REVISTA
DESAFIOS
ISSN: 2359-3652

V.12, n.3, ABRIL/2025 – *DOI:* http://dx.doi.org/10.20873/2025_abr_19534

MINERAÇÃO DE DADOS EDUCACIONAIS: ANÁLISE DO PERFIL SOCIOECONÔMICO DOS ALUNOS DO CURSO DE SISTEMAS DE INFORMAÇÃO COM BASE NOS DADOS DO INEP

EDUCATIONAL DATA MINING: SOCIOECONOMIC PROFILE ANALYSIS OF INFORMATION SYSTEMS STUDENTS BASED ON INEP DATA

MINERÍA DE DATOS EDUCATIVOS: ANÁLISIS DEL PERFIL SOCIOECONÓMICO DE LOS ESTUDIANTES DEL CURSO DE SISTEMAS DE INFORMACIÓN BASADO EN LOS DATOS DEL INEP

ADRIANO VAZ DE ALMEIDA

Mestre em Estudos Antrópicos na Amazônia pelo Programa de Pós-graduação em Estudos Antrópicos na Amazônia (PPGEAA). Universidade Federal do Pará (UFPA). E-mail: adriano.vaz@ufra.edu.br | Orcid.org/0009-0009-2131-5953

SANDIO MACIEL DOS SANTOS

Doutorando em Engenharia Elétrica pelo Programa em Engenharia Elétrica (PPGEE). Universidade Federal do Pará (UFPA). E-mail: sandio.santos@castanhal.ufpa.br | Orcid.org/0000-0002-9487-4977

YOMARA PINHEIRO PIRES

Professora no Programa de Pós-graduação em Estudos Antrópicos na Amazônia (PPGEAA). Universidade Federal do Pará (UFPA). E-mail: yomara@ufpa.br | Orcid.org/0000-0001-7724-6082

MARCOS CÉSAR DA ROCHA SERUFFO

Professor no Programa de Pós-graduação em Estudos Antrópicos na Amazônia (PPGEAA). Universidade Federal do Pará (UFPA). E-mail: seruffo@ufpa.br | Orcid.org/0000-0002-8106-0560

REVISTA DESAFIOS ISSN: 2359-3652

RESUMO:

No contexto educacional atual, a investigação de variáveis relacionadas às diferenças socioeconômicas entre estudantes do ensino superior contribui para a compreensão de fatores relacionados ao sucesso acadêmico. Assim, este trabalho aborda o processo de análise do perfil socioeconômico dos alunos do curso de Sistema de Informação (SI) da Universidade Federal do Pará (UFPA) utilizando a Mineração de Dados Educacionais (MDE) na base de dados pública disponibilizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) (INEP, 2021) referente ao Exame Nacional de Desempenho de Estudantes (ENADE) (ENADE, 2021) realizado em 2021, visando diagnosticar e compreender as características socioeconômicas desses estudantes e bem como identificar os fatores intra e extraescolares que podem influenciar a permanência e finalização do referido curso de graduação. A metodologia adotada engloba a MDE como processo de descoberta de conhecimento em base de dados. Dentre as informações processadas, verificou-se a predominância do gênero masculino nos cursos ofertados; constatou-se que a maioria dos estudantes não trabalha, reside e depende financeiramente dos pais, não receberam bolsa científica e estudaram todo o ensino médio em escola pública.

PALAVRAS-CHAVE: Perfil Socioeconômico, Mineração de Dados Educacionais (MDE), Exame Nacional de Desempenho de Estudantes (ENADE).

ABSTRACT:

In the current educational context, investigating variables related to socioeconomic differences among higher education students is crucial for understanding factors associated with academic success. This study analyzes the socioeconomic profile of students in the Information Systems course at the Federal University of Pará (UFPA) using Educational Data Mining (EDM) based on the publicly available dataset from the 2021 National Student Performance Exam (ENADE) (ENADE, 2021), provided by the National Institute for Educational Studies and Research Anísio Teixeira (INEP) (INEP, 2021). The aim is to diagnose and understand the socioeconomic characteristics of these students, identifying both in-school and out-of-school factors that may influence their continuation and completion of the undergraduate course. The results showed that the majority of students in the analyzed courses are male. Additionally, most students do not work, live with their parents, and are financially dependent on them. They have not received scientific scholarships and attended public schools throughout their high school education. These findings are essential for understanding students' contexts and the challenges they face during their university studies.

KEYWORDS: *Socioeconomic Profile, Educational Data Mining (EDM), National Student Performance Exam (ENADE).*

Artigo Submetido em: 23/05/2024 - Aceito em: 27/03/2025 - Publicado em 30/04/2025

RESUMEN:

En el contexto educativo actual, la investigación de variables relacionadas con las diferencias socioeconómicas entre los estudiantes de educación superior contribuye a la comprensión de los factores asociados al éxito académico. Así, este estudio aborda el proceso de análisis del perfil socioeconómico de los estudiantes del curso de Sistemas de Información (SI) de la Universidad Federal de Pará (UFPA), utilizando la Minería de Datos Educacionales (MDE) en la base de datos pública proporcionada por el Instituto Nacional de Estudios y Investigaciones Educativas Anísio Teixeira (INEP) (INEP, 2021) referente al Examen Nacional de Desempeño de Estudiantes (ENADE) (ENADE, 2021), realizado en 2021. El objetivo principal es diagnosticar y comprender las características socioeconómicas de estos estudiantes, así como identificar los factores intra y extraescolares que pueden influir en su permanencia y finalización del mencionado curso de grado. La metodología adoptada incluye la MDE como un proceso de descubrimiento de conocimiento en bases de datos. Entre los datos procesados, se observó una predominancia del género masculino en los cursos ofrecidos; se constató que la mayoría de los estudiantes no trabaja, reside y depende económicamente de sus padres, no ha recibido becas científicas y cursó toda su educación secundaria en escuelas públicas.

Palabras clave: *Perfil Socioeconómico, Minería de Datos Educacionales (MDE), Examen Nacional de Desempeño de Estudiantes (ENADE).*

REVISTA DESAFIOS ISSN 2359-3652

INTRODUÇÃO

O perfil socioeconômico dos estudantes de graduação é essencial para compreender as dinâmicas educacionais e planejar ações que promovam equidade e qualidade no ensino superior, especialmente no Brasil, sobre fatores econômicos, sociais e culturais que influenciam a trajetória acadêmica, desde a permanência até a conclusão dos cursos pelos estudantes. Nesse cenário, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), vinculado ao Ministério da Educação, desempenha um papel estratégico ao coletar e divulgar indicadores educacionais por meio de avaliações, exames, pesquisas estatísticas e gestão do conhecimento na área.

Por meio da aplicação de exames e questionários, o INEP gera um volume considerável de dados, os quais são disponibilizados ao público por meio de uma base de dados (microdados) em seu *site*. Esses dados, contêm informações detalhadas e podem ser acessados e utilizados por gestores, pesquisadores, educadores e pela sociedade em geral para diversos fins, contribuindo para a análise e a melhoria da educação no Brasil.

No que se refere à avaliação do ensino superior no país, ou seja, observação do desempenho dos estudantes dos cursos de graduação (bacharelados, licenciaturas e tecnólogos), o INEP aplica o Exame Nacional de Desempenho de Estudantes (ENADE), objeto de estudo deste trabalho e que visa:

> "Avaliar o rendimento dos concluintes dos cursos de graduação em relação aos conteúdos programáticos previstos nas diretrizes curriculares dos cursos, o desenvolvimento de competências e habilidades necessárias ao aprofundamento da formação geral e profissional, e o nível de atualização dos estudantes com relação à realidade brasileira e mundial" (BRASIL, 2024).

Dessa forma, as avaliações realizadas pelo INEP, a exemplo do ENADE, além de mensurarem o aprendizado dos conteúdos propostos, procuram obter informações socioeconômicas dos estudantes por meio de aplicação de testes e questionários que geram um grande volume de dados e estes, por sua vez, são armazenados em diferentes arquivos contendo abundância de informações que podem ser utilizadas como ferramenta para auxiliar no processo de tomada de decisão educacional.

Artigo Submetido em: 23/05/2024 - Aceito em: 27/03/2025 - Publicado em 30/04/2025

Em relação à importância dessas bases de dados, Araújo (2005) destaca que “os dados produzidos a partir da avaliação educacional podem subsidiar, de forma efetiva, ações em prol de melhorias na qualidade do aprendizado e das oportunidades educacionais oferecidas à sociedade brasileira” (ARAÚJO, 2005, p. 10). "É notório, portanto, que há um conjunto de informações embutido nos dados coletados pelo INEP que possibilita a compreensão dos fatores que afetam a qualidade do aprendizado" (FONSECA e NAMEN, 2016, p. 3). Para mais, os dados disponibilizados pelo INEP incluem informações socioeconômicas internas e externas à realidade dos estudantes, revelando variáveis que podem influenciar, positiva ou negativamente, sua permanência ou abandono no curso de graduação.

Entretanto, Fonseca e Namen (2016) levantam questões cruciais sobre o aproveitamento dessas bases de dados: esses dados estão efetivamente sendo utilizados para promover melhorias no processo de ensino-aprendizagem? A comunidade educacional tem explorado esse valioso banco de informações para identificar correlações significativas com o desempenho dos estudantes? E, por fim, estão sendo realizadas análises mais aprofundadas que vão além de abordagens descritivas?.

Uma possível resposta a esse último questionamento é apresentada no estudo de Araújo *et al.* (2019), que discute a abordagem adotada nos relatórios anuais divulgados pelo INEP à sociedade. Segundo os autores, esses relatórios se limitam a análises estatísticas descritivas, como a média das notas por região, curso ou instituição. Essas análises são classificadas como métodos que buscam resumir, organizar e descrever os aspectos mais relevantes de um conjunto de características observadas, e, em seguida, realiza comparações entre diferentes conjuntos de dados (ARAÚJO *et al*., 2019). Contudo, não há utilidade no acesso a grandes bases de dados sem a utilização de ferramentas que auxiliem na sua análise e interpretação, explicam Gottardo *et al*. (2014) e Machado *et al.* (2015).

"Diante dos fatos expostos, fica evidente a exigência de que o estudo aprofundado dos dados educacionais se torne uma vertente cada vez mais crescente" (FONSECA e NAMEN, 2016, p. 5). Para atender a essas demandas é

necessário utilizar tecnologia que possa analisar grandes quantidades de dados para obter informações e conhecimentos relevantes (FONSECA e NAMEN, 2016). Nesse contexto, os avanços tecnológicos do século XXI, especialmente na área de Tecnologia da Informação (TI), têm proporcionado novas oportunidades para processar, organizar e analisar esses dados de maneira mais eficiente (MARTUCCI, 2000). Com isso, instituições públicas e privadas passaram a gerar, armazenar e organizar seus dados (imagens, textos, áudios, etc.) de forma mais eficiente (GOLDSCHMIDT *et al*., 2015; TAN *et al*., 2009). Em contrapartida, este aumento na geração de dados também trouxe consigo uma maior complexidade acerca do processo de extração de informações úteis (MITRA e ACHARYA, 2005). Dentre as ferramentas de análise de dados em grande escala, destaca-se a técnica de mineração de dados (MD), que permite extrair conhecimento de um vasto número de registros (ROIGER, 2017), a exemplo de dados relacionados à educação, conhecida como Mineração de Dados Educacionais (MDE).

Sob esse viés, o novo contexto trazido pela revolução tecnológica e pela globalização das relações econômicas e culturais levou as pessoas a reexaminarem os propósitos e os meios da educação para garantir o cultivo de talentos que possam lidar com múltiplos desafios do futuro (MARTUCCI, 2000). Nesta perspectiva, as organizações governamentais em diferentes áreas comprometem-se a tomar medidas para melhorar o ensino e a aprendizagem a todos os níveis, desde a alfabetização até aos níveis profissionais mais elevados. Os sistemas de avaliação que levam ao desenvolvimento de diagnósticos para melhorar a educação pública e a gestão dos recursos disponíveis começam a ocupar um lugar importante na agenda da política educacional (BAUER, 2012; FONSECA e NAMEN, 2016).

Em vista disto, aplicar o processo de descoberta do conhecimento (KDD) e mais especificamente a técnica MDE em bases de dados advindas de contextos educacionais como, por exemplo, o ENADE, visando obter informações detalhadas e estabelecer padrões, mostra-se totalmente viável e necessária. Em relação aos dados analisados neste estudo, é válido ressaltar que os mesmos foram coletados no questionário socioeconômico respondido pelos estudantes no

Artigo Submetido em: 23/05/2024 - Aceito em: 27/03/2025 - Publicado em 30/04/2025

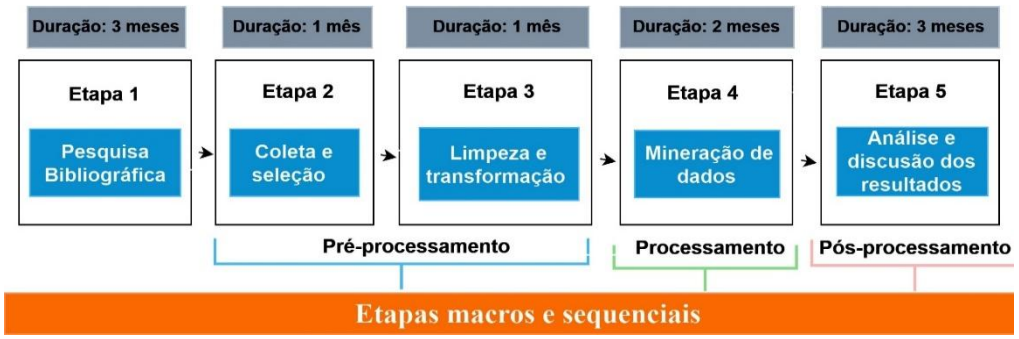
ato da inscrição do ENADE, o qual contém perguntas que captam informações de caráter econômico, cultural e social da realidade de vida deles e bem como do contexto de seus processos formativos, preservando-se o sigilo da identidade dos respondentes.

Diante disto, este trabalho visa a aplicação de técnicas de MDE para análise do perfil socioeconômico dos alunos do curso de Sistema de Informação (SI) da UFPA. Espera-se que o presente trabalho possa contribuir fornecendo entrega de análises personalizadas capazes de auxiliar tanto no processo de tomada de decisão quanto nas medidas de qualidade e estruturação do curso, ou seja, a prática e a gestão educacional. O estudo está divido da seguinte forma: a segunda seção descreve a metodologia empregada no trabalho; já a terceira aborda os resultados e discussão; e por fim, na quarta são apresentadas as considerações finais, evidenciando as limitações, objetivos/resultados alcançados e os trabalhos futuros. A próxima seção trata sobre a identificação das etapas e procedimentos adotados para a execução da pesquisa.

METODOLOGIA

Este trabalho foi desenvolvido conforme o estudo sobre o conhecimento proposto por Fayyad et al. (1996), Goldschmidt e Passos (2015) e Castro e Ferrari (2017). Por conseguinte, a metodologia adotada engloba a aplicação do processo KDD como principal procedimento de extração de conhecimento em base de dados, que tem como um dos objetivos a utilização da informação adquirida para auxiliar no processo de tomada de decisão (SANTOS, 2020). Dessa forma, o percurso metodológico aplicado foi dividido em cinco etapas macros e sequenciais, conforme ilustra a Figura 1.

Figura 1 – Percurso metodológico adotado



Fonte: Elaborada pelo autor, 2024.

A primeira etapa concentrou-se no entendimento e definição dos problemas percebidos no contexto da MD em bases de dados educacionais e a partir disso buscou-se construir a estrutura do trabalho e os objetivos que se deseja alcançar com os resultados da pesquisa. Dessa forma, realizou-se a pesquisa no indexador *Google* acadêmico em busca de materiais dos últimos seis anos (de 2018 a 2023) que trabalham na área da ciência de dados com foco na utilização de técnicas para extração de dados educacionais como, por exemplo, a MDE.

A escolha do *Google* acadêmico deve-se pela sua abrangência e relevância no ambiente acadêmico, amplamente utilizado por pesquisadores e acadêmicos de diversas áreas do conhecimento, sendo a maior fonte de materiais acadêmicos, segundo Longen (2024) conteudista no *site Hostinger* Brasil. Além disso, permite realizar buscas avançadas e filtrar os dados, o que facilita a busca por materiais mais recentes, como foi o caso da pesquisa dos últimos seis anos.

É na etapa de coleta e seleção que se inicia o processo KDD, pois envolve a obtenção e escolha dos dados a serem utilizados na análise. É válido enfatizar que no pré-processamento utilizou-se a base de dados do INEP referente ao ENADE 2021. Sob outra perspectiva, é importante considerar os fatores qualitativos e quantitativos do trabalho, ou seja, dentro do contexto de análise de um total de 125 registros (quantitativo), não é possível garantir a fidelidade dos dados coletados, já que os estudantes que realizaram o exame podem ter respondido o questionário de inscrição de maneira aleatória (qualitativo) e isso influência diretamente na análise real dos atributos inerentes ao perfil socioeconômico e das características intra e extraescolares, em todo caso, ainda

Artigo Submetido em: 23/05/2024 - Aceito em: 27/03/2025 - Publicado em 30/04/2025

assim, são esses os dados oficiais utilizados pelo governo para diagnosticar o ensino superior.

Após isso, se fez necessário realizar a leitura e entendimento da estrutura, dos arquivos, dos atributos e do manual da referida base de dados, a fim de identificar com maior clareza o subconjunto de características comuns e assim formar hipóteses sobre as informações ocultas. Posteriormente a esse momento, empregou-se um *script* desenvolvido na linguagem de programação *Python* juntamente com a intraface de usuário baseado na *Web Jupyter Lab*[1] para acessar a base de dados e selecionar somente os atributos de interesse e consequentemente desconsiderar aqueles que não fazem parte do objeto desta pesquisa e assim foi possível extrair apenas os dados (40 atributos) do curso de SI da UFPA. Os atributos utilizados neste estudo estão referenciados no endereço eletrônico (*GOOGLE DRIVE*, 2024).

Quanto ao processo KDD e suas fases já mencionadas na Figura 1 (etapas 2, 3, 4 e 5), considerando que tanto a aquisição quanto a seleção já foram realizadas nas fases anteriores, a limpeza caracterizou-se pela eliminação dos dados inválidos ou pouco influenciáveis dentro do contexto de análise do trabalho. Já a transformação concentrou-se na conversão do formato (.xls) para o formato (.arff) visando realizar o processamento dos dados. Após isso, utilizou-se a ferramenta de MD *Weka* versão 3.8.6 para a tarefa de agrupamento.

O agrupamento é um: "algoritmo de agrupamento de dados multivariados que utiliza métodos numéricos, objetivando agrupar de maneira automática uma série de dados N em grupos ou *clusters* k de acordo com sua similaridade" (Silva *et al*, 2019). A grande vantagem de se usar o agrupamento refere-se à possibilidade de descrever de forma mais clara e eficiente as características de cada grupo de dados e isso favorece a sua compreensão e a descoberta de relações interessantes entre os atributos (CASSIANO, 2014).

[1] É uma aplicação *WEB* interativa que permite a criação, edição e execução de documentos *Jupyter* de forma colaborativa e eficiente. Disponível gratuitamente, ela oferece uma *intraface* amigável e recursos avançados para análise de dados e desenvolvimento de projetos computacionais.

As escolhas de *Python* e *Weka* para este estudo foram baseadas em suas respectivas eficiências, utilização gratuita, código aberto e ampla aceitação na comunidade de ciência de dados. *Python* é uma escolha popular devido à sua flexibilidade, vasta gama de bibliotecas para análise de dados, facilidade de uso, linguagem intuitiva e capacidade de integração com outras ferramentas. Segundo à análise da *Initiative for Analytics and Data Science Standards* (IADSS) 100% dos cientistas de dados mencionam conhecimento de *Python* em seus perfis do *LinkedIn* (USAMA, 2019).

Quanto ao *Weka*, é uma ferramenta robusta e amplamente utilizada para MD, oferecendo uma variedade de algoritmos de agrupamento e outras técnicas de análise, além de possuir uma *interface* amigável e sua capacidade de lidar com conjuntos de dados de diferentes tamanhos e formatos. Ao longo dos anos se consolidou como a ferramenta de MD mais utilizada em ambiente acadêmico (GONÇALVES, 2011).

Ambas as ferramentas são comumente utilizadas em trabalhos semelhantes a este, devido à sua eficácia comprovada e ampla adoção pela comunidade científica de análise de dados, como explorado no trabalho de Araújo *et al*. (2019) que analisou os microdados do ENADE e propôs uma ferramenta de exploração de dados utilizando a MD para predizer o desempenho dos estudantes. Assim como, foram utilizadas no estudo de Vieira *et al*. (2022) que examinou o desempenho no ENADE dos estudantes concluintes dos cursos de computação das edições do exame realizadas entre 2008 e 2017.

Na confecção deste trabalho, o agrupamento foi parametrizado com três *clusters* de dados usando a técnica do "*Simple K-means*" e a função de distância escolhida foi o método euclidiano. A escolha por três *clusters* se deu pelo fato do curso de SI da UFPA ser ofertado em três municípios paraenses, sendo desejável dessa forma estabelecer o perfil socioeconômico dos alunos de diferentes localidades que apresentem realidades econômicas e sociais específicas, porém, que estudem na mesma instituição de ensino superior.

Levando em consideração as pesquisas realizadas dentro da literatura da MDE e bem como do estudo realizado por Silva *et al*. (2019) que tratou da

Artigo Submetido em: 23/05/2024 - Aceito em: 27/03/2025 - Publicado em 30/04/2025

identificação do perfil docente das instituições de ensino superior avaliado em 2016 no estado do Espírito Santo, optou-se pela utilização do método "*Simple K-means*" bastante usado, tal como em Araújo *et al*. (2019), Vieira *et al*. (2022) e Maia e Fernandes (2021) que usaram o mesmo método para encontrar, calcular e ajustar os centros de proximidade ou similaridade entre os pontos, de acordo com a menor distância entre os membros dos grupos. Kodinariya e Makwana, (2013, p. 5), afirmam que este método "é um dos mais simples algoritmos de aprendizagem não supervisionada que resolvem o conhecido problema de agrupamento". Os autores concluem dizendo que o referido método é sem dúvida o mais popular e o estudo de suas propriedades são de interesse de diferentes áreas (KODINARIYA e MAKWANA, 2013).

No *software Weka* 3.8.6 foi carregado uma base de dados em formato (.arff) e posteriormente aplicado o algoritmo de mineração denominado agrupamento para reunião dos dados através do método "*Simple K-means*". A referida base de dados contém 40 atributos e 125 registros, sendo 46 do município de Belém, 68 de Castanhal e 11 de Cametá.

RESULTADOS E DISCUSSÃO

No processo de MD foram gerados como resultado 3 *clusters e* 9 buscas de melhores características com os atributos definidos na Tabela 1, que estão descritos na etapa de coleta e seleção de dados. Com o intuito de descobrir as características gerais do perfil socioeconômico dos alunos, (gênero, idade, raça, estado civil, turno de estudo, ano do término do ensino médio e ano de início da graduação) houve a necessidade de ignorar atributos não inseridos neste contexto, como, por exemplo, o turno de estudo, se o estudante recebeu bolsa acadêmica, a quantidade de horas de estudo dedicadas ao curso, escolaridade e renda dos pais e dentre outros, visto que, para realizar a mineração dos dados na ferramenta *Weka* é preciso especificar os atributos conforme o objetivo a ser alcançado. Este procedimento foi adotado também nas demais análises, tais como: identificação dos fatores intra e extraescolares.

Após isso, foram realizadas diferentes abordagens analíticas, a exemplo da descoberta sobre o percentual de incidência da variável gênero no curso de SI ofertado pela UFPA. O algoritmo mostrou que apenas o município de Belém contém discentes do gênero feminino, além disso, do total de 125 estudantes, têm-se somente 26 alunas, o que corresponde a 20,8%, evidenciando assim a predominância do gênero masculino nos interiores e no curso em geral, conforme apresentado na Tabela 1.

Tabela 1 – Tipo de gênero dos estudantes do curso de SI da UFPA

Município	**Quantidade**	**Masculino**	**Feminino**
Belém	46 estudantes	43,48%	56,52%
Castanhal	68 estudantes	100%	0%
Cametá	11 estudantes	100%	0%
Total geral	**125 estudantes**		

Fonte: Elaborada pelo autor, 2024.

Nesse mesmo viés, o estudo de Pires *et al*. (2021) sobre o diagnóstico da presença feminina nos cursos de tecnologia da informação (TI) no estado do Pará, revelou que considerando a totalidade de instituições de ensino públicas e privadas no país de 2003 a 2018, aproximadamente 85,5% dos ingressantes são do gênero masculino, e desses, 78,4% são concluintes. Para o gênero feminino, têm-se 14,5% de ingressantes e apenas 21,6% de concluintes, dados estes similares ao encontrado neste trabalho. A partir disso, diferentes discussões podem ser abordadas sobre os fatores que levam a essa discrepância de gênero nos cursos de TI como, por exemplo, questões socioculturais que diferenciam a qualificação da mulher, diferenças salariais, estigmas e dentre outros.

No trabalho de Costa *et al*. (2020) inerente a participação feminina nos cursos de computação em uma universidade no norte do país, as referidas autoras afirmam que os estudos que visam analisar as características das mulheres que se matriculam em cursos de computação destacam a necessidade de compreender as alunas dessa área, a fim de identificar as ações que devem ser tomadas tanto no setor público quanto no privado para garantir a igualdade de

Artigo Submetido em: 23/05/2024 - Aceito em: 27/03/2025 - Publicado em 30/04/2025

oportunidades, sendo fundamental que as mulheres ocupem seu lugar na ciência, principalmente nos cursos de computação (COSTA *et al*., 2020).

Outra análise elaborada foi a definição do perfil socioeconômico dos estudantes considerando as seguintes variáveis: gênero, idade, turno de estudo, ano do término do ensino médio, ano de início da graduação, estado civil e raça. Foram definidos na ferramenta *Weka* 3 *clusters*, onde o primeiro agrupou o registro de 21 alunos (17%), o segundo 55 alunos (44%) e o terceiro 49 alunos (39%). Na interpretação dos dados, foi levado em consideração a maioria dos dados obtidos, onde se verificou que o perfil do aluno que concluiu o curso de SI da UFPA é do gênero masculino, pardo, solteiro, 24 anos, estuda pela manhã ou tarde, iniciou a graduação no ano de 2015 e finalizou o ensino médio em 2009. Para ilustrar a visualização dos resultados, a Tabela 2 mostra o estabelecimento do perfil geral dos alunos em cada *cluster*.

Tabela 2 – Estabelecimento do perfil socioeconômico geral dos alunos (as)

Conteúdos / Atributos analisados	**Cluster 0 (17%)**	**Cluster 1 (44%)**	**Cluster 2 (39%)**
Ano de finalização do ensino médio	2014	2009	2007
Ano de início da graduação	2016	2015	2014
Turno de estudo	Matutino	Matutino	Noturno
Gênero	Masculino	Masculino	Masculino
Idade	29 anos	24 anos	24 anos
Estado civil	Casado (a)	Solteiro (a)	Solteiro (a)
Raça	Pardo (a)	Pardo (a)	Pardo (a)

Fonte: Elaborada pelo autor, 2024.

A dimensão sociodemográfica encontrada no estudo de Mello e Finger, (2020) sobre as diferenças e semelhanças dos perfis dos discentes do curso de Ciência da computação (CC) e Engenharia de *Software* (ES) da Universidade Federal do Pampa *(*UNIPAMPA*)* revelou resultados parecidos ao deste trabalho,

onde o perfil do estudante independente do curso é majoritariamente do gênero masculino, solteiro, não possui filhos e tem menos de 30 anos.

A predominância de estudantes solteiros pode ser influenciada pela média de idade encontrada, onde a maioria deles com menos de 30 anos ainda não tem um relacionamento estável ou opta por adiar o casamento e a formação de uma família para se concentrar em sua carreira. Para mais, o ambiente acadêmico possivelmente é um fator que dificulta a manutenção de relacionamentos de longo prazo. A ausência de filhos entre os estudantes pode ser explicada pelo fato de que muitos estão em uma fase da vida em que estão focados em sua formação acadêmica e profissional, priorizando o investimento de tempo e recursos nessas áreas.

A inclusão da dimensão racial na análise dos resultados deste trabalho é de extrema importância, pois permite identificar possíveis desigualdades e desafios enfrentados por estudantes pertencentes a diferentes grupos étnico-raciais. Segundo dados do Instituto Brasileiro de Geografia e Estatística (IBGE) (IBGE, 2024) referente ao censo 2022, 69,9% dos paraenses se autodeclaram pardos. Dessa forma, o preconceito, o medo e os estigmas associados à identificação como pretos podem influenciar a autodeclaração racial, fazendo com que as pessoas se identifiquem como pardas em vez de pretas. Conforme o Censo da Educação Superior, o Brasil formou, em 2020, 51 mil profissionais da área da computação e da Tecnologia da Informação e comunicação (TIC) (INEP, 2024). Destes, apenas 6% (3.060) identificaram-se como pretos. Essa questão é relevante para a análise dos perfis dos discentes do curso de SI da UFPA, pois, pode haver uma sub-representação de estudantes pretos no curso. A falta de representatividade étnico-racial talvez seja resultado de barreiras sociais e econômicas enfrentadas pelos estudantes pretos, como a falta de acesso à educação de qualidade e oportunidades de qualificação profissional.

Ademais, os microdados do ENADE 2021, contém informações socioeconômicas dos estudantes que podem ser classificadas como intra e extraescolares. Dessa forma, foram selecionados somente os atributos relacionados a vida acadêmica dos estudantes/formandos (Auxílio permanência,

bolsa acadêmica, oferta de programas ou atividades no exterior, acesso a curso de idiomas e rotina de estudo). O agrupamento pré-definido gerou 3 *clusters*, onde o primeiro agrupou o registro de 37 alunos (30%), o segundo 70 alunos (56%) e o terceiro 18 alunos (14%). Os resultados podem ser conferidos na Tabela 3.

Tabela 3 – Resultados encontrados nos atributos intraescolares

Conteúdos / Atributos analisados	**Cluster 0 (30%)**	**Cluster 1 (56%)**	**Cluster 2 (14%)**
Ao longo da trajetória acadêmica, você recebeu algum tipo de auxílio permanência?	Sim	Nenhum	Nenhum
Ao longo da trajetória acadêmica, você recebeu algum tipo de bolsa acadêmica?	Outro tipo de bolsa acadêmica	Nenhuma	Bolsa de iniciação cientifica
Durante o curso de graduação, você participou de programas e/ou atividades curriculares no exterior?	Não	Não	Não
Excetuando-se os livros indicados na bibliografia do seu curso, quantos livros leu neste ano?	De 3 a 5	Nenhum	1 ou 2
Quantas horas por semana, aproximadamente, você dedicou aos estudos, excetuando as horas de aula?	De 4 a 7h	De 1 a 2h	De 4 a 7h
Você teve oportunidade de aprendizado de idioma estrangeiro na Instituição?	Não	Não	Não

Fonte: Elaborada pelo autor, 2024.

Os dados mostram que do total de alunos, 88 (70%) não receberam nenhum tipo de auxílio permanência, ou seja, há indícios que tal fato foi um fator de dificuldade enfrentado pelos alunos (as) principalmente para custear despesas básicas como: alimentação, transporte, moradia, compra de livros e entre outros. Sob outra perspectiva, foi possível perceber que 70 discentes (56%) não receberam nenhum tipo de bolsa acadêmica e isso pode ter comprometido a qualidade da formação profissional e acadêmica dos estudantes, além disso, as bolsas são uma ajuda financeira para custear as despesas.

Outra análise mostra, que 100% dos alunos não participaram de programas e/ou atividades curriculares no exterior e bem como não tiveram oportunidade de aprender outro idioma e com isso perderam a possibilidade de aprimorar o seu currículo, aprender outro idioma e viver novas experiências.

A ausência da leitura de livros relacionados a literatura do curso de SI por mais da metade dos alunos formandos (56%) apresenta-se como uma variável de ensino preocupante, pois, pode comprometer a formação do profissional e a produção de trabalhos científicos de qualidade. Contudo, os dados mostraram que aproximadamente metade dos estudantes (44%) tem um perfil autodidata e estudaram de 4 a 7h por semana durante a graduação e isso com certeza é benéfico para o curso, já que permite absorver novos conhecimentos e complementar os conteúdos estudados em sala de aula.

Em relação às informações extraescolares (Renda familiar, situação financeira e ocupação do estudante, incentivo de familiares, escolaridade dos pais, etc.). O agrupamento previamente especificado criou 3 *clusters* contendo 39 registros (31%) no primeiro, 64 (51%) no segundo e 22 (18%) no terceiro. O estudo gerou a Tabela 4 com os seguintes resultados:

Tabela 4 – Resultados encontrados nos atributos extraescolares

Conteúdos / Atributos analisados	**Cluster 0 (31%)**	**Cluster 1 (51%)**	**Cluster 2 (18%)**
Até que etapa de escolarização seu pai concluiu?	Ensino Médio	Ensino Fundamental	Ensino Superior
Até que etapa de escolarização sua mãe concluiu?	Ensino Médio	Ensino Médio	Ensino Superior
Onde e com quem você mora atualmente?	Casa com os pais	Casa com os pais	Casa com cônjuge / filhos
Quantas pessoas da sua família moram com você?	4	2	5

Qual a renda total de sua família, incluindo seus rendimentos?	De R$ 1.650,01 a 3.300,00	Até R$ 1.650,00	De R$ 4.950,01 a 6.600,00
Qual alternativa melhor descreve sua situação financeira?	Tenho renda e contribuo com o sustento da família	Não tenho renda e sou sustentando pela família ou outras pessoas	Sou o principal responsável pelo sustento da família
Qual alternativa a seguir melhor descreve sua situação de trabalho (exceto estágio ou bolsas)?	Trabalho 40 horas semanais ou mais	Não estou trabalhando	Trabalho 40 horas semanais ou mais
Em que tipo de escola você cursou o ensino médio?	Todo em escola pública	Todo em escola pública	Todo em escola particular
Qual modalidade de ensino médio você concluiu?	Ensino médio tradicional	Ensino médio tradicional	Profissionalizante técnico
Quem lhe deu maior incentivo para cursar a graduação?	Pais	Pais	Outros membros da família que não os pais
Quem foi determinante para você enfrentar dificuldades durante seu curso superior e concluí-lo?	Colegas de curso ou amigos.	Pais	Professores do curso
Alguém em sua família concluiu um curso superior?	Sim	Sim	Sim
Qual o principal motivo para você ter escolhido este curso?	Vocação	Inserção no mercado de trabalho	Outro motivo

Fonte: Elaborado pelo autor, 2024.

Com base nos resultados, conclui-se que a maioria dos atributos apresenta relação direta com o estudante e podem contribuir para a permanência e finalização do curso de graduação. Por exemplo, do total de 125 discentes, 103 (82%) residem na casa dos pais e são oriundos de escola pública; 64 (51%) não tem renda e nem trabalham. Diante desse cenário e aliado ao fato da escolaridade

máxima dos pais de 103 (82%) alunos (as) ser o ensino médio e 51% desses genitores ganharem até um salário mínimo e meio (R$ 1.650,00), nota-se uma situação propicia a desistência e extremamente difícil para estudar e finalizar o curso, sendo desejável a ampliação de políticas de assistência estudantil que leve em consideração os dados deste trabalho.

No entanto, mesmo diante desse contexto, a UFPA entregou, no ano de 2021, 125 recursos humanos para o desenvolvimento de sistemas de informação nas diferentes áreas do conhecimento, conforme a base de dados do INEP estudada. Além disso, buscou formar profissionais conscientes do seu papel social e da sua contribuição no avanço científico e tecnológico da região amazônica e como consequência, do país, por meio de ações antrópicas positivas como, por exemplo, entrega de profissionais e produção de novas tecnologias.

CONSIDERAÇÕES FINAIS

Este trabalho realizou uma análise exploratória visando identificar o perfil socioeconômico educacional e os fatores intra e extraescolares dos alunos concluintes do curso de SI da UFPA que realizaram o ENADE em 2021. Os resultados mostraram que há uma predominância do gênero masculino nos cursos oferecidos nos campus do interior e bem como a incidência de apenas 20,8% do gênero feminino no curso. Essa disparidade específica levanta a necessidade de analisar os motivos que levam à baixa demanda das mulheres pelo curso e os fatores de evasão do gênero.

Em relação às características intra e extraescolares dos discentes, constatou-se que o cenário de vida pessoal e acadêmico destes é limitado e dificultoso, onde a maioria não trabalha; reside e depende financeiramente dos pais; não recebeu bolsa científica; estudaram todo o ensino médio em escola pública e os seus pais receberam mensalmente até R$ 1.650,00 reais, conforme detalhado na seção de resultados e discussão. Ademais, muitas pesquisas encontraram resultados semelhantes ao deste estudo em relação a variáveis como gênero, raça, idade e estado civil, visto que, podem impactar as oportunidades

de acesso e permanência no ensino superior. A disparidade de gênero, por exemplo, é um tema recorrente em estudos sobre cursos de tecnologia e ciências exatas, onde geralmente há uma predominância masculina. Historicamente, essas áreas têm sido dominadas por homens, o que pode criar barreiras e estereótipos que desencorajam a participação feminina. Isso pode levar a uma menor representação de mulheres nessas áreas, como observado no caso do curso de SI da UFPA.

A influência da raça também pode ser observada em pesquisas similares. Minorias étnicas muitas vezes enfrentam desafios adicionais no acesso e na permanência no ensino superior. Questões como desigualdade socioeconômica, preconceito e discriminação podem afetar a representatividade de determinados grupos raciais em cursos específicos. A idade dos alunos também pode ser um fator relevante, onde os discentes mais velhos podem enfrentar desafios complementares, como conciliar estudos com responsabilidades familiares ou profissionais. Além disso, a idade pode influenciar a perspectiva e as motivações dos alunos em relação ao curso. O estado civil dos alunos também pode ter impacto no perfil socioeconômico e nas demandas pessoais dos estudantes, pois, os casados (as) ou com filhos podem ter obrigações extras e isso pode influenciar sua disponibilidade de tempo e recursos para investir nos estudos. É importante ressaltar que esses resultados não são uma justificativa para a manutenção dessas desigualdades, mas sim uma base para a reflexão e implementação de políticas de inclusão e equidade no ensino superior. Os resultados obtidos, são importantes para entender o contexto em que os alunos do curso de SI da UFPA se encontram e podem contribuir para a implementação de ações e políticas futuras que visem a inclusão de todos e o apoio socioeconômico aos estudantes.

Quanto às limitações do trabalho, teve-se a problemática a respeito da fidelidade dos dados coletados na base de dados, ou seja, não há como garantir que as respostas dos estudantes no formulário de inscrição do ENADE são fidedignas a sua realidade de vida pessoal e acadêmica durante o período da graduação e em visto disso pode ocorrer alternância ou influência nos resultados

obtidos. Como trabalhos futuros, sugere-se estender este estudo para a aplicação e combinação de outros algoritmos como, por exemplo, aprendizagem de máquina e linguagem R, verificando também a problemática da retenção, evasão e desempenho dos estudantes. Considera-se que seria possível também a aplicação para outros cursos da universidade, verificando se os resultados adquiridos se repetem para alunos de graduação de outras unidades acadêmicas da UFPA.

Agradecimentos

À equipe deste trabalho. Ao Programa de Pós-Graduação em Estudos Antrópicos na Amazônia (PPGEAA) da Universidade Federal do Pará (UFPA).

Referências Bibliográficas

ARAÚJO, C. H.; LUZIO, N. Avaliação da Educação Básica: em busca da qualidade e equidade no Brasil. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2005.

ARAÚJO, R. A. et al. Análise dos microdados do Enade: proposta de uma ferramenta de exploração utilizando mineração de dados. 2019. Dissertação (Mestrado) – Universidade Federal de Goiás, Goiás, 2019.

BAUER, A. É possível relacionar avaliação discente e formação de professores? A experiência de São Paulo. Educação em Revista, SciELO Brasil, v. 28, p. 61–82, 2012.

BRASIL. Ministério da Educação. Exame Nacional de Desempenho dos Estudantes (Enade). 2023. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade. Acesso em: 1 set. 2023.

CASSIANO, K. M. Análise de séries temporais usando análise espectral singular (SSA) e agrupamento de suas componentes baseada em densidade. 2014. Tese (Doutorado) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2014.

CASTRO, L. N.; FERRARI, D. G. Introdução à mineração de dados. 1. ed. São Paulo: Saraiva Educação, 2017.

COSTA, R. H. F. et al. Análise da participação feminina na faculdade de computação do campus Castanhal da Universidade Federal do Pará. In: ANAIS DO XIV WOMEN IN INFORMATION TECHNOLOGY, 2020. p. 174-178.

DA SILVA VIEIRA, A.; BERTOLINI, D.; SCHWERZ, A. L. Análise do desempenho no Enade dos concluintes de computação usando técnica de agrupamento. In: ANAIS DO XXXIII SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2022. p. 834-845.

DE MELLO, A. V.; FINGER, A. F.; BORDIN, A. S. Ciência da computação e engenharia de software: semelhanças e diferenças a partir da realidade dos egressos. In: ANAIS DO XXXI SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2020. p. 1773-1782.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. AI Magazine, v. 17, n. 3, p. 37–37, 1996.

FONSECA, S. O. D.; NAMEN, A. A. Mineração em bases de dados do Inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. Educação em Revista, v. 32, p. 133–157, 2016.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. Data mining. São Paulo: Elsevier Brasil, 2015.

GONÇALVES, C. Data mining com a ferramenta Weka. In: FÓRUM DE SOFTWARE LIVRE DE DUQUE DE CAXIAS, 2011.

GOOGLE DRIVE. Tabela 1 – Atributos selecionados da plataforma INEP e suas descrições. Disponível em: https://docs.google.com/document/d/1vDGY8SbBahlae94Ayj0LSyZAdrCV58sl/edit. Acesso em: 9 maio 2024.

GOTTARDO, E.; KAESTNER, C. A. A.; NORONHA, R. V. Estimativa de desempenho acadêmico de estudantes: análise da aplicação de técnicas de mineração de dados em cursos a distância. Revista Brasileira de Informática na Educação, v. 22, n. 1, 2014.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Censo 2022: pela primeira vez, desde 1991, a maior parte da população do Brasil se declara parda. Disponível em: https://agenciadenoticias.ibge.gov.br. Acesso em: 26 jan. 2024.

INEP – INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Censo da educação superior. Brasília: Inep, 2022. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior. Acesso em: 23 jan. 2024.

KODINARIYA, T. M.; MAKWANA, P. R. Revisão sobre como determinar o número de clusters no clustering k-means. ResearchGate, v. 1, 2013.

LONGEN, S. A. Google Acadêmico: o que é e como usar a plataforma de literatura acadêmica. 2024. Disponível em: https://www.hostinger.com.br. Acesso em: 23 fev. 2024.

MACHADO, R. D. et al. Estudo bibliométrico em mineração de dados e evasão escolar. In: CONGRESSO NACIONAL DE EXCELÊNCIA EM GESTÃO, 2015. p. 1-21.

MAIA, M. M.; DE ANDRADE, L. H. F.; FERNANDES, S. K-means na análise de características socioeconômicas de candidatos ao ensino superior. In: ENCONTRO DE COMPUTAÇÃO DO OESTE POTIGUAR (ECOP/UFERSA), n. 5, 2021.

MARTUCCI, E. M. Informação para educação: os novos cenários para o ensino fundamental. Informação & Sociedade, Universidade Federal da Paraíba, v. 10, n. 2, 2000.

MITRA, S.; ACHARYA, T. Data mining: multimedia, soft computing and bioinformatics. John Wiley & Sons, 2003.

PIRES, Y. P. et al. Diagnóstico da presença feminina nos cursos superiores e no mercado de trabalho em tecnologia da informação no Estado do Pará. In: COMPUTER ON THE BEACH, 2021, v. 12, p. 428-434.

ROIGER, R. J. Data mining: a tutorial-based primer. Chapman and Hall/CRC, 2017.

SANTOS, S. M. dos. Data science aplicado a dados abertos do governo federal: estudos de caso sobre a economia dos municípios brasileiros. 2020. Dissertação (Mestrado) – Universidade Federal do Pará, Belém, 2020.

SILVA, A. B. da; PAULA, D. M. D.; GOMES, G. R. R. Mineração de dados: um estudo para identificação do perfil docente das IES com conceito 3 ou superior no IGC avaliado em 2016

no Estado do Espírito Santo. In: SIMPÓSIO DE PESQUISA OPERACIONAL E LOGÍSTICA DA MARINHA (SPOLM), 2019. p. 1673-1688.

Artigo Submetido em: 23/05/2024 - Aceito em: 27/03/2025 - Publicado em 30/04/2025