



**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS TERESINA ZONA SUL
CURSO LICENCIATURA EM COMPUTAÇÃO**

ALAN ALEXANDRE SAMPAIO

**CONJUNTO DE DADOS PARA ANÁLISE DA EVASÃO ACADÊMICA NO CURSO
DE LICENCIATURA EM COMPUTAÇÃO DO IFPI-CATZS**

TERESINA

2026

ALAN ALEXANDRE SAMPAIO

CONJUNTO DE DADOS PARA ANÁLISE DA EVASÃO ACADÊMICA NO CURSO
DE LICENCIATURA EM COMPUTAÇÃO DO IFPI-CATZS

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso Licenciatura em Computação do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Teresina Zona Sul.

Orientador: Prof^a. Msc. Kelson Carvalho Santos.

TERESINA

2026

CONJUNTO DE DADOS PARA ANÁLISE DA EVASÃO ACADÊMICA NO CURSO DE LICENCIATURA EM COMPUTAÇÃO DO IFPI-CATZS.

Alan Alexandre Sampaio¹
Kelson Carvalho Santos.²

RESUMO

A evasão acadêmica é um fenômeno multifatorial que compromete o desempenho das Instituições de Ensino Superior, gerando déficit institucional e desperdício de recursos públicos. No curso de Licenciatura em Computação do Instituto Federal do Piauí – Campus Teresina Zona Sul (IFPI-CATZS), os índices históricos revelam taxas de evasão frequentemente superiores a 60%, evidenciando a urgência de investigar suas causas. Diante desse cenário, este trabalho teve como objetivo elaborar um conjunto de dados estruturado para entender a análise da evasão acadêmica no referido curso. A pesquisa apoiou-se no processo de Descoberta de Conhecimento em Banco de Dados (KDD), com ênfase nas etapas de seleção e pré-processamento. Os dados socioeconômicos e acadêmicos dos ingressantes foram extraídos dos sistemas SUAP e Q-Acadêmico e integrados em uma modelagem bidimensional, utilizando a linguagem *Python*, a biblioteca *Pandas* e o ambiente *Google Colab*. Após a coleta, os dados passaram por limpeza, tratamento de valores nulos e técnicas de codificação, normalização e padronização. Como resultado, foram construídos diferentes conjuntos de dados, um contendo apenas dados reais e outro com valores imputados pela moda, totalizando 504 instâncias, que permitiram identificar o perfil predominante de estudantes concluintes e evadidos. Constatou-se que a imputação pela moda agravou disparidades em alguns atributos. Conclui-se que os conjuntos produzidos representam base consistente para futuras aplicações em mineração de dados e aprendizado de máquina, fornecendo auxílio para políticas institucionais de enfrentamento à evasão.

Palavras-chave: Evasão acadêmica; Descoberta de conhecimento em banco de dados (KDD); Conjunto de dados; Licenciatura em computação.

ABSTRACT

Academic dropout is a multifactorial phenomenon that undermines the performance of Higher Education Institutions, generating institutional deficits and the waste of public resources. In the Computing Teaching Degree program at the Federal Institute of Piauí – Teresina Zona Sul Campus (IFPI-CATZS), historical figures reveal dropout rates frequently exceeding 60%, highlighting the urgency of investigating its causes. In light of this scenario, this study aimed to develop a structured dataset to support the analysis of academic dropout in the aforementioned program. The research was based on the Knowledge Discovery in Databases (KDD) process, with emphasis on the selection and pre-processing stages. The socioeconomic and academic data of

¹ Graduando em Licenciatura em Computação. Instituto Federal do Piauí. catzs.2022112linf0014@aluno.ifpi.edu.br.

² Professor Orientador. Doutorando em Ciência da Computação na Faculdade de Computação - FACOM da Universidade Federal de Uberlândia - UFU. Instituto Federal do Piauí. kelson@ifpi.edu.br.

the incoming students were extracted from the SUAP and Q-Acadêmico systems and integrated into a two-dimensional model, using the Python programming language, the *Pandas* library, and the Google Colab environment. After collection, the data underwent cleaning, treatment of null values, and encoding, normalization, and standardization techniques. As a result, different datasets were built, one containing only real data and another with values imputed by the mode, totaling 504 instances, which made it possible to identify the predominant profile of graduating and dropout students. It was found that imputation by the mode aggravated disparities in some attributes. It is concluded that the datasets produced represent a consistent basis for future applications in data mining and machine learning, providing support for institutional policies to address dropout.

Keywords: Academic dropout; Knowledge Discovery in Databases (KDD); Dataset; Licentiate degree in Computing.

Data de aprovação: 24/07/2026

1. INTRODUÇÃO

A análise dos dados das Instituições de Ensino Superior (IES) é uma tarefa importante que deve ser utilizada para entender o que está ocorrendo na instituição. Nesse contexto, para identificar ocorrências na instituição é necessário analisar os dados de forma agrupada, assim, é necessário a elaboração de um conjunto de dados, que é uma forma simples de examinar uma grande quantidade de dados, sendo representada em um modelo bidimensional.

Essa abordagem estruturada em um modelo bidimensional, caracterizado como uma matriz onde as colunas indicam as variáveis e as linhas representam os indivíduos(alunos), viabiliza o estudo das relações diretamente ou inversamente proporcionais existentes entre as variáveis, conforme apontado por Bussab e Morettin (2017).

Desse modo, ao realizar uma análise da literatura sobre a temática da pesquisa, observa-se que trabalhos como os de Josende Paz e Cazella (2017), Oliveira (2023) e Kantorski et al. (2023) já exploram a estruturação de bases de dados para fins de treinamento e utilização analítica. Da mesma forma, Manhães et al. (2011) adotam uma abordagem preditiva ao avaliar o risco de evasão na Universidade Federal do Rio de Janeiro (UFRJ), tomando como critério as primeiras notas obtidas pelos acadêmicos no semestre de ingresso. Contudo, a presente pesquisa diferencia-se das abordagens tradicionais e do exemplo citado ao adotar

como base as características econômico-sociais e acadêmicas dos ingressantes obtidas no exato momento da matrícula, extraídas dos questionários socioeconômicos preenchidos no Sistema Unificado de Administração Pública (SUAP) e, anteriormente, no QAcadêmico.

A urgência dessa investigação no âmbito do IFPI-CATZS é explicitada pelos dados disponibilizados pelo Controle Acadêmico da instituição (2024), detalhados no panorama histórico do curso de Licenciatura em Computação, apresentado na Tabela 1.

Tabela 1 - Panorama das turmas dos Evadidos e Concluídos

Ano de Ingresso	Matriculados	Evadidos	Concluídos	Ano de Conclusão	Previsão de Conclusão
2012	44	37 (84,09%)	7 (15,91%)	2016	-
2013	35	25 (71,43%)	10 (28,57%)	2017	-
2014	26	20 (76,92%)	6 (23,08%)	2018	-
2015	36	31 (86,11%)	5 (13,89%)	2019	-
2016	35	27 (77,14%)	8 (22,86%)	2020	-
2017	42	36 (85,71%)	6 (14,29%)	2021	-
2018	39	26 (66,67%)	13 (33,33%)	2022	-
2019	47	44 (93,62%)	3 (6,38%)	2023	-
2020	41	39 (95,12%)	2 (4,88%)	2024	-
2021	32	25	Em andamento	-	2025
2022	35	17	Em andamento	-	2026
2023	47	30	Em andamento	-	2027
2024	45	21	Em andamento	-	2028

Fonte: Adaptado de Controle Acadêmico, IFPI - CATZS (2024).

Uma análise desses indicadores revela que as turmas ingressantes entre 2012 e 2017 apresentam taxas de evasão severas, sempre superiores a 70%. Já em contrapartida, 2018 apresentou a menor taxa de abandono do período analisado, com 66,67%, demonstrando uma melhoria na retenção dos estudantes e um número proporcionalmente maior de concluintes.

Contudo, após esse declínio temporário, os índices de abandono voltaram a subir de maneira alarmante nos anos subsequentes, atingindo 93,62% em 2019 e 95,12% em 2020. Conforme discute Silva Filho (2017), essa recorrência é altamente preocupante, pois gera um severo déficit institucional e desperdício de recursos

públicos, acarretando a ociosidade de professores, funcionários, equipamentos e espaços físicos.

Diante desse panorama crítico revelado pelos dados apresentados pelo controle acadêmico, destaca-se a necessidade de elaborar um conjunto de dados bidimensional para a estruturação das informações extraídas dos questionários socioeconômicos e acadêmicos presentes no Sistema Unificado de Administração Pública (SUAP) e no Q-Acadêmico, respondidos pelos discentes ao ingressar no curso de Licenciatura em Computação do Instituto Federal do Piauí – Campus Teresina Zona Sul (IFPI-CATZS), agregando assim para mais pesquisas que busquem compreender, através de um conjunto de dados, as reais razões que levam os acadêmicos a se evadirem do curso de Licenciatura em Computação do IFPI-CATZS. Assim, esta pesquisa buscou validar as seguintes hipóteses:

- H1: Fatores socioeconômicos dos discentes (como renda familiar, necessidade de trabalho remunerado, ter filhos ou ser cotista) estão associados à ocorrência de evasão no curso de Licenciatura em Computação.
- H2: Variáveis acadêmicas (escola de origem, nível técnico profissionalizante, ter cursado outro nível superior) influenciam a ocorrência de evasão no curso de Licenciatura em Computação.
- H3: A quantidade de instâncias disponíveis será suficiente para a geração de um conjunto de dados íntegro, garantindo a qualidade das informações e minimizando a ocorrência de ruídos e perdas;
- H4: A modelagem bidimensional é a abordagem mais apropriada para a estruturação do conjunto de dados, assegurando sua usabilidade para a análise dos dados.

Com base nessas hipóteses, o estudo analisou como a construção de um conjunto de dados contribui para a compreensão da evasão no curso de Licenciatura em Computação do IFPI – Campus Teresina Zona Sul. Os dados foram extraídos dos sistemas SUAP e Q-Acadêmico e integrados em uma abordagem bidimensional. A pesquisa foi motivada pela relevância da evasão como um desafio para o ensino superior. Foram analisadas informações socioeconômicas e acadêmicas para identificar padrões e características dos estudantes. Após a coleta, os dados passaram por etapas de pré-processamento para remover inconsistências, ruídos e

valores nulos. Como resultado, foram construídos dois conjuntos de dados. Esses conjuntos permitiram identificar o perfil predominante dos estudantes concluintes e evadidos. Também foram aplicadas técnicas de codificação, padronização e escalonamento dos atributos. Embora essas transformações não tenham sido utilizadas na análise deste estudo, elas prepararam os dados para futuras pesquisas. Assim, os conjuntos de dados produzidos representam uma base consistente para aplicações em mineração de dados e aprendizado de máquina.

2. REFERENCIAL TEÓRICO

Para elaborar um conjunto de dados voltado à análise da evasão acadêmica, é fundamental conhecer alguns conceitos relacionados à temática. Assim, este capítulo apresenta o referencial teórico estruturado em subseções essenciais para a compreensão da pesquisa: Descoberta de Conhecimento em Banco de Dados (KDD), Codificação, Normalização, Padronização dos dados, Ferramentas para manipulação de dados, com ênfase na linguagem de programação *Python*, na biblioteca *Pandas* e no *Google Colab*, Evasão Acadêmica no Ensino Superior, o Sistema Unificado de Administração Pública (SUAP), o formato CSV, as diretrizes da Lei Geral de Proteção de Dados (LGPD) e, por fim, os Trabalhos Relacionados a pesquisa.

2.1 DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS (KDD)

Iniciando os conceitos relacionados à temática, o processo KDD (*Knowledge Discovery in Databases*) é um método desenvolvido por Fayyad, Piatetsky-Shapiro e Smyth (1996) para substituir e facilitar a análise de dados tradicional. Quando realizada de forma manual, a análise convencional mostra-se lenta, cara e altamente subjetiva, resultando em um trabalho exaustivo. Com o crescimento exponencial do volume de dados, essa abordagem manual tornou-se impraticável, uma vez que a quantidade de informações a serem processadas excede a capacidade humana, o que dificulta a interpretação eficiente dos dados e a obtenção de informações precisas em um curto período de tempo.

Nesse contexto, o processo KDD surgiu como uma solução para lidar com grandes volumes de dados, permitindo a extração automática de conhecimento útil. Para tanto, o método consiste em cinco etapas de manipulação de dados, iniciando

pela escolha dos dados brutos que estão sendo utilizados durante a pesquisa. Após essa seleção, ocorre a segunda etapa, responsável pelo pré-processamento dos dados escolhidos, assegurando a precisão e a integridade das informações fornecidas pelas instâncias. O fluxo completo inclui ainda as etapas de mineração de dados e interpretação das informações geradas pelo modelo.

O método KDD possui três etapas descritas a seguir:

- **Seleção de dados:** Esta é a primeira etapa do processo KDD, na qual ocorre a seleção dos dados brutos a serem utilizados ao longo de todo o processo de utilização dos dados. A escolha dos dados é feita com base no objetivo da construção do conjunto de dados.
- **Pré-processamento:** Na segunda etapa do processo KDD, realiza-se a limpeza e o tratamento dos dados, um passo essencial para garantir a qualidade do conjunto de dados. Nesta fase, são aplicadas operações visando a análise para a remoção de ruídos nas instâncias, utilizando técnicas de eliminação específicas através de métodos de seleção de atributos com codificação em *Python* e uso da biblioteca *Pandas*. Além disso, são identificados e analisados os campos de dados ausentes, considerando a quantidade de instâncias apresentadas no conjunto de dados. A fim de solucionar os dados nulos (NaN) no conjunto de dados, é realizada uma análise do impacto da eliminação de todas as instâncias com dados ausentes, tomando-se a decisão pela exclusão da instância ou pela substituição, utilizando técnicas que aplicam valores de tendência (média, mediana ou moda) nas células com dados nulos, garantindo a integridade do modelo."
- **Transformação dos dados:** A terceira etapa destaca a necessidade de transformar os dados em um formato mais adequado para análise e que também contribua para sua utilização futura no treinamento de modelos de aprendizado de máquina. Essa modificação é realizada seguindo dois fatores: o processo de discretização, no qual os valores das instâncias são convertidos em categorias, facilitando a interpretação e o processamento dos dados; e a normalização dos dados, visando que eles estejam padronizados em um intervalo adequado para sua análise.

Após a realização dessa etapa, é assegurada a possibilidade de retorno às fases anteriores, sempre que necessário, com o objetivo de aperfeiçoar o processo e garantir melhorias contínuas na análise e no tratamento dos dados trabalhados (Fayyad, Piatetsky-Shapiro e Smyth, 1996).

Essa etapa do processo KDD tem como objetivo principal a limpeza e a organização dos dados, incluindo a eliminação de ruídos, valores inconclusivos e inconsistências presentes no conjunto de dados, sendo fundamental para assegurar a qualidade das informações antes das etapas seguintes.

2.2. FERRAMENTAS UTILIZADAS NA MANIPULAÇÃO DOS DADOS

A manipulação de dados é um processo essencial na ciência de dados que permite a organização, análise e transformação das informações de forma eficiente. Para que essas operações sejam otimizadas, é necessário o uso de ferramentas computacionais apropriadas. Diante disso, este tópico apresenta três aspectos tecnológicos fundamentais para o desenvolvimento do fluxo de trabalho: a linguagem de programação *Python*, a biblioteca *Pandas* e o ambiente *Google Colab*.

2.2.1. Linguagem de programação *Python*

Para a manipulação direta dos dados, incluindo as atividades de importação, leitura, alterações e tratamentos, é necessário empregar uma linguagem de programação específica para a ciência de dados. A linguagem contribui diretamente para a otimização da análise e a redução do fluxo de trabalho, permitindo a transformação das informações por meio de linhas de código (Vicentiner, Mattedi, Mello, 2020).

Desse modo, o *Python* torna-se uma ferramenta essencial para ser utilizada em pesquisa envolvendo dados, pois conta com um ecossistema rico em diversas bibliotecas voltadas ao processamento dessas informações.

2.2.2. Biblioteca *Pandas*

Com base na linguagem de programação *Python*, possui a biblioteca *Pandas*, considerada uma das principais ferramentas para utilização com ciência de dados devido à sua ampla utilidade, e variados tipos de funções que contribuem para manipular os dados. Segundo o autor, o *Pandas*:

Trata-se essencialmente de uma biblioteca para Data Wrangling, ou seja, uma biblioteca especialmente projetada para oferecer suporte às atividades de seleção, exploração, integração, limpeza e transformação de dados (GONÇALVES, 2020, p.18).

Dessa forma, a *Pandas* se destaca por possibilitar a manipulação eficiente dos dados, facilitando a alteração de informações e a análise dos dados de maneira simplificada, por meio da utilização de linhas de código para realizar as modificações necessárias.

2.2.3. Google Colab

O *Google Colaboratory*, mais conhecido como *Google Colab*, é um serviço em nuvem seguro hospedado pelo próprio *Google*, com foco voltado à pesquisa em Aprendizagem de Máquina e Inteligência Artificial. Ele possui um ambiente de desenvolvimento interativo parecido ao *Jupyter Notebook*, buscando a facilitação da execução de códigos e a análise de dados de forma simples (Alura, 2020).

Ao integrar o *Python* e a biblioteca *Pandas*, o *Google Colab* possibilita o acesso e a manipulação de grandes volumes de dados com eficiência, dispensando configurações locais de máquina, oferecendo acesso gratuito a GPUs e permitindo o compartilhamento dinâmico entre colaboradores. Ademais, o ambiente suporta documentação em blocos de texto formatados em *Markdown* e a inserção de imagens, proporcionando a visualização imediata dos resultados dos códigos executados, o que o torna ideal para análises precisas no fluxo de ciência de dados deste projeto.

2.3. CODIFICAÇÃO DOS DADOS

Conforme Nazarudeen e Sundar (2022) e a biblioteca do *Scikit-learn* (2026), o *LabelEncoder* consiste em uma técnica na qual variáveis categóricas são transformadas e convertidas em valores numéricos, iniciando a codificação a partir do valor 0. Dessa forma, cada categoria recebe um número específico para representá-la. Para facilitar o entendimento, destaca-se o seguinte exemplo de transformação das variáveis ['gato', 'cachorro', 'cavalo', 'gato'] utilizando o *LabelEncoder*: ['gato': 0, 'cachorro': 1, 'cavalo': 2], assim no conjunto de dados é apresentado dessa forma: [0, 1, 2, 0].

Assim, essa técnica é importante por garantir maior confiabilidade e automação no processo de transformação de variáveis categóricas em valores numéricos, reduzindo a ocorrência de erros humanos durante a codificação dos atributos. Dessa forma, cada categoria recebe uma representação numérica específica, iniciando em 0 e seguindo até a quantidade total de categorias existentes em cada atributo.

2.4. NORMALIZAÇÃO DOS DADOS

Segundo Asha *et al.* (2025), o método *MinMaxScaler* regulariza os valores desbalanceados para um intervalo fixo entre 0 e 1, visando melhorar o desempenho durante o treinamento dos dados. Cruvinel (2024) afirma que os dados originais apresentam características distintas, assim esse método é aplicado para garantir que todos os dados do conjunto estejam em uma escala compatível agregando para sua utilização no modelo. Nesse sentido, destaca-se a importância da utilização dessa técnica de normalização, pois ela permite que os dados estejam organizados de forma uniforme no conjunto de dados, contribuindo para análises mais aprofundadas sobre a causa da evasão acadêmica.

Nesse sentido, conforme a documentação do Scikit-learn (2026) para que essa análise seja realizada, é necessária a utilização do método que é executado através de um cálculo, cuja fórmula é apresentada abaixo na Equação 1, sendo representada pela letra X_{scaled} , que corresponde ao valor da amostra original utilizada. Já X_{min} e X_{max} representam os valores mínimos e máximos ajustados do conjunto de dados original. O resultado X_{scaled} representa o valor calculado após a aplicação dessa fórmula, que corresponde ao método *MinMaxScaler*, visando converter o valor de X , com base nos limites mínimo e máximo apresentados no conjunto de dados, para uma escala entre 0 e 1.

(1)

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2.5. PADRONIZAÇÃO DOS DADOS

De acordo com Raju et al. (2020), a técnica de padronização *StandardScaler*, disponibilizada pela biblioteca *Scikit-learn*, destaca-se por transformar as variáveis dos atributos em valores concentrados na distribuição padrão, centralizada em 0 e com desvio padrão igual a 1. Assim, para que essa transformação seja aplicada, é necessário utilizar a média (μ) e o desvio padrão (σ), calculados individualmente para cada atributo da base de dados, conforme ilustra a Equação 2 a seguir:

$$Z_{scaled} = \frac{X - \mu}{\sigma} \quad (2)$$

Dessa forma, essa técnica de padronização é importante para realizar um contraponto à normalização utilizando o *MinMaxScaler*, pois, ao padronizar os dados, é realizada uma recentralização e um reescalonamento em torno da média igual a 0 e do desvio padrão igual a 1. Já o *MinMaxScaler* possui uma abordagem diferente, pois visa comprimir os dados dentro de um intervalo fixo, entre 0 e 1, preservando as relações e proporções originais entre os valores.

2.6. EVASÃO ACADÊMICA NO ENSINO SUPERIOR

A evasão acadêmica é ocorrente em todos os níveis de ensino, seja na educação básica ou superior. Dessa forma, para entender a evasão no âmbito educacional, é necessário determinar as suas causas de ocorrência, conforme se expressa que:

A evasão é um fenômeno multifatorial, que ocorre em todos os contextos socioeconômicos, culturais e modalidades de ensino. É preciso compreendê-lo para criar alternativas de retenção, apoiando os estudantes na permanência e êxito em seus cursos (TEIXEIRA; MENTGES; KAMPFF, 2019, p. 1).

Assim, entende-se que a evasão, ocorrente no ensino superior, não depende apenas do que ocorre dentro do ambiente acadêmico, mas também é determinada por diversos outros fatores internos e externos ligados ao acadêmico que influenciam nessa decisão e que para combater essa evasão é necessário realizar

análises do contexto socioeconômico, cultural e entender os tipos de modalidades de ensino que estavam sendo utilizadas para que chegue na conclusão de entendimento da decisão de abandono do acadêmico.

Com o objetivo de compreender com maior profundidade a evasão e identificar o número de alunos que abandonaram uma turma, as Instituições de Ensino Superior (IES) devem organizar as informações disponíveis nos próprios setores de registros e controle acadêmico de suas instituições, pois com esses dados é possível realizar a quantificação do número de alunos evadidos em uma turma por meio da comparação do número de aluno ingressantes com o número de alunos concluintes (Silva Filho *et al.*, 2007).

2.7. SISTEMA UNIFICADO DE ADMINISTRAÇÃO PÚBLICA (SUAP)

Como principal fonte de dados para mapear o perfil desses estudantes, destaca-se o Sistema Unificado de Administração Pública (SUAP). Implementado originalmente em 2006 como uma atualização do sistema de ponto eletrônico do Instituto Federal do Rio Grande do Norte (IFRN), o SUAP foi expandido para abranger as áreas administrativas e acadêmicas de diversas instituições. Atualmente, o sistema é utilizado por 28 instituições federais mediante convênios, oferecendo flexibilidade para a inclusão de novas funcionalidades e personalização de *layouts* (Portal IFRN, 2023).

O SUAP apresenta um módulo de gestão acadêmica que é responsável pelo gerenciamento das atividades de ensino, abrangendo os diferentes níveis educacionais oferecidos pela instituição. Assim, destaca-se que o sistema está em constante evolução, recebendo atualizações regulares para a correção de falhas, aprimoramento de funcionalidades e desenvolvimento de novos módulos, visando à modernização e eficiência da gestão acadêmica e administrativa das instituições que o utilizam.

2.8. FORMATO DE ARQUIVO CSV

Para viabilizar a extração e a análise dos dados armazenados no SUAP e no QAcadêmico, utiliza-se o formato CSV (*Comma-Separated Values*). O CSV é um formato amplamente adotado no ecossistema de Dados Abertos devido à sua simplicidade, armazenando informações tabuladas em arquivos de texto simples. Nele, cada linha representa um registro e todas as linhas possuem a mesma

quantidade de campos separados por um delimitador, tipicamente a vírgula, o ponto e vírgula ou aspas duplas, sendo os campos nulos representados pela ausência de caracteres entre os delimitadores (Mittlöhner *et al.*, 2016).

Essa estrutura simplificada e sem formatação fixa agregada torna o processamento mais eficiente e reduz o risco de inconsistências. Logo, justifica-se a escolha por esse formato por ser prático e de fácil utilização no ambiente Google Colab, facilitando o carregamento, a leitura e a modelagem necessários para a criação do sistema preditivo de evasão.

2.9. LEI GERAL DE PROTEÇÃO DE DADOS (LGPD)

O manuseio desses arquivos contendo informações estudantis exige rigor ético e legal, visto que a utilização de dados pessoais em pesquisas envolve variáveis que caracterizam diretamente os indivíduos, como nome, gênero, idade, endereço, informações de saúde e origem racial. Quando coletados para fins científicos, tais dados devem ser preservados e anonimizados para assegurar a privacidade e a segurança dos participantes. De acordo com o artigo 5º, inciso I, da Lei nº 13.709/2018 (Lei Geral de Proteção de Dados - LGPD), o dado pessoal é definido como qualquer informação relacionada a uma pessoa natural identificada ou identificável; por sua vez, o inciso II tipifica o dado pessoal sensível como aquele que descreve origem racial, convicção religiosa, opinião política, filiação sindical ou dados referentes à saúde, vida sexual, genética e biometria.

Desse modo, com os dados pessoais é necessário o tratamento dos dados para que os mesmos sejam utilizados da forma correta. Nesse sentido, para garantir o tratamento adequado e ético dos dados, a LGPD destaca no seu artigo 6º, que os trabalhos acadêmicos devem seguir: a finalidade e a adequação, asseguram que os dados sejam utilizados apenas para o propósito de análise da evasão escolar; a necessidade, inicialmente realizando uma coleta de todos os dados descritivos dos alunos e após realizando tratamento de dados visando limitar a quantidade de dados de informações descritivas dos participantes; o livre acesso e transparência, que os participantes possuem o direito de saber informações importante sobre a utilização dos dados, com no caso do acesso a informação de onde será armazenado os dados, por quanto tempo, quem são os pesquisadores que terão acesso e como será realizado esse descarte dos dados; a qualidade dos dados é crucial para o

desenvolvimento da pesquisa, bem como a necessidade de ter dados fidedignos com a realidade para que as informações presentes no produto da pesquisa seja verídica; a segurança e a prevenção, com a adoção de medidas técnicas e administrativas, como a anonimização das informações, visando a proteção dos dados; a não discriminação, os dados serão utilizados e tratados para fins acadêmicos e não sendo aplicados para quaisquer fins discriminatório.

Além disso, a LGPD destaca também que o princípio de responsabilização deve ser cumprido mediante a assinatura e documentação do Termo de Compromisso de Utilização de Dados (TCUD) e do Termo de Confidencialidade. Ao colocar em prática as exigências do artigo 7º da LGPD por meio da desvinculação das identidades dos estudantes, a pesquisa científica deve avançar em total conformidade ética, salvaguardando a privacidade dos titulares.

3. TRABALHOS RELACIONADOS

Nesta seção, são destacados trabalhos relacionados à temática do estudo, com foco em pesquisas voltadas para a Mineração de Dados e evasão escolar. Dentre os estudos abordados, destacam-se as contribuições de Digiampietri, Nakano e Lauretto (2016); Josende Paz e Cazella (2017); Oliveira (2023); e Kantorski et al. (2023).

O estudo de Digiampietri; Nakano; Lauretto (2016) teve como seu principal objetivo desenvolver um estudo focado na identificação precoce de alunos com alto risco de evasão, utilizando mineração de dados para analisar os históricos escolares e as notas das disciplinas do primeiro ano do curso de Sistema de Informação da Escola de Artes, Ciências e Humanidade - Universidade de São Paulo (EACH-USP). A metodologia adotada foi estruturada em seis etapas: obtenção de dados brutos, organização dos dados, identificação das informações de interesse, seleção de amostra, execução de experimentos para seleção de atributos e classificação, e análise dos resultados. Com base na análise de mais de mil alunos do curso, os autores destacaram nos resultados duas abordagens principais: a seleção de atributos, considerando o desempenho inicial dos alunos nas disciplinas do curso e selecionando as disciplinas mais informativas com chance do aluno se formar, e a predição de desempenho por meio da técnica Rotation Forest, que alcançou mais de 90% de acurácia na previsão da conclusão do curso pelos alunos.

A pesquisa de Josende Paz e Cazella (2017) investigou a identificação de perfis de alunos com potencial de evasão em uma universidade comunitária, utilizando o processo de Descoberta de Conhecimento em Base de Dados (KDD) e mineração de dados educacionais. A aplicação do KDD contou com etapas de seleção da base de dados, pré-processamento, mineração de dados e, pós-processamento dos dados. O estudo analisou três hipóteses: a primeira, comprovada, indicou que alunos dos primeiros semestres apresentam maior tendência à evasão, agregando à ausência de incentivos financeiros; a segunda, que sugeria que estudantes de outros municípios seriam mais propensos a abandonar o curso, essa hipótese não foi confirmada; já a terceira, validada, demonstrou que a concessão de bolsas influencia diretamente na decisão de permanência dos alunos.

A pesquisa referente dissertação de mestrado de Oliveira (2023) tem como objetivo analisar os dados educacionais do Instituto Federal da Paraíba (IFPB), extraídos das bases da Plataforma Nilo Peçanha (PNP) e do Sistema Unificado de Administração Pública (SUAP), com foco na identificação de atributos relevantes para a predição da evasão escolar. A metodologia envolve seis etapas para o desenvolvimento da pesquisa: revisão sistemática da literatura, coleta e pré-processamento dos dados, análise dos dados, seleção de atributos, construção dos modelos de predição e avaliação de desempenho. Os resultados foram divididos em três partes análise dos dados, resultados das seleção de atributos e resultados dos classificadores, destacando-se a análise detalhada das variáveis do PNP e SUAP, sendo que as análises dos dados da PNP, são relacionados ao Ano, Sexo, Cor/Raça, Faixa etária, Renda familiar e Turno, já relacionado aos dados do SUAP, são destacados o Turno, Cota Ministério da Educação (MEC), Cota Sistema Nacional de Informações da Educação Profissional e Tecnológica do Ministério da Educação (SISTEC), Zona de residência, Origem da escola, Faixa etária, Cor/Raça e Estado civil. Na base da PNP, o melhor desempenho foi obtido com o algoritmo Perceptron Multicamadas (MLP) e o seletor Embedded, atingindo 0,95 de desempenho no curso de Licenciatura. Já no SUAP, o algoritmo Floresta Aleatória combinado com o seletor Wrappers GB alcançou 0,99 de desempenho no curso de Bacharelado. A pesquisa conclui que os dados do SUAP apresentaram maior

eficácia na predição, evidenciando seu potencial como ferramenta de apoio à prevenção da evasão escolar no IFPB.

O estudo de Kantorski et al. (2023) tem sua abordagem voltada à análise e predição da evasão em cursos de graduação presenciais, utilizando métodos de Mineração de Dados Educacionais e algoritmos de aprendizado supervisionado para classificar estudantes com maior ou menor risco de evasão. A metodologia adotada foi uma adaptação do modelo Cross Industry Standard Process for Data Mining (CRISP-DM), estruturada em seis etapas cíclicas, sendo realizado a seleção de maneira manual dos atributos para construção do conjunto de dados: entendimento da evasão, compreensão e preparação dos dados, modelagem, ocorrendo a separação em duas partes, uma parte para treinamento e outra para teste, avaliação e aplicação dos resultados, para . A pesquisa foi aplicada a 106 cursos da instituição, considerando dados de aplicação em três semestres, 2022/1, 2022/2 e 2023/1, com análise experimental focada no semestre de 2022/1, o conjunto de dados para treinamento totaliza 61.644 instâncias, sendo os seus dados de 2010 a 2021/2, já o conjunto de dados de teste possui 15.367 instâncias com os dados de 2022/1. As métricas utilizadas incluíram acurácia, precisão, recall e f-measure, alcançando resultados como 53% de acurácia, 2% de precisão, 88% de recall e 92% de recall médio. A partir desses dados, foi possível identificar o perfil dos estudantes com risco de evasão, permitindo que a instituição de ensino superior utilize essas informações para implementar estratégias de intervenção mais eficazes. A Tabela 2 resume os trabalhos discutidos, durante o tópico.

Tabela 2 - Resumo dos trabalhos relacionados.

Autor(es)/ Ano	Objetivo	Metodologia	Técnica utilizada	Resultados
Digiampietri; Nakano; Lauretto (2016)	Identificação precoce de alunos com alto risco de evasão, analisando históricos escolares e notas das disciplinas do 1º ano do curso de Sistema de Informação da EACH-USP.	Seis etapas: obtenção de dados brutos, organização dos dados, identificação das informações de interesse, seleção de amostra, execução de experimentos para seleção de atributos e classificação, e análise dos resultados. Base com mais de mil alunos.	Seleção de atributos; Rotation Forest.	Seleção das disciplinas mais informativas quanto à chance de formatura; Rotation Forest alcançou mais de 90% de acurácia na previsão de conclusão do curso.
Josende Paz; Cazella (2017)	Identificação de perfis de alunos com potencial de evasão em uma	Processo KDD (Descoberta de Conhecimento em Base de Dados) e mineração de dados	KDD; Mineração de Dados	1ª hipótese (confirmada): alunos dos primeiros semestres têm maior

	universidade comunitária.	educacionais, com etapas de seleção da base, pré-processamento, mineração e pós-processamento. Análise de três hipóteses.	Educacionais.	tendência à evasão; 2ª hipótese (não confirmada): estudantes de outros municípios seriam mais propensos a evadir; 3ª hipótese (confirmada): a concessão de bolsas influencia a permanência.
Oliveira (2023)	Analisar dados educacionais do IFPB (bases PNP e SUAP) para identificar atributos relevantes na predição da evasão escolar.	Seis etapas: revisão sistemática da literatura, coleta e pré-processamento, análise dos dados, seleção de atributos, construção dos modelos de predição e avaliação de desempenho.	Perceptron Multicamadas (MLP) com seletor Embedded; Floresta Aleatória com seletor Wrappers GB.	PNP: MLP + Embedded atingiu 0,95 de desempenho (Licenciatura). SUAP: Floresta Aleatória + Wrappers GB alcançou 0,99 (Bacharelado). Os dados do SUAP mostraram maior eficácia na predição.
Kantorski et al. (2023)	Análise e predição da evasão em cursos de graduação presenciais, classificando estudantes com maior ou menor risco de evasão.	Adaptação do modelo CRISP-DM em seis etapas cíclicas, com seleção manual dos atributos. Aplicada a 106 cursos (2022/1, 2022/2 e 2023/1); treino: 61.644 instâncias (2010–2021/2); teste: 15.367 instâncias (2022/1).	Mineração de Dados Educacionais; algoritmos de aprendizado supervisionado.	Métricas: 53% de acurácia, 2% de precisão, 88% de recall e 92% de recall médio. Permitiu identificar o perfil dos estudantes com risco de evasão para intervenções mais eficazes.

Fonte: Elaborada pelo autor, 2026.

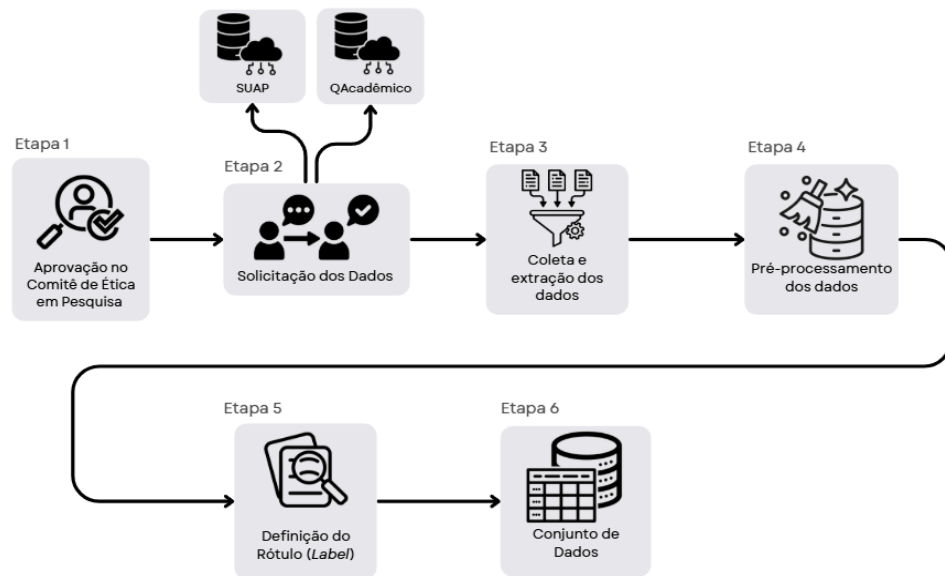
4. METODOLOGIA

A pesquisa adota uma abordagem qualitativa, visando compreender e explorar as causas da evasão no curso de Licenciatura em Computação do IFPI - Campus Teresina Zona Sul (IFPI-CATZS). Para isso, foi construído um conjunto de dados contendo informações dos alunos matriculados no curso no período de 2012 a 2024. A delimitação desse intervalo foi estabelecida considerando o período inicial da pesquisa e a disponibilidade dos dados, não sendo incluídas, portanto, as informações referentes ao ano de 2025.. Assim, a investigação da problemática da evasão caracteriza-se como um estudo de caso, por buscar analisar as razões que contribuem para a decisão de desistência dos estudantes do curso de Licenciatura em Computação.

Após a explanação da problemática destacada e as abordagens necessárias para solucionar ela, a Figura 1 ilustra as seis etapas do desenvolvimento da pesquisa que contribuíram para a criação do conjunto de dados, com o objetivo de

analisar as causas que contribuem para a evasão dos alunos. Este tópico está organizado em ordem sequencial de execução, seguindo as seis etapas apresentadas na figura abaixo, o detalhamento de cada etapa é apresentado nos parágrafos a seguir, destacando o desenvolvimento da pesquisa.

Figura 1 - Fases para o desenvolvimento da pesquisa.



Fonte: Próprio autor, 2026.

4.1. APROVAÇÃO NO COMITÊ DE ÉTICA EM PESQUISA

Antes do início da solicitação, da coleta e da análise dos dados, o projeto foi submetido à Plataforma Brasil, por se tratar de uma pesquisa que envolve dados pessoais de seres humanos. Essa etapa assegura a conformidade da pesquisa com as diretrizes do Comitê de Ética em Pesquisa (CEP), garantindo o anonimato, a privacidade e a confidencialidade dos participantes envolvidos.

4.2. SOLICITAÇÃO DOS DADOS

Após a aprovação pelo Comitê de Ética em Pesquisa (CEP), foi realizada a solicitação de acesso aos dados, acompanhada de uma explicação sobre os objetivos da pesquisa e da garantia de preservação do anonimato dos discentes, por meio de e-mail encaminhado à Coordenação de Controle Acadêmico do IFPI-CATZS. O objetivo da solicitação foi obter acesso aos dados socioeconômicos e acadêmicos dos alunos matriculados nas turmas ingressantes entre os anos de 2012 e 2024, a fim de viabilizar o desenvolvimento da pesquisa.

4.3. COLETA E EXTRAÇÃO DOS DADOS

Após a realização da solicitação, a pesquisa enfrentou um empecilho relacionado ao acesso aos dados, uma vez que eles não foram disponibilizados de forma tabulada em resposta à requisição encaminhada por e-mail. Assim, os dados foram obtidos mediante autorização do Diretor e da Coordenação de Controle Acadêmico do IFPI-CATZS, que concederam acesso aos questionários impressos preenchidos pelos alunos no ato da matrícula no curso, bem como aos arquivos contendo as informações acadêmicas armazenadas no Sistema Unificado de Administração Pública (SUAP).

Dessa forma, a coleta dos dados foi realizada manualmente a partir de duas bases de dados, o SUAP e o QAcadêmico. O processo consistiu no registro fotográfico individual dos questionários socioeconômicos e acadêmicos dos alunos referentes ao período de 2012 a 2020, bem como baixando os arquivos contendo as informações dos alunos disponíveis no SUAP, correspondentes ao período de 2021 a 2024.

Os dados extraídos manualmente e adicionados a uma planilha compõem um conjunto de dados bruto, contendo informações ainda não tratadas. A extração dos dados foi realizada por pessoas previamente autorizadas. Essa restrição tem como objetivo garantir a privacidade, a confidencialidade e a segurança dos dados utilizados na pesquisa.

4.4. PRÉ-PROCESSAMENTO DOS DADOS

Com os dados brutos tabulados em formato bidimensional, iniciou-se a Etapa 4, correspondente ao pré-processamento dos dados, cujo objetivo é tratar as informações extraídas para assegurar a qualidade das instâncias do conjunto de dados. Nesse contexto, o processo de tratamento teve início por meio de scripts desenvolvidos na linguagem de programação *Python*, utilizando a biblioteca *Pandas* para a manipulação dos dados, com o objetivo de identificar instâncias que apresentavam valores nulos, ruídos e inconsistências provenientes do processo de extração. Após a identificação das inconsistências, foram aplicadas técnicas do processo de KDD para eliminar os ruídos e corrigir os problemas encontrados, visando garantir a qualidade das instâncias do conjunto de dados. Esse tratamento foi realizado utilizando as tecnologias apresentadas anteriormente.

Ao final dessa etapa, foi criado um atributo denominado "codigo" para promover a anonimização dos dados. A cada instância foi atribuído um identificador único, composto por um código alfanumérico, garantindo o anonimato dos participantes da pesquisa, preservando sua privacidade e impedindo sua identificação por meio desse código.

Em seguida ao processo de pré-processamento dos dados, foi realizada a etapa de transformação dos dados por meio da aplicação de diversas técnicas, como o uso do *LabelEncoder* para codificar as variáveis em valores numéricos, do método *StandardScaler* para padronizar os dados, e do método *MinMaxScaler* para normalizar os atributos. Todas essas técnicas foram aplicadas por meio da biblioteca *Scikit-learn*, utilizando o ambiente de desenvolvimento *Google Colab*.

4.5. DEFINIÇÃO DO RÓTULO

Nesse sentido, após a transformação dos dados e sua preparação para análise, foi atribuído individualmente a cada instância um valor de rótulo(*label*), correspondente à classificação do aluno quanto à conclusão do curso de Licenciatura em Computação do IFPI-CATZS. Desse modo, essa coluna contém apenas valores inteiros, inseridos manualmente, distribuídos em dois níveis: o valor 0, que representa o aluno que não concluiu o curso, e o valor 1, que representa o aluno que concluiu o curso.

Essa classificação foi realizada com base nos atributos relacionados às características socioeconômicas e acadêmicas de cada aluno, a fim de identificar se o discente não concluiu o curso (0) ou o concluiu (1). Dessa forma, tornou-se possível identificar o perfil dos alunos com maior propensão à evasão, evidenciando a relevância da pesquisa.

4.6. CONJUNTO DE DADOS

Após a análise, avaliação e adição dos rótulos em cada instância, que representou a finalização da construção do conjunto de dados em formato CSV (Valores Separados por Vírgula), contendo informações descritivas, consistentes, íntegras e com anomalias tratadas, priorizando a anonimização dos dados dos discentes. Com a conclusão de todas essas etapas, o conjunto de dados elaborado passou a servir para a análise da evasão no curso de Licenciatura em Computação

do IFPI-CATZS, além de ser utilizado como conjunto de dados base para outras pesquisas que tenham como objetivo analisar a evasão acadêmica.

Com a conclusão da pesquisa e a elaboração do conjunto de dados, todas essas informações passaram por um procedimento de backup realizado em um Hard Disk (HD) externo. O dispositivo de armazenamento está localizado na própria instituição de ensino onde a pesquisa foi desenvolvida, que ficará depositado no período de 5 anos, conforme estabelecido no Termo de Confidencialidade aprovado pelo CEP, sob a responsabilidade e proteção do pesquisador responsável. Após o término desse período de armazenamento dos dados, o HD será formatado, visando ao descarte correto e seguro das informações armazenadas.

5. EXPERIMENTOS E DISCUSSÕES DOS RESULTADOS

Neste capítulo, são apresentados os experimentos realizados de maneira sequencial ao longo do desenvolvimento da pesquisa, conforme destacado na Figura 1, apresentada na Seção 3, até a obtenção dos resultados por meio da construção do conjunto de dados. Dessa forma, este capítulo está organizado em subseções que abordam os experimentos realizados e os resultados obtidos.

5.1. COLETA E EXTRAÇÃO DOS DADOS

As bases de dados selecionadas para esta pesquisa são as disponibilizadas pelo Instituto Federal do Piauí - Campus Teresina Zona Sul (IFPI-CATZS), relacionadas aos alunos do curso de Licenciatura em Computação. As duas bases de dados utilizadas foram o QAcadêmico e o Sistema Unificado de Administração Pública (SUAP). Esses dois conjuntos de dados foram coletados de maneiras distintas, sendo que os dados do QAcadêmico foram obtidos por meio de uma solicitação impressa, aprovada pelo Diretor e pelo setor de Controle Acadêmico do campus. Dessa forma, o acesso ocorreu por meio dos questionários impressos respondidos pelos alunos no momento da matrícula, abrangendo o período de 2012 a 2020.

Assim, para coletar todos os dados, foram realizados registros fotográficos de todas as folhas dos questionários socioeconômicos de cada aluno referente ao período destacado anteriormente. Já as informações dos alunos do período de 2021 a 2024 foram coletadas por meio da plataforma SUAP. Com o acesso disponibilizado pelo professor orientador da pesquisa, foi realizada a extração dos questionários

baixando em Formato de Documento Portátil (PDF), de todos os arquivos relacionados às características dos alunos.

Após toda a etapa de coleta dos dados, os arquivos foram organizados em pastas por ano, com o objetivo de facilitar a localização e a extração das informações. Em seguida, foi realizado o processo de extração dos dados, iniciando com a tentativa de extraí-los por meio de um script de programação em *Python* que utiliza a tecnologia OCR (Reconhecimento Óptico de Caracteres), uma tecnologia capaz de converter textos presentes em imagens, PDFs ou fotografias em texto editável. No entanto, a tentativa não obteve êxito, uma vez que os textos extraídos apresentavam inconsistências decorrentes de falhas no reconhecimento dos caracteres, ocasionadas pela escrita dos alunos no preenchimento dos documentos. Dessa forma, a extração de todos os dados, tanto dos registros fotográficos quanto dos arquivos em PDF baixados, foi realizada manualmente e adicionada a uma planilha da ferramenta Google Sheets.

Destaca-se que todos os dados foram unificados durante o processo de extração, resultando em um total de 504 instâncias. Além disso, foram extraídos 26 atributos que se mantiveram presentes nas diferentes variações do formulário ao longo dos anos, os quais são apresentados na Tabela 3, juntamente com a descrição de cada um deles. Todos esses atributos compõem as colunas do conjunto de dados bruto.

Tabela 3: Atributos extraídos das bases de dados.

Nome do atributo	Tipo	Descrição	Exemplo dos Dados
ano_letivo	Data	Ano em que a matrícula foi efetivada na instituição. O número "1" representa o primeiro semestre letivo.	2012/1, 2013/1, 2014/1,...
matricula	Texto	Código de identificação único atribuído ao ingressante no momento da efetivação da matrícula. Os quatro últimos dígitos foram substituídos por "xxxx" para preservar a identidade do discente.	..., 2012112LINFxxxx, 2024112LCOMxxxx.
sexo	Texto	Informa o sexo do estudante.	Masculino, Feminino.
data_de_nascimento	Data	Data de nascimento do(a) ingressante, informada no formato dd/mm/aaaa.	dd/mm/aaaa.

bairro	Texto	Nome do bairro em que reside o(a) ingressante.	Porto Alegre, Dirceu, Itararé, Santo Antônio,...
cidade	Texto	Nome do município onde o(a) estudante mora.	Teresina, Timon, Barras, ...
uf	Texto	Unidade da Federação onde o(a) ingressante reside.	PI, MA, RN.
profissao	Texto	Nome da profissão exercida pelo ingressante no momento da efetivação da matrícula.	Professor, Estudante, Chefe Administrativo, Técnico de Informática, ...
cor_raca	Texto	Cor/Raça do(a) ingressante declarado no momento da matrícula.	Amarela, Branca, Indígena, Parda e Preta.
idade	Número inteiro	Idade do(a) ingressante calculada com base no ano de ingresso no curso e na data de nascimento do estudante.	..., 18, 44, 52, 23, 19, 26, ...
portador_de_necessidades_especiais	Texto	Indica se o(a) ingressante possui necessidades especiais.	Sim, Não.
estado_civil	Texto	Estado civil do(a) ingressante no momento da matrícula.	Solteiro(a), União Estável, Casado, Divorciado(a), Viúvo.
filhos	Texto	Indicação se o(a) ingressante possui filhos ou não.	Sim, Não.
transporte_escolar_publico	Texto	Informa se o(a) ingressante utiliza transporte público para se deslocar até o campus da instituição de ensino.	Sim, Não.
meio_de_transporte	Texto	Corresponde ao meio de transporte até chegar no campus.	Ônibus, A pé, bicicleta, Carro ou moto.
oriundo_escola	Texto	Classificação do tipo de escola da qual o(a) ingressante era oriundo.	Pública, Particular, Filantrópica.
atualmente_reside	Texto	Informa com quem o(a) estudante reside em sua residência atual.	Casa do Estudante, Com amigos, Com os pais, Com parentes, Com cônjuge, Sozinho(a), Casa própria.
acesso_ao_computador	Texto	Indica se o(a) ingressante possui acesso ao computador.	Sim, Não.
tipo_de_moradia	Texto	Categorização do tipo de moradia que o(a) ingressante reside.	Cedida, Alugada, Programa de Apartamento Residencial, Financiada, Própria.
estado_civil_dos_pais	Texto	Estado civil dos pais do(a) ingressante no momento da matrícula.	Solteiros, Casados, Separados, Divorciado(a), Viúvo(a).

renda_mensal	Texto	Faixa da renda mensal do(a) estudante.	Até 1 salário, De 1 a 3 salários, De 3 a 5 salários, De 5 a 10 salários, Mais de 10 salários.
participacao_economica	Texto	Classificação da participação econômica do(a) estudante na composição da renda familiar.	Não trabalha e é sustentada pela família ou por outras pessoas, Trabalha, mas recebe ajuda financeira da família ou de outras pessoas, Trabalha e é responsável apenas pelo seu próprio sustento, Trabalha e é responsável pelo sustento, além de contribuir parcialmente para o sustento da família, Trabalha e é o principal responsável pelo sustento da família.
zona_residencial	Texto	Nome da zona residencial onde está localizada a residência do(a) estudante.	Urbana, Rural.
cota	Texto	Classificação se o(a) ingressante adentrou por meio do programa de cota ou não.	Sim, Não.
formacao_nivel_medio	Texto	Identificação se a formação do ensino médio foi profissionalizante ou não.	Sim ou Não, profissionalizante.
cursou_outro_nivel_superior	Texto	Informa se o(a) estudante possui outra graduação de nível superior ou não.	Sim, Não.

Fonte: Próprio autor, 2026.

5.2. PRÉ-PROCESSAMENTO DOS DADOS

Com a finalização da extração de todos os dados, iniciou-se o processo de pré-processamento com o objetivo de eliminar os ruídos presentes na base de dados, como valores inconclusivos, nomes inválidos e dados nulos. Para resolver essa problemática, inicialmente foram identificados os valores inconclusivos e os nomes inválidos por meio de scripts em *Python*, utilizando a biblioteca *Pandas* no ambiente de desenvolvimento *Google Colab*. Após essa identificação, as correções foram realizadas diretamente na planilha, visando eliminar inconsistências e padronizar as informações. Para o tratamento dos valores nulos, foram aplicadas duas abordagens, gerando duas versões de conjunto de dados.

A primeira versão consistiu na exclusão de toda a instância quando identificado um valor nulo na linha, reduzindo o conjunto de dados de 504 para 196 instâncias, formando uma base com os dados 100% originais. Dessa forma, o conjunto passou a representar aproximadamente 38,9% do total de instâncias extraídas, conforme destacado no Gráfico 1.

Gráfico 1 - Divisão das instâncias do conjunto de dados original.



Fonte: Próprio autor, 2026.

A segunda abordagem consistiu na aplicação de uma medida de tendência central, utilizando a moda de cada atributo para preencher os valores ausentes presentes nas instâncias, que correspondiam a 61,1% do conjunto de dados, como destacado no Gráfico 1. Assim, essa versão do conjunto de dados manteve sua quantidade total de 504 instâncias, porém com a presença de dados reais e dados imputados.

Dessa forma, obteve-se dois conjuntos de dados livres de ruídos, permitindo a comparação entre as duas abordagens e a adoção dessas duas abordagens distintas teve como finalidade permitir a comparação do entendimento do perfil dos alunos em cada cenário.

Com os dados tratados e após a remoção das inconsistências, foram realizadas algumas transformações nos dados, sendo elas a codificação e o escalonamento das variáveis, sendo destacada a padronização e a normalização das variáveis de cada atributo, as quais serão explicadas nos subtópicos seguintes.

5.2.1 Codificação dos dados

A codificação é uma etapa importante do pré-processamento, pois converte variáveis categóricas em formato de texto para valores numéricos, o que facilita a manipulação, o cruzamento e a análise dos atributos, além de deixar o conjunto de dados pronto para reuso em pesquisas futuras que venham a aplicar algoritmos de aprendizado de máquina, os quais em geral exigem entradas numéricas. Para isso, foram utilizadas duas abordagens: a codificação manual e a técnica *LabelEncoder*.

A primeira forma foi realizada por meio da análise de todos os atributos presentes em cada coluna, sendo atribuídos manualmente valores numéricos conforme as categorias identificadas. Assim, na variável “oriundo_escola”, foi realizado um mapeamento manual que definiu as seguintes categorias: { 'Pública': 1, 'Particular': 2, 'Filantrópica': 3 }, sendo esse mesmo procedimento aplicado às demais categorias. Já nas variáveis de condição binária, como o atributo “sexo”, foi estabelecida a seguinte codificação: { 'Masculino': 1, 'Feminino': 0 }.

Por outro lado, utilizando a técnica *LabelEncoder*, as categorizações foram realizadas de maneira automática conforme as variáveis apresentadas em cada atributo, por meio do seguinte código: `df['nome_atributo'] = le.fit_transform(df['nome_atributo'])`. Nesse processo, o método “fit” identifica quais categorias únicas existem na coluna e associa um número a cada uma delas em ordem alfabética, enquanto o método “transform” substitui cada valor pelo número correspondente. Utilizando o mesmo exemplo da coluna “oriundo_escola”, a aplicação dessa técnica destaca que o índice inicial é 0, resultando na seguinte codificação: { 'Filantrópica': 0, 'Particular': 1, 'Pública': 2 }.

Diante disso, optou-se pela codificação por meio do *LabelEncoder*, por proporcionar maior uniformidade no tratamento do conjunto de dados, o que contribui para a confiabilidade dos dados codificados na etapa de análise. Ressalta-se que a diferença nos valores atribuídos por cada método é apenas de

rótulo numérico, não afetando as proporções e distribuições das categorias e, portanto, os resultados da análise de perfil.

Vale evidenciar que o *LabelEncoder* atribui valores ordinais às categorias, o que poderia introduzir uma noção artificial de ordem em variáveis nominais caso fossem aplicados modelos sensíveis à magnitude. Como este trabalho se restringe à construção e à análise exploratória do conjunto de dados, sem etapa de modelagem preditiva, essa característica não compromete os resultados.

5.2.2 Escalonamento das variáveis

Com os dados codificados seguindo a técnica *LabelEncoder* destacada anteriormente, foram construídas duas versões do conjunto: uma preservando os dados originais e outra com a imputação dos valores nulos pela moda. Sobre essas versões foram aplicados dois métodos de escalonamento da biblioteca *Scikit-learn*, a padronização (*StandardScaler*), que reescala os atributos para média 0 e desvio padrão 1, sem restringi-los a um intervalo fixo, e a normalização (*MinMaxScaler*), que comprime os valores para o intervalo [0, 1]. O objetivo desta transformação dos dados não é o treinamento de modelos preditivos, mas a disponibilização de conjuntos de dados tratados e prontos para reuso, permitindo que pesquisas futuras selecionem a versão mais adequada aos seus algoritmos, com base em testes e validações. A partir desses conjuntos, esta pesquisa realiza uma análise exploratória dos atributos e instâncias, com o intuito de caracterizar o perfil dos alunos que evadem e dos que concluem o curso.

5.3. DISCUSSÕES DOS RESULTADOS

Com os experimentos realizados, foi possível analisar e identificar o perfil dos alunos que concluem e dos que evadem o curso. Essa constatação é fundamentada na análise dos dois conjuntos de dados desenvolvidos nesta pesquisa, sendo um composto apenas por dados reais e o outro por dados reais e valores imputados pela moda nas células com valores nulos. Com a extração das duas fontes de dados SUAP e QAcadêmico, e a criação desses dois conjuntos de dados, foram atingidas as duas hipóteses: H1 e H2. Essa análise foi realizada a partir dos atributos socioeconômicos e dos atributos agrupados, utilizando o filtro na coluna "label" para selecionar apenas as instâncias classificadas como "Formado" e "Evadido".

Além disso, nos apêndices, são apresentadas duas tabelas comparativas entre os alunos que concluíram o curso e os que evadiram, comparando-se cada perfil em cada conjunto de dados (dados reais e dados com valores imputados), com o objetivo de evidenciar as características predominantes de cada perfil.

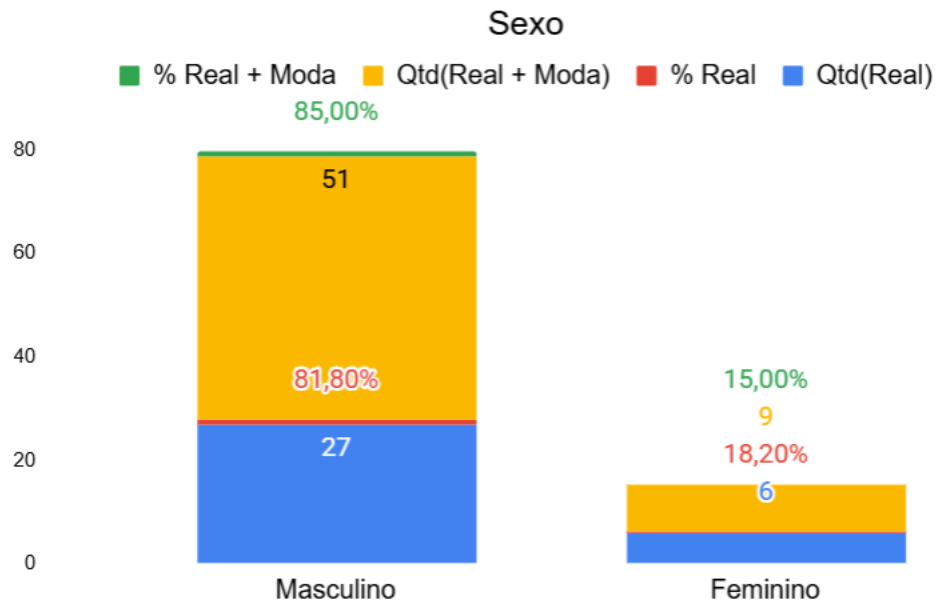
Assim, nos próximos dois tópicos, serão abordados os perfis dos alunos que concluíram e dos que evadiram o curso, aplicando-se um filtro na coluna "label" para analisar apenas as instâncias correspondentes aos alunos classificados como "Formado" e "Evadido".

5.3.1 Comparação do perfil do aluno que conclui o curso

A análise do perfil dos alunos que concluem o curso, ilustrada no Apêndice A, foi realizada com base no conjunto de dados composto apenas por informações reais, aplicando-se um filtro na coluna "label" para selecionar somente os alunos classificados como "Formado", no conjunto de dados [dataset Real](#), que resultou em 33 instâncias. No outro conjunto de dados, composto por valores reais e valores imputados pela técnica da moda ([dataset_Moda](#)), foi aplicado o mesmo filtro, sendo selecionadas, ao todo, 60 instâncias.

A análise revelou que, em ambos os conjuntos, o egresso típico caracteriza-se como um indivíduo do sexo masculino, correspondendo a 81,8% no conjunto de dados real, conforme apresentado no Gráfico 2, pela cor azul, e a 85,0% no conjunto com imputação pela moda, conforme apresentado no mesmo Gráfico pela cor amarela. Essa comparação evidencia a aplicação do princípio da utilização de uma medida de tendência central, em que o valor mais frequente (moda) é utilizado para preencher os dados ausentes, preservando o perfil predominante da variável e não ocasionando disparidades significativas em relação aos dados reais.

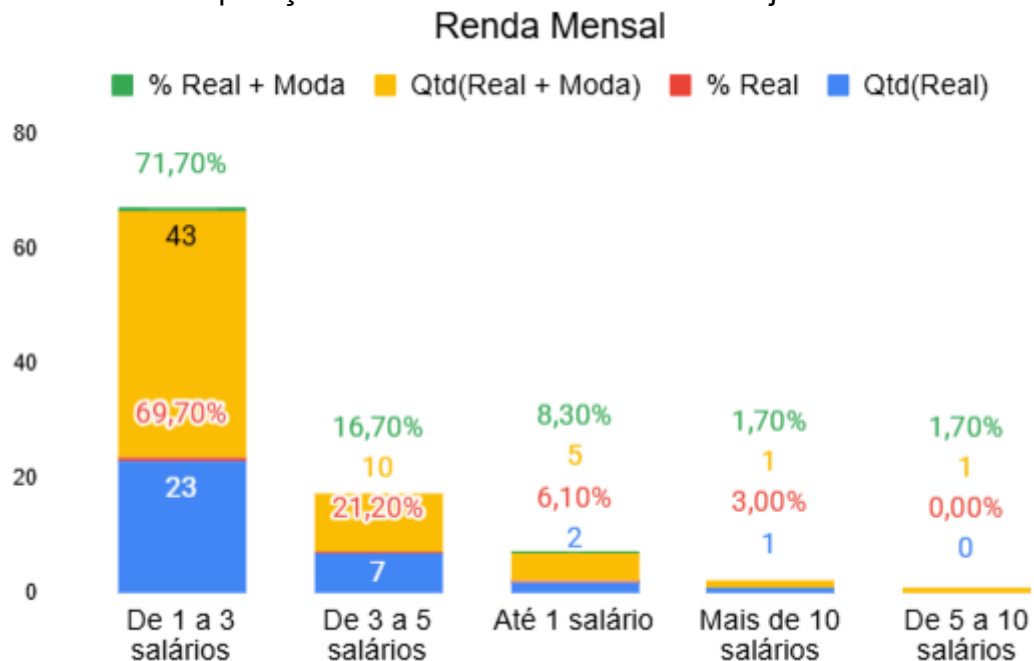
Gráfico 2 - Distribuição dos ingressantes formados por sexo nos conjuntos de dados.



Fonte: Próprio autor, 2026.

Nesse sentido, as demais características predominantes do egresso correspondem à autodeclaração de cor parda (63,6% no conjunto de dados real e 61,7% no conjunto de dados com valores imputados pela moda), ao estado civil solteiro (63,6% e 71,7%), à origem em escola pública (57,6% e 66,7%), à residência com os pais (63,6% e 65,0%), à moradia própria (84,8% e 85,0%), à renda familiar entre um e três salários mínimos (69,7% e 71,7%) como destacado no Gráfico 3 e ao acesso praticamente universal ao computador (97,0% e 98,3%). A análise relacionada a essas variáveis encontram-se apresentadas no Apêndice A.

Gráfico 3 - Comparação da renda mensal nos dois conjuntos de dados.

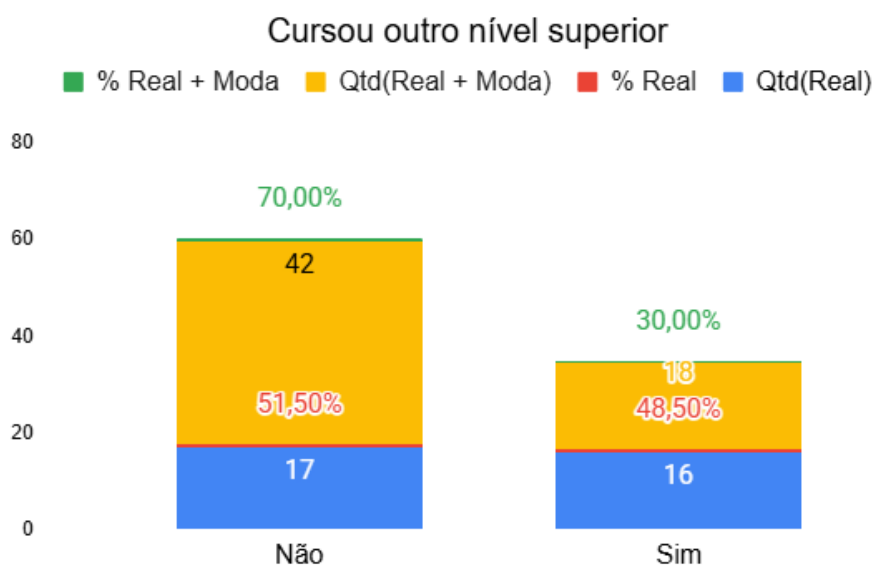


Fonte: Próprio autor, 2026.

A proximidade percentual observada nesses atributos, como moradia, renda e acesso tecnológico, indica que a imputação preservou a base estrutural dos dados, conferindo consistência à caracterização geral do concluinte e evidenciando atributos com valores equilibrados. Essa estabilidade decorre do fato de que, nesses atributos, os valores predominantes já eram fortemente dominantes no conjunto original, de modo que a substituição dos valores ausentes pelo valor mais frequente apenas reiterou uma tendência preexistente, sem distorcer a análise.

No entanto, apesar dessa convergência, verificaram-se disparidades relevantes em atributos originalmente mais equilibrados, nos quais o efeito da imputação mostrou-se mais evidente, como no caso do atributo “cursou outro nível superior”. O gráfico que ilustra esse comportamento encontra-se apresentado no Gráfico 4.

Gráfico 4 - Resultado do atributo nos conjuntos de dados.



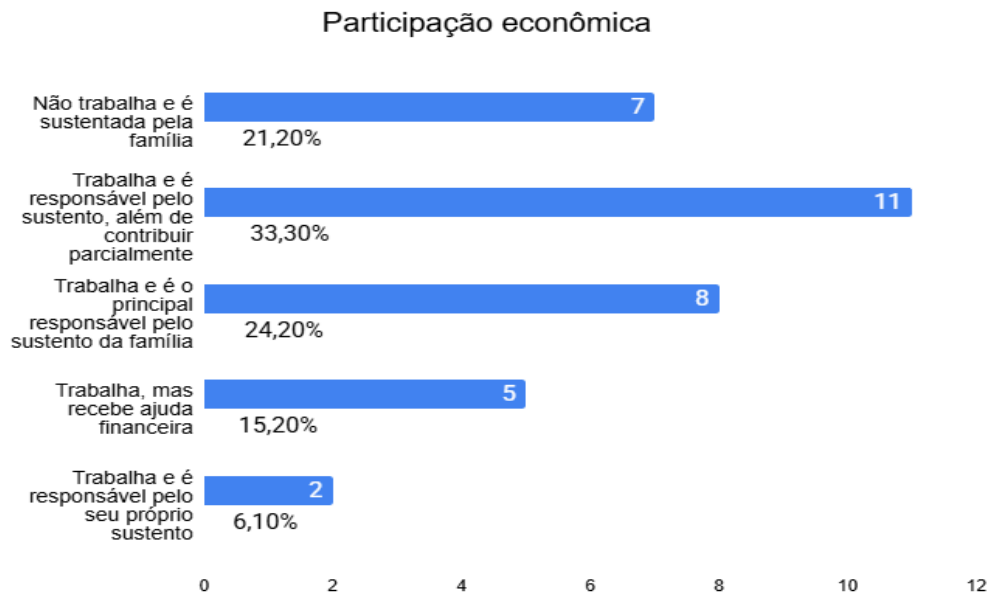
Fonte: Próprio autor, 2026.

O atributo "cursou outro nível superior" passou de um cenário de equilíbrio nos dados reais (51,5% (Sim) e 48,5% (Não)), como apresentado no Gráfico 4, para uma nítida assimetria após a imputação pela moda (30,0% (Sim) e 70,0% (Não)) como destacado também no mesmo gráfico. Essa alteração ocorreu porque, nos dados reais, a categoria "Não" era a mais frequente e, por esse motivo, foi utilizada

para preencher os valores ausentes presentes no conjunto de dados com imputação pela moda, ampliando sua representatividade. Comportamento semelhante foi observado no atributo "filhos", que apresentava uma distribuição relativamente equilibrada nos dados reais (57,6% sem filhos e 42,4% com filhos), mas passou a apresentar uma distribuição assimétrica após a imputação (75,0% sem filhos e 25,0% com filhos). Situação semelhante também foi identificada no atributo "participação econômica", no qual a categoria predominante foi alterada em decorrência da imputação pela moda.

Dando continuidade à análise, outro caso observado refere-se à mudança da categoria predominante no atributo "participação econômica". No conjunto de dados composto apenas por valores reais, uma categoria apresentou maior representatividade. Entretanto, após a aplicação da imputação pela moda, a categoria predominante foi alterada. Essa diferença é observada nos Gráficos 5 e 6.

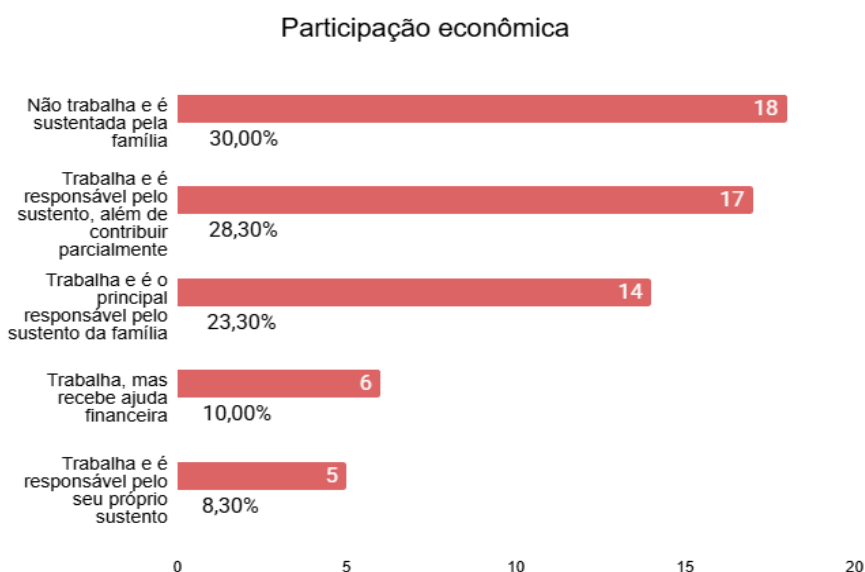
Gráfico 5 - Distribuição da participação econômica no conjunto de dados real.



Fonte: Próprio autor, 2026.

Com base no Gráfico 5, observa-se que a categoria predominante é a "trabalha e contribui parcialmente para a família" (33,3% no conjunto de dados real) e depois, analisando o Gráfico 6, passou a ser "não trabalha e é sustentado pela família" (30,0% no conjunto de dados com imputação pela moda).

Gráfico 6 - Representação da participação econômica no conjunto de dados com imputação pela moda.



Fonte: Próprio autor, 2026.

Essa alteração observada quando comparamos os dois gráficos ocorreu porque, embora o atributo "participação econômica" estivesse preenchido em determinadas instâncias, essas mesmas instâncias continham valores ausentes em outros atributos e, por esse motivo, foram excluídas do conjunto de dados real. Com a imputação dos valores ausentes pela moda, essas instâncias passaram a integrar novamente o conjunto de dados, aumentando a quantidade de informações disponíveis e tornando a categoria "não trabalha e é sustentado pela família" a mais representativa. Ainda, ocorreu uma pequena revitalização do grupo, com redução da idade média de 29,8 para 27,6 anos e da mediana de 30 para 25 anos.

Esses resultados indicam que os atributos secundários e mais sensíveis, relacionados à constituição familiar, à trajetória econômica e ao histórico acadêmico prévio, são justamente os mais vulneráveis à imputação, uma vez que, por não apresentarem uma categoria dominante robusta, absorveram de forma desproporcional os valores imputados, reforçando artificialmente a categoria mais frequente e reduzindo a variabilidade interna dos dados.

Nesse sentido, o perfil do aluno concluinte do curso caracteriza-se como um indivíduo do sexo masculino (81,8% nos dados reais e 85,0% nos dados com imputação pela moda), autodeclarado pardo (63,6% e 61,7%), solteiro (63,6% e

71,7%) e oriundo de escola pública (57,6% e 66,7%). Quanto às condições de moradia e socioeconômicas, esse egresso reside com os pais (63,6% e 65,0%), em moradia própria (84,8% e 85,0%), possui renda familiar entre um e três salários mínimos (69,7% e 71,7%) e dispõe de acesso praticamente universal ao computador (97,0% e 98,3%). Predominantemente sem filhos (57,6% nos dados reais), apresenta idade média entre 27,6 e 29,8 anos. Esses atributos mantiveram-se estáveis em ambos os conjuntos de dados, demonstrando que constituem as dimensões centrais e consistentes na caracterização do perfil predominante do concluinte, independentemente da técnica de tratamento de dados adotada.

Os efeitos observados nos conjuntos de dados relacionam-se diretamente à natureza da técnica empregada, pois a imputação pela moda, ao substituir sistematicamente os dados ausentes pelo valor de maior frequência, tende a intensificar a assimetria da distribuição, reduzir a representatividade das categorias minoritárias e, em casos-limite, alterar a própria categoria modal dos atributos originalmente equilibrados. Como implicação, conclui-se que a análise comparativa é confiável para a identificação do perfil predominante em suas dimensões centrais, que se mantêm estáveis independentemente da técnica de tratamento adotada. Entretanto, deve ser interpretada com cautela no que se refere aos atributos submetidos à imputação pela moda dos valores nulos, sob o risco de conduzir a inferências enviesadas caso o conjunto imputado seja analisado isoladamente.

Como limitações deste estudo, reconhece-se o reduzido tamanho da amostra do conjunto de dados real, em razão da exclusão das instâncias que apresentavam valores ausentes na análise realizada dos dados dos alunos concluintes. Essa redução na quantidade de alunos analisados, nesse caso dos alunos que concluíram o curso, impossibilitou a validação da hipótese H3, uma vez que comprometeu a representatividade da amostra e a robustez das análises realizadas, por se tratar de uma quantidade pequena para ser analisada. Além disso, essa limitação aumenta a sensibilidade dos resultados ao processo de imputação. Destaca-se, ainda, o próprio caráter da técnica de imputação pela moda, que privilegia os valores mais frequentes em detrimento da heterogeneidade dos dados, reduzindo a representatividade de características menos recorrentes no conjunto de dados.

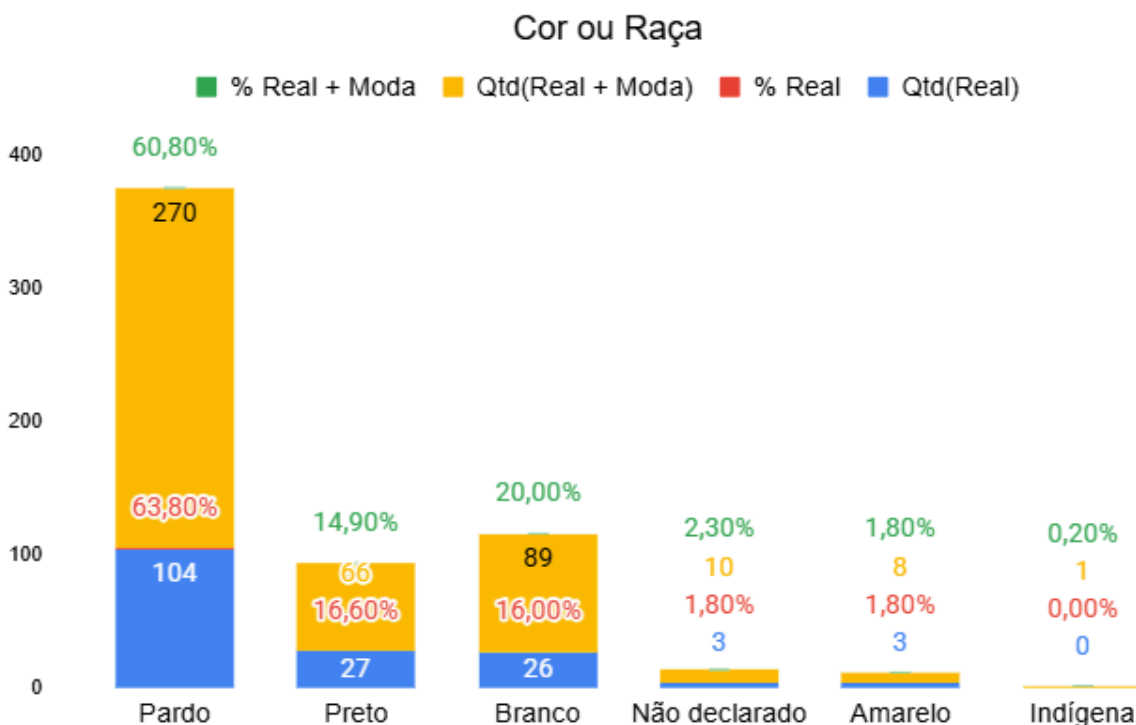
Recomenda-se, portanto, que estudos futuros contemplem a comparação com técnicas alternativas de imputação, a fim de mensurar a robustez das inferências obtidas e mitigar o viés introduzido no tratamento dos valores ausentes.

5.3.2. Comparação do perfil do aluno que não conclui o curso

Com os mesmos conjuntos de dados utilizados no tópico anterior, foi realizada a comparação entre os dois conjuntos de dados, destacada no Apêndice B. Nesses dois conjuntos, também foi aplicado um filtro na coluna "label" com o valor "Evasão". No conjunto de dados composto apenas por dados reais, denominado [dataset Real](#), foram identificadas 163 instâncias. Esse mesmo procedimento foi realizado no conjunto de dados composto por dados reais acrescidos de valores imputados pela técnica da moda, denominado [dataset Moda](#), que apresentou 444 instâncias. A comparação entre os conjuntos de dados evidencia uma convergência estrutural no perfil do aluno evadido, acompanhada de divergências socioeconômicas pontuais.

Em ambos os conjuntos de dados, o aluno evadido caracteriza-se predominantemente como do sexo masculino (93,3% no conjunto de dados real e 85,4% no conjunto de dados com imputação pela moda) e autodeclarado pardo (63,8% e 60,8%). A comparação desses percentuais da autodeclaração indica que, no conjunto de dados composto apenas por valores reais, a proporção de estudantes autodeclarados pardos é superior à observada no conjunto de dados com imputação pela moda, como destaca no Gráfico 7.

Gráfico 7 - Resultado da comparação dos dois conjuntos de dados com base.



Fonte: Próprio autor, 2026.

Essa diferença decorre da disparidade no número de instâncias entre os dois conjuntos de dados, uma vez que, no conjunto real, 104 dos 163 estudantes são autodeclarados pardos, enquanto, no conjunto com imputação pela moda, 270 dos 444 estudantes pertencem a essa categoria. Embora o número absoluto de estudantes autodeclarados pardos seja maior no conjunto com imputação, o maior número de instâncias faz com que sua representatividade percentual seja menor.

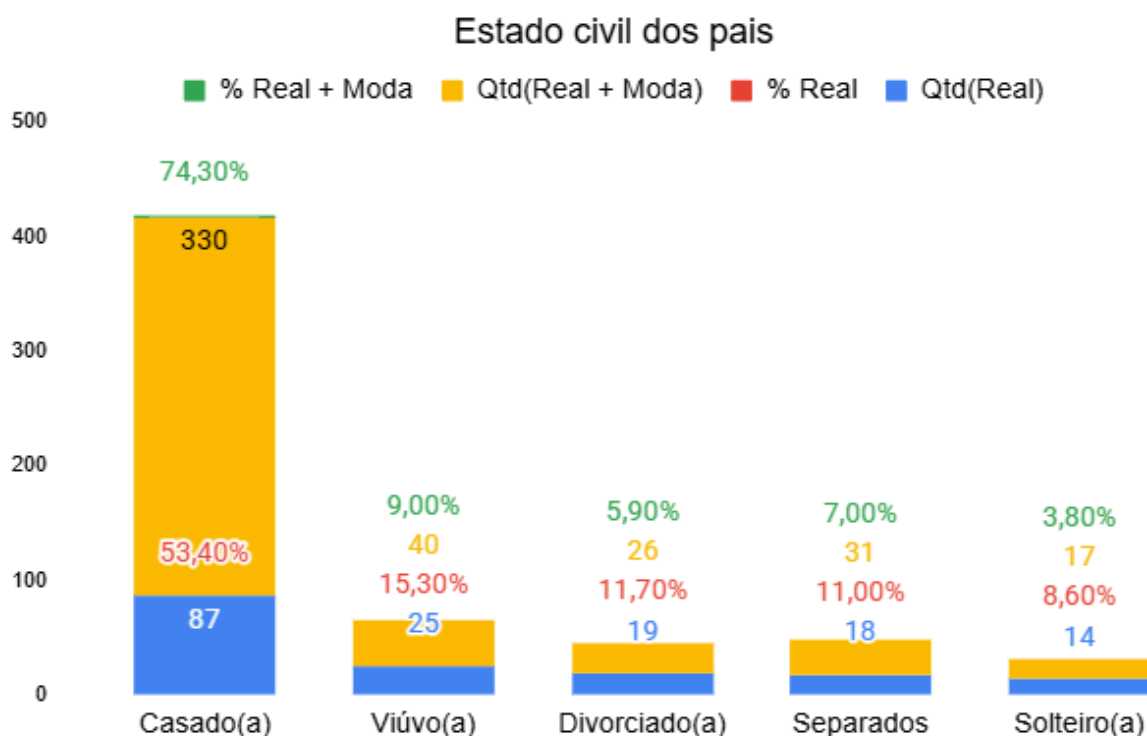
Esse comportamento também é observado na categoria "preto", na qual a quantidade de estudantes no conjunto de dados real é inferior à do conjunto de dados com imputação pela moda. No entanto, o percentual dessa categoria é maior no conjunto de dados real.

As demais características predominantes do aluno evadido correspondem a jovens, solteiros (67,5% no conjunto de dados real e 74,3% no conjunto de dados com imputação pela moda), oriundos de escola pública (75,5% em ambos os conjuntos), residentes em moradia própria (82,2% e 81,5%) e com renda familiar entre um e três salários mínimos (64,4% e 65,3%). Os dados correspondentes a essas variáveis encontram-se apresentados no Apêndice B.

A estabilidade desses atributos indica que os valores imputados coincidiram com a tendência real dominante e que as conclusões estruturais da pesquisa são robustas e pouco sensíveis ao processo de imputação. Esse comportamento é esperado precisamente porque a moda reproduz a categoria mais frequente, de modo que, em variáveis já fortemente concentradas na realidade, a imputação apenas ratifica o padrão preexistente.

Em contrapartida, atributos que originalmente apresentavam maior dispersão de dados faltantes sofreram deslocamentos expressivos, superiores a dez pontos percentuais, revelando uma consequência da técnica empregada: o reforço artificial da categoria modal em detrimento das categorias minoritárias, como analisado o atributo "estado_civil_dos_pais" no Gráfico 8.

Gráfico 8 - Distribuição das quantidades em cada conjunto de dados.



Fonte: Próprio autor, 2026.

Nesse atributo, a categoria "casado" passou de 53,4% para 74,3%, evidenciando que a imputação pela moda ocasionou uma disparidade mais acentuada quando comparada aos demais atributos. Esse comportamento também foi observado em outros atributos, como a formação de nível médio não profissionalizante, que passou de 74,2% para 87,2%, e a ausência de curso superior anterior, que aumentou de 60,7% para 81,1%.

O caso mais crítico, contudo, é o da participação econômica, em que a categoria modal foi invertida, ocorrendo de forma parecida no tópico anterior, passando de "trabalha e é o principal responsável pelo sustento da família" (29,4% no conjunto real) para "não trabalha e é sustentado pela família" (44,8% no conjunto imputado), evidenciando o impacto da imputação pela moda sobre a interpretação socioeconômica desse perfil. Essas alterações indicam que a imputação pela moda, ao atribuir sistematicamente o valor mais frequente às lacunas, reduz a variabilidade e homogeneiza a distribuição dos dados. Como consequência, não apenas amplia a frequência da categoria modal, mas também reconfigura a própria interpretação do

fenômeno, deslocando a imagem do aluno evadido de um sujeito economicamente ativo e sobrecarregado para um sujeito dependente e financeiramente sustentado.

Assim, com toda essa análise realizada, o perfil do aluno que não conclui o curso caracteriza-se, de forma consistente nos dois conjuntos de dados, predominantemente como do sexo masculino (93,3% no dataset real e 85,4% no imputado), autodeclarado pardo (63,8% e 60,8%), jovem e solteiro (67,5% e 74,3%), oriundo de escola pública (75,5% em ambos), residente em moradia própria (82,2% e 81,5%) e com renda familiar de um a três salários mínimos (64,4% e 65,3%). Seus pais são, majoritariamente, casados (53,4% no conjunto real e 74,3% no imputado), possui formação de nível médio não profissionalizante (74,2% e 87,2%) e não realizou curso superior anterior (60,7% e 81,1%).

Infere-se, portanto, que a técnica da moda mostrou-se adequada para preservar o perfil demográfico geral dos alunos evadidos e para viabilizar o aumento do tamanho amostral, fortalecendo a validade estatística da estrutura do conjunto de dados. Todavia, ela introduz um viés de concentração que limita a precisão das inferências socioeconômicas mais refinadas, sobretudo aquelas relativas à situação de trabalho, à configuração familiar e ao histórico educacional, nas quais a distribuição real é mais heterogênea.

A principal implicação metodológica é que a interpretação dos resultados deve ocorrer em dois níveis distintos: com maior confiabilidade para as características demográficas centrais, que se mantêm consistentes entre os datasets, e com maior cautela para as dimensões econômicas e educacionais secundárias, cujas conclusões ter sido artificialmente influenciadas pela imputação. Reconhece-se, assim, como limitação do estudo, o fato de que a homogeneização promovida pela moda subestimar a diversidade real da população evadida, recomendando-se, em pesquisas futuras, o emprego complementar de técnicas de imputação múltipla ou baseadas em modelos preditivos, capazes de preservar melhor a variabilidade natural dos dados.

Como parte do produto final da construção do conjunto de dados, visando à sua utilização em pesquisas futuras, este trabalho também realizou, durante os experimentos, a etapa de codificação dos atributos utilizando a técnica *LabelEncoder* e a transformação dos dados por meio da aplicação das técnicas

StandardScaler, para padronização, e *MinMaxScaler*, para normalização, nos mesmos conjuntos de dados compostos por valores reais e por valores reais acrescidos da imputação pela moda.

Como resultado desses experimentos, foram gerados quatro conjuntos de dados que estão passíveis de utilização como base de dados para pesquisas futuras no treinamento, teste e validação de modelos. Os quatro conjuntos de dados são: [dataset_valoresReais_Multiclasse_LabelEncoder_MinMaxScaler_label_binario](#), [dataset_valoresModa_Multiclasse_LabelEncoder_MinMaxScaler_label_binario](#), [dataset_valoresReais_Multiclasse_LabelEncoder_StandardScaler_label_binario](#) e [dataset_valoresModa_Multiclasse_LabelEncoder_StandardScaler_label_binario](#).

Dessa forma, com a disponibilização desses conjuntos de dados, a presente pesquisa contribui para estudos futuros, fornecendo uma base de dados que está apta à utilização no treinamento, teste e validação de modelos desenvolvidos em novas pesquisas.

6. CONSIDERAÇÕES FINAIS

Conclui-se que, com os resultados obtidos após a realização de todas as etapas da pesquisa, foram construídos não apenas um, mas diferentes conjuntos de dados para a análise do perfil dos alunos, utilizando técnicas distintas de tratamento dos dados. Todos os conjuntos foram estruturados por meio da modelagem bidimensional, que contribuiu para uma melhor compreensão da relação entre os atributos e as instâncias, possibilitando uma análise mais direta e detalhada do perfil dos alunos, confirmando a hipótese H4. Por meio desses conjuntos de dados, foi possível compreender os fatores socioeconômicos e acadêmicos que influenciam a conclusão ou a evasão dos estudantes no curso. Os resultados evidenciaram informações relevantes que estão disponíveis para servir de base a coordenação do curso na elaboração de políticas institucionais voltadas ao acompanhamento e ao suporte dos alunos com perfil de evasão, contribuindo para a implementação de ações que visem mitigar a evasão acadêmica e favorecer a permanência e a conclusão dos estudantes.

Com a análise dos conjuntos de dados, também foi possível perceber o quanto a aplicação da técnica de imputação pela moda nos valores nulos agravou a disparidade em alguns atributos, como "cursou_outro_nivel_superior",

"participacao_economica", "formacao_de_nivel_medio" e "estado_civil_dos_pais". Esses atributos apresentaram diferenças significativas quando realizada a comparação entre os conjuntos de dados (Real × Moda).

Com base nas disparidades apresentadas pelos conjuntos de dados, este trabalho também apresenta algumas limitações. A primeira refere-se ao acesso aos dados, que não foram disponibilizados após solicitação via e-mail, o que implicou a necessidade de um processo manual de coleta e ocasionou atrasos nas demais etapas do desenvolvimento da pesquisa. Outra limitação foi a presença de dados inconsistentes durante todo o processo de extração das duas bases de dados, SUAP e QAcadêmico, que continham muitos valores nulos, tornando necessária a aplicação das técnicas de pré-processamento utilizadas. Além disso, observou-se uma quantidade reduzida de dados disponíveis desde a criação do curso, considerando a primeira turma (2012) até o período delimitado pela pesquisa, correspondente à última turma analisada (2024). Essa limitação resultou em uma análise baseada em um número reduzido de registros, principalmente em relação aos alunos concluintes. Nos conjuntos de dados, foram identificados apenas 33 alunos concluintes no conjunto contendo somente dados reais e 60 no conjunto com dados reais e valores imputados. Em relação aos alunos evadidos, foram identificadas 163 instâncias no conjunto contendo apenas dados reais e 444 no conjunto com dados reais e valores imputados. Assim, considerando ambos os perfis, a pesquisa foi realizada com um total de 504 instâncias, evidenciando uma quantidade limitada de dados para a realização das análises. Essas foram as principais limitações e implicações identificadas durante a execução da pesquisa.

Nesse sentido, recomenda-se que trabalhos futuros avancem para as etapas subsequentes do processo de KDD, empregando os conjuntos de dados produzidos nesta pesquisa no treinamento, teste e validação de modelos preditivos de evasão. Além disso, sugere-se a comparação entre diferentes técnicas de tratamento de valores ausentes, como a imputação múltipla ou métodos baseados em modelos preditivos, capazes de preservar melhor a variabilidade natural dos dados. Dessa forma, esta pesquisa contribui para uma compreensão mais aprofundada da evasão no curso de Licenciatura em Computação do IFPI-CATZS, fornecendo subsídios para que a instituição transforme dados em conhecimento e, conseqüentemente, em ações efetivas de enfrentamento à evasão acadêmica.

REFERÊNCIAS

- ASHA, V. Evaluation of ML Models using SMOTE and Feature Scaling for Intrusion Detection System(IDS), 04 mar. 2025, Tirunelveli. Anais eletrônico: International Conference on Visual Analytics and Data Visualization(ICVADV). Índia: IEEE, 2025. p. 243-249. Disponível em: <https://ieeexplore.ieee.org/document/10961506>. Acesso em: 16 jun. 2026.
- Brasil. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República; 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Acesso em 17 set. 2025.
- BUSSAB, Wilton de Oliveira; MORETTIN, Pedro. Alberto Estatística Básica. 9º. ed. São Paulo: Saraiva, 2017.
- CANCHEN, Li. Preprocessing Methods and Pipelines of Data Mining: An Overview. 7 p. 20 jun. 2019 . DOI: <https://doi.org/10.48550/arXiv.1906.08510>.
- CRUVINEL , H. H. Previsão de Ações com Modelo Híbrido LSTM e Algoritmo Genético. 22 jun. 2024, 12 f. TCC (Graduação) - Curso de Ciência da Computação, Pontifícia Universidade Católica de Goiás, Goiás, 2024. Disponível em: <https://repositorio.pucgoias.edu.br/jspui/handle/123456789/7900>. Acesso em: 16 jun. 2026.
- DIGIAMPIETRI, Luciano Antonio; NAKANO, Fabio; LAURETTO, Marcelo de Souza. Mineração de Dados para Identificação de Alunos com Alto Risco de Evasão: Um Estudo de Caso. Revista de Graduação da Universidade de São Paulo, vol. 1, n. 1, p. 17 - 23, 18 jul. 2016. DOI: <https://doi.org/10.11606/issn.2525-376X.v1i1p17-23>.
- FAYYAD, Usama; PIATETSKY-SHAPIO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI Magazine, v. 17, n. 3, p. 37 - 54, 15 mar. 1996. DOI: <https://doi.org/10.1609/aimag.v17i3.1230>.
- GIL, Raul Sidnei. Como elaborar projetos de pesquisa. 4 ed. São Paulo: Atlas, 2002. Disponível em: https://files.cercomp.ufg.br/weby/up/150/o/Anexo_C1_como_elaborar_projeto_de_pesquisa_-_antonio_carlos_gil.pdf. Acesso em: 20 jan. 2025.
- GONÇALVES, Eduardo Corrêa. Pandas Python: Data Wrangling para Ciências de Dados. Casa do Código, 235 p. 2020. Disponível em: [https://www.kufunda.net/publicdocs/Pandas%20Python%20Data%20Wrangling%20Para%20Ciencia%20De%20Dados%20\(Eduardo%20Corr%C3%AAa\).pdf](https://www.kufunda.net/publicdocs/Pandas%20Python%20Data%20Wrangling%20Para%20Ciencia%20De%20Dados%20(Eduardo%20Corr%C3%AAa).pdf). Acesso em: 20 jan. 2025.
- INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ. Comitê de Ética em Pesquisa com Seres Humanos (CEP). Piauí: Site IFPI, 2018. Disponível em: <https://www.ifpi.edu.br/cep>. Acesso em: 1 abr. 2025.
- INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ. Regimento do Comitê de Ética em Pesquisa com Seres Humanos do Instituto Federal de Educação, Ciência e Tecnologia do Piauí. Teresina: Site IFPI, abr. 2018.

Disponível em: https://www.ifpi.edu.br/cep/arquivos/cep_regimentointerno.pdf. Acesso em: 1 abr. 2025.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO NORTE. Sistema Unificado de Administração Pública (SUAP). Rio Grande do Norte: Portal IFRN, 2023. Disponível em: <https://portal.ifrn.edu.br/institucional/tecnologia-da-informacao/servicos/suap/>. Acesso em: 25 fev. 2025.

JUPYTER. Jupyter: Software livre, padrões abertos e serviços web para computação interativa em todas as linguagens de programação. 2025. Disponível em: <https://jupyter.org/>. Acesso em: 25 fev. 2025.

KANTORSKI, G. Z.; MARTINS, R. Z.; BALEJO, A.; F. M. Mineração de dados educacionais para predição da evasão em cursos de graduação presenciais no ensino superior. CBIE 2023 - Congresso Brasileiro de Informática na Educação, Rio Grande do Sul. p. 1133-1142, nov. 2023. DOI: <https://doi.org/10.5753/sbie.2023>.

LIU, Huan; MOTODA, Hiroshi. Computational Methods of Feature Selection. Chapman & Hall/CRC data mining and knowledge discovery. p. 355 - 376. 2008. Disponível em: <https://lings.org/assets/resources/islamaj-inbook08.pdf>. Acesso em 20 fev. 2025.

MANHÃES, L. M. B.; CRUZ, S. D.; COSTA, R. J. M.; ZAVALITA, J.; ZIMBRÃO, G. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. Anais do XXII SBIE-XVII WIE, Aracaju. p. 150 - 159. 21 a 25 nov. 2011. DOI: <https://doi.org/10.5753/cbie.sbie.2011.%25p>.

OLIVEIRA, Isleimar de Souza. Análise de dados aplicada à evasão escolar: um estudo de caso do IFPB. 2023. 135 f. Dissertação (Mestrado em Tecnologia da Informação) - Instituto Federal de Educação, Ciência e Tecnologia Paraíba, Programa de Pós-Graduação em Tecnologia da Informação, Paraíba, 2023.

PAZ, Fábio Josende, CAZELLA, Silvio César. Identificando o perfil de evasão de alunos de graduação através da Mineração de dados Educacionais: um estudo de caso de uma Universidade Comunitária. VI Congresso Brasileiro de Informática na Educação, 6., 2017, Porto Alegre. Anais do WCBIE, Porto Alegre: ResearchGat, 2017. p. 624 - 633. DOI: <https://doi.org/10.5753/cbie.wcbie.2017.624>.

Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020). Study the Influ-ence of Normalization/Transformation process on the Accuracy of Supervised Classifica-tion [GS Search]. 2020 Third International Conference on Smart Systems and InventiveTechnology (ICSSIT), 729–735. <https://doi.org/10.1109/ICSSIT48917.2020.9214160>.

SANTOS, Georgina Gonçalves dos; SILVA, Lélia Custódio da. A evasão na educação superior: entre debate social e objeto de pesquisa. Salvador: EDUFBA, 2011, p. 249-262. Disponível em: <https://books.scielo.org/id/n656x/pdf/sampaio-9788523212117-14.pdf>. Acesso em: 17 dez. 2025.

SCIKIT-LEARN. MinMaxScaler. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>. Acesso em: 17 jun. 2026.

SCIKIT-LEARN. LabelEncoder. Disponível em:

<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>. Acesso em: 10 jun. 2026.

SILVA FILHO, R. L. L. e; MOTEJUNAS, P. R.; HIPÓLITO, O.; MELO LOBO, M. B. de C. A evasão no ensino superior brasileiro. Cadernos de Pesquisa, São Paulo, v. 37, n. 132, p. 641-659, set./dez. 2007.

TEIXEIRA, Rita Petrarca; MENTGES, Manuir José; KAMPFF, Adriana Justin Cerveira. Evasão no Ensino Superior: Um estudo sistemático. Permanência na Educação Superior. 12 p. out. 2019. URI: <https://hdl.handle.net/10923/15080>.

VICENTAINER, Denis; MATTEDI, Marcos; MELLO, Bruno. Aplicação das Bibliotecas Python para tratamento de dados em tempo real: A análise dos dados de isolamento social em Santa Catarina. Fundação Universidade Regional de Blumenau. p. 206 - 217. out. 2020. DOI: <https://doi.org/10.21166/metapre.v3i0.1392>.

APÊNDICE A – TABELA COMPARATIVA DOS ATRIBUTOS DOS CONJUNTOS DE DADOS REAL E COM IMPUTAÇÃO PELA MODA DOS ALUNOS FORMADOS.

Atributo	Variáveis	Quant(Real)	% Real	Quant(Moda)	% Moda
sexo	Masculino	27	81,8%	51	85,0%
	Feminino	6	18,2%	9	15,0%
cor_raca	Pardo	21	63,6%	37	61,7%
	Branco	5	15,2%	13	21,7%
	Preto	5	15,2%	8	13,3%
	Amarelo	2	6,1%	2	3,3%
portador_de_necessidade_especiais	Não	30	90,9%	57	95,0%
	Sim	3	9,1%	3	5,0%
estado_civil	Solteiro(a)	21	63,6%	43	71,7%
	Casado(a)	11	33,3%	16	26,7%
	Divorciado(a)	1	3,0%	1	1,7%
filhos	Não	19	57,6%	45	75,0%
	Sim	14	42,4%	15	25,0%
meio_de_transporte	Ônibus	17	51,5%	35	58,3%
	Carro ou moto	16	48,5%	24	40,0%
	A pé	0	0,0%	1	1,7%
oriundo_escola	Pública	19	57,6%	40	66,7%
	Particular	14	42,4%	20	33,3%
atualmente_reside	Com os pais	21	63,6%	39	65,0%
	Com parentes	8	24,2%	11	18,3%
	Com amigos	2	6,1%	3	5,0%
	Com cônjuge	0	0,0%	3	5,0%
	Sozinho(a)	1	3,0%	2	3,3%
	Casa do Estudante	1	3,0%	1	1,7%
	Casa própria	0	0,0%	1	1,7%
acesso_ao_computador	Sim	32	97,0%	59	98,3%

	Não	1	3,0%	1	1,7%
tipo_de_moradia	Própria	28	84,8%	51	85,0%
	Alugada	3	9,1%	6	10,0%
	Financiada	1	3,0%	1	1,7%
	Cedida	1	3,0%	2	3,3%
estado_civil_dos_pais	Casado(a)	17	51,5%	39	65,0%
	Separados	5	15,2%	8	13,3%
	Divorciado(a)	5	15,2%	5	8,3%
	Viúvo(a)	3	9,1%	4	6,7%
	Solteiro(a)	3	9,1%	4	6,7%
renda_mensal	De 1 a 3 salários	23	69,7%	43	71,7%
	De 3 a 5 salários	7	21,2%	10	16,7%
	Até 1 salário	2	6,1%	5	8,3%
	De 5 a 10 salários	0	0,0%	1	1,7%
	Mais de 10 salários	1	3,0%	1	1,7%
participacao_economica	Não trabalha e é sustentada pela família	7	21,2%	18	30,0%
	Trabalha e é responsável pelo sustento, além de contribuir parcialmente	11	33,3%	17	28,3%
	Trabalha e é o principal responsável pelo sustento da família	8	24,2%	14	23,3%
	Trabalha, mas recebe ajuda financeira	5	15,2%	6	10,0%
	Trabalha e é responsável pelo seu próprio sustento	2	6,1%	5	8,3%
cota	Não	22	66,7%	44	73,3%
	Sim	11	33,3%	16	26,7%
formacao_nivel_medio	Não profissionalizante	30	90,9%	54	90,0%
	Sim, profissionalizante	3	9,1%	6	10,0%
cursou_outro_nivel_superior	Não	17	51,5%	42	70,0%
	Sim	16	48,5%	18	30,0%

idade	Média	29,8	27,6
	Mediana	30	25

APÊNDICE B – TABELA COMPARATIVA DOS ATRIBUTOS DOS CONJUNTOS DE DADOS REAL E COM IMPUTAÇÃO PELA MODA DOS ALUNOS FORMADOS.

Atributo	Variáveis	Quant(Real)	% Real	Quant(Moda)	% Moda
sexo	Masculino	152	93,3%	379	85,4%
	Feminino	11	6,7%	65	14,6%
cor_raca	Pardo	104	63,8%	270	60,8%
	Preto	27	16,6%	66	14,9%
	Branco	26	16,0%	89	20,0%
	Não declarado	3	1,8%	10	2,3%
	Amarelo	3	1,8%	8	1,8%
	Indígena	0	0,0%	1	0,2%
portador_de_necessidade_especiais	Não	154	94,5%	426	95,9%
	Sim	9	5,5%	18	4,1%
estado_civil	Solteiro(a)	110	67,5%	330	74,3%
	Casado(a)	48	29,4%	97	21,8%
	União Estável	2	1,2%	5	1,1%
	Divorciado(a)	2	1,2%	11	2,5%
	Viúvo	1	0,6%	1	0,2%
filhos	Não	105	64,4%	327	73,6%
	Sim	58	35,6%	117	26,4%
meio_de_transporte	Ônibus	87	53,4%	177	39,9%
	Carro ou moto	72	44,2%	252	56,8%
	A pé	3	1,8%	8	1,8%
	Bicicleta	1	0,6%	7	1,6%
oriundo_escola	Pública	123	75,5%	335	75,5%
	Particular	34	20,9%	99	22,3%

	Filantrópica	6	3,7%	10	2,3%
atualmente_resi de	Com os pais	86	52,8%	271	61,0%
	Com parentes	46	28,2%	81	18,2%
	Com amigos	2	1,2%	3	0,7%
	Com cônjuge	1	0,6%	22	5,0%
	Sozinho(a)	27	16,6%	52	11,7%
	Casa do Estudante	1	0,6%	2	0,5%
	Casa própria	0	0,0%	13	2,9%
acesso_ao_com putador	Sim	157	96,3%	401	90,3%
	Não	6	3,7%	43	9,7%
tipo_de_moradia	Própria	134	82,2%	362	81,5%
	Alugada	23	14,1%	52	11,7%
	Financiada	2	1,2%	16	3,6%
	Cedida	4	2,5%	13	2,9%
	Programa de Apartamento Residencial	0	0,0%	1	0,2%
estado_civil_dos _pais	Casado(a)	87	53,4%	330	74,3%
	Viúvo(a)	25	15,3%	40	9,0%
	Divorciado(a)	19	11,7%	26	5,9%
	Separados	18	11,0%	31	7,0%
	Solteiro(a)	14	8,6%	17	3,8%
renda_mensal	De 1 a 3 salários	105	64,4%	290	65,3%
	De 3 a 5 salários	33	20,2%	83	18,7%
	Até 1 salário	11	6,7%	40	9,0%
	De 5 a 10 salários	9	5,5%	21	4,7%
	Mais de 10 salários	5	3,1%	10	2,3%
participacao_eco nomica	Não trabalha e é sustentada pela família	47	28,8%	199	44,8%
	Trabalha e é o principal responsável pelo sustento da família	48	29,4%	99	22,3%
	Trabalha e é responsável pelo sustento, além de contribuir	40	24,5%	77	17,3%

	parcialmente				
	Trabalha e é responsável pelo seu próprio sustento	19	11,7%	53	11,9%
	Trabalha, mas recebe ajuda financeira	9	5,5%	16	3,6%
cota	Não	82	50,3%	270	60,8%
	Sim	81	49,7%	174	39,2%
formacao_nivel_medio	Não profissionalizante	121	74,2%	387	87,2%
	Sim, profissionalizante	42	25,8%	57	12,8%
cursou_outro_nivel_superior	Não	99	60,7%	360	81,1%
	Sim	64	39,3%	84	18,9%
idade	Média	28,3		26,8	
	Mediana	27		25	

formado

Características	% Real	% Moda
Masculino	81,8%	85,0%
Pardo	63,6%	61,7%
Não possui necessidades especiais	90,9%	95,0%
Solteiro(a)	63,6%	71,7%
Não tem filhos	57,6%	75,0%
Ônibus	51,5%	58,3%
Escola pública	57,6%	66,7%
Reside com os pais	63,6%	65,0%
Tem acesso ao computador	97,0%	98,3%
Possui moradia própria	84,8%	85,0%
Os pais são casado(a)	51,5%	65,0%
Renda mensal de 1 a 3 salários	69,7%	71,7%
Participação econômica: Não trabalha e é sustentada pela família	21,2%	30,0%
Participação econômica: Trabalha e é responsável pelo sustento, além de contribuir parcialmente	33,3%	28,3%
Não entrou com cota	66,7%	73,3%
Formação de nível médio não profissionalizante	90,9%	90,0%
Não cursou outro nível superior	51,5%	70,0%
Sim cursou outro nível superior	48,5%	30,0%
Idade média	29,8	27,6

Evadido

Características	% Real	% Moda
Masculino	93,3%	85,4%
Pardo	63,8%	60,8%
Não possui necessidades especiais	94,5%	95,9%
Solteiro(a)	67,5%	74,3%
Não tem filhos	64,4%	73,6%
Ônibus	53,4%	39,9%
Carro ou moto	44,2%	56,8%
Escola pública	75,5%	75,5%
Reside com os pais	52,8%	61,0%
Reside com parentes	28,2%	18,2%
Tem acesso ao computador	96,3%	90,3%
Possui moradia própria	82,2%	81,5%
Os pais são casado(a)	53,4%	74,3%
Renda mensal de 1 a 3 salários	64,4%	65,3%
Não trabalha e é sustentada pela família	28,8%	44,8%
Trabalha e é o principal responsável pelo sustento da família	29,4%	22,3%
Não entrou com cota	50,3%	60,8%
Sim, entrou com cota	49,7%	39,2%
Formação de nível médio não profissionalizante	74,2%	87,2%
Não cursou outro nível superior	60,7%	81,1%
Idade média	28,3	26,8