



**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS PEDRO II
TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS**

JOÃO VITOR PEREIRA DA SILVA

**ANÁLISE EXPLORATÓRIA DOS MICRODADOS DO CENSO DA EDUCAÇÃO
SUPERIOR: Ênfase em cursos de computação**

**PEDRO II, PIAUÍ
2025**

JOÃO VITOR PEREIRA DA SILVA

ANÁLISE EXPLORATÓRIA DOS MICRODADOS DO CENSO DA EDUCAÇÃO
SUPERIOR: Ênfase em cursos de computação

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Pedro II.

Orientador: Prof. Dr. Manuel Goncalves da Silva Neto.

PEDRO II, PIAUÍ

2025

ANÁLISE EXPLORATÓRIA DOS MICRODADOS DO CENSO DA EDUCAÇÃO SUPERIOR: Ênfase em cursos de computação

João Vitor Pereira da Silva¹

RESUMO

A educação em Computação tem ganhado destaque nas últimas décadas, impulsionada pelo crescimento da demanda por profissionais qualificados e pelo reconhecimento da área como uma das principais disciplinas científicas da atualidade. Apesar destes avanços, existem desafios relacionados à qualificação na área de computação, como os elevados índices de evasão nestes cursos. Nesse contexto, a Mineração de Dados Educacionais possibilita a análise de grandes volumes de informações estudantis para embasar decisões estratégicas e prover arcabouços para a tomada de decisões e otimizar processos educacionais. A qualidade dos resultados de pesquisas na área de mineração de dados está relacionada à confiabilidade da base de dados empregada, a qual pode apresentar entradas inconsistentes ou ausentes em decorrência de erros na aquisição das informações. Diante desse cenário, este trabalho realizou a sanitização e análise exploratória dos microdados do Censo da Educação Superior do INEP, com foco nos cursos da área de Computação no Brasil, os quais são bem estabelecidos para uso em mineração de dados. A partir da análise exploratória, foi possível identificar inconsistências na base oficial. No decorrer deste trabalho tais inconsistências foram corrigidas por meio da aplicação de técnicas de pré-processamento e sanitização de dados. Como contribuição, este trabalho disponibiliza duas versões da base de dados sanitizadas e documentadas, visando apoiar estudos em classificação e mineração de dados educacionais. Além disso, acredita-se que a metodologia aplicada possa servir de referência para a sanitização de bases em outros contextos, contribuindo para a melhoria da qualidade dos dados utilizados em pesquisas.

Palavras-chave: Censo da Educação Superior; Sanitização de Dados; Mineração de Dados Educacionais; Análise de Dados.

ABSTRACT

Computer science education has gained prominence in recent decades, driven by the growing demand for qualified professionals and the field's recognition as one of today's leading scientific disciplines. Despite these advances, there are challenges related to qualifications in computing, such as high dropout rates in these courses. In this context, Educational Data Mining enables the analysis of large volumes of student data to support strategic decisions, provide frameworks for decision-making, and optimize educational processes. The quality of data mining research results is related to the reliability of the database used, which may contain inconsistent or missing entries due to errors in data acquisition. Given this scenario, this study

¹ Graduando em Análise e Desenvolvimento de Sistemas. IFPI. vitpereira27@gmail.com

performed the sanitization and exploratory analysis of microdata from the INEP Higher Education Census, focusing on Computing programs in Brazil, which are well-established for use in data mining. Through this exploratory analysis, it was possible to identify inconsistencies in the official database. During this work, these inconsistencies were corrected through the application of data preprocessing and sanitization techniques. As a contribution, this work provides two sanitized and documented versions of the database, aiming to support studies in classification and mining of educational data. Furthermore, it is believed that the applied methodology can serve as a reference for database sanitization in other contexts, contributing to improving the quality of data used in research.

Keywords: Higher Education Census; Data Sanitization; Educational Data Mining; Data Analysis.

Data de aprovação:24/07/2025.

1 INTRODUÇÃO

A educação pode ser definida como o ato de ensinar e promover o desenvolvimento de conhecimentos, habilidades e capacidades intelectuais e morais, visando favorecer a integração pessoal e social dos indivíduos na sociedade (Costa Júnior, 2023). Sendo garantida por lei, conforme estabelecido no artigo 205 da Constituição Federal de 1988:

A educação, direito de todos e dever do Estado e da família, será promovida e incentivada com a colaboração da sociedade, visando ao pleno desenvolvimento da pessoa, seu preparo para o exercício da cidadania e sua qualificação para o trabalho(Brasil, 1988).

Segundo o artigo 21 da Lei 9.394/96 (Brasil, 1996), denominada Lei de Diretrizes e Bases da Educação (LDB), a educação escolar é dividida em dois níveis principais:

1. Educação básica;
2. Educação superior.

O ensino superior tem como finalidades principais a formação de profissionais qualificados, a promoção da pesquisa científica e a difusão do conhecimento, buscando, por meio dessas atividades, integrar o saber adquirido às necessidades da sociedade, contribuindo tanto para o desenvolvimento contínuo dos indivíduos quanto para a resolução de problemas de ordem nacional e regional (Brasil, 2008).

Indivíduos com formação em nível superior geram conhecimento científico e promovem o aumento do capital humano, impulsionando a produtividade e o crescimento econômico (Li *et al.*, 2024). Entretanto, esse nível enfrenta desafios que impactam o desempenho e o bem-estar dos estudantes (Pante-Salas *et al.*, 2025). Entre os desafios enfrentados pelo ensino superior estão a sobrecarga acadêmica, a má gestão do tempo, a adaptação a novas metodologias e conteúdos complexos, além de questões pessoais como estresse, ansiedade e baixa motivação, especialmente na transição do ensino médio para a universidade (Fook *et al.*, 2022).

Nas últimas décadas, a área da computação vem apresentando crescimento acelerado, sendo considerada uma disciplina com influência direta sobre o trabalho, a vida cotidiana e o avanço de novas tecnologias (Chen, 2023; Yaseen *et al.*, 2022). E que contribui para a inovação, abrangendo campos como inteligência artificial, ciência de dados, computação quântica e cibersegurança, com aplicações em setores como saúde, finanças e transporte (Gupta, 2023; RAJKUMAR *et al.*, 2024; Yaseen *et al.*, 2022). A educação em computação vem ganhando destaque, com o aumento da procura por cursos na área e sendo considerada uma das principais disciplinas científicas da atualidade (Nager; Atkinson 2016). Esse crescimento evidencia a importância da formação na área, para preparar profissionais qualificados diante das transformações tecnológicas e da economia digital (Sanati-Mehrziy; Minaie, 2022). Entretanto, essa formação de profissionais enfrenta problemas como altos índices de evasão (Alvim; Bittencourt; Duran., 2024). A literatura apresenta tentativas de identificar os casos prováveis de evasão, na tentativa de mitigar ou criar soluções para atuar de forma precoce nestes casos (Brito; Mello; Alves, 2020; Yaacob *et al.*, 2020; Teodoro; Kappel, 2020).

Diante desses problemas, a mineração de dados acadêmicos (Educational Data Mining - EDM) possibilita a análise de grandes volumes de informações estudantis para embasar decisões estratégicas e promover a melhoria de processos educacionais (Ventura; Romero, 2020). Dessa forma, instituições de ensino conseguem encontrar padrões e fatores que contribuem ou não para o sucesso acadêmico, bem como mapear perfis de risco e identificar estudantes propensos à evasão ou baixo rendimento, viabilizando assim intervenções precoces e orientações direcionadas, permitindo a adoção de estratégias personalizadas e a alocação eficiente de recursos (Yagci, 2022).

No entanto, a presença de dados bem estruturados, sem erros e reutilizáveis

torna-se necessária, pois a qualidade dos resultados está relacionada à confiabilidade da base de dados, que pode ser comprometida por entradas inconsistentes ou ausentes em decorrência de erros no preenchimento pelas instituições (Gibert; Izquierdo; Sànchez-Marrè, 2016). Nesse contexto, são necessárias etapas de pré-processamento e sanitização dos dados para assegurar a qualidade e a utilidade das análises subsequentes (Kang; Tian, 2018). Esses processos têm como finalidade remover registros irrelevantes, corrigir valores faltantes ou inconsistentes, eliminar duplicatas e identificar valores atípicos (Outliers), garantindo que apenas dados válidos e coerentes sejam considerados, aumentando a confiabilidade das estatísticas geradas e dos modelos produzidos (Maharana; Mondal; Nemade, 2022).

Neste cenário, o Censo da Educação Superior do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) é uma base de dados pública que serve como fonte de informação sobre o ensino superior no Brasil (Brasil, 2024). Trata-se de um levantamento estatístico sobre Instituições de Ensino Superior (IES) brasileiras, reunindo dados sobre instituições, cursos, matrículas e perfis dos estudantes (Brasil, 2024).

1.1 OBJETIVOS

Este trabalho objetivou a análise exploratória e sanitização da base de microdados do Censo da Educação Superior, com ênfase nos cursos da área de Computação no Brasil, a fim de disponibilizar uma versão documentada e estruturada da mesma, adequada para aplicações em mineração e ciência de dados no contexto educacional.

A base resultante contempla informações sobre os cursos de Computação no Brasil entre os anos de 2021 e 2023.

1.2 OBJETIVOS ESPECÍFICOS

1. Levantar e compreender a estrutura da base de dados do Censo da Educação Superior, identificando atributos relevantes relacionados aos cursos de Computação;
2. Detectar e documentar inconsistências, valores anômalos e ausências nos dados, como categorias inválidas, valores fora do domínio e campos nulos;

3. Executar processos de sanitização dos dados, padronizando formatos, removendo ou corrigindo registros inconsistentes e tratando valores faltantes;
4. Gerar a versão sanitizada da base de dados;
5. Discutir as limitações encontradas na base original, apontando implicações para pesquisas futuras.

Este trabalho está organizado da seguinte maneira: a Seção 2 apresenta os conceitos teóricos essenciais para o entendimento da pesquisa. A Seção 3 descreve a metodologia empregada, detalhando os procedimentos de coleta, tratamento e análise dos dados. Na Seção 4, são expostos os resultados obtidos. Na Seção 5, são discutidas as implicações e contribuições do trabalho para estudos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A presente seção possui como objetivo apresentar os principais conceitos teóricos que sustentam a investigação.

2.1 A ESTRUTURA DO ENSINO SUPERIOR NO BRASIL

A Lei de Diretrizes e Bases (LDB) estabelece que “a educação superior será ministrada em instituições de ensino superior, públicas ou privadas, com variados graus de abrangência ou especialização” (Brasil, 1996). A avaliação desse nível educacional é de responsabilidade do INEP, principal órgão vinculado ao Ministério da Educação (Brasil, 2025).

No intuito de promover melhorias na qualidade dos cursos de graduação e das instituições de ensino superior (IES), foi instituído, em 2004, o Sistema Nacional de Avaliação da Educação Superior (Sinaes), que integra a avaliação institucional, de cursos e de desempenho dos estudantes (Brasil, 2004).

Além disso, o INEP elabora também o Censo da Educação Superior, um levantamento anual que compila estatísticas sobre cursos de IES no Brasil, servindo de base para a formulação de políticas públicas e indicadores de qualidade do ensino superior (Brasil, 2024).

No âmbito da avaliação institucional, cada IES mantém uma Comissão Própria de Avaliação (CPA) para coordenar a autoavaliação institucional, conforme

previsto no Sinaes (Brasil, 2004). Em paralelo, ocorrem avaliações externas periódicas pelo INEP/MEC em que comissões de especialistas verificam instalações e planos de ensino, resultando em indicadores como o Conceito de Curso.

2.2 CIÊNCIA DE DADOS

A Ciência de Dados é um campo multidisciplinar dedicado à extração de conhecimento a partir da análise de grandes conjuntos de dados (Cao, 2017). Ela combina estatística, aprendizado de máquina, ciência da computação e conhecimento de domínio para identificar padrões, construir modelos preditivos e apoiar a tomada de decisão (Blei; Smyth, 2017). Essa abordagem envolve todo o ciclo de vida dos dados, desde a coleta e preparação (incluindo limpeza e transformação) à análise exploratória, modelagem e interpretação dos resultados (Akerkar; Sajja, 2016).

2.2.1 Processos em Ciência de Dados

Os projetos de ciência de dados seguem modelos estruturados, em geral, definem-se fases como compreensão do problema, obtenção e preparo dos dados, modelagem, avaliação e implantação (Martinez; Viles; Olaizola, 2021). A fase de preparação dos dados costuma envolver tarefas como seleção de registros relevantes, limpeza (tratamento de erros, valores ausentes e duplicações), integração de fontes e formatação dos atributos para modelagem (Fernandes *et al.*, 2023).

2.2.2 Problemas Comuns em Bases de Dados

Dados do mundo real costumam vir com inconsistências ou incompletos, e problemas de qualidade podem comprometer a análise (Raina; Krishnamurthy, 2021). Entre os problemas mais comuns estão:

- **Valores ausentes ou inconsistentes:** campos em branco, dados no formato errado ou respostas “não informadas” dificultam a análise;
- **Registros duplicados:** entradas repetidas que contam a mesma informação várias vezes e podem distorcer estatísticas;

- **Formatação inconsistente:** falta de padronização como datas em formatos mistos, unidades de medida variadas, causa erros no processamento.

2.2.3 Pré-processamento de dados

O pré-processamento de dados é uma etapa na análise de dados cujo objetivo é preparar os dados brutos para torná-los apropriados à modelagem e à extração de conhecimento (Kang; Tian, 2018). Esse processo envolve a limpeza dos dados (remoção de erros, ruídos e tratamento de valores ausentes), a transformação (normalização, discretização e codificação de variáveis), a engenharia de atributos (criação ou seleção de características relevantes), a integração (unificação de dados oriundos de múltiplas fontes) e a redução de dados (diminuição da dimensionalidade ou volume, mantendo a integridade da informação) (Han; Kamber; Pei, 2012).

2.3 MINERAÇÃO DE DADOS EDUCACIONAIS

A Mineração de Dados Educacionais (EDM) refere-se à utilização de técnicas de mineração de dados no contexto educacional, configurando-se como um subcampo da ciência de dados voltado à identificação de padrões e fatores preditivos em dados relacionados ao ensino e à aprendizagem (Ventura; Romero, 2020). Entre seus objetivos estão entender como os alunos aprendem, identificar práticas pedagógicas eficazes e melhorar a qualidade do ensino por meio da personalização da aprendizagem (Alzobaidy, 2025).

As técnicas mais utilizadas na EDM incluem agrupamento, classificação, mineração de regras de associação, predição e mineração de processos, com aplicações como previsão de desempenho, identificação de alunos em risco e análise de trajetórias de aprendizagem (Dutt; Ismail; Herawan., 2017).

2.4 TRABALHOS RELACIONADOS

Alvim, Bittencourt e Duran (Alvim; Bittencourt; Duran., 2024) utilizaram microdados do Censo da Educação Superior entre 2015 e 2019 para analisar a evasão nos cursos de computação no Brasil, centrado nas diferenças entre

modalidades, turnos, uso de financiamento estudantil e outros fatores demográficos. Apesar de utilizarem a base do INEP, não há menção a etapas detalhadas de pré-processamento ou sobre a verificação de inconsistências na base original. No estudo de Teodoro e Kappel (Teodoro; Kappel, 2020), aplicaram-se cinco algoritmos de aprendizado de máquina, incluindo Random Forest e Redes Neurais, sobre dados do INEP, visando identificar os principais atributos relacionados à evasão em universidades públicas; entretanto, não se discute de forma aprofundada o processo de sanitização dos dados utilizados.

Em contraste com os trabalhos mencionados acima, este estudo concentra-se na etapa de análise exploratória e sanitização dos dados do Censo da Educação Superior do INEP, buscando ao fim desse processo entregar uma base consolidada, tratada e apta para análises futuras, buscando também fornecer uma metodologia replicável que possibilite a outros pesquisadores utilizarem a mesma abordagem em diferentes bases de dados.

3 METODOLOGIA

Este trabalho trata-se de uma pesquisa exploratória e experimental, adotando uma abordagem voltada ao pré-processamento e análise preliminar dos microdados do Censo da Educação Superior, com foco nos cursos da área de Computação. Esta base de dados foi escolhida por se tratar da principal fonte de dados sobre o ensino superior no Brasil, coletando informações de todas as instituições de educação superior (públicas e privadas) do país (Hoffmann; Nunes; Muller, 2019). Sendo uma pesquisa realizada anualmente pelo INEP, este levantamento consolida dados sobre cursos de graduação presenciais e a distância, número de vagas, matrículas, ingressos, concluintes, bem como dados dos docentes (formação, titulação, vínculo institucional etc.) (Brasil, 2024).

Os experimentos foram realizados em um computador com sistema operacional Windows, versão 11, processador Intel Core i7-13620H (10 Núcleos) e 32GB de RAM.

3.1 FERRAMENTAS DE ANÁLISE

Para a realização da análise de dados, adotou-se a linguagem de

programação Python ²(versão 3.10.11), em razão de seu amplo suporte a bibliotecas especializadas e sintaxe simples, além de ser atualmente uma das linguagens de programação mais utilizadas no mundo, especialmente em áreas como Ciência de Dados, computação científica, análise de dados e Aprendizado de Máquina (Castro *et al.*, 2023).

A gestão de dependências e a criação de ambientes isolados foram efetuadas por meio da distribuição Anaconda³. A utilização dessa ferramenta permite a instalação e atualização de bibliotecas com maior facilidade. Adicionalmente, o Anaconda Navigator providenciou uma interface gráfica para controle dos ambientes e pacotes, agilizando a configuração do projeto.

Os experimentos e as análises foram conduzidos utilizando o Jupyter Notebook⁴, que é uma plataforma interativa que combina código executável, documentação em Markdown e visualizações gráficas em um único documento, permitindo também a execução célula a célula e a inspeção imediata de resultados, o que permitiu registrar o fluxo de trabalho a cada ponto. Para a manipulação e tratamento dos dados, utilizou-se a biblioteca Pandas, que oferece estruturas de dados como o DataFrame, otimizadas para operações vetorizadas e de indexação eficiente.

3.2 ETAPAS DE ANÁLISE E PRÉ-PROCESSAMENTO

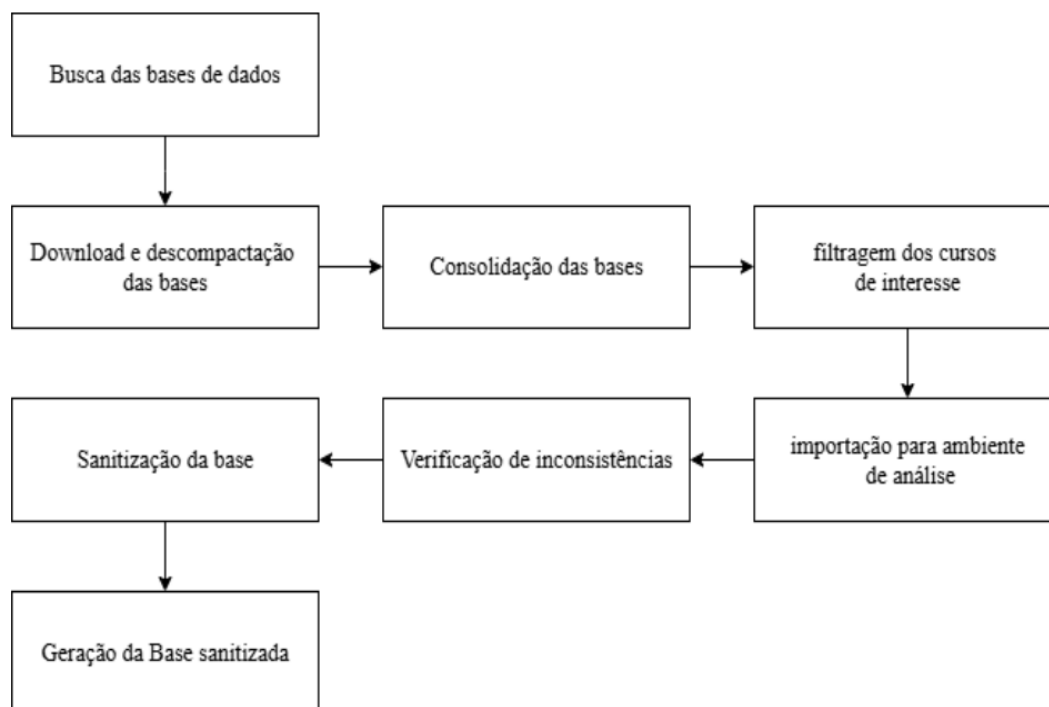
O processo de análise e pré-processamento dos dados consistiu em uma sequência de etapas para garantir a integridade, qualidade e usabilidade da base de dados. Essas ações buscaram preparar o conjunto de dados para futuras aplicações envolvendo classificação e mineração de dados estudantis. Um resumo das etapas é apresentado na figura 1.

² <https://www.python.org/>

³ <https://anaconda.org/>

⁴ <https://jupyter.org/>

Figura 1 – Etapas da metodologia



Fonte: Produzido pelo Autor (2025)

No site oficial do INEP⁵, os censos são disponibilizados anualmente. Esta pesquisa utilizou os dados dos anos de 2021 a 2023, sendo este último o mais recente disponível no momento da elaboração do trabalho. Inicialmente, realizou-se o download e a descompactação dos arquivos correspondentes a cada ano. Em seguida, efetuou-se a consolidação e padronização das tabelas provenientes de diferentes anos do Censo. Essa etapa envolveu a integração dos três arquivos correspondentes a cada ano, objetivando a formação de uma única base consolidada. Para isso, ocorreu a remoção de colunas inconsistentes entre as tabelas originais e a reorganização da estrutura das tabelas para manter a mesma ordem das colunas em todos os registros.

Na sequência, procedeu-se à filtragem dos cursos de interesse, etapa que restringiu o escopo da análise aos cursos pertencentes à área de computação e tecnologias da informação e comunicação. Posteriormente, houve a remoção de variáveis consideradas não essenciais para as possíveis análises subsequentes, como informações duplicadas de localização, cujos códigos já estavam presentes e eram mais adequados ao tratamento, identificação institucional e de classificação de

⁵ <https://www.gov.br/inep/pt-br>

cursos.

Com o conjunto consolidado e filtrado, procedeu-se à leitura dos dados no ambiente de análise, utilizando a biblioteca Pandas da linguagem Python. Para permitir a análise, demandou-se a conversão da base de dados em um DataFrame, uma estrutura de dados que permite operações tabulares com suporte à indexação, filtragem, agregações e manipulações estatísticas.

Após as etapas anteriores, para obter uma compreensão estatística inicial da base de dados, gerou-se um resumo estatístico dos atributos numéricos, fornecendo uma visão preliminar do comportamento dos dados, permitindo assim visualizar medidas como média, desvio padrão, valores mínimos e máximos e os principais percentis.

Na etapa seguinte, realizou-se a verificação de inconsistências nos dados, focando em três tipos principais de problemas:

- Códigos fora do intervalo esperado, como identificadores que excedem os limites definidos pela documentação do INEP;
- Presença de valores não numéricos em campos que deveriam ser numéricos;
- Valores ausentes.

Com base nas inconsistências identificadas, efetuou-se o tratamento dos dados inválidos, e então, procedeu-se à geração de uma versão da base de dados limpa, para eventuais pesquisas. Duas versões distintas foram salvas: a base original consolidada, contendo todos os dados sem alterações e uma com os dados devidamente tratados.

4 RESULTADOS

4.1 O DATASET

Os microdados do Censo da Educação Superior são disponibilizados no site do INEP em um arquivo compactado para cada ano correspondente, que possui diretórios organizados da seguinte forma:

- DADOS: contém os arquivos principais no formato CSV (Comma-Separated Values), que são compatíveis com softwares

estatísticos, planilhas eletrônicas e ferramentas de análise de dados. Os arquivos estão divididos em dois cadastros: um referente às Instituições de Ensino Superior (IES) e outro aos cursos de graduação;

- LEIA-ME: encontram-se instruções técnicas para leitura e extração dos dados;
- ANEXOS que inclui os arquivos: Anexo I: Dicionário de Dados (formato .xlsx), com a descrição detalhada de todas as variáveis e Anexo II: Questionários do Censo 2023 (formato .pdf).

Para os fins desta pesquisa, o principal arquivo utilizado foi o cadastro de cursos, localizado no diretório DADOS. A partir da análise desse arquivo em conjunto com seu respectivo dicionário de dados, observa-se que ele reúne informações detalhadas sobre os cursos de graduação ofertados pelas Instituições de Ensino Superior (IES) do Brasil. Cada registro representa um curso específico vinculado a uma unidade institucional, contendo variáveis que descrevem características institucionais, acadêmicas, geográficas, entre outras.

No que tange à quantidade de dados, a Tabela 1 mostra os valores indicados durante a análise preliminar:

Tabela 1 – Quantidade de dados por ano

Ano	Colunas	Registros
2021	200	444.786
2022	200	573.019
2023	202	671.610

Fonte: Produzido pelo autor (2025)

Observa-se, além do volume crescente de dados a cada ano, uma alteração na estrutura da base em 2023, com a adição de duas novas colunas em relação aos anos anteriores, que apresentavam uma estrutura padronizada. A análise do dicionário de dados revelou que essas novas variáveis correspondem a **IN_COMUNITARIA** e **IN_CONFESSIONAL**. A primeira indica se a instituição é comunitária, ou seja, de iniciativa privada, sem fins lucrativos e com atuação voltada para o interesse público. A segunda identifica se a instituição possui caráter confessional, estando vinculada ou mantida por uma organização religiosa.

4.2 TABELA CONSOLIDADA

Após a consolidação das tabelas, como descrito na metodologia deste trabalho, efetuou-se a padronização das variáveis e a filtragem dos registros referentes aos cursos da área de Computação. Como resultado, a base final passou a conter 185 (cento e oitenta e cinco) colunas e 167.046 (cento e sessenta e sete mil, quarenta e seis) registros; com isso, foi feita a importação da base para o ambiente de análise.

A Tabela a seguir apresenta as variáveis removidas durante o processo de preparação dos dados, acompanhadas das respectivas justificativas para sua exclusão.

Tabela 2 – Variáveis removidas durante o pré-processamento

Categoria	Variáveis Removidas	Justificativa
Dados de Classificação Institucional	IN_COMUNITARIA, IN_CONFESSIONAL	Existe apenas na tabela de 2023, inviabilizando a consolidação com os demais anos e não é essencial para os objetivos da pesquisa.
Dados Geográficos	NO_REGIAO, NO_UF, SG_UF, NO_MUNICIPIO, CO_MUNICIPIO	Redundância com variáveis com códigos numéricos
Identificação	CO_IES, CO_CURSO	Foco na análise macro, sem necessidade de identificação institucional
Classificação de cursos	CO_CINE_ROTULO, NO_CINE_ROTULO, CO_CINE_AREA_GERAL, NO_CINE_AREA_GERAL, CO_CINE_AREA_ESPECIFICA, NO_CINE_AREA_ESPECIFICA, CO_CINE_AREA_DETALHADA, NO_CINE_AREA_DETALHADA	Irrelevante para o estudo, repetindo informações sobre a área do curso

Fonte: Produzido pelo autor (2025)

4.3 VERIFICAÇÃO DE REGISTROS INVÁLIDOS

Com os passos anteriores concluídos, deu-se início à busca por possíveis anomalias no dataframe. A análise revelou que as variáveis **CO_REGIAO** e **IN_CAPITAL** possuíam um número de linhas inferior ao esperado, conforme ilustrado na Figura 2.

Figura 2 – Contagem de registros na base de dados

df.describe()					
	NU_ANO_CENSO	CO_REGIAO	IN_CAPITAL	TP_DIMENSAO	TP_ORGANIZACAO_ACADEMICA
count	167046.00	165123.00	165123.00	167046.00	167046.00
mean	2022.12	3.04	0.16	2.02	1.52
std	0.69	1.02	0.40	0.40	0.59
min	2021.00	0.00	0.00	1.00	1.00
25%	2022.00	3.00	0.00	2.00	1.00
50%	2022.00	3.00	0.00	2.00	1.00
75%	2023.00	4.00	0.00	2.00	2.00
max	2023.00	5.00	4.00	5.00	5.00

Fonte: Produzido pelo Autor (2025)

Nota-se, na figura apresentada, que a primeira linha, que corresponde à contagem de valores por coluna, enquanto o total de registros deveria ser 167.046 (cento e sessenta e sete mil, quarenta e seis), as variáveis **CO_REGIAO** e **IN_CAPITAL** contêm apenas 165.123 (cento e sessenta e cinco mil, cento e vinte e três) registros, indicando a ausência de 1.923 (mil novecentos e vinte e três) observações. Tal fato ocorre porque o método utilizado desconsidera valores nulos na contagem, o que evidencia a presença de ausências nessas colunas. Diante disso, procedeu-se para a verificação de quais campos possuíam valores nulos na base de dados.

4.3.1 Verificação de valores nulos

Em seguida, verificou-se se apenas as variáveis mencionadas anteriormente continham valores nulos. A análise revelou, entretanto, a presença de dados ausentes em um total de cinco colunas, sendo estas:

1. CO_REGIAO;
2. CO_UF;
3. IN_CAPITAL;
4. QT_MAT_MOB_ACADEMICA;
5. QT_CONC_MOB_ACADEMICA.

Três dessas colunas estão diretamente relacionadas a informações geográficas, o que, após uma consulta ao dicionário de dados, verificou-se que essa ausência está relacionada ao fato de que, para cursos na modalidade a distância, os campos que correspondem a variáveis regionais permanecem nulos. Já as duas colunas restantes referem-se à mobilidade acadêmica, correspondendo a estudantes que se vincularam temporariamente a outras instituições de ensino.

4.3.2 Verificação de valores nulos

No decorrer da análise, observou-se ainda que as variáveis **CO_UF** e **TP_REDE** não estavam presentes na saída do método que descrevia o dataframe. Levantou-se então a hipótese de que tal comportamento ocorreu devido ao fato de o método utilizado, por padrão, retornar estatísticas apenas para colunas numéricas. Diante disso, verificou-se a existência de colunas com valores não numéricos (classificados como object no Pandas). Como ilustrado na Figura 3, ambas as variáveis em questão foram identificadas como do tipo object, o que contraria as definições do dicionário de dados. Pois, de acordo com esse documento, **CO_UF** e **TP_REDE** deveriam conter apenas valores numéricos, sendo que esta última deveria apresentar exclusivamente os códigos 1 (um) e 2 (dois), representando, respectivamente, instituições da rede pública e privada.

Figura 3 – Verificação de colunas com valores não numéricos

```
df.describe(include='object')
```

	CO_UF	TP_REDE
count	165123	167046
unique	107	44
top	35	2
freq	44215	159971

Fonte: Produzido pelo Autor (2025)

Para identificar os valores inconsistentes nessas variáveis, buscou-se todos os valores únicos presentes em cada coluna. Os valores encontrados estão ilustrados na tabela 3:

Tabela 3 – Valores únicos das variáveis

Variável	Valores
CO_UF	0, 53, 52, 51, 50, 27, 29, 21, 25, 26, 24, 28, 12, 13, 11, 14, 17, 32, 31, 33, 35, 43, 42, Acaraãf»2300200, Ararendãf»2301257, Assarãf»2301604, Banabuiãf»2301851, Baturitãf»2302107, Canindãf»2302800, Carirãf»2303105, Icapuãf»2305357, Icãf»2305407, Itapajãf»2306306, Maracanaãf»2307650, Quixadãf»2311306, Quixerãf»2311504, Tauãf»2313302, Tianguãf»2313401, Viè¼Ÿsa do Cearãf»2314102, Alegrete do Piauãf»2200277, Angical do Piauãf»2200608, Cajazeiras do Piauãf»2202075, Castelo do Piauãf»2202604, Monte Alegre do Piauãf»2206605, Sç,Š Joç,Š do Piauãf»2210003, Uruè¼¹ãf»2211209, Valenè»« do Piauãf»2211308, Macapãf»1600303, Acarãf»1500206, Afuãf»1500305, Aurora do Parãf»1500958, Cametãf»1502103, ConciŸ%o!dia do Parãf»1502756, GoianéŠ,ia do Parãf»1503093, Gurupãf»1503101, Ipixuna do Parãf»1503457, Jacundãf»1503804, Marabãf»1504208, Muanãf»1504901, Nova Esperanè»« do Piriãf»1504950, Oeiras do Parãf»1505205, Oriximinãf»1505304, Pacajãf»1505486, Rondon do Parãf»1506187, Santa Bç–µbara do Parãf»1506351, Santa Izabel do Parãf»1506500, Santa Maria do Parãf»1506609, Santo AntíŸ°nio do Tauãf»1507003, Sç,Š Miguel do Guamãf»1507607, Tucumãf»1508084, Tucuruãf»1508100, Uruarãf»1508159, Abatiãf»4100103, Almirante Tamandarãf»4100400, Alto Paranãf»4100608, Andirãf»4101101, Assaãf»4101903, Cambarãf»4103602,

	Cambãf»4103701, Carambeãf»4104659, Floraãf»4107801, Goioerãf»4108601, Ibiporãf»4109807, Imbaãf»4110078, Inajãf»4110300, Iporãf»4110607, Ivaiporãf»4111506, Ivatãf»4111555, Ivaãf»4111407, Jaguapitãf»4111902, Japurãf»4112405, Mamborãf»4114005, Maringãf»4115200, Paranaguãf»4118204, Paranavaãf»4118402, Pontal do Paranãf»4119954, Quatiguãf»4120705, Rio Branco do Ivaãf»4122172, Rosç-µio do Ivaãf»4122651, Santa Fãf»4123402, Sç,Š Joç,Š do Ivaãf»4125001, Sç,Š Pedro do Ivaãf»4125803, Ubiratãf»4128005, Uraãf»4128401, Verãf»4128609, nan, Santa Isabel do Ivaãf»4123709, Tré»OE Barras do Paranãf»4127858
TP_REDE	2, 1, Cié°©cia De Dados, Jogos Digitais, Anç-x9dise E Desenvolvimento De Sistemas, Cié°©cias De Dados, Desenvolvimento Para Web, Engenharia De Software, Gestç,Š Da Tecnologia Da Informaé^o, Seguranè»« Da Informaé^o, Banco De Dados, Cié°©cia Da Computaé^o, Sistemas De Informaé^o, Cié°©cias Da Computaé^o, Redes De Computadores, Engenharia De Computaé^o, Big Data E Inteligé°©cia Analâ'œica, Sistemas Para Internet, Defesa Ciberné"x81ica, Ciberseguranè»«, Agrocomputaé^o, Data Science, Desenvolvimento Full Stack, Desenvolvimento Mobile, Devops, Cié°©cias De Dados E Anç-x9dise De Comportamento, Computaé^o Em Nuvem, Coding, Digital Security, Game Design, Internet Das Coisas, Design Digital, Desenvolvimento Back-End, Gestç,Š Da Seguranè»« E Defesa Ciberné"x81ica, Arquitetura De Dados, Inteligé°©cia De Mercado E Anç-x9dise De Dados, Inteligé°©cia Artificial, Desenvolvimento Web, Té!—nicas Para Telemedicina, Sistema De Informaé^o, Informç-,ica, Inteligé°©cia Artificial Aplicada, Desenvolvimento De

	Aplicativos Para Dispositivos Móveis, Análise De Dados
--	--

Fonte: Produzido pelo autor (2025)

No caso da variável **CO_UF**, embora os valores apresentem caracteres ilegíveis devido a possíveis problemas de codificação, é possível identificar referências a localidades, o que sugere um cruzamento incorreto com dados geográficos. Já na variável **TP_REDE**, os valores inconsistentes correspondem claramente a nomes de cursos, como “Ciência de Dados”, “Engenharia de Software” e “Sistemas para Internet”, evidenciando um erro de estruturação em que dados pertencentes a outras colunas foram indevidamente alocados.

Importante destacar que, ao consultar os arquivos originais (anteriores ao processo de consolidação das bases), observou-se a presença de erros similares. Isso indica que as inconsistências encontradas nesta fase não foram introduzidas durante a consolidação (junção dos dados).

4.3.3 Verificação de Valores Fora do Intervalo Permitido

Identificou-se também uma inconsistência nos valores presentes na coluna **TP_CATEGORIA_ADMINISTRATIVA** (Tipo de Categoria Administrativa da instituição). De acordo com o dicionário de dados, os valores válidos para essa variável devem variar de 1 (um) a 9 (nove). No entanto, a análise revelou a presença de valores superiores a esse intervalo, sendo o valor máximo registrado igual a 24.553 (vinte e quatro mil, quinhentos e cinquenta e três), como mostra a Tabela 4.

Tabela 4 – Valores da coluna TP_CATEGORIA_ADMINISTRATIVA

Medida	Valor
Total	167.046,00
Média	30,13
Desvio padrão	328,96
Maior valor	24.553,00

Fonte: Produzido pelo autor (2025)

Para quantificar o número de registros fora do padrão esperado, foi aplicada uma filtragem com base nessa variável, considerando apenas valores superiores a 9 (nove). Foram identificadas 3.416 (três mil, quatrocentos e dezesseis) linhas com valores inconsistentes, o que representa aproximadamente 2% do total de registros.

4.3.4 Considerações sobre problemas encontrados

Dessa forma, observa-se que a tabela demanda uma etapa de sanitização antes de ser utilizada em processos de mineração de dados, uma vez que apresenta problemas como valores nulos, tipos de dados inconsistentes e registros fora dos intervalos esperados. Portanto, de posse das informações providas pela análise exploratória, foi iniciada a etapa de sanitização, a fim de corrigir tais inconsistências e garantir a qualidade dos dados.

4.4 SANITIZAÇÃO DOS DADOS

4.4.1 Tratamento de Dados Geográficos Ausentes

Realizaram-se verificações referentes à presença de valores nulos nas colunas de localização no que trata dos cursos na modalidade a distância. Esta etapa exigiu análise e conhecimento do domínio de aplicação, tendo em vista que cursos a distância ainda necessitam do registro de sua localidade principal. Identificou-se 1.923 (mil novecentos e vinte e três) linhas com valores nulos e, posteriormente, examinou-se se estes registros pertencem à modalidade de ensino a distância.

A verificação foi realizada por meio da correlação de registros com valores ausentes nas variáveis geográficas com a variável **TP_MODALIDADE_CURSO** (a qual indica se o curso é presencial ou a distância). Observou-se que a totalidade dos dados ausentes nas variáveis de localização está associada exclusivamente aos cursos na modalidade a distância.

Para tratar os valores ausentes nas variáveis de localização, foi atribuído um novo valor previamente inexistente na base de dados. Para as colunas **CO_UF** e **CO_REGIAO**, utilizou-se o valor 0 (zero), uma vez que esse dígito não era utilizado por nenhuma das categorias existentes. Assim, o valor 0 passa a representar a condição "não aplicável", associada aos cursos na modalidade a distância. Já para a variável **IN_CAPITAL**, cujos valores originais são 0 (não) e 1 (sim), foi introduzido o

valor 2 (dois) para indicar a mesma condição de "não aplicável", evitando ambiguidade com os significados já atribuídos.

4.4.2 Tratamento de dados nulos de mobilidade acadêmica

Após as correções, restaram apenas duas colunas com valores ausentes. Ambas estão relacionadas à mobilidade acadêmica: **QT_MAT_MOB_ACADEMICA** e **QT_CONC_MOB_ACADEMICA**. De acordo com o dicionário de dados, essas colunas representam, respectivamente, a quantidade de alunos regularmente matriculados e de concluintes que participaram de programas de mobilidade acadêmica, vinculando-se temporariamente a outra instituição, seja nacional ou internacional.

A análise das duas colunas relacionadas à mobilidade acadêmica revela, conforme apresentado na tabela 5, que a maioria de seus registros possui valor igual a 0 (zero):

Tabela 5 – Valores das colunas referentes a mobilidade acadêmica

Medida	Valor
Total	167.046,00
Média	0
Desvio padrão	0,02
Maior valor	3,00

Fonte: Produzido pelo autor (2025)

Como visto, a média dos valores observados é 0 (zero) e seu desvio padrão é de apenas 0,02, indicando baixa incidência de mobilidade acadêmica nos cursos analisados. Adicionalmente, verificou-se a porcentagem de valores 0 (zero) presentes nessas colunas e foi confirmado que 99,91% dos dados nessa coluna correspondiam a esse valor. Diante desse cenário, optou-se por preencher os valores ausentes com 0 (zero), uma vez que essa substituição segue o comportamento predominante dos dados.

Com isso, concluiu-se o tratamento dos dados ausentes da base e deu-se início ao tratamento dos dados inconsistentes, ou seja, informações existentes mas

inseridas de forma errônea.

4.4.3 Tratamento de dados inconsistentes

Consulta ao dicionário de dados revelou a presença de valores inconsistentes nas colunas **CO_UF**, **TP_REDE** e **TP_CATEGORIA_ADMINISTRATIVA**. Aplicou-se então um filtro lógico para isolar esses casos, resultando em um subconjunto de 3.427 (três mil quatrocentos e vinte e sete) linhas inválidas, aproximadamente 2% do total da base de dados. Conforme analisado, a maior parte desses registros, 99.68% apresentava erros simultâneos nas três colunas. Diante desse cenário, optou-se por duas abordagens de tratamento dos dados:

1. Exclusão dos registros inválidos, por conta de sua proporção reduzida em relação ao quantitativo total dos dados;
2. Imputação dos valores inconsistentes pelo valor mais frequente de cada coluna, visando a possibilidade de avaliação do impacto dessa técnica de preenchimento em trabalhos futuros sobre mineração de dados.

Dessa forma, esta etapa resultou em 2 (duas) versões da base sanitizada, uma referente à técnica de exclusão dos registros corrompidos e outra base contendo valores imputados. No primeiro caso, após a remoção dos dados inconsistentes, realizou-se uma verificação do tamanho da tabela resultante em comparação com a base de dados consolidada original, a fim de averiguar a porcentagem de dados excluídos durante o processo de limpeza. Verificou-se que o conjunto manteve 97,9% dos dados originais.

4.4.4 Disponibilização dos resultados

Ao final dos trabalhos preliminares de sanitização, esta pesquisa disponibilizou os arquivos resultantes no repositório de dados ⁶contendo também o conjunto consolidado acompanhado do respectivo dicionário de dados e o código fonte.

Estão incluídas tanto as versões sanitizadas, uma contendo registros imputados e outra com exclusão de inconsistências, quanto a base consolidada original, posterior ao processo de junção das tabelas mas anterior ao processo de limpeza. Além dessas, disponibilizou-se também o respectivo dicionário de dados e

⁶ Bases de dados

um documento de apoio que descreve a estrutura dos arquivos.

A seguir, apresenta-se a descrição de cada um dos arquivos disponibilizados:

- `base_completa.xlsx`: Base consolidada contendo exclusivamente registros de cursos da área de Computação, extraídos dos microdados dos anos entre 2021 e 2023. Esta versão representa o conjunto bruto, sem aplicação das rotinas de limpeza;
- `base_sanitizada_remocao.xlsx`: Versão sanitizada obtida por meio da exclusão de registros que apresentavam valores inválidos ou fora dos domínios esperados;
- `base_sanitizada_valores_imputados.xlsx`: Variante sanitizada da base de dados na qual os registros inconsistentes foram mantidos mediante aplicação de técnicas de imputação estatística (moda), com o objetivo de preservar a completude da amostra;
- `dicionário_dados.xlsx`: Documento baseado no dicionário técnico fornecido pelo INEP, contendo adaptações que indicam as colunas mantidas, removidas e recodificadas durante o pré-processamento;
- `código.ipynb`: código python utilizado na pesquisa;
- `estrutura_dos_arquivos.pdf`: Documento complementar que descreve a função de cada arquivo presente na pasta, para eventuais consultas.

Estes arquivos podem ser empregados em trabalhos subsequentes para análise, mineração e classificação.

5 CONCLUSÃO

Este trabalho objetivou a análise exploratória e sanitização da base de microdados do Censo da Educação Superior, com ênfase nos cursos da área de Computação no Brasil, a fim de disponibilizar uma base sanitizada para futuras análises em mineração de dados. A metodologia utilizada envolveu a identificação de inconsistências, a padronização de variáveis e o tratamento de dados ausentes, culminando na geração de versões limpas da base de dados.

A análise exploratória indicou a presença de inconsistências na base de dados oficial, que podem comprometer sua utilização em atividades de mineração de dados. Dentre os problemas observados estão a presença de registros com campos fora do domínio esperado, erros de codificação, presença de valores nulos em variáveis relevantes. Além de divergências na estrutura das tabelas entre os anos analisados. O trabalho realizado demonstrou que, mesmo dados públicos bem estabelecidos, podem necessitar de uma curadoria para garantir a validade e confiabilidade das informações. Observou-se ainda, ausência de dados geográficos e regionais em registros de cursos na modalidade a distância, estes fatos podem indicar a necessidade de complementação do cadastro com informações relevantes sobre a abrangência destes cursos.

Dentre as contribuições deste trabalho pode-se citar a disponibilização de 2 (duas) bases de dados sanitizadas e documentadas, buscando apoiar estudos envolvendo classificação e mineração de dados estudantis. Adicionalmente, o processo de sanitização sistemático e bem documentado configura-se como uma contribuição metodológica que pode ser reutilizada em contextos similares. Acredita-se que o trabalho aqui realizado possa ser empregado como referência para a sanitização de bases pertencentes a outros domínios de aplicação, contribuindo assim com a qualidade dos dados empregados nas pesquisas educacionais.

Pretende-se empregar as bases resultantes desta pesquisa em estudos de mineração e classificação de dados estudantis, em específico os relacionados a evasão escolar e levantamento de perfis acadêmicos.

REFERÊNCIAS

- AKERKAR, R.; SAJJA, P. Introduction to data science. In: **INTELLIGENT TECHNIQUES FOR DATA SCIENCE**. Cham: Springer, 2016. p. 1-30. ISBN 978-3-319-29205-2. DOI: 10.1007/978-3-319-29206-9_1. Disponível em: https://www.researchgate.net/publication/309030407_Introduction_to_Data_Science. Acesso em: 18 jun. 2025.
- ALVIM, Í. V.; BITTENCOURT, R. A.; DURAN, R. S. Evasão nos cursos de graduação em computação no Brasil. In: **SIMPÓSIO BRASILEIRO DE EDUCAÇÃO EM COMPUTAÇÃO (EDUCOMP)**, 4., 2024, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2024. p. 1-11. ISSN 3086-0733. DOI: 10.5753/educomp.2024.237328. Disponível em: <https://sol.sbc.org.br/index.php/educomp/article/view/28168>. Acesso em: 7 jun. 2025.

ALZOBADY, Z. R. Data mining application in education. **World Journal of Advanced Engineering Technology and Sciences**, v. 14, n. 02, p. 173-182, 2025. DOI: 10.30574/wjaets.2025.14.2.0070. Disponível em: <https://journalwjaets.com/node/313>. Acesso em: 22 jun. 2025.

BLEI, D. M.; SMYTH, P. Science and data science. **Proceedings of the National Academy of Sciences**, v. 114, n. 33, p. 8689-8692, 2017. DOI: 10.1073/pnas.1702076114. Disponível em: <https://www.pnas.org/doi/10.1073/pnas.1702076114>. Acesso em: 25 jun. 2025.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. Brasília, DF: Presidência da República, 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 12 jun. 2025.

BRASIL. **Lei de Diretrizes e Bases da Educação Nacional: Lei nº 9.394, de 20 de dezembro de 1996**. Brasília, DF: Presidência da República, 1996. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l9394.htm. Acesso em: 28 maio 2025.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). **Exame Nacional do Ensino Médio: documento básico**. Brasília, DF: Ministério da Educação, 2002.

BRITO, B. C. P.; MELLO, R. F. L.; ALVES, G. Identificação de atributos relevantes na evasão no ensino superior público brasileiro. In: **Simpósio Brasileiro De Informática Na Educação (SBIE)**, 31., 2020, Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 1032-1041. DOI: 10.5753/cbie.sbie.2020.1032. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/12859>. Acesso em: 30 maio 2025.

CAO, L. Data science: a comprehensive overview. **ACM Computing Surveys**, New York, v. 50, n. 3, p. 1–42, 2017. Art. 43. DOI: 10.1145/3076253. Disponível em: <https://doi.org/10.1145/3076253>. Acesso em: 2 jul. 2025.

CASTRO, O.; BRUNEAU, P.; SOTTET, J.-S.; TORREGROSSA, D. Landscape of high-performance Python to develop data science and machine learning applications. **ACM Computing Surveys**, New York, v. 56, n. 3, p. 1–30, 2023. Art. 65. DOI: 10.1145/3617588. Disponível em: <https://doi.org/10.1145/3617588>. Acesso em: 14 jun. 2025.

CHEN, Y. Analysis of the development of computer science and its future trend. **Applied and Computational Engineering**, v. 8, p. 341–345, 2023. DOI: 10.54254/2755-2721/8/20230180. Disponível em: https://www.researchgate.net/publication/372823356_Analysis_of_the_Development_of_Computer_Science_and_its_Future_Trend. Acesso em: 5 jul. 2025.

CIFUENTES GOMEZ, G.; GUZMÁN, P.; SANTELICES, M. V. Transitioning to higher education: students' expectations and realities. **Educational Research**, v. 64, n. 4, p. 424–439, 2022. DOI: 10.1080/00131881.2022.2087712. Disponível em: <https://doi.org/10.1080/00131881.2022.2087712>. Acesso em: 10 jul. 2025.

COSTA JÚNIOR, J. F. A importância da educação como ferramenta para enfrentar os desafios da sociedade da informação e do conhecimento. **Convergências: Estudos em Humanidades Digitais**, v. 1, n. 1, p. 127–144, 2023. DOI: 10.59616/conehd.v1i01.97. Disponível em: <https://periodicos.ifg.edu.br/cehd/article/view/97>. Acesso em: 23 jun. 2025.

DUTT, A.; ISMAIL, M. A.; HERAWAN, T. A systematic review on educational data mining. **IEEE Access**, v. 5, p. 15991–16005, 2017. DOI: 10.1109/ACCESS.2017.2654247. Disponível em: https://www.researchgate.net/publication/312509093_A_Systematic_Review_on_Educational_Data_Mining. Acesso em: 8 jul. 2025.

FERNANDES, A.; KOEHLER, M.; KONSTANTINO, N.; PANKIN, P.; PATON, N.; SAKELLARIOU, R. Data preparation: a technological perspective and review. **SN Computer Science**, v. 4, 2023. DOI: 10.1007/s42979-023-01828-8. Disponível em: https://www.researchgate.net/publication/371253640_Data_Preparation_A_Technological_Perspective_and_Review. Acesso em: 12 jul. 2025.

FOOK, C. Y.; SIDHU, G. K. Investigating learning challenges faced by students in higher education. **Procedia – Social and Behavioral Sciences**, v. 186, p. 604–612, 2015. DOI: 10.1016/j.sbspro.2015.04.001. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877042815022612>. Acesso em: 12 maio 2025.

FUKAO, A. T.; COLANZI, T. E.; MARTIMIANO, L. A. F.; FELTRIM, V. D. Estudo sobre evasão nos cursos de computação da Universidade Estadual de Maringá. In: **Simpósio Brasileiro De Educação Em Computação (EDUCOMP)**, 3., 2023, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 86-96. ISSN 3086-0733. DOI: 10.5753/educomp.2023.228209. Disponível em: <https://sol.sbc.org.br/index.php/educomp/article/view/23878>. Acesso em: 27 jun. 2025.

GIBERT, K.; SÀNCHEZ-MARRÈ, M.; IZQUIERDO, J. A survey on pre-processing techniques: relevant issues in the context of environmental data mining. **AI Communications**, v. 29, p. 1–37, 2016. DOI: 10.3233/AIC-160710. Disponível em: https://www.researchgate.net/publication/309714328_A_survey_on_pre-processing_techniques_Relevant_issues_in_the_context_of_environmental_data_mining. Acesso em: 29 maio 2025.

GUPTA, V. Recent advancements in computer science: a comprehensive review of emerging technologies and innovations. **International Journal for Research Publication and Seminars**, v. 14, p. 329–334, 2023. DOI: 10.36676/jrps.2023-v14i1-42. Disponível em: https://www.researchgate.net/publication/373150293_Recent_Advancements_in_Computer_Science_A_Comprehensive_Review_of_Emerging_Technologies_and_Innovations. Acesso em: 14 jun. 2025.

HAN, J.; KAMBER, M.; PEI, J. Data preprocessing. In: HAN, J.; KAMBER, M.; PEI, J. (ed.). **Data Mining: Concepts and Techniques**. 3. ed. Boston: Morgan Kaufmann, 2012. p. 83–124. (The Morgan Kaufmann Series in Data Management Systems). ISBN 978-0-12-381479-1.

HOFFMANN, I. L.; NUNES, R. C.; MULLER, F. M. As informações do Censo da Educação Superior na implementação da gestão do conhecimento organizacional sobre evasão. **Gestão & Produção**, v. 26, n. 2, p. e2852, 2019. DOI: 10.1590/0104-530X-2852-19. Disponível em: <https://doi.org/10.1590/0104-530X-2852-19>. Acesso em: 14 jul. 2025.

KANG, M.; TIAN, J. Machine learning: data pre-processing. In: PECHT, M.; KANG, M. (ed.). **Prognostics and Health Management of Electronics**. Hoboken: John Wiley & Sons, 2018. cap. 5, p. 111-130. ISBN 9781119515326. DOI: 10.1002/9781119515326.ch5. Disponível em: <https://onlinelibrary.wiley.com/doi/book/10.1002/9781119515326>. Acesso em: 26 jun. 2025.

LI, J.; XUE, E.; WEI, Y.; HE, Y. How popularising higher education affects economic growth and poverty alleviation: empirical evidence from 38 countries. **Humanities and Social Sciences Communications**, v. 11, n. 1, p. 520, 2024. DOI: 10.1057/s41599-024-03013-5. Disponível em: <https://doi.org/10.1057/s41599-024-03013-5>. Acesso em: 25 jun. 2025.

MAHARANA, K.; MONDAL, S.; NEMADE, B. A review: data pre-processing and data augmentation techniques. **Global Transitions Proceedings**, v. 3, n. 1, p. 91–99, 2022. DOI: 10.1016/j.gltp.2022.04.020. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666285X22000565>. Acesso em: 30 jun. 2025.

MARTINEZ, I.; VILES, E.; OLAIZOLA, I. **Data science methodologies: current challenges and future approaches**. 2021. DOI: 10.48550/arXiv.2106.07287. Disponível em: <https://arxiv.org/abs/2106.07287>. Acesso em: 4 jul. 2025.

NAGER, A.; ATKINSON, R. D. **The case for improving U.S. computer science education**. 2016. DOI: 10.2139/ssrn.3066335. Disponível em: <https://ssrn.com/abstract=3066335>. Acesso em: 14 jul. 2025.

PANTE-SALAS, G.; QUIÑONES-COLCHADO, E.; CUYA-CANDELA, E.; LÓPEZ-SÁNCHEZ, M.; GAMARRA, M.; ALVAREZ, C. Demands and challenges in higher education today. **Journal of Ecohumanism**, v. 4, 2025. DOI: 10.62754/joe.v4i1.5919. Disponível em: https://www.researchgate.net/publication/388169564_Demands_and_Challenges_in_Higher_Education_Today. Acesso em: 11 jul. 2025.

RAINA, V.; KRISHNAMURTHY, S. **Building an effective data science practice: a framework to bootstrap and manage a successful data science practice**. 2022. ISBN 978-1-4842-7418-7. DOI: 10.1007/978-1-4842-7419-4. Disponível em: https://www.researchgate.net/publication/357495958_Building_an_Effective_Data_Science_Practice_A_Framework_to_Bootstrap_and_Manage_a_Successful_Data_Science_Practice. Acesso em: 7 jul. 2025.

RAJKUMAR, S.; SUJAY, D. J.; SANGEETHA, R.; SUDHAKAR, G.; MURALIKRISHNA, K.; KUMARASWAMY, D. An overview of computer science and engineering and its latest technologies. **International Research Journal on Advanced Engineering Hub**, v. 2, p. 683–698, abr. 2024. DOI:

10.47392/IRJAEH.2024.0099. Disponível em:
https://www.researchgate.net/publication/379703254_An_Overview_of_Computer_Science_and_Engineering_and_Its_Latest_Technologies. Acesso em: 19 jun. 2025.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: an updated survey. **WIREs Data Mining and Knowledge Discovery**, v. 10, n. 3, e1355, 2020. DOI: 10.1002/widm.1355. Disponível em:
<https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1355>. Acesso em: 2 jul. 2025.

SANATI-MEHRIZY, R.; MINAIE, A. Essential experiences for computer science graduates. In: **ASEE Annual Conference & Exposition**, 2022. DOI: 10.18260/1-2--41887. Disponível em:
https://www.researchgate.net/publication/378029056_Essential_Experiences_for_Computer_Science_Graduates. Acesso em: 8 jul. 2025.

TEODORO, L. A.; KAPPEL, M. A. A. Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no Brasil. **Revista Brasileira de Informática na Educação**, v. 28, p. 838–863, 2020. DOI: 10.5753/rbie.2020.28.0.838. Disponível em:
<https://journals-sol.sbc.org.br/index.php/rbie/article/view/3691>. Acesso em: 14 jul. 2025.

YAACOB, W. F. W.; SOBRI, N.; NASIR, S. A. M.; NORSHAHIDI, N.; HUSIN, W. Z. W. Predicting student drop-out in higher institution using data mining techniques. **Journal of Physics: Conference Series**, v. 1496, p. 012005, 2020. DOI: 10.1088/1742-6596/1496/1/012005. Disponível em:
https://www.researchgate.net/publication/341970081_Predicting_Student_Drop-Out_in_Higher_Institution_Using_Data_Mining_Techniques. Acesso em: 1 jul. 2025.

YAGCI, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. **Smart Learning Environments**, v. 9, 2022. DOI: 10.1186/s40561-022-00192-z. Disponível em:
https://www.researchgate.net/publication/358991518_Educational_data_mining_prediction_of_students'_academic_performance_using_machine_learning_algorithms. Acesso em: 6 jul. 2025.

YASEEN, M. G.; ALJANABI, M.; ALI, A. H.; ABD, S. A. Current cutting-edge research in computer science. **Mesopotamian Journal of Computer Science**, v. 2022, p. 1–4, 2022. DOI: 10.58496/MJCSC/2022/001. Disponível em:
<https://mesopotamian.press/journals/index.php/cs/article/view/9>. Acesso em: 14 jun. 2025.