superior tecnológica, que apontam a consolidação dos cursos tecnólogos como estratégia de ampliação do acesso ao ensino superior e de alinhamento entre formação acadêmica e demandas do mercado.

Outro aspecto relevante refere-se à diferença entre os graus acadêmicos. Ao segmentar os dados entre cursos tecnólogos e cursos de bacharelado, em resposta a Q5, observa-se que os tecnólogos apresentam taxa de conclusão de 14,8%, enquanto os bacharelados registram 9,3%. Essa diferença pode estar relacionada às características estruturais dessas formações. Por possuírem menor duração e currículos mais orientados à prática, os cursos tecnólogos reduzem o tempo de exposição a fatores associados à evasão ao longo da trajetória acadêmica.

Por outro lado, os cursos de bacharelado costumam apresentar maior carga horária e duração mais extensa, ampliando o período em que o estudante está sujeito a desafios acadêmicos, financeiros e profissionais que podem comprometer a permanência até a diplomação. Além disso, a orientação mais aplicada dos tecnólogos pode fortalecer a percepção de utilidade imediata da formação, favorecendo o engajamento discente.

Nesse contexto, a maior taxa de conclusão observada nos cursos tecnológicos pode refletir não apenas diferenças estruturais de duração, mas também uma maior aderência entre o modelo formativo e as expectativas profissionais dos estudantes na área de Tecnologia da Informação.

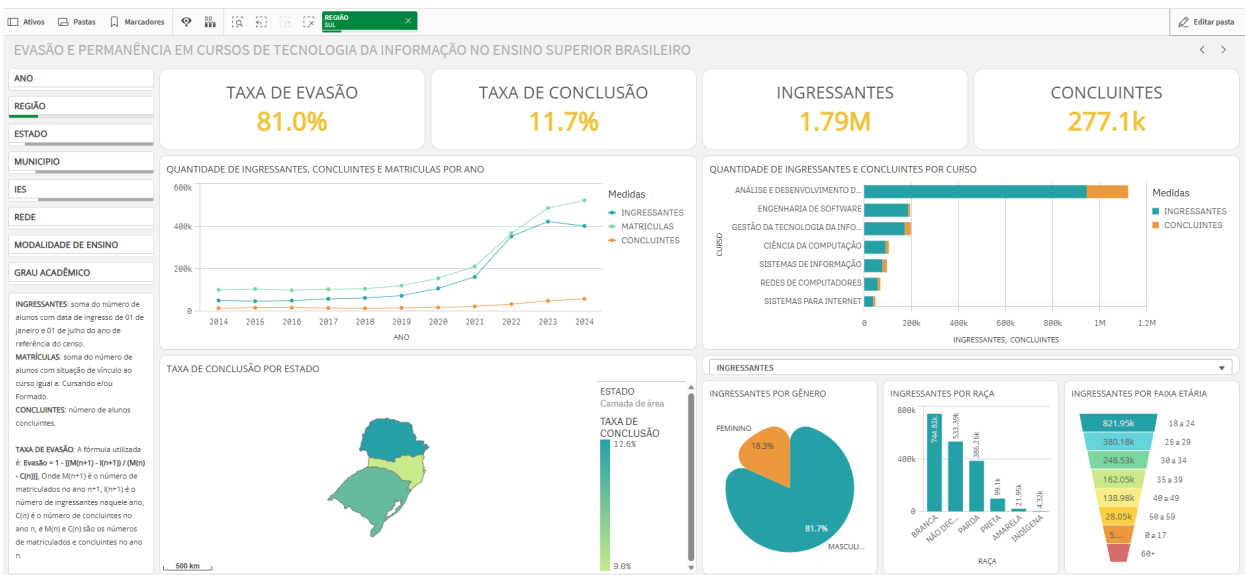
4.3 CONTRASTES REGIONAIS E DESAFIOS DE RETENÇÃO TERRITORIAL

Além das diferenças entre cursos e grau acadêmico, a análise dos dados também permite observar variações significativas quando considerado o recorte territorial.

A análise das taxas de conclusão por unidade da federação evidencia diferenças importantes entre os estados, em resposta à Q6. Os maiores percentuais de conclusão foram observados em São Paulo (15,1%), Pará (14,7%) e Acre (14,1%), indicando um desempenho relativamente mais favorável desses estados quanto à permanência e à conclusão dos estudantes nos cursos de graduação analisados.

No que se refere à Q7, a análise regional revela que o fenômeno da evasão apresenta distribuição heterogênea no território nacional. Entre as regiões brasileiras, destaca-se a Região Sul, que registrou a maior taxa de evasão observada no estudo, atingindo 81%, conforme ilustrado na Figura 19, que apresenta o painel analítico com o filtro aplicado para essa região.

Figura 19 – Painel com filtro aplicado da região Sul.



Fonte: Elaborado pelo autor (2026).

Esse resultado sugere a presença de desafios específicos relacionados à permanência discente, mesmo em um contexto regional caracterizado por dinamismo econômico e pela presença consolidada de polos tecnológicos. Uma interpretação para esse cenário está associada à forte inserção do setor de tecnologia no mercado de trabalho local. Em regiões com maior oferta de oportunidades profissionais, estudantes podem ser absorvidos precocemente pelo mercado antes da conclusão formal da graduação.

Essa dinâmica pode ser compreendida à luz da literatura sobre o mercado de trabalho em Tecnologia da Informação. Estudos indicam que o setor apresenta elevada demanda por profissionais, o que favorece a inserção precoce de estudantes ainda durante a graduação (Silveira; Tonini, 2023). Além disso, a literatura aponta que a conciliação entre trabalho e estudo constitui um dos principais fatores associados à evasão no ensino superior (Oliveira Junior, 2022).

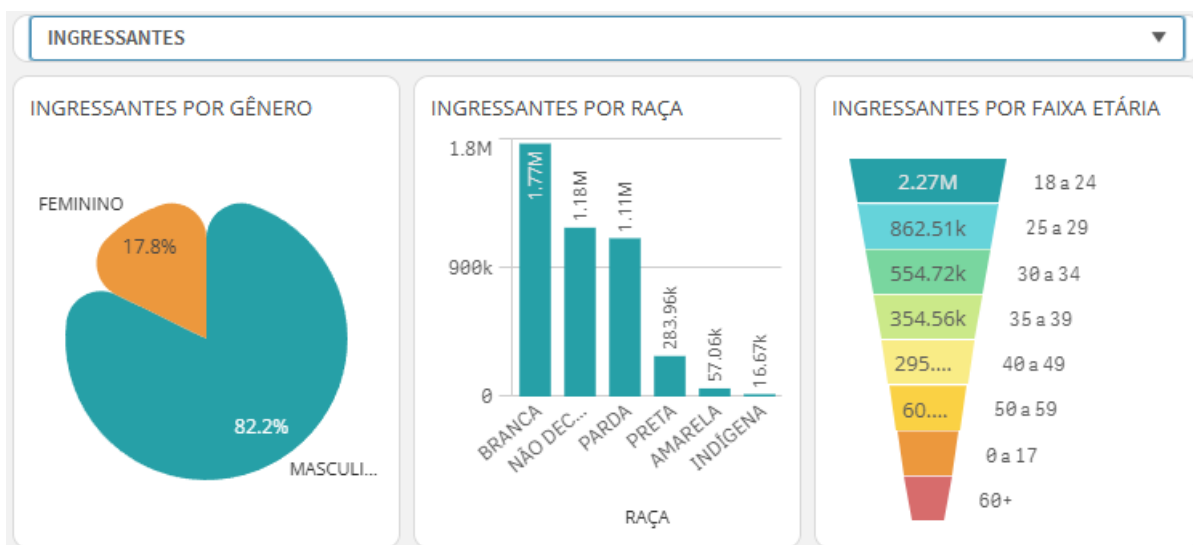
Nesse sentido, a literatura aponta para a coexistência de vetores de expulsão e vetores de atração no processo de evasão acadêmica (Correia; Mendonça; Silva, 2024). Enquanto dificuldades pedagógicas e desafios curriculares podem contribuir para o desengajamento discente, as oportunidades de emprego no setor tecnológico atuam como fatores de atração, estimulando a priorização da inserção profissional em detrimento da conclusão da graduação.

Apesar das diferenças regionais observadas, as taxas de conclusão permanecem relativamente baixas em todo o país, variando entre 8,4% e 15,1%. Esse resultado indica que a baixa diplomação não se restringe a uma região específica, configurando-se como um desafio estrutural do ensino superior na área de Tecnologia da Informação no Brasil.

4.4 PERFIL SOCIODEMOGRÁFICO E DIVERSIDADE NO CAMPO DE TI

Complementarmente à análise territorial, a análise sociodemográfica permite compreender a composição do público que ingressa nos cursos de Tecnologia da Informação e identificar possíveis desigualdades de acesso e permanência. Respondendo à Q8, observa-se forte predominância masculina entre os ingressantes, representando 82,2% do total, enquanto a participação feminina corresponde a 17,8%, conforme apresentado na Figura 20.

Figura 20 – Gráficos de quantidade de ingressantes por gênero, raça e faixa etária.



Fonte: Elaborado pelo autor (2026).

Esse padrão evidencia um histórico desafio de diversidade de gênero na área de tecnologia, amplamente discutido na literatura acadêmica. A baixa participação feminina pode estar associada a fatores socioculturais, estereótipos de gênero relacionados às áreas STEM (Ciência, Tecnologia, Engenharia e Matemática) e à menor exposição de meninas a estímulos educacionais voltados para tecnologia desde as etapas iniciais da formação.

Conforme discutido por Cheryan *et al.* (2017), ambientes acadêmicos fortemente associados a estereótipos masculinos podem afetar negativamente o sentimento de pertencimento de estudantes mulheres, influenciando sua permanência e progressão acadêmica.

No que diz respeito ao recorte racial, observa-se maior concentração de estudantes autodeclarados brancos (1,77 milhão) e pardos (1,11 milhão). Entretanto, destaca-se também a presença significativa de registros sem declaração (1,18 milhão), o que limita análises mais precisas sobre inclusão racial e desigualdades de acesso.

Enquanto na análise de faixa etária, predomina o grupo entre 18 e 24 anos, totalizando 2,27 milhões de ingressantes, o que indica que a maioria dos estudantes ingressa no ensino superior logo após a conclusão do ensino médio. Esse perfil reforça a caracterização do público como majoritariamente jovem, em fase inicial de formação acadêmica e profissional.

4.5 IMPACTO DAS MODALIDADES E REDES DE ENSINO NA EFETIVIDADE FORMATIVA

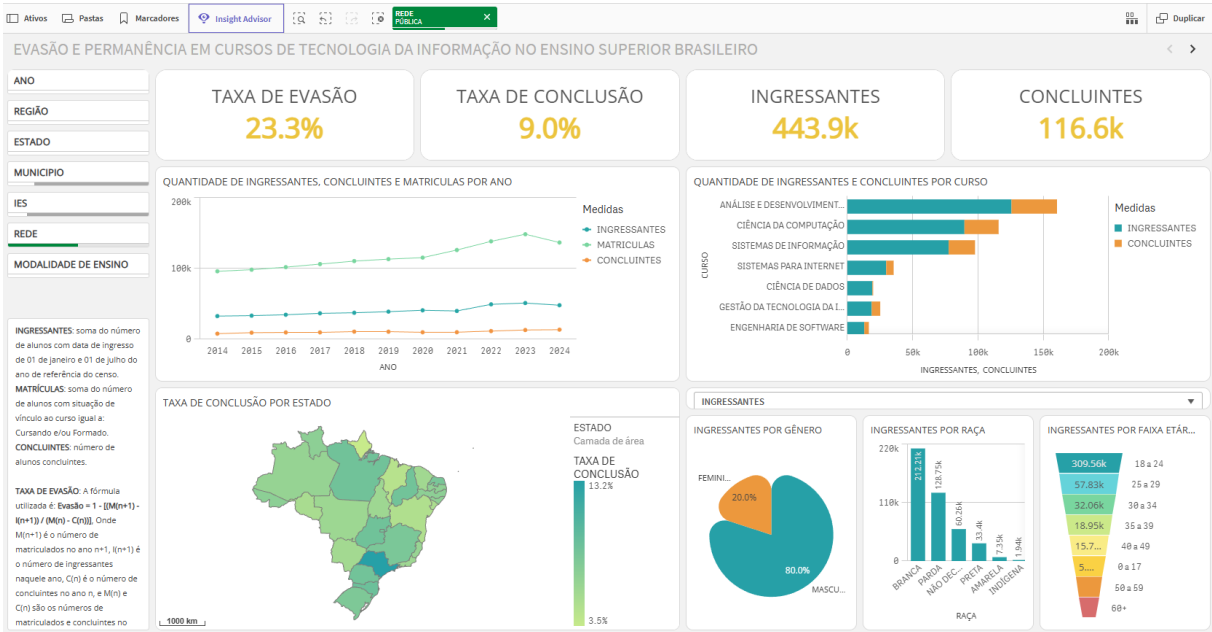
Por fim, além das características individuais e regionais, no que diz respeito a fatores institucionais, a análise comparativa entre modalidades de ensino revela diferenças significativas no comportamento dos indicadores acadêmicos. A modalidade ensino a distância apresentou taxa de evasão de 77,9%, enquanto o ensino presencial registrou 32%, respondendo à Q10.

Apesar dessa diferença expressiva, as taxas de conclusão mostraram-se relativamente próximas entre as duas modalidades, indicando que a menor evasão observada no ensino presencial não se traduz necessariamente em maior efetividade formativa. Esse resultado sugere a presença de um fenômeno de retenção acadêmica, caracterizado pela permanência prolongada do estudante com vínculo ativo, mas sem progressão efetiva no percurso curricular.

Nesse contexto, parte dos estudantes permanece matriculada por períodos superiores ao tempo previsto de integralização, configurando trajetórias acadêmicas mais longas e com menor eficiência na conclusão.

Esse padrão torna-se ainda mais evidente quando analisada a rede de ensino. Em resposta à Q9, conforme apresentado na Figura 21, a rede pública apresentou a menor taxa de evasão do estudo (23,3%), mas também a menor taxa de conclusão (9%).

Figura 21 – Painel analítico com filtro aplicado para a rede pública de ensino.



Fonte: Elaborado pelo autor (2026).

Esse descompasso indica que, embora haja maior persistência institucional, existem barreiras estruturais que dificultam a conclusão dos cursos.

Entre os fatores que podem contribuir para esse cenário destacam-se:

(i) limitações socioeconômicas, que levam estudantes a conciliar trabalho e estudos;

(ii) dificuldades pedagógicas, frequentemente associadas à reprovação em disciplinas estruturantes da área, como programação e cálculo;

(iii) rigidez organizacional e curricular, com estruturas acadêmicas pouco flexíveis para estudantes trabalhadores.

Esses elementos contribuem para o prolongamento da trajetória acadêmica e para o acúmulo de matrículas ativas sem conclusão, caracterizando um quadro de permanência institucional com baixa efetividade formativa. Esse descompasso pode ser interpretado à luz dos fatores estruturais apontados por Correia, Mendonça e Silva, (2024), que destacam a influência de desigualdades socioeconômicas e lacunas na formação básica como determinantes importantes da permanência e do sucesso acadêmico.

5 CONSIDERAÇÕES FINAIS

A evasão no ensino superior brasileiro representa um desafio persistente para a efetividade dos processos formativos, especialmente em áreas estratégicas como a Tecnologia da Informação, que apresentam simultaneamente elevada demanda no mercado de trabalho e dificuldades relacionadas à efetiva formação acadêmica. Nesse contexto, este trabalho teve como objetivo analisar os padrões de evasão e permanência em cursos de Tecnologia da Informação no ensino superior brasileiro, no período de 2014 a 2024, com base nos microdados do Censo da Educação Superior (Inep). A análise foi realizada com o apoio de ferramentas de BI, que permitiram estruturar e explorar os dados de forma sistemática, possibilitando a identificação de padrões e tendências ao longo do período analisado.

De forma geral, os resultados evidenciam um cenário caracterizado por elevados índices de evasão e por uma baixa taxa de conclusão nos cursos analisados. A taxa de evasão observada no período alcança 61,4%, enquanto a taxa de conclusão corresponde a 12,6%. Números que apontam para baixa efetividade formativa e para a ocorrência de trajetórias acadêmicas prolongadas.

Os resultados indicaram ainda que o crescimento expressivo das matrículas a partir de 2021, associado ao contexto pós-pandêmico e, possivelmente, à expansão do ensino a distância, ainda não se refletiu em aumento proporcional no número de concluintes, indicando um possível descompasso entre a expansão do acesso ao ensino superior e a efetividade na conclusão dos cursos, aspecto que deverá ser acompanhado em análises futuras à medida que novos dados dos censos educacionais forem disponibilizados.

A análise demonstrou diferenças significativas nas taxas de conclusão entre os graus acadêmicos, indicando que cursos tecnológicos apresentam maior taxa de conclusão quando comparados aos cursos de bacharelado, possivelmente em função da menor duração e da maior orientação prática desses programas de formação, relacionado à dinâmica do mercado de trabalho em tecnologia, caracterizado por elevada demanda por profissionais qualificados.

A análise também revelou dinâmicas distintas de insucesso acadêmico. Enquanto determinadas modalidades apresentam maior abandono, a rede pública demonstra um padrão de retenção prolongada, caracterizado pela permanência do estudante com vínculo ativo, porém com progressão curricular limitada. Esse padrão sugere que, além da evasão, a retenção prolongada também representa um desafio importante para a efetividade da formação superior.

No que se refere às características sociodemográficas, a análise indicou a predominância masculina entre os estudantes ingressantes na área de Tecnologia da Informação. Esse resultado está em consonância com a literatura que aponta para a persistência de desigualdades de gênero nas áreas relacionadas à computação e à tecnologia, evidenciando que a ampliação da diversidade nesse campo ainda representa um desafio relevante para o ensino superior.

A aplicação de conceitos de BI e modelagem dimensional na organização e exploração dos dados foi um elemento central para o desenvolvimento deste estudo. A estruturação dos microdados em um modelo analítico e a construção de um dashboard interativo possibilitaram uma visualização mais clara dos padrões identificados, facilitando a interpretação dos dados e ampliando as possibilidades de exploração das informações educacionais.

Como contribuição, o estudo disponibiliza uma abordagem analítica orientada por dados, capaz de apoiar gestores e formuladores de políticas públicas na identificação de gargalos temporais, regionais e sociodemográficos. As limitações concentram-se na dependência de dados secundários, os quais apresentam natureza agregada. Dessa forma, não é possível acompanhar individualmente a trajetória acadêmica dos estudantes ao longo do tempo. Como consequência, a evasão não é explicitamente identificada como uma variável direta na base de dados, sendo estimada a partir de indicadores derivados das informações disponíveis. Essa característica impõe limitações à precisão do cálculo da evasão, uma vez que não é possível distinguir, de forma direta, diferentes tipos de trajetória acadêmica, como transferência entre cursos ou instituições.

Adicionalmente, algumas variáveis presentes nos microdados são baseadas em informações autodeclaradas pelos estudantes ou pelas instituições, o que pode introduzir inconsistências ou lacunas de preenchimento. Esse aspecto deve ser considerado na interpretação dos resultados, especialmente em análises relacionadas a variáveis sociodemográficas.

Para pesquisas futuras, sugere-se a implementação de um pipeline de dados que possibilite a atualização automática das informações, à medida que novos microdados forem disponibilizados pelo Inep, garantindo a continuidade e a escalabilidade das análises. A partir dessa base estruturada, recomenda-se a aplicação de métodos inferenciais e preditivos, incluindo técnicas de aprendizado de máquina, com o objetivo de estimar o risco de evasão e identificar fatores associados ao abandono acadêmico, contribuindo para o desenvolvimento de estratégias de intervenção mais precisas e preventivas no contexto do ensino superior.

Ademais, propõe-se a disponibilização do painel analítico para o Instituto Federal de Alagoas e demais IES, de modo a apoiar a tomada de decisão baseada em dados, bem como a ampla divulgação dos resultados da pesquisa, através de publicações, visando ampliar seu alcance e fomentar o debate acadêmico e institucional sobre a evasão no ensino superior.

Em síntese, os resultados deste estudo contribuem para ampliar a compreensão sobre os padrões de evasão nos cursos de Tecnologia da Informação no Brasil, evidenciando tendências estruturais relacionadas à permanência, conclusão e retenção acadêmica ao longo da última década. Além disso, o trabalho demonstra o potencial da utilização de dados educacionais e ferramentas analíticas no apoio à tomada de decisão baseada em evidências no contexto do ensino superior.

6 DISPONIBILIDADE DE DADOS E CÓDIGO

Os microdados utilizados neste estudo são de natureza pública e podem ser obtidos no portal oficial do Inep. Para fins de reprodutibilidade e transparência metodológica, os scripts de ETL desenvolvidos em linguagem Python, bem como o carregamento da modelagem dimensional desenvolvida no ambiente Qlik, foram disponibilizados em repositório público no GitHub, acessível pelo endereço: <https://github.com/irismayara/evasao-ti-bi>.

REFERÊNCIAS

- BRASIL. Lei nº 12.527, de 18 de novembro de 2011. **Regula o acesso a informações previsto na Constituição Federal**. Brasília, DF: Presidência da República, 2011. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/112527.htm. Acesso em: 30 mar. 2026.
- BRASSCOM. **Relatório de Inteligência e Talento Digital**. São Paulo: Brasscom, 2021. Disponível em: <https://brasscom.org.br/relatorio-setorial-2021/>. Acesso em: 17 fev. 2026.
- CHECKLAND, Peter; POULTER, John. **Learning for action: a short definitive account of soft systems methodology**. 2. ed. Chichester: John Wiley & Sons, 2020.
- CHERYAN, Sapna; *et al.* Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, v. 143, n. 1, p. 1–35, 2017.
- CORONEL, Carlos; MORRIS, Steven. **Database Systems: Design, Implementation, & Management**. 13. ed. Boston: Cengage Learning, 2020.
- DATABRICKS. **Visão geral do esquema Snowflake**. [S. l.]: Databricks, 2024. Disponível em: <https://www.databricks.com/br/glossary/snowflake-schema>. Acesso em: 26 fev. 2026.
- DATABRICKS. **Visão geral do esquema em estrela (star schema)**. [S. l.]: Databricks, 2024. Disponível em: <https://www.databricks.com/br/glossary/star-schema>. Acesso em: 26 fev. 2026.
- DAVENPORT, Thomas H.; HARRIS, Jeanne G. **Competing on analytics: the new science of winning**. 2. ed. Boston: Harvard Business Review Press, 2017.
- INEP. **Microdados do Censo da Educação Superior 2014–2024**. Brasília: Inep, 2025. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>. Acesso em: 04 nov. 2025.
- INEP. **Resumo Técnico do Censo da Educação Superior 2022**. Brasília: Inep, 2022. Disponível em: <https://www.gov.br/inep/pt-br/centrais-de-conteudo/acervo-linha-editorial/publicacoes-institucionais/estatisticas-e-indicadores-educacionais/resumo-tecnico-do-censo-da-educacao-superior-2022>. Acesso em: 04 nov. 2025.
- INMON, William H. **Building the data warehouse**. 4. ed. Indianapolis: John Wiley & Sons, 2005.
- KIMBALL, Ralph; ROSS, Margy. **The data warehouse toolkit: the definitive guide to dimensional modeling**. 3. ed. Indianapolis: John Wiley & Sons, 2013.
- KITCHIN, Rob. **The data revolution: big data, open data, data infrastructures and their consequences**. 2. ed. London: SAGE Publications, 2022.
- MARR, Bernard. **Data strategy: how to profit from a world of big data, analytics and the Internet of Things**. London: Kogan Page, 2017.

MCKINNEY, Wes. **Python for data analysis: data wrangling with pandas, NumPy, and IPython**. 3. ed. Sebastopol: O'Reilly Media, 2022.

NONAKA, Ikujiro; TAKEUCHI, Hirotaka. **The wise company: how companies create continuous innovation**. Oxford: Oxford University Press, 2019.

OCDE. **Education at a Glance 2021: OECD Indicators**. Paris: OECD Publishing, 2021. Disponível em:

https://www.oecd.org/en/publications/education-at-a-glance-2021_b35a14e5-en.html. Acesso em: 27 jan. 2026.

PROVOST, Foster; FAWCETT, Tom. **Data science for business: what you need to know about data mining and data-analytic thinking**. Sebastopol: O'Reilly Media, 2013.

QLIK. **Why Qlik?** Radnor, PA: Qlik, 2023. Disponível em:

<https://www.qlik.com/us/why-qlik>. Acesso em: 30 mar. 2026

REDMAN, Thomas C. **Getting in front on data quality**. Massachusetts: Technics Publications, 2016.

SEMESP. **Mapa do Ensino Superior no Brasil 2023**. 13. ed. São Paulo: Semesp, 2023.

Disponível em: <https://www.semesp.org.br/mapa/edicao-13/>. Acesso em: 27 jan. 2026.

SEMESP. **Mapa do Ensino Superior no Brasil 2024**. São Paulo: Semesp, 2024. Disponível

em: <https://www.semesp.org.br/mapa/edicao-14/>. Acesso em: 27 jan. 2026.

SHARDA, Ramesh; DELEN, Dursun; TURBAN, Efraim. **Business Intelligence, Analytics e Data Science: uma perspectiva gerencial**. 4. ed. Porto Alegre: Bookman, 2021.

SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN, S. **Sistema de Banco de Dados**. 7. ed. Rio de Janeiro: Elsevier, 2020.

SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN, S. **Database System Concepts**. 7. ed. New York: McGraw-Hill Education, 2020.

SILVA FILHO, Roberto L. L.; LOBO, Maria Beatriz C. M. **Esclarecimentos metodológicos sobre os cálculos de evasão**. São Paulo: Instituto Lobo, 2012. Disponível em:

https://www.institutolobo.org.br/core/uploads/artigos/art_078.pdf. Acesso em: 27 jan. 2026.

SBC. **Educação superior em computação: estatísticas 2022**. Porto Alegre: SBC, 2022.

Disponível em:

<https://www.sbc.org.br/wp-content/uploads/2024/08/ESTATISTICAS-INEP-2022-compactado.pdf>. Acesso em: 27 jan. 2026.

SBC. **Grandes desafios da educação em computação no Brasil 2025–2035**. Porto Alegre:

SBC, 2024. Disponível em: <https://books-sol.sbc.org.br/index.php/sbc/catalog/book/174> .

Acesso em: 27 jan. 2026.

STAIR, Ralph M.; REYNOLDS, George W. **Princípios de sistemas de informação**. 13. ed. São Paulo: Cengage Learning, 2018.

TAVERA ROMERO, Carlos Andrés; ORTIZ, José Hernández; KHALAF, Osamah Ibrahim; RÍOS PRADO, Alejandro. **Business Intelligence**: Business Evolution after Industry 4.0. *Sustainability*, v. 13, n. 18, p. 10026, 2021.

TINTO, Vincent. Reflections on Student Persistence. *Student Success*, v. 8, n. 2, p. 1–8, 2017.

TINTO, Vincent. Dropout from higher education: a theoretical synthesis of recent research. *Review of Educational Research*, v. 45, n. 1, p. 89–125, 1975.

TURBAN, Efraim; POLLARD, Carol; WOOD, Gregory. **Information Technology for Management**: Driving Digital Transformation to Increase Local and Global Performance, Growth and Sustainability. 12. ed. Hoboken: John Wiley & Sons, 2021.