Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 21/05/2025; 1st round notif.: 23/11/2025; New version: 07/12/2025; 2nd round notif.: 08/07/2026;
Camera ready: 13/07/2026; Edition review: 14/07/2026; Available online: 14/07/2026; Published: 14/07/2026;

Identificando Padrões em Dados Socioeconômicos de Estudantes Universitários: Uma Abordagem de Agrupamento e Regras de Associação na Universidade Federal de Alagoas

Title: Identifying Patterns in Socioeconomic Data of University Students: A Clustering and Association Rules Approach at the Federal University of Alagoas

Título: Identificación de Patrones en Datos Socioeconómicos de Estudiantes Universitarios: Un Enfoque de Agrupamiento y Reglas de Asociación en la Universidad Federal de Alagoas

Wellington Batalha dos Santos
Universidade Federal de Alagoas (UFAL)
ORCID: 0009-0000-8025-186X
wbds@ic.ufal.br

Bruno Almeida Pimentel
Universidade Federal de Alagoas (UFAL)
ORCID: 0000-0001-8792-4588
brunopimentel@ic.ufal.br

Resumo

A Assistência Estudantil desempenha um importante papel no combate à evasão, mas sua eficácia exige maior compreensão dos perfis de vulnerabilidade dos estudantes. Este estudo investiga padrões socioeconômicos de estudantes universitários da Universidade Federal de Alagoas (UFAL) utilizando técnicas de aprendizado não supervisionado, com ênfase em agrupamento e regras de associação. A análise foi conduzida com dados provenientes de inscrições em programas de assistência estudantil, contemplando informações sobre renda, transporte, escolaridade dos pais, identificação racial e participação em programas sociais. Para o agrupamento, empregou-se o algoritmo K-Modes, adequado a dados majoritariamente categóricos, utilizando a distância de Hamming como medida de dissimilaridade. A definição do número de grupos foi orientada por diferentes métricas de validação e por análises visuais, como projeção via t-SNE e dendrogramas. Os resultados apontaram a formação de grupos distintos, com destaque para perfis relacionados a dificuldades de transporte e menor renda familiar. Na análise de regras de associação, a renda per capita, a origem da renda e a inserção em programas sociais emergiram como fatores centrais, reforçando sua relevância na caracterização da vulnerabilidade socioeconômica. Observou-se ainda a associação entre escolaridade dos pais e identificação racial. Os padrões identificados podem subsidiar políticas de assistência estudantil mais focadas e eficazes, contribuindo para a promoção da permanência acadêmica. O trabalho também sugere a ampliação futura da análise com a inclusão de dados acadêmicos e a avaliação da evolução dos perfis ao longo do tempo.

Palavras-chave: *Assistência Estudantil; Aprendizado de Máquina; Aprendizado Não Supervisionado; Ciência de Dados; Agrupamento; Regras de Associação; Redução da Evasão.*

Abstract

Student assistance plays an important role in reducing dropout rates, but its effectiveness requires a deeper understanding of students' vulnerability profiles. This study investigates the socioeconomic patterns of university students at the Federal University of Alagoas (UFAL) using unsupervised learning techniques, focusing on clustering and association rules. The analysis was based on data from applications for student aid programs, including information on income, transportation, parents' education level, racial identification, and participation in social programs. Clustering was performed using the K-Modes algorithm, suitable for predominantly categorical data, with Hamming distance as the dissimilarity measure. The choice of the number of clusters was guided by several validation metrics and visual analyses, such as t-SNE projections and dendrograms. The results revealed distinct groups, highlighting

Cite as: Santos, W. B. & Pimentel, B. A. (2026). Identificando Padrões em Dados Socioeconômicos de Estudantes Universitários: Uma Abordagem de Agrupamento e Regras de Associação na Universidade Federal de Alagoas. Revista Brasileira de Informática na Educação, vol. 34, pp. 1095–1127. https://doi.org/10.5753/rbie.2026.6052.

profiles related to transportation difficulties and lower household income. In the association rules analysis, per capita income, income source, and participation in social programs emerged as central factors, reinforcing their relevance in characterizing socioeconomic vulnerability. Additionally, associations were found between parents' education level and racial identification. The patterns identified can support the development of more targeted and effective student assistance policies, contributing to the promotion of academic persistence. Future work suggests expanding the analysis by incorporating academic performance data and evaluating changes in student profiles over time.
Keywords: *Student Assistance; Machine Learning; Unsupervised Learning; Data Science; Clustering; Association Rules; Dropout Reduction.*

Resumen

La asistencia estudiantil desempeña un papel importante en la reducción de la deserción académica, pero su eficacia requiere una mejor comprensión de los perfiles de vulnerabilidad de los estudiantes. Este estudio investiga los patrones socioeconómicos de los estudiantes universitarios de la Universidad Federal de Alagoas (UFAL) utilizando técnicas de aprendizaje no supervisado, con énfasis en agrupamiento y reglas de asociación. El análisis se basó en datos de inscripciones en programas de asistencia estudiantil, incluyendo información sobre ingresos, transporte, nivel educativo de los padres, identificación racial y participación en programas sociales. El agrupamiento se realizó utilizando el algoritmo K-Modes, adecuado para datos predominantemente categóricos, empleando la distancia de Hamming como medida de disimilitud. La definición del número de grupos fue guiada por varias métricas de validación y análisis visuales, como las proyecciones de t-SNE y dendrogramas. Los resultados revelaron grupos distintos, destacándose perfiles relacionados con dificultades de transporte y bajos ingresos familiares. En el análisis de reglas de asociación, el ingreso per cápita, la fuente de ingreso y la participación en programas sociales surgieron como factores centrales, reforzando su relevancia para caracterizar la vulnerabilidad socioeconómica. También se observaron asociaciones entre el nivel educativo de los padres y la identificación racial. Los patrones identificados pueden orientar políticas de asistencia estudiantil más específicas y eficaces, contribuyendo a la promoción de la permanencia académica. Como trabajo futuro, se sugiere ampliar el análisis incluyendo datos académicos y evaluando la evolución de los perfiles estudiantiles a lo largo del tiempo.
Palabras clave: *Asistencia Estudiantil; Aprendizaje Automático; Aprendizaje No Supervisado; Ciencia de Datos; Agrupamiento; Reglas de Asociación; Reducción de la Deserción.*

1 Introdução

A assistência estudantil desempenha um papel essencial na promoção da inclusão e da equidade no ensino superior, oferecendo suporte a estudantes em situação de vulnerabilidade socioeconômica, fornecendo meios para a superação dos obstáculos que impactam no rendimento acadêmico e permanência no curso (Tinto, 2012; Vasconcelos, 2010). No Brasil, o Plano Nacional de Assistência Estudantil (PNAES) busca garantir condições para a permanência de alunos nas Instituições Federais de Ensino Superior (IFES), fornecendo auxílio em áreas como moradia, alimentação, transporte e saúde. A efetividade dessas políticas, no entanto, depende de uma gestão eficiente dos recursos, que deve considerar a diversidade de perfis dos estudantes e suas diferentes necessidades (Imperatori, 2017; Vasconcelos, 2010).

Na Universidade Federal de Alagoas (UFAL), assim como em outras IFES, há uma crescente demanda por programas assistenciais. Com recursos limitados, torna-se fundamental compreender melhor os perfis socioeconômicos dos estudantes para otimizar a distribuição dos auxílios. Além disso, é importante entender como determinados grupos de características se associam a outros, revelando padrões que possam apoiar a formulação de critérios mais justos e eficazes de seleção. Compreender essas relações não apenas contribui para uma gestão mais eficiente dos recursos, mas também tem o potencial de embasar o desenvolvimento de políticas assistenciais mais sensíveis à diversidade de necessidades de uma comunidade acadêmica heterogênea. Tradicionalmente, a análise dos candidatos a esses programas é feita de forma manual ou baseada em critérios predefinidos. No entanto, avanços na Ciência de Dados (CD) e Inteligência Artificial (IA) oferecem novas possibilidades para tornar esse processo mais preciso e eficiente.

A Inteligência Artificial (IA), que pode ser compreendida como o campo da Ciência da Computação que estuda agentes capazes de perceber seu ambiente e agir de forma a maximizar seu desempenho em tarefas que normalmente exigiriam inteligência humana (Russell & Norvig, 2016), engloba técnicas que possibilitam a análise inteligente de dados. Entre elas, destaca-se o Aprendizado de Máquina Não Supervisionado, que permite examinar grandes volumes de dados sem a necessidade de rótulos prévios, identificando padrões ocultos e estruturas latentes nos dados (Alpaydin, 2020). No contexto da assistência estudantil, técnicas como agrupamento (clusterização) e regras de associação podem contribuir para a segmentação dos estudantes e a descoberta de fatores socioeconômicos correlacionados, fornecendo *insights* para uma tomada de decisão mais embasada, fundamentada em dados.

Dessa forma, este estudo busca explorar como métodos de Aprendizado de Máquina Não Supervisionado podem ser aplicados à análise de dados socioeconômicos de estudantes da UFAL interessados em assistência estudantil. O objetivo é identificar padrões e perfis socioeconômicos, possibilitando uma alocação mais eficiente dos recursos e um aprimoramento dos programas assistenciais da instituição.

Nesse contexto, a contribuição deste trabalho para a área de Informática na Educação reside no uso estratégico de dados para enfrentar desafios de evasão e equidade. Ao utilizar algoritmos de agrupamento e regras de associação para auxiliar a assistência estudantil, a abordagem proposta complementa as aplicações de IA na Educação ao incorporar a análise de fatores socioeconômicos como ferramenta de apoio à tomada de decisão, contribuindo para que estudantes em situação de vulnerabilidade recebam suporte adequado para concluir sua trajetória acadêmica.

Para atingir esse objetivo, a pesquisa realiza uma análise exploratória dos dados coletados, seguida do pré-processamento adequado para normalização e transformação das variáveis. Em seguida, são aplicadas técnicas de agrupamento, como o K-Modes, para segmentar os estudantes com base em suas características socioeconômicas. Além disso, a mineração de regras de associação, utilizando o algoritmo Apriori, é empregada para identificar padrões de correlação entre variáveis como renda, moradia e procedência escolar. Os resultados obtidos são avaliados e discutidos em relação ao seu potencial de contribuir para a assistência estudantil.

Em resumo, esta pesquisa busca responder às seguintes perguntas:

- **QP1**: Quais perfis socioeconômicos podem ser identificados entre os estudantes da UFAL candidatos à assistência estudantil?
- **QP2**: Em que medida os perfis e associações identificados podem subsidiar a gestão e a priorização da concessão de auxílios, contribuindo para decisões mais transparentes e orientadas por dados?

O artigo está estruturado da seguinte forma: a Seção 2 apresenta a revisão da literatura, abordando estudos que investigam a assistência estudantil e o uso de técnicas de aprendizado de máquina na análise de dados educacionais. A Seção 3 traz a fundamentação teórica, detalhando os conceitos de assistência estudantil, agrupamento e mineração de regras de associação. A Seção 4 descreve a metodologia adotada na pesquisa, incluindo a origem dos dados, o pré-processamento e as técnicas aplicadas. Na Seção 5, são apresentados os resultados obtidos, seguidos da Seção 6, que traz uma discussão sobre as descobertas. Por fim, a Seção 7 apresenta as conclusões do estudo e sugestões para trabalhos futuros.

2 Trabalhos Relacionados

Dentre os trabalhos relacionados ao escopo deste artigo, um estudo relevante é o de Villar e Andrade (2024), que investigou a predição de evasão e sucesso acadêmico no ensino superior utilizando algoritmos supervisionados de Aprendizado de Máquina (AM). Foram analisadas 35 variáveis, incluindo atributos socioeconômicos como ocupação dos pais, situação financeira, bolsas de estudo e taxa de desemprego local. O estudo comparou algoritmos tradicionais (como *Decision Tree* e *Support Vector Machine*) com técnicas de *boosting* (*LightGBM*, *CatBoost* e *XGBoost*), otimizadas via *Optuna*, e utilizou o método *Synthetic Minority Over-Sampling Technique* (SMOTE) para tratar o desbalanceamento das classes. Os melhores resultados foram obtidos com *LightGBM* e *CatBoost*. A análise de importância das variáveis, via *SHapley Additive exPlanations* (SHAP), destacou o número de disciplinas aprovadas e o status de bolsista. Embora o foco seja em aprendizado supervisionado, a presença de variáveis socioeconômicas aproxima o estudo da proposta deste trabalho, reforçando sua relevância na compreensão da permanência estudantil.

Outro exemplo de aplicação de AM Supervisionado com dados socioeconômicos no contexto educacional é o estudo de Matz et al. (2023), que analisou 50.095 estudantes de quatro instituições dos Estados Unidos para prever a evasão após o primeiro semestre. O estudo combinou dados institucionais em nível macro, incluindo aspectos acadêmicos, demográficos e socioeconômicos, com dados de engajamento em níveis micro e meso, como interações via aplicativo

universitário e participação em eventos. A mescla dessas variáveis aumentou o desempenho dos modelos, que alcançaram Área Sob a Curva (AUC) de até 88%, com destaque para o *Random Forest*. Embora o foco principal tenha sido a previsão de evasão, o estudo se aproxima deste trabalho ao valorizar os dados socioeconômicos na caracterização dos estudantes, evidenciando a importância de análises mais abrangentes para subsidiar estratégias institucionais.

Utilizando AM Não Supervisionado, Mohamed Nafuri et al. (2022) investigou padrões sociais e acadêmicos em estudantes universitários de baixa renda por meio de técnicas não supervisionadas. O estudo analisou dados de 117.069 estudantes de baixa renda de universidades públicas da Malásia, utilizando atributos como renda familiar, número de atividades extracurriculares, status de emprego, após o fim da graduação, e histórico de evasão. A metodologia comparou três algoritmos de agrupamento: *K-Means*, *Balanced Iterative Reduction and Clustering using Hierarchies* (BIRCH) e *Density Based Spatial Clustering of Application with Noise* (DBSCAN). Os resultados demonstraram que o *K-Means* obteve a melhor separação entre grupos, revelando relações significativas entre participação em atividades extracurriculares, status de emprego e rendimento acadêmico. Esse estudo apresenta similaridades com o presente trabalho no uso de AM não Supervisionado, no entanto, se diferencia por mesclar dados socioeconômicos, acadêmicos e da trajetória pós-graduação.

Em uma abordagem com AM Supervisionado, Sun et al. (2021) propõe um método de classificação para prever o nível de concessão de bolsas estudantis com base em *Ensemble Learning*. O estudo utilizou dados comportamentais de 10.885 estudantes de uma universidade, incluindo despesas, desempenho acadêmico, uso de dormitório e empréstimos de livros, integrados em 21 atributos. A abordagem combinou algoritmos como *Gradient Boosting*, *Random Forest*, *AdaBoost* e *Support Vector Machines*, alcançando acurácia média de 95,4%. Embora o foco principal seja a previsão do nível de auxílio, o estudo se aproxima deste trabalho ao utilizar dados socioeconômicos e comportamentais para embasar decisões em assistência estudantil. Além disso, ao tratar o problema do desbalanceamento dos dados, contribui para uma alocação mais equitativa dos recursos.

No estudo de Bennasar-Veny et al. (2020) os autores investigaram o agrupamento de estilos de vida relacionados à saúde entre estudantes da Universidade das Ilhas Baleares. Com amostra de 444 alunos, utilizaram *K-Means* após padronização das variáveis. Três grupos distintos foram identificados, variando de estilos de vida saudáveis a não saudáveis. O estudo sugere que intervenções integradas podem ser mais eficazes na promoção da saúde no ambiente universitário. O estudo se assemelha a este trabalho por utilizar AM não Supervisionado com dados não-acadêmicos, mas analisa dados referentes à saúde e estilo de vida dos estudantes.

Um exemplo do uso de AM Supervisionado com dados socioeconômicos no contexto educacional nacional é o trabalho de Soares (2020), desenvolvido com dados de estudantes do Instituto Federal do Amazonas (IFAM). O estudo teve como objetivo auxiliar a assistência estudantil por meio de um modelo capaz de classificar os estudantes como aptos ou não ao recebimento de auxílio, utilizando a plataforma Weka. Foram testados 12 algoritmos em sua configuração padrão, além da otimização de 5 algoritmos. O melhor desempenho foi alcançado com o algoritmo *Sequential Minimal Optimization* (SMO), seguido por *Random Forest* e *Instance-Based k* (IBk). O estudo também extraiu regras de classificação que evidenciaram fatores como renda, moradia e condições familiares como determinantes da elegibilidade. A similaridade com o presente traba-

lho ocorre devido ao uso de dados socioeconômicos no contexto educacional, mas se diferencia por utilizar AM Supervisionado.

No contexto do ensino médio, V. A. A. Silva et al. (2020) aplicou AM não supervisionado para agrupar estudantes do Exame Nacional do Ensino Médio (ENEM) de Minas Gerais com base em suas notas, utilizando o algoritmo *K-Means* e identificando dois grupos distintos com base no método da silhueta. As regras de associação extraídas revelaram relações entre variáveis como escolaridade da mãe e tipo de escola, contribuindo para a compreensão do impacto dos fatores socioeconômicos no desempenho. A semelhança com este trabalho pode ser observada principalmente pelo uso de técnicas semelhantes de AM não Supervisionado com dados socioeconômicos, divergindo no contexto por ser relacionado ao ensino médio.

Ainda no escopo do ENEM, Simon e Cazella (2017) aplicaram uma Árvore de Decisão para prever o desempenho médio dos estudantes em Ciências da Natureza, utilizando dados do ENEM 2015. Com validação cruzada e análise de variáveis institucionais, o modelo alcançou 77,02% de acurácia. As variáveis mais relevantes foram o tipo de escola e o nível socioeconômico. Embora também analise dados socioeconômicos, esse estudo difere por usar AM Supervisionado e ser realizado em um contexto de ensino médio.

No âmbito da educação superior, Santos et al. (2015) utilizaram o algoritmo Apriori para extrair padrões de comportamento desanimado em estudantes de uma disciplina da Universidade Federal do Rio Grande do Sul (UFRGS), a partir de dados de questionários e logs de atividade. As regras extraídas revelaram que estudantes desmotivados tendem a não completar as atividades e apresentam pior desempenho final, ressaltando a importância de fatores afetivos no processo de aprendizagem. Mesmo tratando-se de um contexto de um ambiente virtual, esse trabalho apresenta similaridades por utilizar AM não Supervisionado, por meio de Regras de Associação, e englobar dados não acadêmicos.

Por fim, L. A. Silva et al. (2014) também utilizaram o algoritmo Apriori em dados do ENEM 2010, após pré-processamento e categorização das notas. As regras geradas evidenciaram que baixos níveis de renda e escolaridade dos pais, além de número elevado de moradores por residência, impactam negativamente o desempenho estudantil. Novamente, a similaridade ocorre no emprego de Regras de Associação juntamente com o uso de dados socioeconômicos, porém em dados relacionados ao ensino médio.

Para consolidar a análise e evidenciar as contribuições deste trabalho, uma comparação estruturada das principais características e diferenciais metodológicos dos estudos correlatos é apresentada na Tabela 1.

A literatura sobre o uso de AM no contexto educacional tem explorado majoritariamente técnicas de Aprendizado Supervisionado aplicadas a dados acadêmicos, como notas e frequência, com foco principal na previsão de evasão ou desempenho estudantil. Embora essa abordagem seja relevante e amplamente consolidada, observa-se um número reduzido de estudos que consideram dados socioeconômicos como foco central da análise. Nesse sentido, este trabalho contribui ao ampliar esse escopo, investigando como variáveis socioeconômicas podem ser utilizadas para identificar padrões entre estudantes por meio de técnicas de Aprendizado Não Supervisionado, com potencial para subsidiar políticas de assistência estudantil e estratégias de permanência acadêmica.

Tabela 1: Comparação estruturada entre os Trabalhos Relacionados e o Presente Estudo.

Ref.	Dados*	Técnica / Abordagem	Foco / Diferencial
(Villar & Andrade, 2024)	SE+Acad	Superv. (Boosting, SHAP)	Predição de evasão e sucesso.
(Matz et al., 2023)	SE+Acad+Comp	Superv. (Random Forest)	Evasão (1º sem.) e engajamento.
(Mohamed Nafuri et al., 2022)	SE+Acad	N. Superv. (K-Means, BIRCH)	Padrões em baixa renda e pós-grad.
(Sun et al., 2021)	SE+Acad+Comp	Superv. (Ensemble)	Predição de nível de bolsa.
(Bennasar-Veny et al., 2020)	Comp	N. Superv. (K-Means)	Estilos de vida e saúde.
(Soares, 2020)	SE	Superv. (SMO, RF, IBk)	Elegibilidade para auxílio.
(V. A. A. Silva et al., 2020)	SE+Acad	N. Superv. (K-Means, Regras)	Perfil SE e desempenho (EM).
(Simon & Cazella, 2017)	SE+Acad	Superv. (Árvore Decisão)	Desempenho em Ciências (EM).
(Santos et al., 2015)	Comp	N. Superv. (Apriori)	Desmotivação em amb. virtual.
(L. A. Silva et al., 2014)	SE+Acad	N. Superv. (Apriori)	Renda e estrut. familiar (EM).
Atual	**SE**	**N. Superv. (K-Modes, Apriori)**	**Perfis de vuln., regras, apoio à gestão da assistência, prevenção da evasão.**

Legenda: SE: Socioecon.; Acad: Acadêmicos; Comp: Comportamentais; Superv.: Supervisionado; EM: Ensino Médio.
Fonte: Elaboração própria.

3 Fundamentação Teórica

Nesta seção, são apresentados os principais conceitos teóricos utilizados no desenvolvimento deste trabalho. São abordados temas como Assistência Estudantil, Aprendizado de Máquina, Aprendizado de Máquina Não Supervisionado, Agrupamento e Regras de Associação. Além disso, ao final, uma seção discutirá os principais trabalhos relacionados ao tema da pesquisa.

3.1 Assistência Estudantil

A Assistência Estudantil (AE) desempenha um papel central na promoção da inclusão social e na garantia da permanência de estudantes em instituições de ensino superior, especialmente aqueles em situação de vulnerabilidade socioeconômica. De acordo com Costa (2009), as políticas de assistência estudantil na educação superior têm a finalidade de destinar recursos e mecanismos para que os alunos possam permanecer na universidade e concluir seus estudos de modo eficaz. Sendo assim, tais políticas devem se voltar não só para as questões de ordem econômica, como auxílio financeiro, mas também para aspectos pedagógicos e psicológicos.

No entendimento de Vasconcelos (2010), a AE, como um direito social, visa fornecer os recursos necessários para superar obstáculos ao desempenho acadêmico, garantindo o desenvolvimento dos estudantes durante a graduação e reduzindo o abandono e trancamento de matrículas. Ela abrange diversas áreas dos direitos humanos, oferecendo desde condições de saúde e acesso a instrumentos pedagógicos até apoio a necessidades educativas especiais e provisão de recursos básicos como moradia, alimentação e transporte.

3.1.1 Histórico da Assistência Estudantil no Brasil

Historicamente, um dos primeiros marcos da AE no Brasil ocorreu na década de 1930. Durante o governo Getúlio Vargas, a educação passou a ser reconhecida pelo Estado como um direito público, embora esse direito não estivesse completamente assegurado nas constituições da época (Vasconcelos, 2010). Ainda nesse período, como destaca Costa (2009), ocorreu a primeira reforma educacional voltada para o ensino superior, sendo também a primeira tentativa de regulamentação da AE para esse segmento de estudantes.

Outro marco importante ocorreu nos anos 1980, com a promulgação da Constituição Federal de 1988 (Brasil, 1988), resultado do processo de redemocratização do país. Esse documento representou um avanço significativo ao garantir os direitos sociais e fortalecer a democratização do acesso e da permanência na educação superior (Dutra & Santos, 2017).

A consolidação da AE ocorreu em 2007, com a criação do Programa Nacional de Assistência Estudantil (PNAES). O PNAES surgiu em conjunto com o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (REUNI), que visava a expansão das universidades públicas em todo o Brasil. Inicialmente instituído pela Portaria Normativa 39/2007 do Ministério da Educação, o PNAES foi transformado em decreto em 2010, pelo então presidente Luiz Inácio Lula da Silva, dando origem ao Decreto Nº 7.234/2010 (Brasil, 2010), o que fortaleceu a AE como uma política de Estado.

3.1.2 Objetivos e Áreas de Atuação do PNAES

O Decreto Nº 7.234/2010 define os princípios orientadores do PNAES, estabelecendo sua finalidade como a ampliação das condições de permanência dos jovens na educação superior pública federal. O programa busca democratizar as condições de permanência dos estudantes, minimizar desigualdades sociais e regionais e promover a inclusão social por meio da educação (Brasil, 2010).

As ações do PNAES abrangem diversas áreas fundamentais para a permanência estudantil, incluindo moradia estudantil, alimentação, transporte, atenção à saúde, inclusão digital, cultura, esporte, creche, apoio pedagógico e acessibilidade para estudantes com deficiência, transtornos do desenvolvimento e altas habilidades (Brasil, 2010).

O público-alvo do programa são, prioritariamente, estudantes oriundos da rede pública de educação básica ou aqueles com renda familiar per capita de até um salário mínimo e meio, sem prejuízo de demais requisitos estabelecidos pelas Instituições Federais de Ensino Superior (IFES) (Brasil, 2010).

3.1.3 A Assistência Estudantil e a Permanência Acadêmica

A relação entre a AE e a permanência acadêmica tem sido objeto de diversos estudos. Segundo Tinto (2012), a integração social e acadêmica do estudante é um fator essencial para reduzir a evasão universitária. Além disso, Dias e Costa (2015) destacam que a ampliação do acesso ao ensino superior deve ser acompanhada de políticas que promovam a permanência, especialmente para estudantes de baixa renda.

Com a implementação do PNAES e o fortalecimento das políticas assistenciais, a AE tem desempenhado um papel fundamental na redução da evasão acadêmica. Ao oferecer suporte financeiro, pedagógico e psicológico, essas políticas não apenas aumentam as taxas de conclusão dos cursos, mas também promovem um ambiente acadêmico mais inclusivo e equitativo.

Por fim, a continuidade e o aprimoramento das políticas de AE são fundamentais para fortalecer a democratização do ensino superior no Brasil. A alocação eficiente dos recursos e a constante avaliação dos programas assistenciais podem contribuir para garantir que um número crescente de estudantes tenha condições de concluir sua graduação com sucesso. Nesse cenário, as técnicas de Aprendizado de Máquina surgem como soluções promissoras para apoiar a tomada de decisão e otimizar a distribuição e cobertura desses benefícios, oferecendo diferentes abordagens conforme veremos a seguir.

3.2 Aprendizado de Máquina

Dentre as diversas áreas da IA, destaca-se o Aprendizado de Máquina (AM), ou *Machine Learning* (ML), que se dedica ao desenvolvimento de sistemas que aprendem e extraem conhecimento automaticamente a partir de dados (Alpaydin, 2020; Monard & Baranauskas, 2003).

O AM caracteriza-se pela interdisciplinaridade, integrando conceitos da ciência da computação, estatística, engenharia e matemática (Ghahramani, 2003), e é amplamente aplicado em áreas como visão computacional, biologia e processamento de linguagem natural.

Pode ser classificado em três categorias principais: Aprendizado Supervisionado, Aprendizado Não Supervisionado e Aprendizado por Reforço (Janiesch et al., 2021). No Aprendizado Supervisionado, busca-se aprender um mapeamento entre entradas e saídas a partir de exemplos rotulados (Cunningham et al., 2008), envolvendo tarefas de classificação ou regressão. Por sua vez, o Aprendizado por Reforço refere-se ao aprendizado por interação com o ambiente, visando maximizar as recompensas recebidas (Janiesch et al., 2021; Sutton & Barto, 2018).

Considerando a natureza dos dados utilizados neste trabalho, será dada ênfase ao Aprendizado Não Supervisionado. Nesta abordagem, os dados não possuem rótulos ou informações de supervisão explícita. O objetivo é descobrir padrões ou estruturas ocultas nos dados (Ghahramani, 2003). Embora mais desafiador devido à subjetividade na avaliação dos resultados (James et al., 2023), essa abordagem permite obter *insights* valiosos sobre os dados sem a necessidade de supervisão.

Entre suas principais técnicas estão: Agrupamento, Regras de Associação e Redução de Dimensionalidade. As técnicas de Agrupamento e Regras de Associação, essenciais para este estudo, serão detalhadas a seguir.

O Agrupamento consiste em particionar dados em grupos (*clusters*) com alta similaridade interna (Han et al., 2012). Essa técnica é amplamente utilizada em aplicações como segmentação de mercado (James et al., 2023). As abordagens mais comuns incluem o Agrupamento Particional, que divide os dados em K grupos, com destaque para o algoritmo *K-Means* e suas variações, como o *K-Modes*, criado especificamente para tratar atributos qualitativos/categóricos em vez de numéricos (Huang, 1998).

Outras técnicas incluem o Agrupamento Hierárquico, que organiza dados em estruturas de árvore; e o Agrupamento Baseado em Densidade, como o DBSCAN, capaz de lidar com *outliers* (Oyewole & Thopil, 2023).

A Mineração de Regras de Associação busca identificar correlações entre itens em bases de dados transacionais, sendo aplicada em áreas como marketing e apoio à decisão (Zhang & Zhang, 2002; Zhao & Bhowmick, 2003). Essa técnica diferencia-se do agrupamento por investigar relações entre variáveis, não entre as observações. Introduzido por Agrawal et al. (1993), o conceito de Regra de Associação é formalizado como uma implicação $A \Rightarrow B$, avaliada pelas métricas de suporte, confiança e *lift* (Han et al., 2012).

O algoritmo *Apriori*, também de Agrawal et al. (1993), é um dos métodos mais tradicionais dessa abordagem. Apesar de apresentar limitações de desempenho em bases de dados volumosas, destaca-se pela robustez e interpretabilidade ao reduzir o espaço de busca através da propriedade de conjuntos frequentes.

4 Metodologia

Nesta seção, serão descritas as etapas metodológicas adotadas para o desenvolvimento deste trabalho. Inicialmente, será realizada uma análise exploratória dos dados, com o objetivo de compreender as características gerais da base de dados utilizada. Em seguida, serão aplicadas técnicas de Aprendizado de Máquina não supervisionado, com foco em dois métodos principais: o agrupamento com o algoritmo *K-Modes* e a descoberta de Regras de Associação com o algoritmo Apriori. O uso dessas técnicas, ajuda a responder às seguintes questões que guiam a pesquisa: *(i)* quais variáveis tendem a ocorrer em conjunto, e com que frequência? e *(ii)* existem perfis distintos de vulnerabilidade entre os estudantes, e como se caracterizam?

4.1 Coleta e Tratamento de Dados

A base de dados utilizada nesse estudo foi obtida em parceria com a Pró-Reitoria Estudantil (PROEST) da UFAL, que é a pró-reitoria responsável pela execução dos programas de assistência estudantil na instituição. Os dados se originam a partir das inscrições realizadas pelos estudantes da Universidade Federal de Alagoas nos editais de Assistência Estudantil, coordenados pela PROEST. Esses dados consistem em informações socioeconômicas dos estudantes de graduação, sendo parte das variáveis autodeclaradas, como etnia e orientação sexual, e outras validadas pela equipe de assistência social da PROEST, como renda familiar e situação de moradia, o que garante uma boa confiabilidade aos dados. Tais dados são usados pela instituição para calcular o Índice de Vulnerabilidade Social (IVS) dos estudantes, que serve como critério para definir a prioridade de acesso desses alunos aos serviços assistenciais oferecidos.

Na Figura 1, pode ser observada uma representação desse processo. Na etapa de Inscrição, os estudantes se inscrevem por meio de formulários e fornecimento de documentação exigida nos editais. Na etapa de Validação, a equipe de Assistência Social faz a checagem das inscrições e valida as informações por meio das documentações fornecidas. Na última etapa, as inscrições são organizadas numa planilha, o cálculo do IVS é realizado e então divulgado, por meio de publicação do resultado oficial no site da universidade.

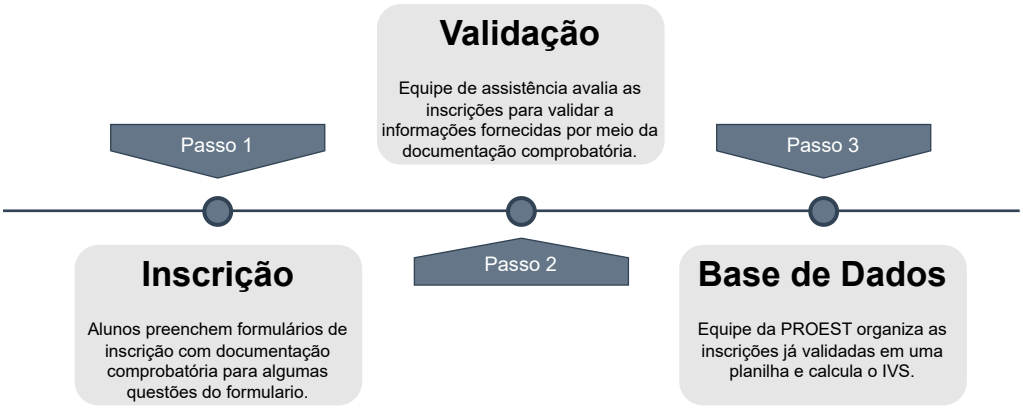


Figura 1: Fluxo de Processo da Coleta de Dados.

São considerados diversos indicadores para a avaliação do IVS dos estudantes, como a renda per capita familiar, procedência escolar e a composição familiar. Também são analisados aspectos como identidade de gênero e orientação sexual, o contexto sócio-familiar marcado por violência ou fragilização de vínculos, além das condições de trabalho do grupo familiar e do próprio estudante. A situação de moradia do estudante ou de sua família, o impacto de doenças graves na organização familiar, e a condição de gestante são outros fatores considerados. A participação em programas de transferência de renda governamentais e a presença de estudantes ou membros da família com deficiência ou transtornos globais do desenvolvimento também entram na análise. Adicionalmente, a condição étnico-racial (como preto, pardo, indígena ou quilombola), condição de refugiado, egressos do sistema prisional, e dificuldades de mobilidade ou transporte são avaliadas para garantir um entendimento completo das vulnerabilidades enfrentadas pelos estudantes.

4.2 Análise Estatística

A base de dados utilizada neste estudo contém informações de 2.604 registros de estudantes inscritos nos editais de assistência estudantil da Universidade Federal de Alagoas, abrangendo um total de 42 variáveis. Essas variáveis refletem o perfil socioeconômico e as condições de vulnerabilidade social dos alunos, como renda per capita, situação de moradia, composição familiar, identidade de gênero, entre outros. Na Tabela 2 é possível observar a lista de variáveis presentes na base de dados em mais detalhes. A variável ID, consiste em um identificador do registro; a Renda per Capita e o IVS são do tipo numérico. Todas as demais variáveis são do tipo categórica/binária, com os valores ’0’ e ’1’ representando os valores lógicos *falso* e *verdadeiro* respectivamente.

Durante o processo preliminar de limpeza de dados, foram removidos 7 registros que apresentavam um Índice de Vulnerabilidade Social (IVS) inválido, o que indicava que o cadastro que esses registros representavam estava incompleto ou com alguma inconsistência. Após a remoção desses registros, o conjunto de dados final utilizado nas análises contém 2.597 registros válidos. Além disso, foi necessário um tratamento inicial para a variável que representa a Renda Per Capita, que estava em formato textual, sendo realizada a remoção/substituição de alguns caracteres e conversão dos dados para o formato numérico. A variável de índice 10, que indica se o estudante está sob a condição de refugiado, foi removida, uma vez que não existem estudantes nessa condi-

Tabela 2: Variáveis do Conjunto de Dados.

ID	Descrição
0	ID
1	Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.
2	O/a estudante é o/a principal mantenedor/a da família.
3	Rompimento e/ou fragilização de vínculos familiares.
4	Situações de violência (física, sexual, psicológica, patrimonial e moral).
5	Sexo / Identidade de Gênero / Orientação Sexual
6	Identificação racial
7	Inserção em programas sociais
8	Gestante
9	Egressos do sistema prisional / em regime semi-aberto
10	Refugiados
11	Transporte intermunicipal integralmente custeado pelo estudante
12	Transporte intermunicipal cedido pela prefeitura com contrapartida do estudante
13	Transporte intermunicipal cedido pela prefeitura sem contrapartida do estudante
14	Transporte urbano municipal (ônibus, moto táxi, etc.)
15	Dificuldades de acesso à universidade considerando a distância
16	Dificuldades de acesso à universidade considerando o tempo de deslocamento
17	Inexistência de transporte público na localidade onde mora
18	Pessoa em sofrimento psíquico ou adoecimento mental na família ou dependência química.
19	Pessoa com deficiência na família
20	Estudante com deficiência
21	Pessoa com doença crônica na família
22	Estudante com doença crônica
23	Possui crianças (até 12 anos) filho/a do estudante
24	Possui crianças (até 12 anos) filho(a) de outro membro da família.
25	Possui idoso (a partir de 60 anos)
26	Ensino Médio em escola pública
27	Escolaridade dos pais ou responsável: Analfabeto/a
28	Escolaridade dos pais ou responsável: Ensino Fundamental ou equivalente incompleto
29	Escolaridade dos pais ou responsável: Ensino Fundamental ou equivalente completo ou Ensino Médio ou equivalente incompleto
30	Escolaridade dos pais ou responsável: Ensino Médio ou equivalente completo ou Superior incompleto
31	Casa do núcleo familiar cedida
32	Casa do núcleo familiar financiada
33	Casa do núcleo familiar alugada
34	Casa do núcleo familiar ocupada
35	Estudante morador de albergue ou em situação de rua
36	Localização do bairro ou comunidade do núcleo familiar: comunidade indígena ou quilombola
37	Localização do bairro ou comunidade do núcleo familiar: área rural
38	Estudante reside em outro município (sem o núcleo familiar) em função dos estudos
39	Estudante reside em república estudantil
40	Renda Per Capita (Registrada)
41	IVS

Fonte: Elaboração própria.

ção. A manipulação e tratamento dos dados foram realizados com a linguagem de programação *Python* através da biblioteca de análise de dados Pandas (McKinney, 2010).

Uma análise de correlação foi inicialmente realizada para averiguar as relações entre as variáveis do conjunto de dados. Para garantir que a variável de renda per capita não influenciasse de forma desproporcional o cálculo das correlações, essa variável foi normalizada, igualando-a às escalas das demais variáveis. Utilizou-se as correlações de Pearson, Spearman e tau de Kendall, acompanhadas de um teste de hipótese para avaliar a significância estatística das associações. Um teste de normalidade, Shapiro-Wilk, foi executado para a variável da renda. Todos os cálculos foram realizados com o auxílio da biblioteca SciPy do Python (Virtanen et al., 2020), que oferece ferramentas robustas para essa análise. Os gráficos apresentados foram criados com a biblioteca *Seaborn* (Waskom, 2021), uma ferramenta do *Python* voltada para a visualização de dados estatísticos.

4.3 Análise Regras de Associação

Essa subseção aborda as técnicas de Mineração de Regras de Associação para explorar padrões e relações entre variáveis socioeconômicas dos estudantes. O objetivo é identificar combinações frequentes de características que possam estar associadas à vulnerabilidade social e ao perfil de candidatos aos programas de assistência estudantil. Para uma melhor visualização do processo, o fluxo completo desta etapa é apresentado na Figura 2.

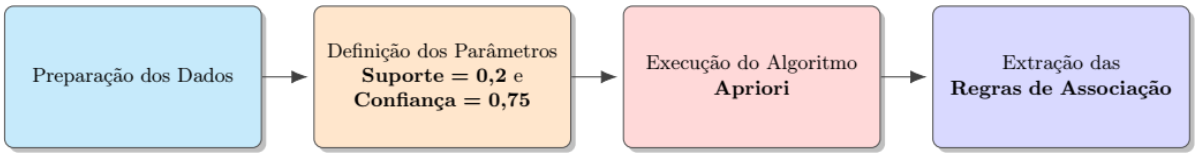


Figura 2: Fluxo Metodológico da Análise de Regras de Associação.

4.3.1 Preparação dos Dados

Para viabilizar a criação de regras de associação, foi necessário transformar a variável contínua de renda per capita em variáveis binárias, resultando na criação de faixas de renda categóricas. Essa transformação foi realizada com base nos quantis da renda, resultando nas seguintes faixas:

- Baixa Renda: Estudantes cuja renda per capita é inferior ou igual ao primeiro percentil (33%), definido em R$ 212,00.
- Renda Média: Estudantes cuja renda per capita está entre o primeiro e o segundo percentil, abrangendo rendas superiores a R$ 212,00 e inferiores ou iguais a R$ 516,77 (percentil 66%).
- Alta Renda: Estudantes cuja renda per capita é superior ao segundo quantil, ou seja, acima de R$ 516,77.

Com essa categorização, cada faixa de renda foi convertida em variável binária separada, representando a presença ou ausência de cada faixa nos registros. Essa transformação ajuda a obter uma melhor interpretação dos dados no contexto das regras de associação, além de possibilitar a análise das relações entre os diferentes níveis de renda e as demais variáveis socioeconômicas.

Além de que o algoritmo Apriori requer que os dados estejam em um formato transacional, onde cada registro deve representar uma transação única com variáveis binárias. As demais variáveis do conjunto de dados previamente se apresentavam no formato binário, o que facilitou sua utilização sem a necessidade de transformação adicional.

4.3.2 Execução do Algoritmo Apriori

Com os dados preparados, foi utilizado o algoritmo Apriori, implementado pela biblioteca *Python Mlxtend* (Raschka, 2018), para identificar padrões frequentes e extrair regras de associação. Esse algoritmo detecta combinações recorrentes de variáveis que ocorrem juntas, proporcionando uma base sólida para analisar como essas variáveis socioeconômicas se relacionam.

Os parâmetros inicialmente escolhidos para a execução do algoritmo foram os seguintes:

- Suporte mínimo (min_support): Definido inicialmente em 0,2, indicando que apenas as combinações de variáveis presentes em pelo menos 20% dos registros seriam consideradas. Esse critério ajuda a focar nas associações mais relevantes.
- Confiança mínima (min_confidence): Estabelecida em 0,75, o que significa que, para uma regra ser aceita, pelo menos 75% das ocorrências da combinação antecedente deve estar associada ao consequente. Essa escolha assegura que as regras identificadas sejam potencialmente úteis para a tomada de decisão.

Esses parâmetros foram ajustados para balancear a quantidade e a qualidade das regras extraídas, garantindo que as associações identificadas sejam significativas e representativas. A escolha do suporte e da confiança foi baseada em uma análise preliminar dos dados, visando filtrar padrões que pudessem revelar informações importantes sobre a relação entre as variáveis.

4.4 Análise de Agrupamento

Uma análise de agrupamento foi realizada com o intuito de identificar padrões entre os estudantes a partir das variáveis socioeconômicas coletadas. Esse método possibilita agrupar indivíduos com características similares, revelando perfis distintos dentro da base de dados. Nesta etapa do estudo, o algoritmo utilizado foi o *K-Modes*, aplicado através da biblioteca *kmodes* do *Python* (de Vos, 2015–2024).

A Figura 3 sintetiza o fluxo metodológico desenvolvido nesta etapa, que integra desde a preparação dos dados, a avaliação comparativa por diversas métricas (estatísticas e visuais), bem como a execução do algoritmo de agrupamento.

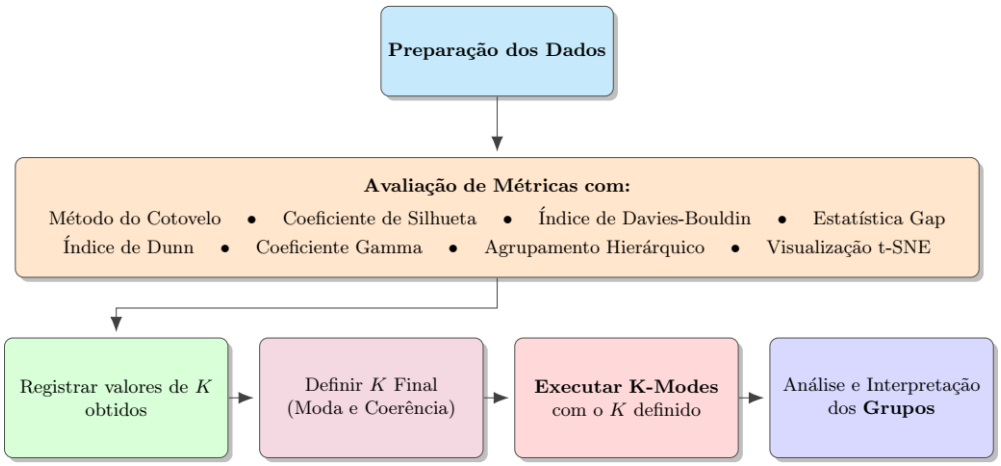


Figura 3: Fluxo Metodológico da Análise de Agrupamento.

4.4.1 Preparação dos Dados

A preparação dos dados para a análise de agrupamento foi feita considerando as características e requisitos específicos do algoritmo *K-Modes*.

O algoritmo K-Modes é adequado para dados categóricos, sendo necessário que todas as variáveis estejam neste formato. Assim, a variável de renda per capita, originalmente numérica, foi transformada em categorias de faixa de renda, sendo aplicada a mesma transformação utilizada para Regras de Associação na Subsubseção 4.3.1, de modo que a variável de renda numérica foi removida, e apenas a versão categorizada foi utilizada no modelo *K-Modes*, garantindo que todas as variáveis fossem categóricas e adequadas ao algoritmo.

4.4.2 Definição do Número de Grupos (K)

A escolha inicial do número de grupos foi guiada por uma combinação de métodos, métricas e análises visuais, com o objetivo de tornar mais robusta a definição do parâmetro **K**, utilizado como ponto de partida para a segmentação dos dados. Para isso, foram aplicadas diversas métricas de validação de grupos, como os índices de Dunn, Cotovelo, Silhueta, Davies-Bouldin (DBI), Gamma e Estatística GAP, além da análise visual de um dendrograma e da projeção dos dados por meio do algoritmo *t-SNE*. Essas abordagens foram utilizadas em conjunto com o algoritmo *K-Modes*, sendo amplamente reconhecidas na literatura por avaliarem tanto a coesão interna quanto a separação entre os grupos.

Inicialmente, foi utilizado o método do cotovelo. Nessa abordagem, a variação interna entre elementos de um mesmo grupo é medida em função do número de grupos, revelando o ponto em que a adição de novos grupos traz benefícios reduzidos na coesão dos dados. Esse ”ponto de cotovelo” representa um possível número ideal de grupos, além do qual os ganhos de segmentação passam a ser mínimos.

Na Figura 4, o gráfico possibilita notar o ”ponto de cotovelo” para o algoritmo *K-Modes*. Para auxiliar na definição desse ponto de inflexão, foi utilizada a biblioteca *kneed* (Satopaa et al.,

2011). O valor 3 apresenta-se como um bom candidato para a escolha da quantidade de grupos, pois a partir desse ponto, não há redução significativa na variação intra-grupo.

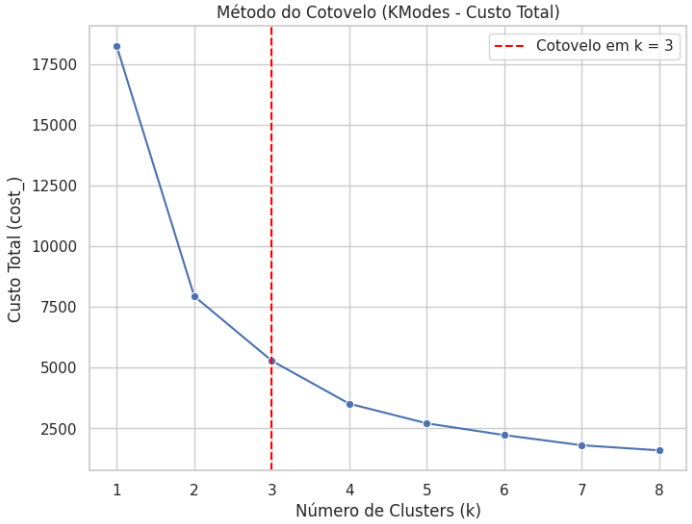


Figura 4: Gráfico do Cotovelo para o *K-Modes*.

O índice de Davies-Bouldin (DBI), que pode ser observado na Figura 5, mede a média da razão entre a dispersão intra-grupo e a separação inter-grupo. Neste trabalho, o DBI foi adaptado para utilizar a medida de dissimilaridade por correspondência (equivalente à distância de Hamming), com o intuito de melhor se adequar à natureza categórica dos dados utilizados no *K-Modes*. Quanto menor o valor do DBI, melhor a configuração dos grupos formados. Observa-se que o menor valor ocorre em $K = 2$, sugerindo uma estrutura de grupos com boa coesão interna e separação adequada entre os grupos.

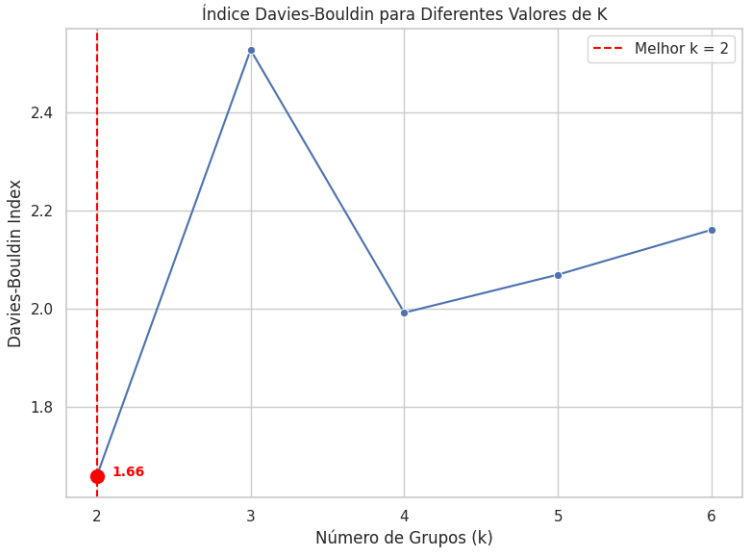


Figura 5: Gráfico do Índice Davies-Bouldin (DBI) *K-Modes*.

A métrica da Estatística GAP, representada na Figura 6, compara a variabilidade intra-grupo observada com a esperada em uma distribuição aleatória. Quanto maior o valor da estatística GAP, mais distinta é a segmentação obtida. No gráfico, nota-se um dos maiores valores em $K = 3$, reforçando essa escolha como a mais adequada para a divisão dos dados com o *K-Modes*.

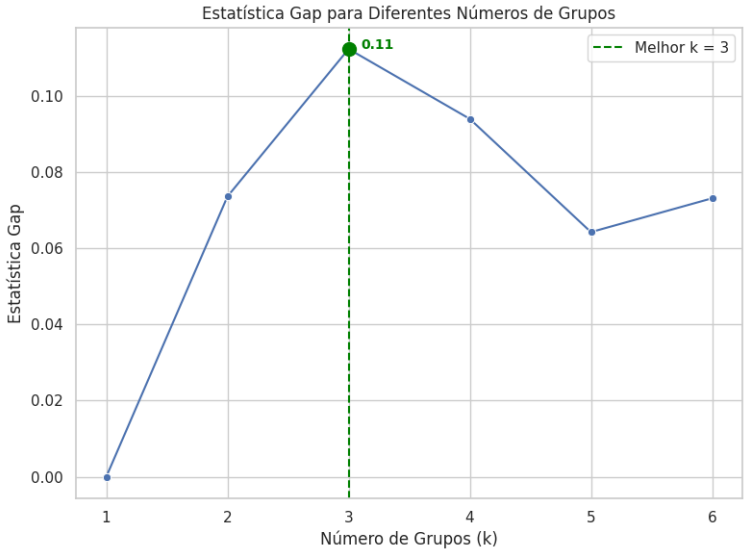


Figura 6: Gráfico do GAP *K-Modes*.

O índice Gamma, mostrado na Figura 7, avalia a concordância entre pares de observações com base nas distâncias intra e intergrupos. Valores mais altos de Gamma indicam melhor organização dos dados em grupos. Nesse caso, para o *K-Modes*, o pico mais acentuado ocorre em $K = 4$, que seria o valor ideal de acordo com essa métrica.

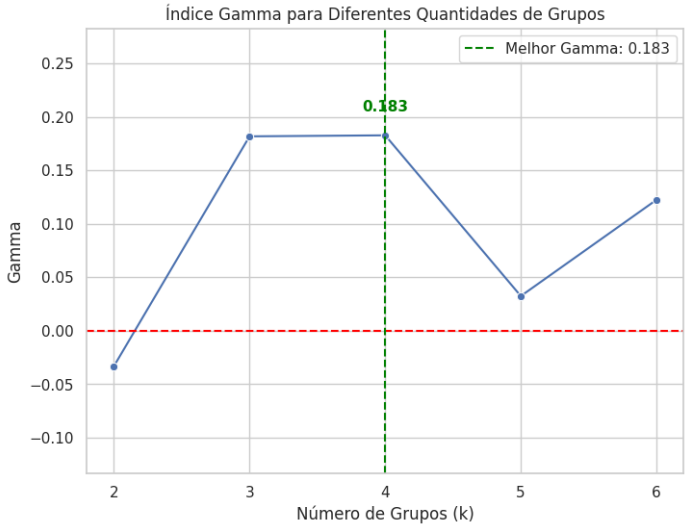


Figura 7: Gráfico do Índice Gamma *K-Modes*.

Por sua vez, o índice de Dunn, visualizado na Figura 8, busca maximizar a menor distância entre grupos e minimizar a maior dispersão interna. Um valor mais alto de Dunn aponta para uma

melhor definição dos grupos. O gráfico demonstra que $K = 6$ proporciona a maior pontuação, fortalecendo sua escolha como valor ideal, mas destoando dos outros métodos.

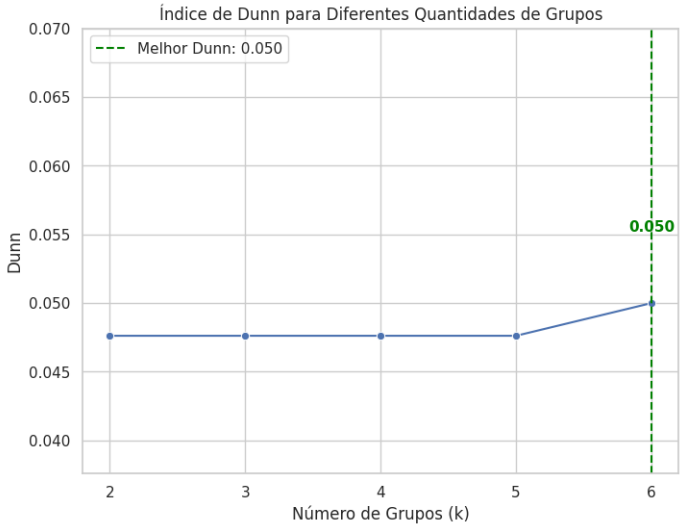


Figura 8: Gráfico do Índice Dunn *K-Modes*.

Embora o índice de Silhueta seja tradicionalmente mais utilizado com dados numéricos, ele também foi empregado como métrica complementar nesta análise, utilizando a distância de Hamming (ou de correspondência) para adequar-se às variáveis categóricas presentes. Como mostra a Figura 9, para o *K-Modes* o maior valor médio de silhueta foi observado em $K = 2$, sugerindo que os estudantes dentro de cada grupo estão, em média, mais bem associados entre si e mais distantes dos demais, indicando uma boa separação entre os grupos.

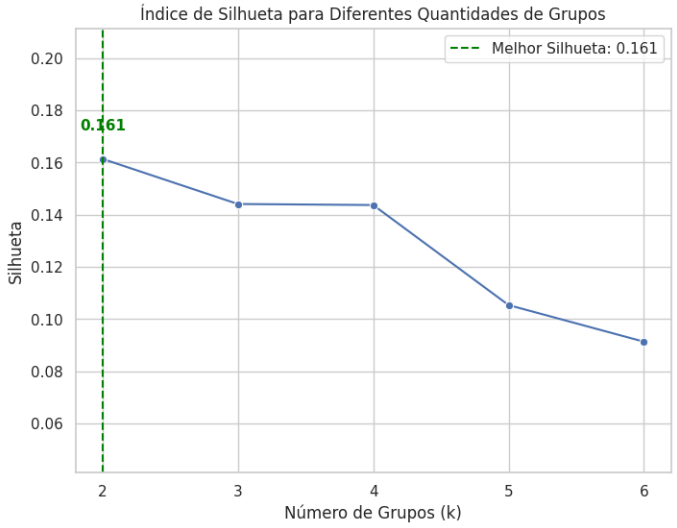
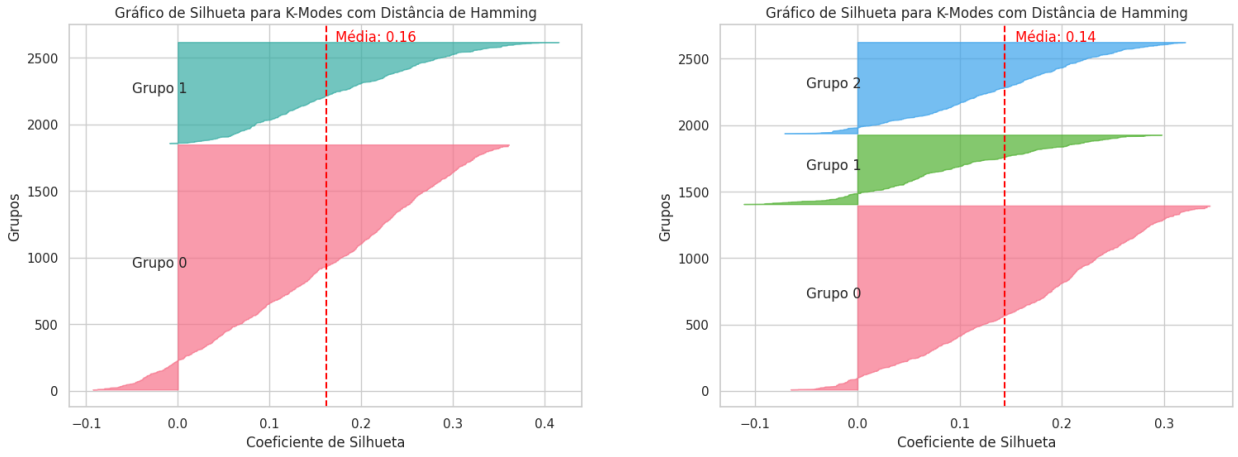


Figura 9: Gráfico do Índice Silhueta *K-Modes*.

Para complementar essa análise, foram gerados gráficos individuais com os valores de silhueta por grupo para $K = 2$ e $K = 3$, apresentados nas Figura 10a e Figura 10b. Esses gráficos

oferecem uma visão mais detalhada sobre a coesão interna e a separação de cada grupo. No caso de $K = 2$, os grupos exibem uma distribuição mais uniforme, com valores de silhueta mais elevados, o que reforça a ideia de que essa configuração oferece melhor qualidade de agrupamento. Por outro lado, para $K = 3$, apesar de ainda haver uma boa distinção entre os grupos, observa-se maior variação interna e a presença de observações com silhuetas mais baixas, o que pode sugerir sobreposição ou indefinição em alguns casos.



(a) Silhueta por grupo para $K = 2$

(b) Silhueta por grupo para $K = 3$

Figura 10: Comparação dos gráficos de Silhueta por grupo para $K = 2$ e $K = 3$ *K-Modes*.

O dendrograma apresentado na Figura 11 foi gerado com o método de agrupamento hierárquico aglomerativo (*bottom-up*), utilizando a distância de Hamming normalizada. Essa visualização mostra como os registros se unem progressivamente em grupos, de acordo com suas semelhanças. Ao traçar uma linha horizontal em torno do valor 0,5 no eixo vertical, é possível identificar três grupos bem definidos antes que os próximos agrupamentos envolvam junções entre grupos mais distantes (ou seja, menos semelhantes). Isso reforça a escolha de $K = 3$, também sugerida por outras métricas.

O *t-Distributed Stochastic Neighbor Embedding* (t-SNE) é um algoritmo de redução de dimensionalidade que preserva relações de vizinhança entre pontos, sendo particularmente útil para visualizar estruturas complexas de dados em duas ou três dimensões. Este método é eficaz para revelar padrões e agrupamentos naturais que podem não ser evidentes em espaços de alta dimensionalidade. No presente estudo, o t-SNE foi aplicado utilizando a distância de Hamming normalizada como métrica de dissimilaridade. A Figura 12 mostra a projeção resultante, onde é possível observar três regiões densas que sugerem uma estrutura de agrupamento subjacente. É importante ressaltar que esta visualização tem caráter exploratório e deve ser interpretada em conjunto com outras análises.

Com base nas métricas analisadas para o algoritmo *K-Modes*, observou-se um empate na escolha do valor ideal de K, com as métricas divididas entre $K = 2$ e $K = 3$, como pode ser observado no Tabela 3. Esse resultado indica que ambas as configurações oferecem segmentações com boa coesão interna e separação entre os grupos, sendo necessário considerar também aspectos qualitativos dos perfis formados para decidir entre essas opções. Dessa forma, considerando

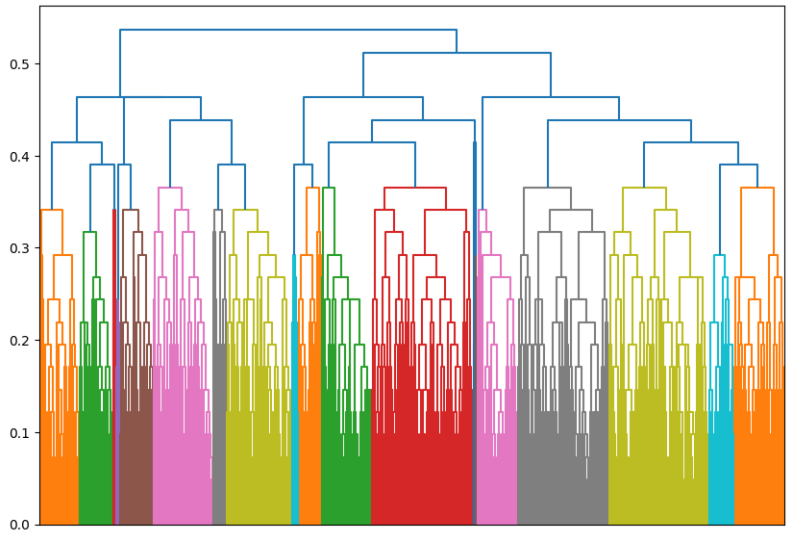


Figura 11: Gráfico do Dendrograma (Agrupamento Hierárquico) *K-Modes*.

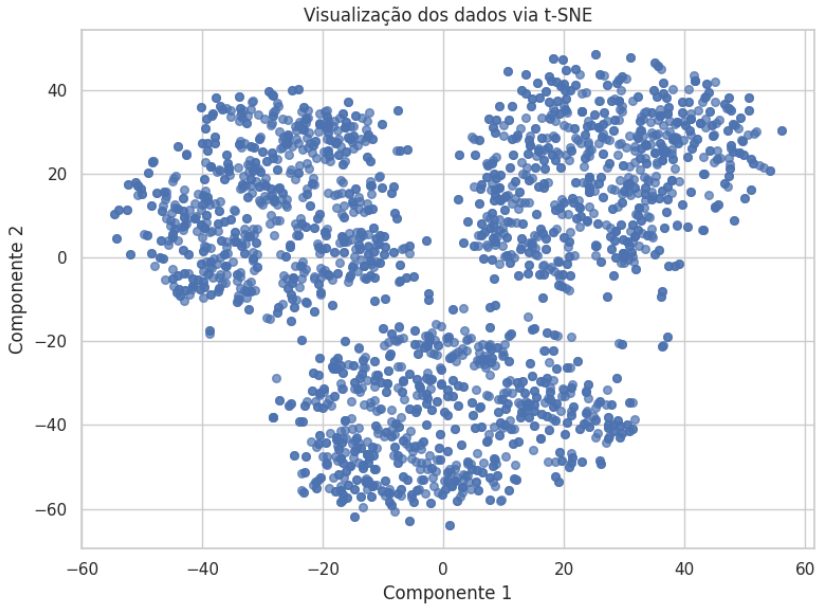


Figura 12: Gráfico de Visualização T-SNE (Bi-Dimensional) *K-Modes*.

também análises qualitativas por meio da visualização dos dados no t-SNE e do dendrograma, o valor $K = 3$ foi escolhido para as próximas análises.

4.5 Aplicação do Algoritmo de Agrupamento

Após a definição do número de grupos, foi realizado o agrupamento dos dados utilizando o algoritmo *K-Modes*. O método de inicialização dos centroides adotado foi o de Cao et al. (2009), que é determinístico e utiliza a distribuição de frequência dos dados para selecionar pontos iniciais representativos para cada grupo, buscando melhorar a estabilidade do agrupamento e reduzir a variabilidade nos resultados.

Tabela 3: Resultados das Métricas de Avaliação para Definição do Número de Grupos do K-Modes (K).

Métrica	**Valor sugerido de K**
Método do Cotovelo	3
Coeficiente de Silhueta	2
Índice de Davies-Bouldin	2
Estatística Gap	3
Índice de Dunn	6
Coeficiente Gamma	4
Agrupamento Hierárquico	3
t-SNE	3

Fonte: Elaboração própria.

Para medir a similaridade entre os dados e os centroides, foi utilizada a distância de Hamming, que avalia a quantidade de variáveis com valores diferentes entre as observações e o centroide ao qual foram atribuídas. Essa medida é adequada para a natureza categórica dos dados, pois permite uma quantificação direta das diferenças entre categorias.

5 Resultados

Nesta seção, são apresentados os resultados obtidos a partir da análise estatística, bem como a aplicação das técnicas de Aprendizado de Máquina não Supervisionado, Agrupamento e Mineração de Regras de Associação. Na Análise Estatística, foram investigadas as correlações entre variáveis, juntamente com a visualização de alguns gráficos. Para a Análise de Agrupamento, foi utilizado o algoritmo *K-Modes*, enquanto para a extração das Regras de associação, o algoritmo aplicado foi o Apriori.

5.1 Resultados Análise Estatística

Um teste de normalidade Shapiro-Wilk foi realizado para a variável contínua de Renda Per Capita, e os resultados, apresentados na Tabela 4, indicam que essa variável não segue uma distribuição normal. As demais variáveis, por serem binárias, também não atendem aos critérios de normalidade.

Tabela 4: Resultado do Teste de Normalidade Shapiro-Wilk para Renda Per Capita.

Estatística	**Valor**
Estatística do teste	0,9145
Valor-p	$5{,}42 \times 10^{-36}$
Distribuição	Não normal

Fonte: Elaboração própria.

Na Tabela 5, são apresentados os resultados obtidos para o coeficiente de correlação de Spearman entre as variáveis do estudo. Em comparação com outras abordagens testadas, a correlação de Spearman apresentou valores idênticos para as associações entre variáveis binárias, mas revelou coeficientes ligeiramente superiores nas correlações entre variáveis binárias e a variável

numérica de renda per capita. Considerando que a correlação de Spearman é mais apropriada para dados que não atendem aos pressupostos de normalidade, essa medida foi adotada na análise. Vale ressaltar que os p-valores associados às correlações ficaram muito próximos de zero, indicando que todas as correlações identificadas são estatisticamente significativas ($p < 0{,}05$).

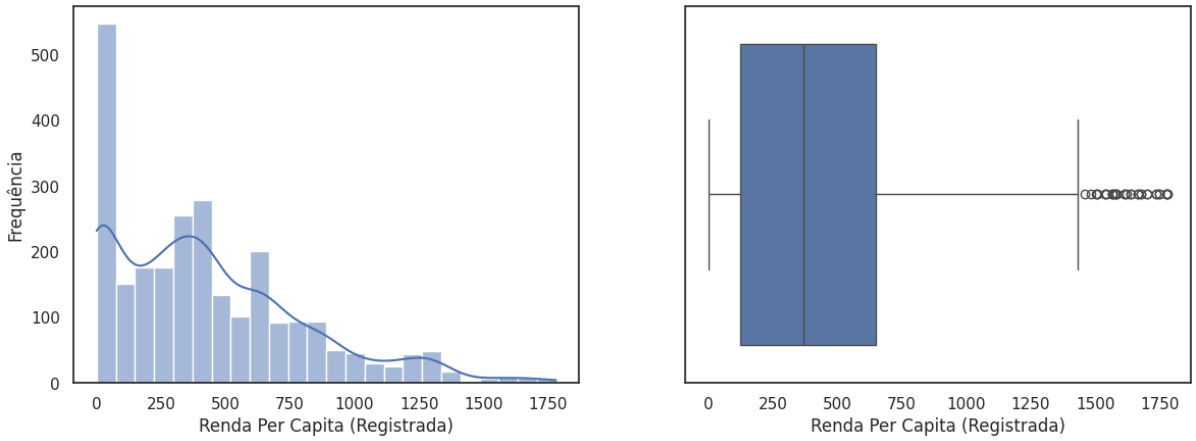
Tabela 5: 10 Maiores Correlações (*Spearman*) entre Variáveis.

Variável 1	**Variável 2**	**Correlação**	**Valor-p**
Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	Renda per capita (registrada)	-0,6249	0,0000
Estudante reside em outro município (sem o núcleo familiar) em função dos estudos	Estudante reside em república estudantil	0,5811	0,0000
Inserção em programas sociais	Renda per capita (registrada)	-0,5342	0,0000
Escolaridade dos pais ou responsável: Ensino Fundamental ou equivalente incompleto	Escolaridade dos pais ou responsável: Ensino Médio ou equivalente completo ou Superior incompleto	-0,5317	0,0000
Transporte intermunicipal cedido pela prefeitura sem contrapartida do estudante	Transporte urbano municipal (ônibus, moto táxi, etc.)	-0,4928	0,0000
Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	Inserção em programas sociais	0,4445	0,0000
Dificuldades de acesso à universidade considerando a distância	Dificuldades de acesso à universidade considerando o tempo de deslocamento	0,4351	0,0000
Escolaridade dos pais ou responsável: Ensino Fundamental ou equivalente completo ou Ensino Médio ou equivalente incompleto	Escolaridade dos pais ou responsável: Ensino Médio ou equivalente completo ou Superior incompleto	-0,3068	0,0000
Transporte urbano municipal (ônibus, moto táxi, etc.)	Dificuldades de acesso à universidade considerando a distância	-0,2994	0,0000
Situações de violência (física, sexual, psicológica, patrimonial e moral).	Sexo / Identidade de Gênero / Orientação Sexual	0,2937	0,0000

Fonte: Elaboração própria.

Ao analisar a Figura 13a, que apresenta o gráfico de distribuição da Renda Per Capita, observa-se uma alta concentração de estudantes, mais de 500, com renda próxima de R$ 0. Em seguida, há uma concentração moderada, com cerca de 250 alunos, cuja renda varia entre R$ 300 e R$ 400. À medida que a renda aumenta, o número de estudantes em cada faixa de renda diminui gradualmente, indicando uma distribuição decrescente nos níveis de renda mais elevados. Na Figura 13b, também se observa uma assimetria positiva, com maior variabilidade entre as rendas mais elevadas, além da presença de diversos *outliers* no limite superior, indicando uma dispersão considerável nos níveis mais altos de renda.

A estratificação do *box plot* de Renda Per Capita pela variável que indica a participação dos estudantes em programas sociais, conforme mostrado na Figura 14a, revela que os alunos não inseridos em tais programas apresentam uma distribuição semelhante à do *box plot* geral, com a mediana em torno de R$ 500. Por outro lado, entre os estudantes que participam de programas sociais, observa-se um deslocamento significativo, com a mediana aproximando-se de R$ 200.



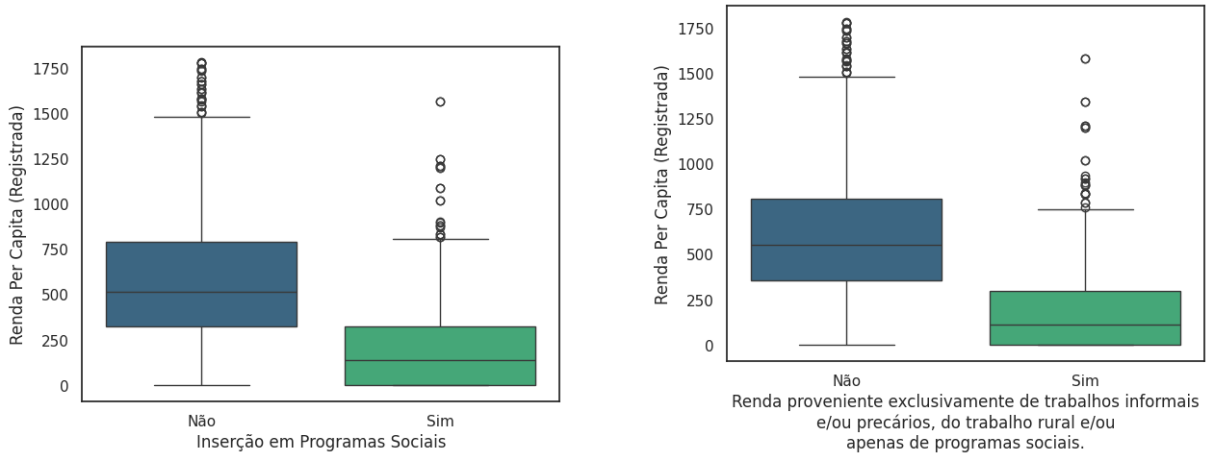
(a) Histograma da Renda Per Capita

(b) *Boxplot* da Renda Per Capita

Figura 13: Distribuição da variável Renda Per Capita.

Isso indica que metade desses alunos possui renda abaixo desse valor. O limite superior, ou seja, o maior valor observado, foi de cerca de R$ 750, o que se assemelha ao terceiro quartil dos alunos não participantes. Esses dados evidenciam a renda mais baixa dos estudantes inseridos em programas sociais. Em seguida, foram gerados alguns gráficos para ajudar a entender melhor as variáveis e suas relações.

De maneira similar, esse contraste entre a renda também ocorre na estratificação da renda pela variável que mede se a origem da renda é precarizada ou informal, como pode ser observado na Figura 14b.



(a) *Boxplot* por inserção em programa social

(b) *Boxplot* por origem da renda

Figura 14: *Boxplots* da Renda Per Capita por características socioeconômicas.

5.2 Resultados Agrupamento

Os grupos resultantes foram analisados com base nas modas das variáveis, que representam as características centrais ou predominantes em cada grupo, funcionando como os “perfis” dos estu-

dantes pertencentes a cada grupo. Para facilitar a interpretação dos resultados, foram destacadas apenas as variáveis com valor 1 (Verdadeiro) em pelo menos um grupo, dado que o valor 0 (Falso) era comum a todos os grupos para as demais variáveis. Além disso, foram examinadas a quantidade de estudantes em cada grupo e a porcentagem de valores 1 (Verdadeiro) para as variáveis mais relevantes em cada grupo, o que permitiu caracterizar os grupos em relação aos principais fatores socioeconômicos. Esse enfoque busca proporcionar uma visão mais clara sobre as diferenças nos perfis estudantis, que podem indicar necessidades específicas de assistência.

5.2.1 Agrupamento com K-Modes

Para esse algoritmo, com $K = 3$, foi obtido um agrupamento com uma quantidade maior de estudantes pertencentes ao Grupo 1, seguido pelos grupos 3 e 2, conforme mostrado na Figura 15.

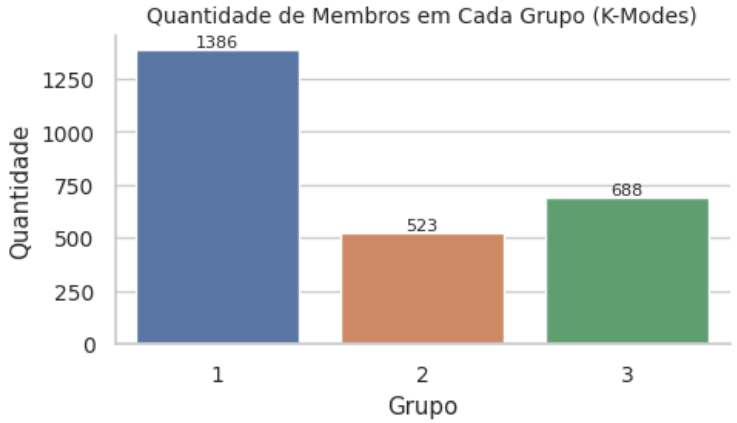


Figura 15: Quantidade de Membros em Cada Grupo (*K-Modes*).

Na Tabela 6, a moda, valores que mais se repetem, para cada variável é mostrada, possibilitando fazer uma diferenciação nos perfis dos grupos obtidos. Percebe-se que o Grupo 1, apresentou renda mais alta que os demais grupos, e se destaca também por utilizar transporte urbano municipal, algo que não ocorre nos outros grupos. Os Grupos 2 e 3 apresentam características parecidas em relação à renda, ambos na faixa mais baixa, e as variáveis com influência nela (origem precária/informal e inserção em programas sociais).

A diferenciação entre os Grupos 2 e 3 se dá pelos indicadores referentes ao transporte. O Grupo 2 apresenta uso de transporte intermunicipal cedido por prefeitura, apresentando dificuldades no transporte tanto quanto à distância quanto ao tempo de deslocamento. A escolaridade dos pais como ensino médio completo ou superior incompleto, também foi um diferencial nesse grupo.

Tabela 6: Moda e Percentual de Ocorrência de 1 (Verdadeiro) dos Grupos *K-Modes*.

Indicador	Grupo 1	Grupo 2	Grupo 3
Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	0 (11.11%)	1 (67.30%)	1 (86.48%)
Identificação racial	1 (67.24%)	1 (72.85%)	1 (74.71%)
Inserção em programas sociais	0 (10.68%)	1 (67.30%)	1 (75.87%)
Transporte intermunicipal cedido pela prefeitura sem contrapartida do estudante	0 (22.58%)	1 (80.31%)	0 (23.84%)
Transporte urbano municipal (ônibus, moto táxi, etc.)	1 (61.04%)	0 (12.05%)	0 (46.22%)
Dificuldades de acesso à universidade considerando a distância	0 (14.86%)	1 (69.22%)	0 (5.09%)
Dificuldades de acesso à universidade considerando o tempo de deslocamento	0 (20.63%)	1 (66.73%)	0 (4.51%)
Ensino Médio em escola pública	1 (69.05%)	1 (79.54%)	1 (80.96%)
Escolaridade dos pais ou responsável: Ensino Médio ou equivalente completo ou Superior incompleto	0 (41.63%)	1 (52.20%)	0 (27.33%)
Faixa de Renda Baixa	0 (3.39%)	1 (57.17%)	1 (74.42%)
Faixa de Renda Alta	1 (60.68%)	0 (7.27%)	0 (0.58%)

Fonte: Elaboração própria.

As diferenças apontadas entre os grupos também podem ser observadas na Figura 16, na qual pode ser observada a porcentagem de valores 1 (Verdadeiro) para cada uma das variáveis analisadas anteriormente.

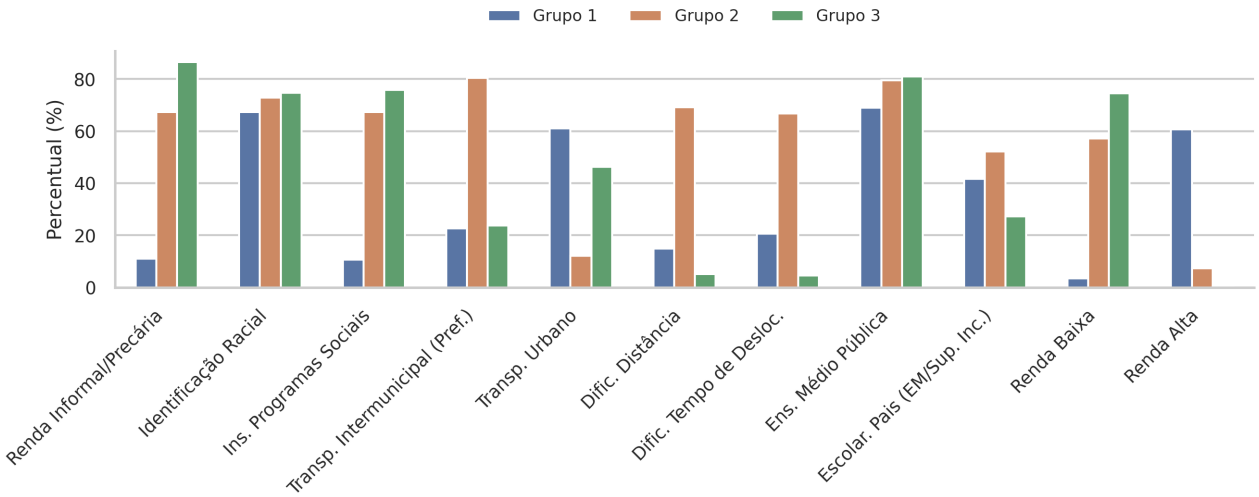


Figura 16: Percentual de Valores 1 (Verdadeiro) *K-Modes*.

5.3 Resultados Regras de Associação

Após execução do algoritmo com os parâmetros mencionados na Subseção 4.3, foram obtidos um total de 20 regras de associação que podem ser observadas no Tabela 7 e no Tabela 8. Nota-se uma grande ocorrência da variável que indica que o estudante cursou ensino médio em escola pública, que se relaciona com diversas outras variáveis, como a renda, escolaridade dos pais, origem da

renda precarizada e inserção em programas sociais. Isoladamente, essa variável tem um suporte de 0,74, indicando que está presente em 74% das observações, o que ajuda a entender a presença dessa variável em diversas regras.

Tabela 7: Regras de Associação Obtidas (Parte 1).

ID	Antecedentes	Consequentes	Ant. Sup.	Cons. Sup.	Sup.	Conf.	*Lift*
0	Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	Ensino Médio em escola pública	0.42	0.74	0.33	0.77	1.04
1	Faixa Baixa	Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	0.33	0.42	0.28	0.83	1.97
2	Identificação racial	Ensino Médio em escola pública	0.70	0.74	0.53	0.76	1.02
3	Escolaridade dos pais ou responsável: Ensino Fundamental ou equivalente incompleto	Identificação racial	0.30	0.70	0.23	0.75	1.07
4	Inserção em programas sociais	Ensino Médio em escola pública	0.39	0.74	0.32	0.80	1.08
5	Transporte intermunicipal cedido pela prefeitura sem contrapartida do estudante	Ensino Médio em escola pública	0.35	0.74	0.29	0.83	1.11
6	Dificuldades de acesso à universidade considerando o tempo de deslocamento	Ensino Médio em escola pública	0.26	0.74	0.20	0.79	1.06
7	Escolaridade dos pais ou responsável: Ensino Fundamental ou equivalente incompleto	Ensino Médio em escola pública	0.30	0.74	0.25	0.82	1.10
8	Faixa Baixa	Ensino Médio em escola pública	0.33	0.74	0.26	0.79	1.06
9	Identificação racial, Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	Ensino Médio em escola pública	0.31	0.74	0.24	0.78	1.04

Fonte: Elaboração própria.

Tabela 8: Regras de Associação Obtidas (Parte 2).

ID	Antecedentes	Consequentes	Ant. Sup.	Cons. Sup.	Sup.	Conf.	*Lift*
10	Faixa Baixa, Identificação racial	Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	0.24	0.42	0.20	0.84	1.97
11	Inserção em programas sociais, Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	Ensino Médio em escola pública	0.27	0.74	0.22	0.80	1.08
12	Faixa Baixa, Inserção em programas sociais	Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	0.24	0.42	0.21	0.90	2.11
13	Faixa Baixa, Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	Inserção em programas sociais	0.28	0.39	0.21	0.77	1.96
14	Inserção em programas sociais, Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	Faixa Baixa	0.27	0.33	0.21	0.78	2.35
15	Faixa Baixa, Ensino Médio em escola pública	Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	0.26	0.42	0.22	0.85	2.00
16	Faixa Baixa, Renda proveniente exclusivamente de trabalhos informais e/ou precários, do trabalho rural e/ou apenas de programas sociais.	Ensino Médio em escola pública	0.28	0.74	0.22	0.80	1.08
17	Identificação racial, Inserção em programas sociais	Ensino Médio em escola pública	0.29	0.74	0.24	0.82	1.11
18	Ensino Médio em escola pública, Inserção em programas sociais	Identificação racial	0.32	0.70	0.24	0.76	1.07
19	Identificação racial, Transporte intermunicipal cedido pela prefeitura sem contrapartida do estudante	Ensino Médio em escola pública	0.25	0.74	0.21	0.83	1.12

Fonte: Elaboração própria.

6 Discussão

Esta seção interpreta os padrões revelados pelos algoritmos de agrupamento e regras de associação, visando compreender o perfil socioeconômico dos estudantes analisados. Discute-se a relevância das variáveis identificadas e as implicações práticas desses achados para o aprimoramento das políticas de permanência estudantil.

6.1 Discussão Análise Estatística

Os resultados da análise de correlação destacam associações relevantes entre variáveis socioeconômicas e a renda dos estudantes. As variáveis relacionadas à renda apresentaram correlações consistentes com diversos indicadores de vulnerabilidade, o que pode indicar a influência significativa desses fatores na condição econômica dos discentes.

Observou-se, ainda, que as variáveis referentes à escolaridade dos pais apresentaram correlações que sugerem uma relação de exclusividade entre os níveis de escolarização, reforçando a importância da trajetória educacional familiar na caracterização dos perfis socioeconômicos.

No que diz respeito às variáveis relacionadas ao transporte, foram identificadas correlações que indicam uma possível associação entre o tipo de transporte utilizado e as dificuldades de acesso à universidade. Esse achado pode ser relevante para a formulação de políticas voltadas à mobilidade estudantil.

Além disso, embora com coeficientes baixos, verificou-se uma correlação entre variáveis relacionadas a situações de violência e aquelas que indicam gênero ou orientação sexual. Essa associação, ainda que sutil, pode apontar para a existência de riscos diferenciados enfrentados por estudantes de determinadas identidades de gênero ou orientações sexuais, sugerindo a necessidade de atenção a essas vulnerabilidades.

Em relação aos métodos utilizados, notou-se que as variáveis binárias apresentaram valores idênticos de correlação entre si, independentemente do método empregado. Por outro lado, as correlações envolvendo a variável contínua de renda per capita variaram entre os métodos: o coeficiente de Spearman apresentou os maiores valores, seguido por Pearson e Tau de Kendall.

Os resultados confirmam que a baixa renda, especialmente quando associada a vínculos informais e programas sociais, demarca perfis socioeconômicos distintos. A correlação observada expõe a disparidade econômica e reitera a assistência estudantil como mecanismo de suporte indispensável para a permanência desses grupos.

6.2 Discussão Resultados do Agrupamento

Os resultados obtidos com o algoritmo *K-Modes* revelaram uma diferenciação significativa entre as variáveis binárias, especialmente em relação às variáveis de transporte. As diferenças entre os grupos foram mais evidentes em aspectos como o uso de transporte urbano municipal e o transporte intermunicipal cedido pela prefeitura, permitindo identificar grupos com perfis distintos de acesso ao transporte, o que pode estar relacionado a fatores socioeconômicos e logísticos. Assim, o *K-Modes* proporcionou uma análise detalhada das variáveis categóricas, capturando as

variações na distribuição dessas características e refletindo a importância desses elementos na composição do perfil dos estudantes.

6.3 Discussão Resultados das Regras de Associação

Os resultados obtidos com as regras de associação reforçam os achados da análise de correlações, evidenciando novamente a importância da renda per capita como variável central na identificação de vulnerabilidades socioeconômicas. A renda aparece nas regras exclusivamente na faixa de baixa renda, frequentemente associada à origem da renda e à participação em programas de assistência.

Regras como a 1, 10, 12, 13, 14 e 15 ilustram de forma clara essas inter-relações. Em especial, a regra 10 apresenta como antecedentes a identificação racial e a faixa de renda baixa, e como consequente a renda de origem precária. O *lift* dessa regra é de 1,97, o que indica que, na presença do antecedente, a chance de ocorrência do consequente quase dobra. Situação semelhante é observada na regra 12, que substitui a variável de identificação racial pela participação em programas sociais no antecedente, apresentando um *lift* de 2,11, ou seja, mais que dobrando a probabilidade de ocorrência do consequente.

Além dessas associações que poderiam ser consideradas esperadas, emergem novas relações que também merecem destaque. A regra 3 aponta que níveis mais baixos de escolaridade dos pais (ensino fundamental incompleto ou equivalente) estão associados à variável de identificação racial, sugerindo um possível ciclo intergeracional de vulnerabilidade. Por outro lado, a regra 18 mostra que a combinação entre ter cursado o ensino médio em escola pública e participar de programas sociais está associada a determinados grupos raciais, revelando como fatores estruturais interagem para compor os perfis de vulnerabilidade entre os estudantes.

6.4 Implicações dos Resultados e Perfis de Vulnerabilidade

A análise dos agrupamentos evidenciou perfis socioeconômicos distintos: enquanto o Grupo 1 apresenta o menor grau de vulnerabilidade, com concentração nas faixas de renda mais altas, os Grupos 2 e 3 reúnem os perfis mais críticos, marcados por baixa renda, vínculos informais e dependência de programas sociais. O Grupo 2 destaca-se especificamente pela forte necessidade de transporte cedido por prefeituras, diferenciando-se do Grupo 3, o que sinaliza o deslocamento como uma barreira adicional à permanência e reforça a necessidade de políticas assistenciais focadas na mobilidade estudantil.

As regras de associação, reforçaram a renda como eixo central das vulnerabilidades ao associá-la a fontes informais ou programas sociais, além de outras variáveis, como cor/raça e tipo de escola cursada.

Ressalta-se, contudo, o viés de seleção inerente ao uso exclusivo de dados dos inscritos. Estudantes vulneráveis excluídos do processo por barreiras informacionais ou tecnológicas não estão representados, o que sugere que a análise pode subestimar a real vulnerabilidade presente na universidade.

Nesse sentido, vale ressaltar o contraponto entre a análise documental tradicional e os *insights* obtidos por essas técnicas de AM. Embora a verificação humana realizada pela equipe

técnica seja indispensável para validar a veracidade das informações e a elegibilidade legal, a abordagem computacional pode ampliar a assertividade da concessão ao revelar padrões de vulnerabilidade que escapam a uma triagem puramente burocrática. Assim, tanto os agrupamentos quanto as regras geradas não devem ser utilizados como critério automático de seleção, mas sim para apoiar a gestão, permitindo identificar perfis variados e orientar decisões mais justas.

Esses achados indicam que a vulnerabilidade estudantil é multifacetada e exige uma abordagem integrada. Em termos práticos, indicam a necessidade de políticas de permanência que articulem o fortalecimento da representatividade racial e programas de transporte subsidiado ou em parceria com prefeituras, com ações voltadas à melhoria e articulação com o ensino médio público, visando não apenas garantir a locomoção, mas também mitigar as lacunas de formação básica que perpetuam a desigualdade no ensino superior.

Por fim, a interpretação dos resultados deve considerar ameaças à validade. Quanto à validade de construção e interna, o uso de variáveis socioeconômicas apenas como indicadores de um fenômeno complexo (vulnerabilidade) e o recorte amostral restrito aos inscritos podem não capturar todas as dimensões do problema ou excluir estudantes invisibilizados (viés de seleção). Em relação à validade externa, a especificidade regional restringe a generalização dos resultados, embora o fluxo metodológico seja replicável em outras IFES. Sobre a validade da conclusão, a ausência de métricas de acurácia em aprendizagem não supervisionada introduz subjetividade na interpretação; contudo, o uso rigoroso de métricas estatísticas de validação interna mitiga essa ameaça, permanecendo a validação experimental como trabalho futuro.

7 Considerações Finais e Trabalhos Futuros

Este trabalho aplicou técnicas de aprendizado de máquina não supervisionado, combinadas com regras de associação, para analisar dados socioeconômicos de estudantes interessados em programas de assistência estudantil. A proposta permitiu identificar grupos distintos de estudantes, assim como padrões de associação entre variáveis que refletem diferentes contextos de vulnerabilidade social.

Nos agrupamentos, as variáveis mais relevantes foram a renda familiar, a dificuldade com transporte e o uso de transporte cedido por prefeituras. Essas características permitiram identificar um grupo com maior vulnerabilidade relacionada à mobilidade, o que pode representar um obstáculo significativo à permanência acadêmica. No que diz respeito às regras de associação, destacaram-se a renda, a origem da renda, a escolaridade dos pais, a identificação racial e a participação em programas sociais como as variáveis mais influentes, evidenciando padrões recorrentes entre estudantes com maior vulnerabilidade socioeconômica.

Como trabalhos futuros, propõe-se incorporar dados acadêmicos — como desempenho em disciplinas, índice de reprovação e tempo médio de integralização — e realizar a validação interpretativa dos clusters com especialistas e gestores da assistência estudantil, a fim de investigar de modo mais robusto possíveis relações entre perfis socioeconômicos e indicadores de rendimento e permanência acadêmica dos estudantes. Outra proposta é realizar uma análise temporal, caso os dados estejam organizados por ano ou semestre, a fim de investigar se há mudanças nos per-

fis identificados ao longo do tempo. Tais avanços podem contribuir para o desenvolvimento de políticas de assistência estudantil mais direcionadas e sensíveis à realidade dos alunos.

Referências

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 207–216. https://doi.org/10.1145/170035.170072 [GS Search].

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press. [GS Search].

Bennasar-Veny, M., Yañez, A. M., Pericas, J., Ballester, L., Fernandez-Dominguez, J. C., Tauler, P., & Aguilo, A. (2020). Cluster analysis of health-related lifestyles in university students. *International journal of environmental research and public health*, *17*(5), 1776. https://doi.org/10.3390/ijerph17051776 [GS Search].

Brasil. (1988). *Constituição da República Federativa do Brasil de 1988*. Recuperado agosto 25, 2024, de https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm

Brasil. (2010). Decreto Nº 7.234, de 19 de julho de 2010. *Diário Oficial [da] República Federativa do Brasil*. Recuperado agosto 25, 2024, de https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2010/decreto/d7234.htm

Cao, F., Liang, J., & Bai, L. (2009). A new initialization method for categorical data clustering. *Expert Systems with Applications*, *36*(7), 10223–10228. https://doi.org/10.1016/j.eswa.2009.01.060 [GS Search].

Costa, S. G. (2009). A permanência na educação superior no Brasil: uma análise das políticas de assistência estudantil. *Colóquio Internacional sobre Gestão Universitária na América do Sul*, *9*, 1–13. http://repositorio.ufsc.br/xmlui/handle/123456789/37031 [GS Search].

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. Em *Machine learning techniques for multimedia: case studies on organization and retrieval* (pp. 21–49). Springer. https://doi.org/10.1007/978-3-540-75171-7_2 [GS Search].

de Vos, N. J. (2015–2024). kmodes categorical clustering library. https://github.com/nicodv/kmodes

Dias, S. M. B., & Costa, S. L. (2015). A permanência no ensino superior e as estratégias institucionais de enfrentamento da evasão. *Jornal de políticas educacionais*, *9*(17/18). https://doi.org/10.5380/jpe.v9i17/18.38650 [GS Search].

Dutra, N. G. R., & Santos, M. F. S. (2017). Assistência estudantil sob múltiplos olhares: a disputa de concepções. *Ensaio: avaliação e políticas públicas em educação*, *25*, 148–181. https://doi.org/10.1590/S0104-40362017000100006 [GS Search].

Ghahramani, Z. (2003). Unsupervised learning. Em *Summer school on machine learning* (pp. 72–112). Springer. https://doi.org/10.1007/978-3-540-28650-9_5 [GS Search].

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques* (3ª ed.). Morgan Kaufmann. https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques [GS Search].

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, *2*(3), 283–304. https://doi.org/10.1023/A:1009769707641 [GS Search].

Imperatori, T. K. (2017). A trajetória da assistência estudantil na educação superior brasileira. *Serviço Social & Sociedade*, 285–303. https://doi.org/10.1590/0101-6628.109 [GS Search].

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. Springer Nature. https://doi.org/10.1007/978-3-031-38747-0 [GS Search].

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2 [GS Search].

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, *13*(1), 5705. https://doi.org/10.1038/s41598-023-32484-w [GS Search].

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Em S. van der Walt & J. Millman (Ed.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a [GS Search].

Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. *Applied Sciences*, *12*(19), 9467. https://doi.org/10.3390/app12199467 [GS Search].

Monard, M. C., & Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. Em S. O. Rezende (Ed.), *Sistemas inteligentes: Fundamentos e aplicações*. Manole. https://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf [GS Search].

Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: application and trends. *Artificial Intelligence Review*, *56*(7), 6439–6475. https://doi.org/10.1007/s10462-022-10325-y [GS Search].

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *The Journal of Open Source Software*, *3*(24). https://doi.org/10.21105/joss.00638 [GS Search].

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson. [GS Search].

Santos, F. D., Bercht, M., Wives, L., & Cazella, S. (2015). Análise de evidências do estado de ânimo desanimado de alunos de um AVEA: uma proposta a partir da aplicação de regras de associação. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, *4*(1), 1054. http://milanesa.ime.usp.br/rbie/index.php/wcbie/article/view/6213 [GS Search].

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "kneedle" in a haystack: Detecting knee points in system behavior. *31st International Conference on Distributed Computing Systems Workshops*, 166–171. https://doi.org/10.1109/ICDCSW.2011.20 [GS Search].

Silva, L. A., Morino, A. H., & Sato, T. M. C. (2014). Prática de mineração de dados no Exame Nacional do Ensino Médio. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, *3*(1), 651. http://milanesa.ime.usp.br/rbie/index.php/wcbie/article/view/3289 [GS Search].

Silva, V. A. A., Moreno, L. L. O., Gonçalves, L. B., Soares, S. S. R. F., & Souza Júnior, R. R. (2020). Identificação de Desigualdades Sociais a partir do desempenho dos alunos do Ensino Médio no ENEM 2019 utilizando Mineração de Dados. *Anais do XXXI Simpósio*

Brasileiro de Informática na Educação, 72–81. https://doi.org/10.5753/cbie.sbie.2020.72 [GS Search].

Simon, A., & Cazella, S. (2017). Mineração de Dados Educacionais nos Resultados do ENEM de 2015. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, *6*(1), 754. http://milanesa.ime.usp.br/rbie/index.php/wcbie/article/view/7461 [GS Search].

Soares, G. C. (2020). *SAM: uma abordagem específica de mineração de dados socioeconômicos de alunos do IF Amazonas para apoio ao processo de concessão de assistência estudantil* [diss. de mestr., Universidade Federal de Pernambuco]. https://repositorio.ufpe.br/handle/123456789/38104 [GS Search].

Sun, Y., Li, Z., Li, X., & Zhang, J. (2021). Classifier selection and ensemble model for multi-class imbalance learning in education grants prediction. *Applied Artificial Intelligence*, *35*(4), 290–303. https://doi.org/10.1080/08839514.2021.1877481 [GS Search].

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press. [GS Search].

Tinto, V. (2012). *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago Press. [GS Search].

Vasconcelos, N. B. (2010). Programa Nacional de Assistência Estudantil: uma análise da evolução da assistência estudantil ao longo da história da educação superior no Brasil/National Student Assistance Program: an analysis of the evolution of student assistance along the history of. *Ensino em Re-vista*. https://seer.ufu.br/index.php/emrevista/article/view/11361 [GS Search].

Villar, A., & Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, *4*(1), 2. https://doi.org/10.1007/s44163-023-00079-z [GS Search].

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2 [GS Search].

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021 [GS Search].

Zhang, C., & Zhang, S. (2002). *Association rule mining: models and algorithms*. Springer. https://doi.org/10.1007/3-540-46027-6 [GS Search].

Zhao, Q., & Bhowmick, S. S. (2003). Association rule mining: A survey. *Nanyang Technological University, Singapore*, *135*, 18. [GS Search].