Refas
11 anos
Revista Fatec Zona Sul
1ª. Edição Outubro 2014
ISSN 2359-182X

ANÁLISE DE DADOS EDUCACIONAIS: APLICAÇÃO DE TÉCNICA DE MINERAÇÃO DE DADOS PARA O ESTUDO DA EVASÃO ESCOLAR

EDUCATIONAL DATA ANALYSIS: APPLICATION OF DATA MINING TECHNIQUES TO THE STUDY OF SCHOOL DROPOUT

ANÁLISIS DE DATOS EDUCATIVOS: APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA ESTUDIAR LA DESERCIÓN ESCOLAR

Isabella Lie Oshima[1]
Maria das Graças Junqueira Machado Tomazela[2]

DOI: 10.26853/Refas_ISSN-2359-182X_v12n05_01

Artigo recebido em dezembro de 2025
Artigo aceito em abril de 2026

RESUMO

A evasão escolar é um problema social persistente que afeta o desenvolvimento da educação no Brasil. Resolver esse problema exige enfrentar outros desafios relacionados, que influenciam na permanência ou não da criança nos estudos. Nesse contexto, as tecnologias de análise de dados e inteligência artificial oferecem ferramentas poderosas para identificar fatores que influenciam o abandono dos estudos. Assim, o objetivo deste trabalho foi utilizar técnicas de mineração de dados para identificar padrões e analisar os principais fatores associados à evasão escolar. Para isso, realizou-se uma pesquisa quantitativa, exploratória e descritiva, com posterior interpretação qualitativa dos resultados. Inicialmente, foi realizada uma pesquisa bibliográfica para levantamento dos conceitos-chave: Descoberta de Conhecimento em Bases de Dados, *Clusterização* e algoritmo *k-means*; e de trabalhos relacionados. Os dados foram coletados da PNAD Contínua de 2021 a 2024, em seguida foi realizada a limpeza e transformação desses dados, a *clusterização* e, por fim, a análise dos *clusters*. Os resultados demonstraram que, por meio das etapas de preparação dos dados e da aplicação de uma técnica de mineração de dados, foi possível identificar padrões relevantes relacionados à evasão escolar. Entre os resultados obtidos, destacou-se a relação entre a evasão escolar e a necessidade de conciliar trabalho e estudo, evidenciando esse fator como um elemento relevante no abandono escolar. Nesse sentido, essa pesquisa pode contribuir para que gestores educacionais e instituições de ensino desenvolvam estratégias voltadas à permanência escolar, como ações de acompanhamento ou medidas de apoio aos estudantes que trabalham.

Palavras-chave: Educação. Inteligência artificial. *Clusterização*. Abandono escolar.

[1] Graduanda em Análise e Desenvolvimento de Sistemas pela Faculdade de Tecnologia de Indaiatuba. E-mail: isabella.oshima@fatec.sp.gov.br. Lattes: http://lattes.cnpq.br/6810603107798675. OrcId: https://orcid.org/0009-0005-1422-5125.

[2] Doutora em Engenharia de Produção pela Universidade Metodista de Piracicaba. Docente da Fatec Sorocaba. E-mail: graca.tomazela@fatec.sp.gov.br. Lattes: http://lattes.cnpq.br/9809874053943708. OrcId: https://orcid.org/0000-0002-5471-2658.

ABSTRACT

School dropout is a persistent social problem that affects the development of education in Brazil. Solving this problem requires addressing other related challenges that influence whether children remain in school or not. In this context, data analysis technologies and artificial intelligence offer powerful tools to identify factors that influence school dropout. Thus, the objective of this study was to use data mining techniques to identify patterns and analyze the main factors associated with school dropout. To achieve this, a quantitative, exploratory, and descriptive study was conducted, followed by a qualitative interpretation of the results. Initially, a bibliographic review was carried out to survey the key concepts: Knowledge Discovery in Databases, Clustering and the k-means algorithm; as well as related works. The data were collected from PNAD Contínua from 2021 to 2024, after which data cleaning and transformation were performed, followed by clustering and, finally, the analysis of the clusters. The results showed that, through the stages of data preparation and the application of a data mining technique, it was possible to identify relevant patterns related to school dropout. Among the results obtained, the relationship between school dropout and the need to balance work and study stood out, highlighting this factor as a relevant element in school abandonment. In this sense, this research may contribute to helping educational managers and educational institutions develop strategies aimed at school retention, such as monitoring actions or support measures for students who work.

Keywords: Education. Artificial intelligence. Clustering. Early school leaving.

RESUMEN

La deserción escolar es un problema social persistente que afecta el desarrollo de la educación en Brasil. Resolver este problema requiere enfrentar otros desafíos relacionados que influyen en la permanencia o no de los niños en los estudios. En este contexto, las tecnologías de análisis de datos y la inteligencia artificial ofrecen herramientas poderosas para identificar factores que influyen en el abandono escolar. Así, el objetivo de este trabajo fue utilizar técnicas de minería de datos para identificar patrones y analizar los principales factores asociados a la deserción escolar. Para ello, se realizó una investigación cuantitativa, exploratoria y descriptiva, con una posterior interpretación cualitativa de los resultados. Inicialmente, se llevó a cabo una revisión bibliográfica para el levantamiento de los conceptos clave: Descubrimiento de Conocimiento en Bases de Datos, Clusterización y el algoritmo k-means; además de trabajos relacionados. Los datos fueron recolectados de la PNAD Continua de 2021 a 2024; posteriormente se realizó la limpieza y transformación de estos datos, la clusterización y, finalmente, el análisis de los clusters. Los resultados demostraron que, mediante las etapas de preparación de los datos y la aplicación de una técnica de minería de datos, fue posible identificar patrones relevantes relacionados con la deserción escolar. Entre los resultados obtenidos, se destacó la relación entre la deserción escolar y la necesidad de conciliar trabajo y estudio, evidenciando este factor como un elemento relevante en el abandono escolar. En este sentido, esta investigación puede contribuir a que gestores educativos e instituciones de enseñanza desarrollen estrategias orientadas a la permanencia escolar, como acciones de seguimiento o medidas de apoyo para los estudiantes que trabajan.

Palabras clave: Educación. Inteligencia artificial. Clustering. Abandono escolar.

1 INTRODUÇÃO

O ambiente escolar desempenha um papel fundamental na formação dos indivíduos e na preparação para a vida em sociedade. Ele não apenas transmite conhecimentos, mas também promove valores, habilidades e atitudes necessárias para uma participação ativa e consciente na comunidade (Silva, 2024).

Entretanto, a evasão escolar é um desafio persistente que afeta significativamente o sistema educacional do Brasil (Silva, 2024). Diversas situações podem ocorrer simultaneamente na vida do aluno, dificultando seu engajamento nos estudos e aumentando o risco de abandonar a escola (Ramos; Gonçalves Junior, 2024). Resolver esse problema exige enfrentar outros desafios relacionados, que influenciam na permanência ou não da criança nos estudos (Silva, 2024).

Nesse contexto, a Mineração de Dados Educacionais surge como uma disciplina emergente que tem como objetivo desenvolver metodologias para compreender os alunos em seus ambientes de aprendizagem, explorando os dados provenientes de cenários educacionais (Couto, 2017). De acordo com Nunes (2023), essa é uma área de estudo em constante crescimento, capaz de encontrar as falhas no sistema educacional, prever a evasão e aumentar a retenção de alunos.

Com isso, essa pesquisa justifica-se pela necessidade de compreender as variáveis relacionadas à evasão escolar. Com a utilização de técnicas de mineração de dados, é possível detectar tendências e elementos de risco que influenciam a evasão, o que permite uma compreensão mais profunda desse problema e facilita a implementação de soluções.

A questão norteadora deste trabalho foi: “Como as técnicas de mineração de dados são capazes de identificar fatores significativos que influenciam a evasão escolar, revelando correlações entre diversas variáveis?”.

O objetivo geral dessa pesquisa foi utilizar essas técnicas para identificar padrões e analisar os principais fatores associados à evasão escolar. Dessa forma, os objetivos específicos foram: identificar dados adequados para serem minerados; preparar esses dados de forma apropriada para a aplicação da técnica de mineração de dados; identificar os melhores parâmetros para a obtenção de resultados mais efetivos na *clusterização*; realizar *clusterização* e analisar os resultados obtidos a partir da aplicação dessa técnica.

Para o desenvolvimento deste trabalho foi aplicado o processo de KDD (*Knowledge Discovery in Databases*), executando as etapas de pré-processamento, mineração de dados e pós-processamento. O estudo abordou a evasão escolar entre jovens de 14 a 18 anos, compreendida, neste trabalho, a partir da variável de frequência escolar da PNAD (Pesquisa Nacional por Amostra de Domicílios) Contínua, considerando como evadidos os indivíduos que declararam não estar frequentando a escola no momento da coleta. A análise contemplou os níveis de ensino fundamental e médio, realizando *clusterização* sobre os dados da PNAD Contínua de 2021 a 2024.

Embora diversos estudos investiguem a evasão escolar com foco em indicadores acadêmicos, este trabalho direciona a análise para fatores socioeconômicos associados ao fenômeno. Nesse sentido, esta pesquisa busca contribuir para essa discussão ao investigar padrões relacionados às características sociais e à inserção no mercado de trabalho entre jovens que não estão frequentando a escola.

A delimitação da faixa etária entre 14 e 18 anos considera um período que abrange o final do ensino fundamental e do ensino médio, etapa decisiva para a conclusão da educação básica. O recorte temporal de 2021 a 2024 foi definido por contemplar os dados mais recentes disponíveis no momento da pesquisa, permitindo uma análise atualizada do fenômeno. A escolha da base de dados da PNAD Contínua baseia-se na disponibilidade pública dos microdados, na sua abrangência nacional e na presença de variáveis relacionadas à educação, trabalho e aspectos sociais, adequadas aos objetivos da pesquisa. Por fim, ressalta-se que o estudo não contempla outras faixas etárias, nem níveis de ensino fora do escopo definido.

2 REFERENCIAL TEÓRICO

Inicia-se o referencial teórico pela descoberta de conhecimento em base de dados.

2.1 Descoberta de conhecimento em base de dados

O volume de dados gerados e armazenados tem aumentado exponencialmente nos últimos anos, superando a possibilidade de compreendê-los sem o auxílio de ferramentas poderosas. Esse cenário cria desafios para a extração de conhecimento significativo a partir de dados brutos, o que exige métodos e técnicas especializadas (Han; Kamber; Pei, 2012).

Nesse cenário, foi desenvolvido o processo de Descoberta de Conhecimento em Base de Dados, também conhecido como KDD. O processo de KDD refere-se à extração ou mineração de conhecimento a partir de grandes volumes de dados (Han; Kamber; Pei, 2012). Conforme Fayyad, Piatetsky-Shapiro e Smyth (1996), o KDD é um processo interativo (atuação humana no processo), iterativo (possibilidade de repetição das etapas) e não trivial (complexo), destinado a descobrir padrões compreensíveis, válidos, novos e potencialmente úteis, com o objetivo de transformar dados brutos em conhecimento prático e aplicável. Ele pode ser dividido em três etapas principais: pré-processamento, mineração de dados e pós-processamento.

O pré-processamento prepara os dados brutos para serem analisados nas próximas fases. Nesse processo, realizam-se os seguintes procedimentos: 1) Limpeza dos dados: consiste em preencher valores faltantes, remover ruídos e eliminar inconsistências; 2) Integração: etapa em que várias fontes de dados podem ser combinadas; 3) Seleção: extração dos dados relevantes para a análise; 4) Transformação: consolidação dos dados em formas apropriadas para a mineração, executando operações de agregação e normalização, por exemplo.

A mineração de dados tem como objetivo a busca efetiva por conhecimentos novos e úteis. Amo (2004) destaca que ela é essencial para o processo de KDD. Os algoritmos de mineração de dados podem ter aprendizagem supervisionada ou não-supervisionada. No aprendizado supervisionado, o pesquisador espera uma saída de dados correta para cada objeto de entrada (Han; Kamber; Pei, 2012). Já no aprendizado não-supervisionado, as classes não estão definidas e os algoritmos verificam os dados, procurando estabelecer relacionamentos entre eles (Amo, 2004; Goldschimdt, Passos, 2005).

A descoberta pode ser dividida em duas atividades: previsão, que aplica o aprendizado supervisionado, e descrição, que aplica o aprendizado não supervisionado.

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), as tarefas de mineração preditiva buscam encontrar padrões para prever o comportamento de algumas entidades. Já as tarefas de mineração descritiva caracterizam as propriedades gerais dos dados armazenados (Han; Kamber; Pei, 2012). De acordo com Amo (2004), as tarefas preditivas são: Classificação e Predição, enquanto as tarefas descritivas incluem: Associação, *Clusterização*, Padrões Sequenciais e Detecção de *Outliers.*

No pós-processamento, são avaliados os resultados obtidos nas etapas anteriores. De acordo com Goldschmidt e Passos (2005), essa etapa abrange o tratamento do conhecimento gerado na mineração de dados, com o objetivo de facilitar a interpretação e a avaliação da utilidade do conhecimento descoberto. Eles definiram que as principais funções para a etapa de pós-processamento são: a) Elaboração e organização, que inclui a simplificação de gráficos,

diagramas, ou relatórios demonstrativos; b) Transformação do modelo de representação do conhecimento obtido.

2.2 Clusterização

Clusterização é o processo de agrupar um conjunto de objetos em classes semelhantes, de forma que os objetos de um *cluster* são similares a outros do mesmo *cluster* e são diferentes de objetos em outros *clusters* (Han; Kamber; Pei, 2012).

O processo de *clusterização* requer que o usuário determine qual o número de grupos a ser considerado. Uma vez formados os grupos, é possível fazer uma análise dos elementos que compõem cada um deles, identificando as características comuns aos seus elementos. Desta forma, poderá criar-se um rótulo que represente cada grupo (Goldschmidt; Passos, 2005).

Han, Kamber e Pei (2012) dividiram os métodos de *clusterização* em 5 categorias: a) Baseados em particionamento: classifica em grupos com base na distância entre objetos, de forma que cada grupo deve conter pelo menos um objeto e cada objeto deve pertencer a exatamente um grupo; b) Hierárquicos: o objeto se inicia em um grupo separado e posteriormente se mescla com grupos próximos, até que todos são mesclados em um; c) Baseados em densidade: a ideia geral é que o *cluster* continue crescendo, até que sua densidade (número de objetos) exceda um determinado limite; d) Baseados em grade: quantifica o espaço do objeto em um número finito de células e executa o agrupamento nessa estrutura de grade; e) Baseados em modelos: levanta a hipótese de um modelo para cada *cluster* e encontra o melhor ajuste dos dados para esse modelo.

2.3 K-means

O algoritmo *k-means* é um dos algoritmos mais utilizados para a tarefa de *clusterização*. Ele implementa o método de particionamento. Seu funcionamento ocorre da seguinte forma:

a) Seleciona aleatoriamente *k* objetos do banco de dados, em que cada um deles representa, inicialmente, o centro do *cluster*;

b) Os objetos restantes são atribuídos no *cluster* ao qual são mais semelhantes, baseado em uma medida de distância, geralmente a euclidiana, entre o objeto e o centro do *cluster* (Han; Kamber; Pei, 2012);

c) Após este processo, calcula-se a média dos elementos de cada *cluster*, gerando o seu centro de gravidade. Este ponto será o novo representante do *cluster*;

d) Em seguida, a etapa b) é repetida. Para cada objeto do banco de dados, calcula-se a distância entre este objeto e os novos centros dos *clusters*. Ele será realocado para o *cluster* em que a distância entre o objeto o centro do *cluster* é a menor possível;

e) Quando todos os objetos forem devidamente realocados entre os *clusters*, calcula-se novamente os centros dos *clusters* (Amo, 2004);

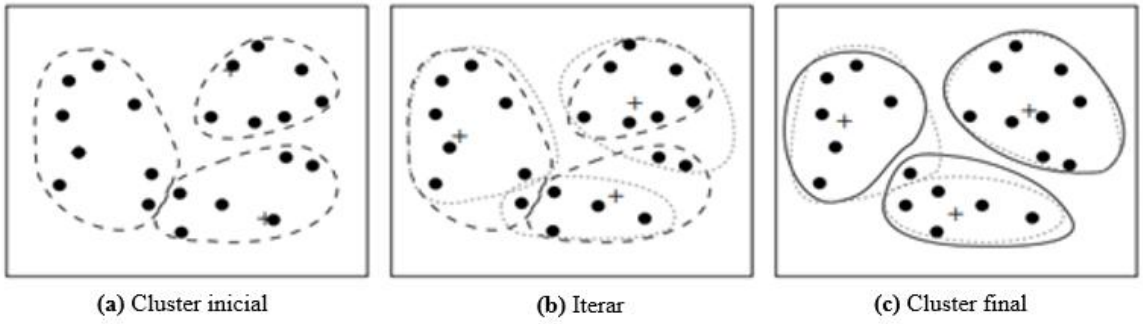
f) Esse processo se repete até que os *clusters* formados na rodada atual sejam os mesmos formados na rodada anterior (Han; Kamber; Pei, 2012).

A Figura 1 exemplifica o processo do método *k-means* quando $k = 3$, ou seja, quando três *clusters* devem ser formados. Conforme o funcionamento descrito anteriormente, três objetos são escolhidos aleatoriamente como os centros iniciais dos *clusters*, representados pelo

símbolo +. Cada objeto é atribuído a um *cluster* com base no centro do *cluster* do qual está mais próximo, como mostra a Figura 1 (a).

Em seguida, o ponto central é recalculado com base na média dos objetos atuais no *cluster*. Na Figura 1 (b), os objetos são redistribuídos considerando os novos centros, formando novos *clusters* circundados por curvas tracejadas. Esse processo se repete, levando à Figura 1 (c) (Han; Kamber; Pei, 2012).

Figura 1- O algoritmo de particionamento k-means



Fonte: Adaptado de Han, Kamber e Pei, 2012, p. 453

O método exige a definição do valor de *k* e é sensível a ruídos, já que os valores discrepantes podem distorcer o centro de gravidade dos *clusters*.

2.4 Trabalhos relacionados

O trabalho de Fida (2020) teve como objetivo prever quais estudantes apresentariam baixo desempenho, para melhorar a retenção e o desempenho acadêmico. A autora coletou dados de um Sistema de Gerenciamento de Aprendizagem (LMS) na plataforma *Kaggle*. Foram utilizados algoritmos de agrupamento e classificação. Os resultados do agrupamento indicaram que, dos 478 registros de alunos, o *k-means* conseguiu agrupar corretamente 357 alunos em categorias de desempenho alto, médio ou baixo, o *PAM (Partitioning Around Medoids)* agrupou 338 alunos corretamente, e o *EM (Expectation-Maximization)* agrupou 274 alunos corretamente. Em seguida, utilizaram o *cluster* gerado pelo *k-means* nos testes de classificação. O *Random Forest* apresentou o melhor desempenho, com uma acurácia de 93%.

Souza (2020) buscou gerar modelos de predição com algoritmos de classificação que apoiassem os gestores em ações de combate à evasão escolar. Para isso, o autor coletou informações no Sistema Acadêmico de uma instituição de Belo Horizonte, e utilizou a metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Os resultados mostraram que o melhor algoritmo foi *Decision Tree*, com uma acurácia de 96,4% e uma curva AUC (*Area Under the Curve*) de 0,994. O segundo que se destacou foi o *Neural Network*, com uma acurácia de 94,4% e uma curva AUC de 0,988.

Kunchala (2021) propôs a predição da evasão em uma instituição pública de ensino superior nos Estados Unidos, por meio de modelos de classificação. Os dados foram coletados nas bases de dados da faculdade e na *National Student Clearinghouse*. A partir dos resultados, observou-se que as características mais influentes no abandono estão relacionadas com o

desempenho acadêmico e a ajuda financeira. O modelo com melhor resultado foi o *Random Forest*, que alcançou uma *F-measure* de 0,903 e uma precisão de 90,8%.

Costa (2021) buscou identificar os atributos que mais afetam a evasão de alunos em Ciência da Computação e Engenharia da Universidade Federal de Pelotas e determinar os melhores algoritmos de classificação para prever alunos em risco de evasão. Foram utilizadas etapas da metodologia CRISP-DM e os dados foram coletados no sistema Cobalto, de 2000 a 2018. Os resultados mostraram que a média de notas e o coeficiente de dificuldade do terceiro semestre foram os atributos mais influentes, e os melhores modelos foram o *Naive Bayes*, com acurácia de 87,57% e o *Random Forest*, com acurácia de 85,58%.

O trabalho de Silva (2021) teve por objetivo desenvolver e analisar modelos preditivos da evasão escolar no ensino médio do Instituto Federal de Mato Grosso. Os dados foram coletados no sistema Q-Seleção e Q-Acadêmico, e foram aplicados algoritmos de classificação por árvore de decisões. Os índices de precisão dos modelos, em seus melhores resultados, variaram de 80,79% a 84,40%. O *Random Forest* apresentou o maior índice entre os algoritmos analisados.

Tamada (2022) combinou técnicas de agrupamento e classificação para prever o risco de evasão em cursos EaD (Educação a Distância). A autora coletou dados dos estudantes no *Moodle* em diferentes estágios de conclusão do curso. O agrupamento revelou que estudantes no início do curso, cuja frequência e nota são baixas, têm pouca chance de conclusão. Os resultados da classificação mostraram que *Logistic Regression* foi o melhor modelo quando o experimento foi realizado com alunos que haviam concluído 10% do curso, enquanto o *Naive Bayes* se destacou no experimento com os alunos que já haviam completado o primeiro semestre até aqueles que estavam no final do curso.

Bineid (2022) buscou identificar os alunos em risco de abandono, determinar as variáveis mais influentes e verificar a precisão dos modelos de classificação. A autora utilizou um conjunto de dados de 1.272 alunos do *Central Higher Education Data Store* (CHEDS). Os resultados mostraram que os alunos com baixa média de notas e baixa média de horas de registro (frequência) são os mais propensos a desistir do curso. Entre os algoritmos testados, o *Random Forest* apresentou o maior desempenho em previsão e pontuou 0,878 em AUC, com uma precisão de 84,82%.

Singer (2023) explorou a mineração de dados educacionais e aprendizado de máquina para prever as decisões de matrícula dos aprovados na *Washington State University* (WSU). A técnica de aprendizado de máquina utilizada foi a *Logistic Regression*, que categorizou com precisão 81% dos resultados de matrícula. Os resultados mostraram que, dentre as 24 variáveis analisadas, as variáveis mais influentes foram: residência (verifica se o estudante mora em Washington ou não), período de admissão, tipo de graduação, tipo do estudante (se é um ex-aluno, transferido, primeiro ano estudando ou se é um estudante não graduado) e nível de escolaridade dos pais. Além disso, o autor propôs a realização de um workshop para explorar as discussões sobre os resultados obtidos.

Nunes (2023) desenvolveu um software para auxiliar gestores educacionais na previsão de alunos com risco de evasão, por meio da mineração de dados e métodos de classificação. Para o processo de KDD, aplicou-se a metodologia CRISP-DM. O autor conduziu dois experimentos, um com uma base de dados de escolas portuguesas e outro com uma base de dados fictícia. Na base de dados reais, o *Random Forest* obteve a melhor acurácia, com 97,46% de precisão. Na base de dados fictícia, o *Random Forest* e o *MLP (Multilayer Perceptron)*

obtiveram a melhor acurácia, com 99% de precisão. Além disso, o software atingiu uma pontuação SUS[3] (*System Usability Scale*) de 91,125, na validação de usabilidade.

O trabalho de O'Neill (2024) teve como objetivo identificar os alunos do *Educational Opportunity Fund* (EFO) com a maior probabilidade de abandonar a faculdade no primeiro ano do curso. Foi analisada a retenção de alunos pré-covid e pós-covid, com técnicas de agrupamento e classificação. Os resultados mostraram que estudantes matriculados em matemática, biologia, química e informática são os mais propensos à evasão. O algoritmo de classificação que apresentou o melhor desempenho foi o *Random Forest* implementado com a técnica SMOTE (*Synthetic Minority Over-sampling Technique*), atingindo 96% de precisão.

3 MÉTODO

Inicia-se o método pela classificação da pesquisa.

3.1 Classificação da pesquisa

Esse trabalho teve uma abordagem quantitativa, pois utilizou dados numéricos para analisar fenômenos sociais, permitindo identificar padrões e tendências por meio da análise e comparação das informações. Posteriormente, os resultados obtidos foram interpretados qualitativamente, buscando compreender seus significados e implicações no contexto analisado. A interpretação qualitativa preocupa-se em analisar aspectos mais profundos dos fenômenos sociais, descrevendo a complexidade do comportamento humano (Lakatos; Marconi, 2010).

Quanto aos objetivos, o estudo classificou-se como exploratório e descritivo. A pesquisa exploratória tem como finalidade proporcionar maior familiaridade com o problema, tornando-o mais explícito e possibilitando uma compreensão inicial do fenômeno investigado. Já a pesquisa descritiva busca descrever as características de determinada população ou fenômeno, apresentando suas distribuições e comportamentos a partir dos dados analisados (Gil, 2002).

3.2 Ferramentas da pesquisa

Nessa pesquisa, foram utilizadas as seguintes ferramentas:

1) Pandas: uma biblioteca de código aberto e gratuita, escrita para a linguagem de programação Python. Ela fornece estruturas para armazenar grandes quantidades de dados com eficiência e ferramentas para analisar e manipular esses dados. Foi utilizada principalmente na etapa de pré-processamento.
2) Scikit-learn: é uma biblioteca de aprendizado de máquina que oferece uma variedade de algoritmos para análise de dados e modelagem preditiva. Possui código aberto e é implementado em Python, possuindo boa integração com outras bibliotecas, como Pandas e Matpotlib. Foi utilizada na tarefa de *clusterização*.

[3] O questionário SUS consiste em 10 perguntas cuidadosamente elaboradas para avaliar a facilidade de uso e a satisfação do usuário com o software em questão.

3) Matpotlib: é uma biblioteca para visualização de dados e plotagem gráfica em Python. Permite a criação e personalização de gráficos de linha, barras, pizza, histogramas, dispersão, entre outros. É essencial para comunicar as descobertas da análise dos dados. Auxiliou na etapa de pós-processamento.

3.3 Experimento da pesquisa

Para o desenvolvimento deste trabalho foi utilizado o processo de KDD, realizando as etapas de pré-processamento, mineração de dados e pós-processamento.

Na etapa de pré-processamento, foram extraídos dados sobre o nível de escolaridade e a situação de trabalho dos jovens de 14 a 18 anos. Com a biblioteca Pandas, também foram realizados os procedimentos de normalização, seleção de atributos, limpeza, integração e transformação dos dados, para facilitar a análise realizada nas próximas etapas.

Após o pré-processamento, iniciou-se a mineração de dados. Para isso, a biblioteca Scikit-learn foi utilizada para implementar o algoritmo *k-means*, a fim de realizar as tarefas de *clusterização*. Nessa etapa, foi aplicado o método Elbow e as métricas Silhoeutte, Davies Bouldin e Calinski-Harabasz, buscando definir a quantidade ideal de *clusters* para essa pesquisa. Por último, foi usada a biblioteca Matpotlib para auxiliar na tarefa de análise dos *clusters* gerados.

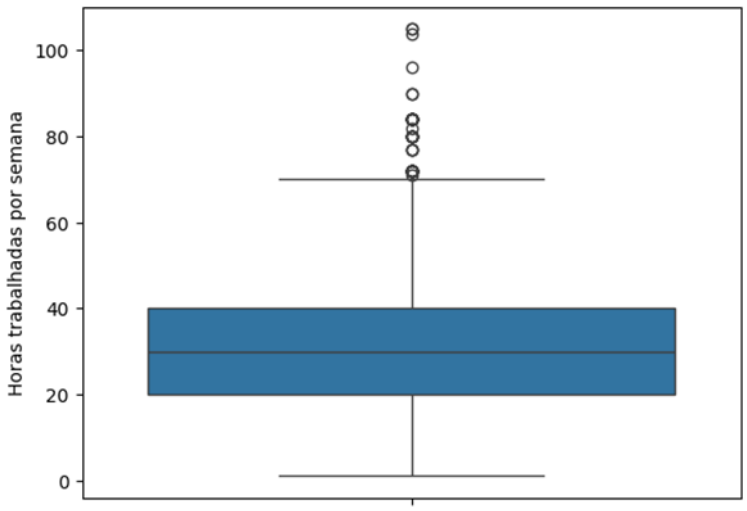
3.4 Pré-processamento

Na etapa de pré-processamento, os dados foram obtidos da PNAD Contínua de 2021 a 2024, inicialmente em formato TXT e posteriormente convertidos para o formato CSV, com o uso da biblioteca Pandas. Como o foco do estudo é o abandono escolar entre jovens, os dados foram filtrados para incluir apenas indivíduos na faixa etária entre 14 e 18 anos e foram extraídos atributos relacionados à situação escolar, às características sociais (cor/raça, sexo, estado de residência) e à situação de trabalho.

As bases de dados estavam organizadas separadamente por ano e foi necessário consolidá-las em um único arquivo, adicionando uma nova coluna denominada ‘Ano’.

Realizou-se a limpeza dos dados, retirando-se os *outliers* e os registros inconsistentes. Foram identificados *outliers* somente no atributo “Quantidade de horas trabalhadas por semana”. A Figura 2 apresenta o gráfico Box Plot dessa variável, no qual os registros com mais de 60 horas semanais foram considerados como *outliers* (94 registros), por representarem valores muito acima do esperado para a faixa etária analisada. Além disso, foram encontrados 39 registros com nível de instrução “Superior Completo”, mesmo após o filtro de idade de até 18 anos. Esses registros também foram removidos da base por serem inconsistentes com o recorte de idade definido para o estudo.

Figura 2 - Outliers na variável “Horas Trabalhadas por semana”



Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

Para os indivíduos que não estavam trabalhando, os atributos relacionados à situação de trabalho apresentavam valores ausentes. Nesses casos, os campos foram preenchidos com o valor 0, representando a opção “Não aplicável”. Dessa forma, os valores ausentes nas variáveis relacionadas ao trabalho passaram a indicar que aquela informação não se aplicava ao indivíduo. Para todas as demais variáveis da base de dados, os valores que representavam a opção “Não aplicável” também foram padronizados com o valor 0.

Com base na coluna UF (Unidade Federativa), foi criada uma nova coluna correspondente à região. As demais colunas foram renomeadas com nomes mais simples, e as categorias dos atributos ‘Posição no emprego’ e ‘Tipo de atividade’ foram agrupadas em opções mais resumidas, visando a facilitar a análise e reduzir a dimensionalidade das categorias.

A variável ‘Horas trabalhadas por semana’ passou por um processo de discretização por frequência, utilizando a biblioteca Pandas, de modo a dividir os valores em intervalos com quantidades semelhantes de registros. Esse procedimento permite reduzir o impacto de valores muito altos ou muito baixos e facilita a identificação de padrões nos dados. Além disso, as colunas ‘Ano’ e ‘Idade’ foram ajustadas para uma escala numérica menor, com o objetivo de padronizar os valores utilizados na análise. Para isso, os anos foram representados da seguinte forma: 2021 = 1, 2022 = 2, 2023 = 3 e 2024 = 4. Da mesma forma, as idades foram codificadas como: 14 = 1, 15 = 2, 16 = 3, 17 = 4 e 18 = 5. Esse procedimento foi adotado para manter os valores em uma escala semelhante aos demais atributos utilizados no modelo.

Por último, foi aplicado o método de *One Hot Encoding* para transformar as variáveis categóricas ‘Sexo’, ‘Região’ e ‘Cor/raça’ em colunas binárias, com o uso da biblioteca Pandas. Esse método evita que os modelos interpretem essas categorias como se houvesse uma hierarquia ou ordem entre elas. Após a transformação, a variável ‘Sexo’ passou a ser representada por duas colunas (Sexo_1 e Sexo_2), ‘Região’ por cinco colunas (Regiao_1 a Regiao_5) e ‘Cor/raça’ por seis colunas (Cor_Raca_0 a Cor_Raca_5).

Ao final das atividades de pré-processamento dos dados, a base de dados ficou com 116.350 instâncias e 26 atributos, conforme descrito no Quadro 1.

Quadro 1 - Atributos da base de dados e suas características

Nome do campo	Descrição	Tipo
Ano	Ano de referência (2021 a 2024)	Ordinal
Sexo_1	É do sexo masculino? (0: Não; 1: Sim)	Nominal
Sexo_2	É do sexo feminino? (0: Não; 1: Sim)	Nominal
Idade	Faixa etária (14 a 18 anos)	Ordinal
Cor/raca_0 a Cor/raca_5	Cor/raça (Ignorado, Branca, Preta, Amarela, Parda ou Indígena)	Nominal
Regiao_1 a Regiao_5	Região de residência (Norte, Nordeste; etc.)	Nominal
Frequenta escola?	Está estudando atualmente? (0: Não; 1: Sim)	Nominal
Nível de instrução	Grau de escolaridade (Sem instrução a Superior incompleto)	Ordinal
Trabalhou por dinheiro?	Trabalhou em atividade remunerada em dinheiro?	Nominal
Trabalhou por produtos?	Trabalhou em atividade remunerada em produtos?	Nominal
Ajudou no trabalho?	Ajudou no trabalho remunerado de terceiros?	Nominal
Afastado do trabalho?	Tinha algum trabalho do qual estava afastado?	Nominal
Faixa de rendimento	Rendimento mensal (de 0 até 20 salários mínimos)	Ordinal
Tempo no trabalho	Tempo no trabalho atual (menos de 1 mês até mais de 2 anos)	Ordinal
Horas trabalhadas	Horas trabalhadas por semana	Ordinal
Tipo de atividade	Setor econômico e nível de especialização do trabalho.	Ordinal
Posição no emprego	Categoria do trabalho (Autônomo, Carteira Assinada, etc.)	Ordinal

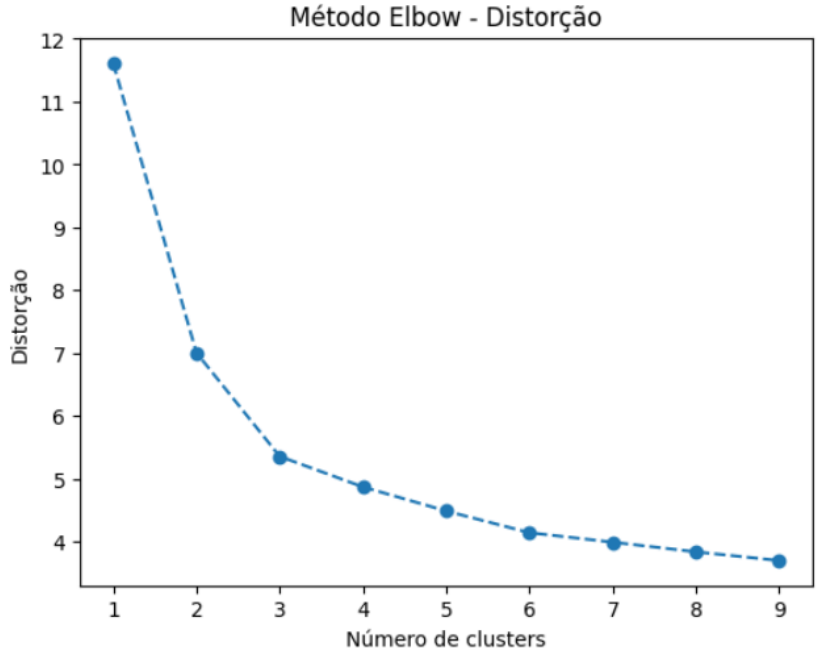
Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

3.5 Mineração de Dados

Finalizado o pré-processamento, foi realizada a etapa de mineração de dados. Nessa etapa, foram utilizadas as bibliotecas Matplotlib e Scikit-learn, para análise dos dados e para aplicar a *clusterização*.

Para definir o número ideal de *clusters*, utilizou-se o método Elbow (Figura 3) e três métricas de avaliação: Silhoeutte Score, índice Davies-Bouldin e índice Calinski-Harabasz (Tabela 1). Com isso, foi possível analisar a qualidade dos *clusters*, considerando aspectos como coesão, separação e variabilidade.

Figura 3 - Método de Elbow



Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

Tabela 1 - Índices de medidas avaliação dos clusters

	Silhoeutte	Davies Bouldin	Calinski-Harabasz
Cluster 3	0.27	1.27	67937.42
Cluster 4	0.22	1.40	53680.67
Cluster 5	0.21	1.56	46174.43
Cluster 6	0.19	1.79	41958.73
Cluster 7	0.18	1.89	37019.38

Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

O método Elbow mede a inércia, que calcula a soma das distâncias quadráticas entre os pontos de cada *cluster* e o seu centroide. Quanto menor a inércia, mais próximos os pontos estão do centroide, o que indica que os *clusters* são mais compactos e bem definidos. O objetivo é encontrar o ponto em que a redução na inércia começa a desacelerar, formando um "cotovelo" no gráfico. Na Figura 3, esse ponto ocorre em torno de k=3, mas k=4 e k=5 também mostram uma queda considerável antes da estabilização.

O Silhoeutte Score avalia a qualidade da separação entre os *clusters*, variando entre -1 e 1, sendo que valores mais altos indicam uma separação mais clara. Nos resultados apresentados, o maior valor ocorre para k=3, mas k=4 e k=5 ainda apresentam bons índices de separação, com uma diminuição mais notável em k=6 e k=7.

O índice Davies-Bouldin mede a compactação e separação dos *clusters*, sendo que valores menores indicam uma melhor divisão. Para os resultados analisados, k=3 apresenta o menor valor, mas um aumento significativo só ocorre a partir de k=6.

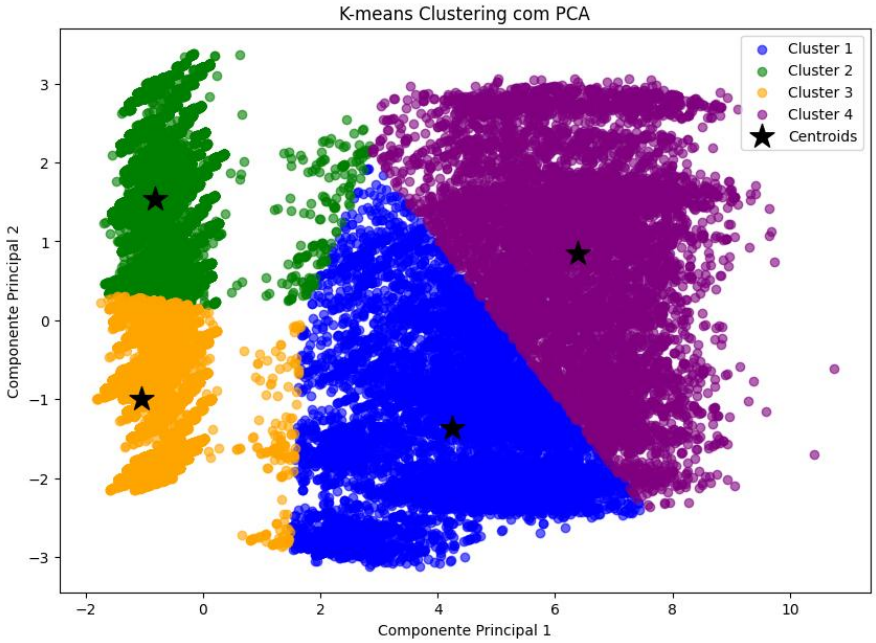
O índice Calinski-Harabasz avalia a relação entre a dispersão entre os *clusters* e a dispersão dentro de cada *cluster*, com valores mais altos indicando melhores agrupamentos. Para os dados analisados, o valor mais alto ocorre em k=3, mas k=4 também tem um bom resultado antes de uma queda mais expressiva.

Com base nesses critérios, k=3 apresentou os melhores resultados, mas k=4 e k=5 também são opções viáveis. Considerando um equilíbrio entre coesão e separação dos grupos, sem uma perda significativa de qualidade, definiu-se que seriam utilizados 4 *clusters* neste trabalho.

Além disso, o método de inicialização do algoritmo k-means utilizado foi o k-means++, que realiza uma escolha inicial mais adequada dos centroides. O algoritmo foi executado com 10 inicializações diferentes (n_init = 10), sendo selecionado o resultado com melhor desempenho entre essas execuções. O número máximo de iterações adotado foi 300, valor padrão da biblioteca utilizada. O critério de parada ocorre quando os centroides deixam de apresentar mudanças significativas entre as iterações ou quando o limite máximo de iterações é atingido.

Por último, utilizou-se um gráfico de dispersão para fornecer uma representação visual dos *clusters* formados (Figura 4).

Figura 4 - Gráfico de dispersão com 4 clusters



Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

Os *clusters* gerados possuem, em relação ao seu número de registros e porcentagem de distribuição, os valores apresentados na Tabela 2.

Tabela 2 - Distribuição de frequência de registros com 4 clusters

Cluster	Registros	Porcentagem
0	40483	34,79%
1	28528	24,52%
2	17405	14,96%
3	29934	25,73%

Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

3.6 Limitações da pesquisa

Este estudo apresentou limitações decorrentes das variáveis de controle estabelecidas. A primeira limitação seria com relação à limitação técnica da *clusterização*, em que os grupos formados dependem das variáveis selecionadas. A segunda se refere aos atributos da base de dados da PNAD Contínua, que não contém variáveis qualitativas específicas sobre os motivos da evasão escolar. Por fim, a terceira está relacionada à necessidade de definir previamente o número de *clusters*, o que pode influenciar nos resultados.

Esses aspectos podem ser considerados em trabalhos futuros, utilizando bases de dados com variáveis qualitativas relacionadas diretamente à evasão escolar, testando outros métodos, como modelos de classificação, ou outros números de *clusters*, a fim de complementar os resultados dessa pesquisa.

4 RESULTADOS E DISCUSSÃO

Com os quatro *clusters* gerados, foi feito a análise detalhada de cada um. Após as análises, apresenta-se a discussão do resultado desse trabalho em relação às pesquisas relacionadas.

4.1 Cluster 0

O *cluster* 0 é composto por 40483 instâncias. Em relação às características gerais deste *cluster*, Figura 5, observa-se que a maior parte dos registros (29,9%) corresponde ao ano de 2024, seguido por 2022 e 2023.

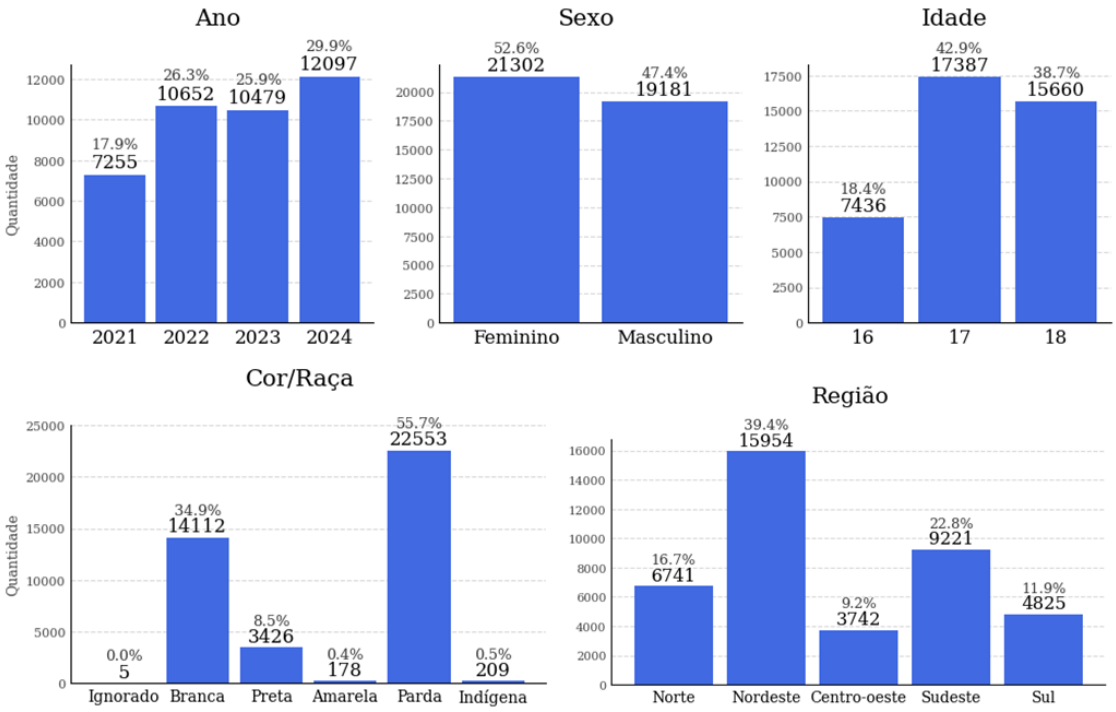
A respeito do sexo, o feminino é ligeiramente predominante, representando 52,6% dos registros.

O *cluster* é composto exclusivamente por indivíduos com idades entre 16 e 18 anos, sendo a idade de 17 anos a mais frequente, abrangendo 42,9% dos casos. Não há registros de indivíduos com idades de 14 ou 15 anos.

Quanto à cor/raça, a maior quantidade de pessoas pertence à categoria "parda", representando 55,7% dos registros.

A região com maior ocorrência é o Nordeste, com 39,4% das instâncias.

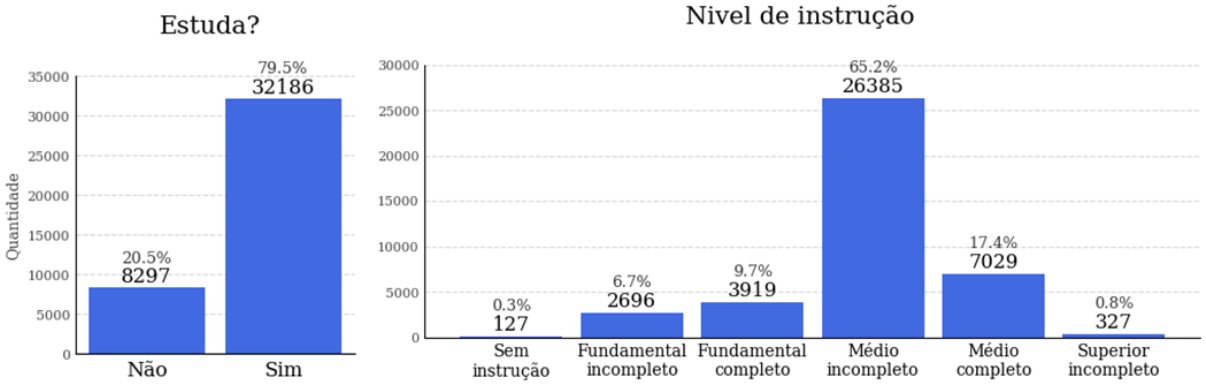
Figura 5 - Características sociais do cluster 0



Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

Em relação às variáveis relacionadas aos dados escolares, Figura 6, 79,5% dos indivíduos do *cluster* frequentam a escola. O nível de instrução predominante é o "Ensino Médio Incompleto", representando 65,2% dos casos.

Figura 6 - Características escolares do cluster 0



Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

Em relação às variáveis associadas à situação de trabalho, o *cluster* não apresenta ninguém que tenha trabalhado em atividades remuneradas em dinheiro ou em produtos. Apenas 92 indivíduos (0,23%) ajudaram no trabalho remunerado de terceiros, e 14 indivíduos (0,03%) estavam afastados de alguma ocupação.

4.2 Cluster 1

O *cluster* 1 é composto por 28528 instâncias. Em relação às características gerais deste *cluster*, Figura 7, observa-se que ele possui registros somente dos anos de 2021 (52,9% dos registros) e 2022 (47,1% dos registros).

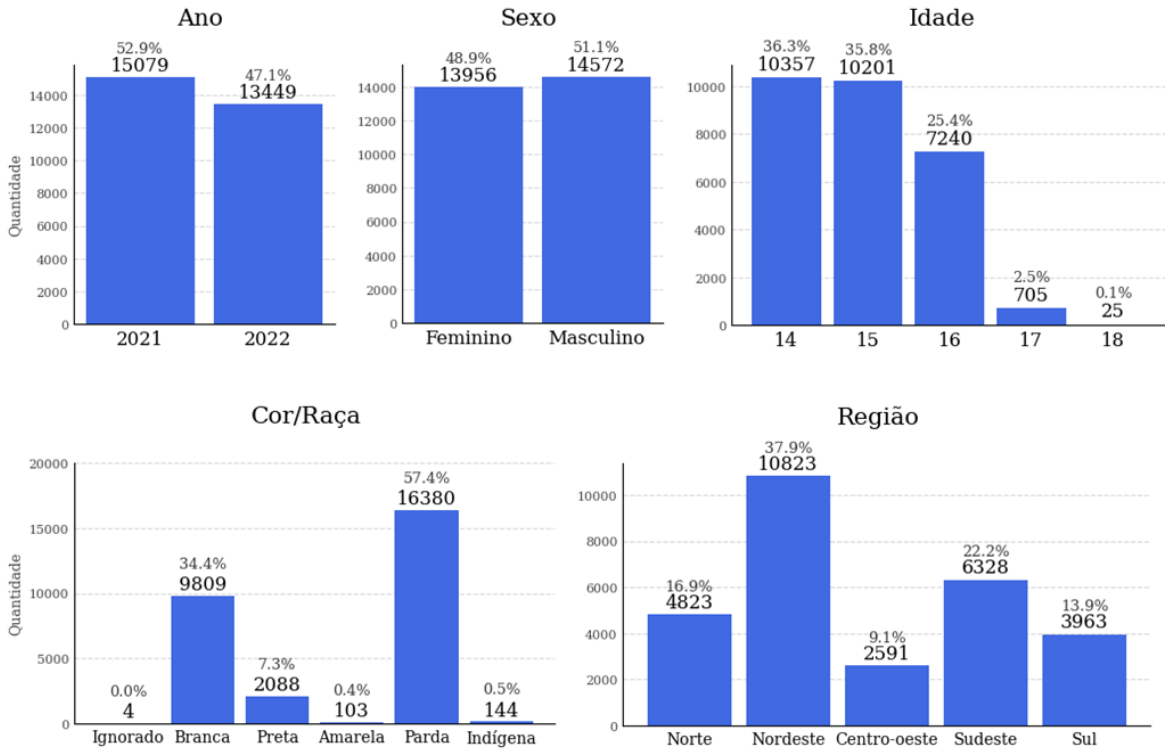
A respeito do sexo, o masculino é ligeiramente predominante, representando 51,1% dos registros.

O *cluster* é composto principalmente por indivíduos com idades entre 14 e 15 anos (39,6% e 38,2% dos casos, respectivamente).

Quanto à cor/raça, a maior quantidade de pessoas pertence à categoria "parda", representando 57,4% dos registros.

A região com maior ocorrência é o Nordeste, com 37,9% das instâncias.

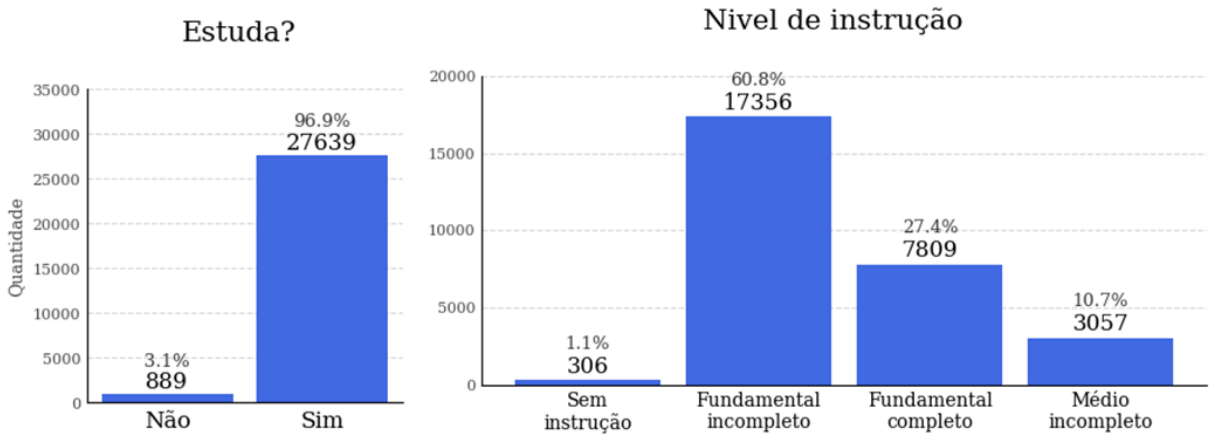
Figura 7 - Características sociais do cluster 1



Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

Em relação às variáveis relacionadas aos dados escolares, Figura 8, 96,9% dos indivíduos do *cluster* frequentam a escola. O nível de instrução predominante é o “Ensino Fundamental Incompleto”, representando 60,8% dos casos, seguido do “Ensino Fundamental Completo”, com 27,4% dos casos, o que é coerente com a faixa etária do *cluster*.

Figura 8 - Características escolares do cluster 1



Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

Em relação às variáveis associadas à situação de trabalho, o *cluster* apresenta apenas 34 indivíduos que trabalharam em atividades remuneradas em dinheiro, nenhum indivíduo que trabalhou em atividades remuneradas em produtos, 270 indivíduos (0,9%) que ajudaram no trabalho remunerado de terceiros e 4 indivíduos que estavam afastados de alguma ocupação.

4.3 Cluster 2

O *cluster* 2 é composto por 17405 instâncias. Em relação às características gerais deste *cluster*, Figura 9, observa-se que ele possui registros bem divididos entre todos os anos, sendo o ano de 2024 mais recorrente, com 32,3% dos casos.

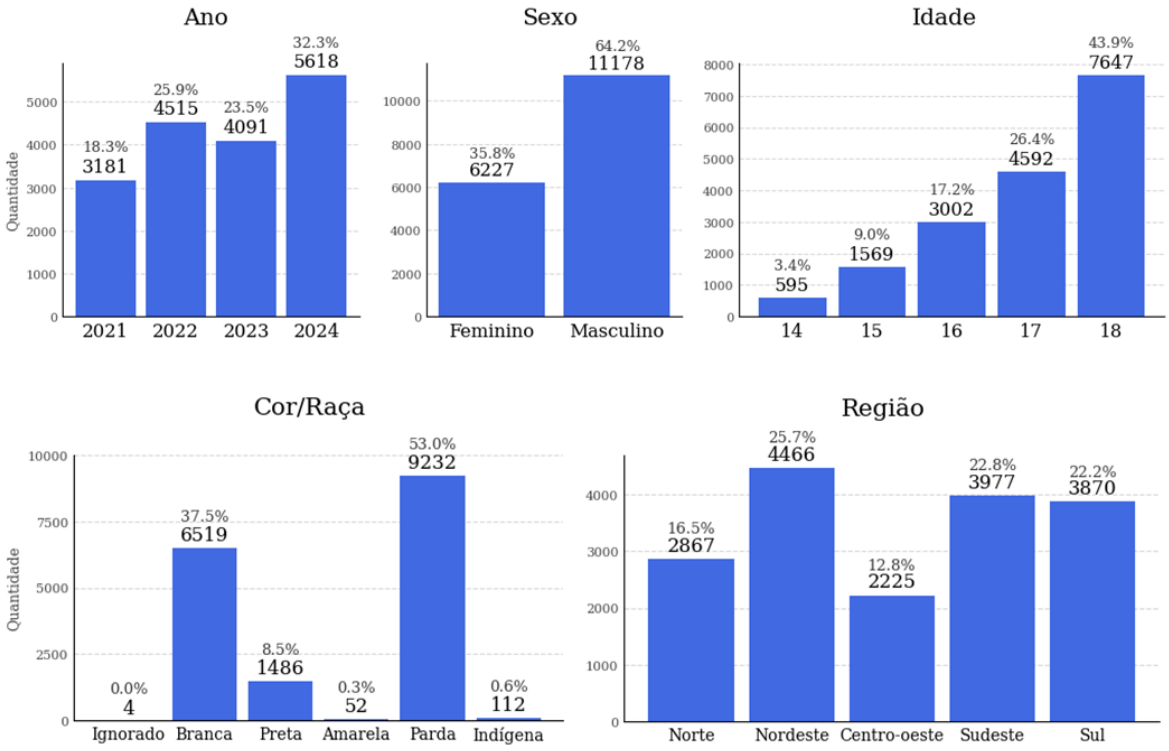
A respeito do sexo, o masculino é predominante, representando 64,2% dos registros.

A idade de 18 anos é a mais frequente, abrangendo 43,9% dos casos.

Quanto à cor/raça, a maior quantidade de pessoas pertence à categoria "parda", representando 53,0% dos registros.

O *cluster* possui três regiões mais recorrentes, sendo elas: Nordeste com 25,7% dos registros, Sudeste com 22,8% dos registros e Sul com 22,2% dos registros.

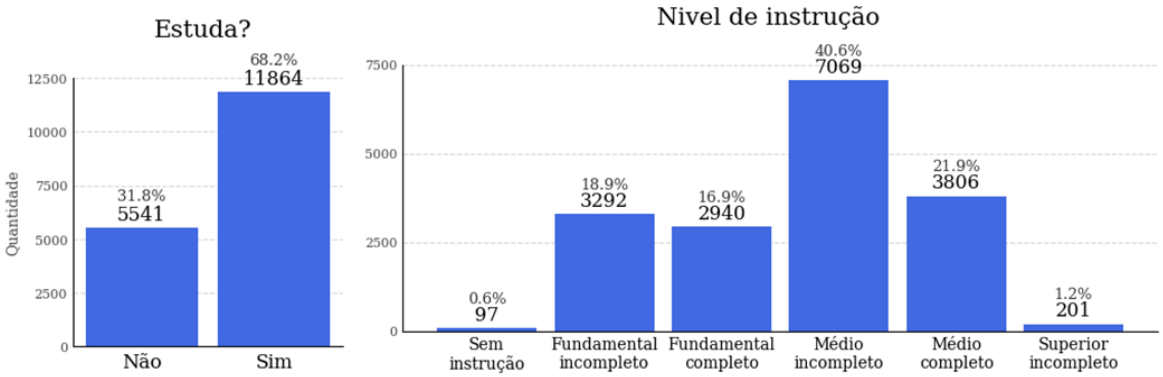
Figura 9 - Características sociais do cluster 2



Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

Em relação às variáveis relacionadas aos dados escolares, Figura 10, 68,2% dos indivíduos do *cluster* frequentam a escola. O nível de instrução predominante é o "Ensino Médio Incompleto", representando 40,6% dos casos. Além disso, 36,4% dos registros possuem um nível de instrução abaixo desse.

Figura 10 - Características escolares do cluster 2



Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

Em relação às variáveis associadas à situação de trabalho, o *cluster* apresenta 83,6% de indivíduos que trabalharam em atividades remuneradas em dinheiro. Apenas 53 (0,3%) indivíduos trabalharam em atividades remuneradas em produtos. 15,3% dos indivíduos

ajudaram no trabalho remunerado de terceiros, e 143 indivíduos (0,8%) estavam afastados de alguma ocupação. Somando as porcentagens, praticamente todos desse *cluster* estavam em alguma situação de trabalho.

As principais faixas de rendimento foram a faixa até 0,5 salário mínimo (34,0% dos casos) e a faixa de mais de 0,5 até 1 salário mínimo (32,5% dos casos).

A maior parte encontra-se nesse trabalho há um período de 1 ano a 2 anos (49,8% dos casos), e está classificado como empregado sem carteira de trabalho assinada (47,4% dos casos).

A quantidade de horas mais recorrente foi de 1 a 20 horas, com 36,4% dos registros.

Por último, a atividade principal dos trabalhos foi “Comércio e Serviços Gerais” com 44,8% dos registros.

4.4 Cluster 3

O *cluster* 3 é composto por 29934 instâncias. Em relação às características gerais deste *cluster*, Figura 11, ele possui registros dos anos de 2024 (56,0% dos registros) e de 2023 (44,0% dos registros).

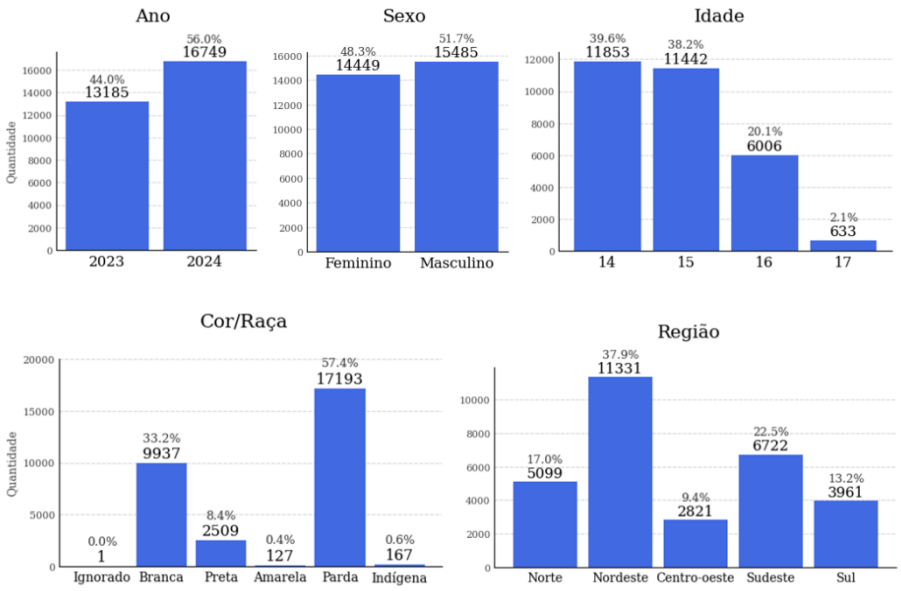
A respeito do sexo, o masculino é ligeiramente predominante, representando 51,7 % dos registros.

O *cluster* é composto principalmente por indivíduos com idades entre 14 e 15 anos (36,3% e 35,8% dos casos, respectivamente),

Quanto à cor/raça, a maior quantidade de pessoas pertence à categoria "parda", representando 57,4% dos registros.

A região com maior ocorrência é o Nordeste, com 37,9% das instâncias.

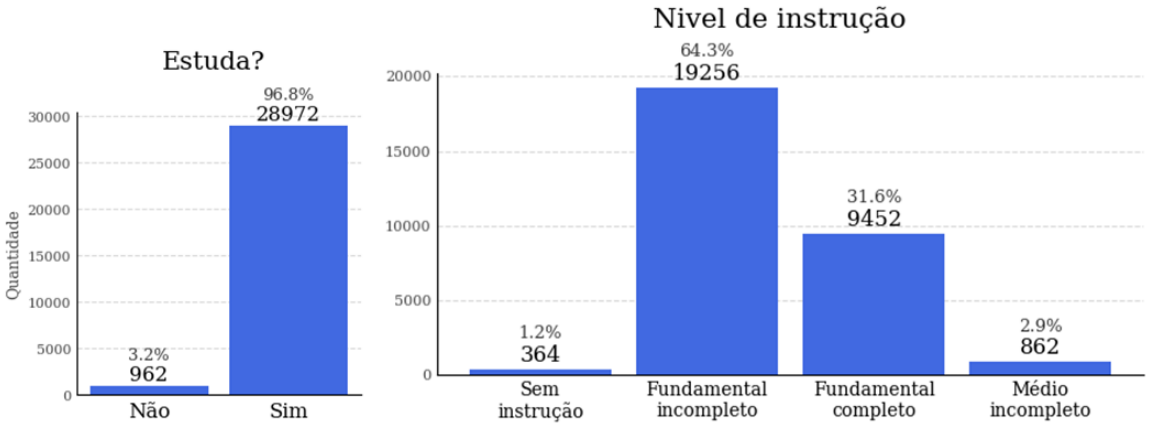
Figura 11 - Características sociais do cluster 3



Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

Em relação às variáveis relacionadas aos dados escolares, Figura 12, 96,8% dos indivíduos do *cluster* frequentam a escola. O nível de instrução predominante é o "Ensino Fundamental Incompleto", representando 64.3% dos casos.

Figura 12 - Características escolares do cluster 3



Fonte: Elaboração própria, com dados da PNAD Contínua de 2021 a 2024

Em relação às variáveis associadas à situação de trabalho, o *cluster* apresenta apenas 44 indivíduos que trabalharam em atividades remuneradas em dinheiro, nenhum indivíduo que trabalhou em atividades remuneradas em produtos, 234 indivíduos (0,8%) que ajudaram no trabalho remunerado de terceiros e 1 indivíduo que estava afastado de alguma ocupação.

4.5 Síntese das Características dos Clusters

O *cluster* 0 apresenta jovens entre 16 e 18 anos, dos quais 20,5% não estão frequentando a escola, um índice relevante comparado aos outros *clusters*. Apesar da predominância da cor/raça "parda" e da região Nordeste, essas características também se destacam em *clusters* em que a frequência escolar é alta, o que mostra que elas, isoladamente, não são suficientes para explicar o abandono escolar. Além disso, praticamente nenhum desses jovens está inserido no mercado do trabalho, o que sugere que o abandono, neste caso, não está associado diretamente à necessidade de trabalhar. Isso reforça a complexidade do fenômeno, apontando para outras possíveis causas, como desmotivação, barreiras socioeconômicas, questões familiares ou dificuldades escolares — aspectos que podem ser investigados em análises futuras.

O *cluster* 1 é formado principalmente por adolescentes de 14 e 15 anos, com alta taxa de frequência escolar (96,9%) e pouco envolvimento com o trabalho. Os registros estão concentrados nos anos de 2021 e 2022. Desta forma, trata-se de um grupo com forte presença escolar e ainda pouco impactado por fatores externos, como o trabalho.

O *cluster* 2 é o *cluster* com maior presença de jovens no mercado de trabalho, sendo que 83,6% possuem um trabalho remunerado, e praticamente todos os demais estão envolvidos em algum outro tipo de trabalho. Ao mesmo tempo, é o *cluster* com a menor taxa de frequência escolar (68,2%). Isso evidencia uma relação direta entre o abandono escolar e a necessidade de trabalhar, revelando um contexto em que o trabalho acaba se sobrepondo à continuidade dos estudos. A maioria está em empregos de nível básico, sem carteira assinada, e trabalhando de 1

a 20 horas por semana. Predominam jovens de 18 anos, do sexo masculino e de cor/raça parda, distribuídos entre as regiões Nordeste, Sudeste e Sul.

O *cluster* 3, assim como o *cluster* 1, é formado por adolescentes de 14 e 15 anos, com elevado índice de frequência escolar (96,8%) e baixa participação no mercado de trabalho. A principal diferença é que os registros são mais recentes (2023 e 2024), indicando que esse padrão de forte vínculo com a escola entre os mais jovens foi mantido ao longo do tempo.

4.6 Discussão

O presente trabalho se destaca por investigar a evasão escolar entre jovens de 14 a 18 anos, utilizando dados da PNAD Contínua dos anos de 2021 a 2024. A partir de técnicas descritivas, como a *clusterização*, foi possível identificar padrões relacionados ao abandono escolar, incluindo fatores como inserção no mercado de trabalho, região, cor/raça e sexo. Entre os resultados obtidos, destacou-se a relação entre a evasão escolar e a necessidade de trabalhar, evidenciando a inserção precoce no mercado de trabalho como um fator relevante associado ao afastamento da escola.

Esse resultado pode ser analisado em comparação com outros estudos da área de mineração de dados educacionais. Souza (2020), Costa (2021), Silva (2021) e Bineid (2022), por exemplo, desenvolveram modelos de classificação a partir de dados institucionais, enfatizando principalmente indicadores relacionados ao desempenho acadêmico, frequência e histórico escolar para identificar estudantes em risco de evasão. De forma semelhante, Fida (2020) e Tamada (2022) combinaram técnicas de agrupamento e classificação utilizando dados de ambientes educacionais, com foco na análise de desempenho e na identificação do risco de evasão a partir de indicadores acadêmicos. Estudos como os de Kunchala (2021) e O'Neill (2024), voltados ao ensino superior, também identificaram fatores relacionados ao desempenho acadêmico, às características do curso ou ao apoio financeiro como elementos relevantes para explicar o abandono.

Nesse contexto, os resultados deste estudo indicam que, no caso de jovens entre 14 e 18 anos, fatores socioeconômicos, especialmente a necessidade de trabalhar, podem ter influência significativa na evasão escolar. Do ponto de vista teórico, isso reforça a importância de considerar não apenas indicadores acadêmicos, mas também condições sociais e econômicas dos estudantes na análise do fenômeno. Do ponto de vista prático, os resultados indicam a necessidade de políticas e ações que auxiliem estudantes que precisam conciliar trabalho e estudo, contribuindo para o desenvolvimento de estratégias voltadas à permanência escolar.

5 CONSIDERAÇÕES FINAIS

Este trabalho teve como pergunta norteadora "Como as técnicas de mineração de dados são capazes de identificar fatores significativos que influenciam a evasão escolar, revelando correlações entre diversas variáveis?". Os resultados demonstraram que, por meio das etapas de preparação dos dados e da aplicação de uma técnica de mineração de dados, foi possível identificar padrões relevantes relacionados à evasão escolar, evidenciando relações significativas entre as variáveis analisadas.

Foi observado que os indivíduos mais jovens, de 14 e 15 anos, agrupados nos *clusters* 1 e 3, apresentam altas taxas de frequência escolar e baixa inserção no mercado de trabalho, o

que indica um forte vínculo com o ambiente escolar nos primeiros anos da juventude. O *cluster* 3 concentra os registros mais recentes, dos anos de 2023 e 2024, indicando que esse padrão de permanência escolar se tem mantido ao longo do tempo.

O *cluster* 2 revela um outro cenário, marcado pela predominância de jovens de 18 anos, do sexo masculino e de cor/raça parda, com distribuição significativa nas regiões Nordeste, Sudeste e Sul. Esse grupo apresentou a menor taxa de frequência escolar entre os *clusters* e a maior taxa de inserção no mercado de trabalho, geralmente em ocupações informais e empregos de nível básico, evidenciando que muitos abandonaram os estudos para trabalhar.

Por fim, o *cluster* 0, composto por jovens de 16 a 18 anos, majoritariamente da região Nordeste e de cor/raça parda, apresenta uma taxa relevante de evasão, mas com pouca inserção no mercado de trabalho, sugerindo que o afastamento da escola também pode estar associado a outros fatores, como desmotivação, dificuldades no processo de aprendizagem ou questões familiares.

Dessa forma, este trabalho conseguiu caracterizar os diferentes perfis de jovens de 14 a 18 anos, demonstrando suas principais características com relação à frequência escolar, inserção no mercado de trabalho e aspectos sociais, como região, sexo e cor/raça. Entre os resultados obtidos, destacou-se a relação entre a evasão escolar e a necessidade de conciliar trabalho e estudo, indicando esse fator como um elemento relevante no abandono escolar. Esse resultado representa uma das principais contribuições deste estudo, ao evidenciar, a partir da análise de dados, a associação entre a inserção precoce no mercado de trabalho e o afastamento da escola.

Nesse sentido, essa pesquisa pode contribuir para que gestores educacionais e instituições de ensino desenvolvam estratégias voltadas à permanência escolar, como ações de acompanhamento ou medidas de apoio aos estudantes que trabalham. Além disso, do ponto de vista metodológico, o trabalho demonstra o potencial da mineração de dados como ferramenta de apoio à análise de grandes bases de dados, permitindo identificar padrões e relações que podem auxiliar na criação de políticas educacionais.

Como sugestão para pesquisas futuras, recomenda-se aprofundar a investigação sobre os fatores que influenciam a evasão escolar em situações nas quais ela não está diretamente associada à inserção no mercado de trabalho. Os resultados observados em um dos grupos identificados indicaram jovens com taxas relevantes de evasão, mas com baixa participação no trabalho, sugerindo a influência de outros aspectos, como fatores familiares, motivacionais ou dificuldades no processo de aprendizagem. Dessa forma, estudos futuros podem explorar esses aspectos com maior detalhamento, contribuindo para ampliar a compreensão do fenômeno da evasão escolar.

6 REFERÊNCIAS

AMO, Sandra de. **Técnicas de mineração de dados**. 2004. 43 f. Tese (Doutorado) - Curso de Computação, Universidade Federal de Uberlândia, Uberlândia, 2004.

BINEID, Ahmad Abdulla. **Predicting student withdrawal from UAE CHEDS repository using data mining methodology**. 2022. 72 f. Dissertação (Mestrado) – Mestrado em Gestão de Tecnologia da Informação, The British University, Dubai, 2022.

COSTA, Alexandre Gomes da. **Aplicação de técnicas de mineração de dados e learning analytics para predição de evasão de alunos nos cursos de Ciência da Computação e Engenharias da UFPel**. 2021. 91 f. Dissertação (Mestrado) – Mestrado em Ciência da Computação, Universidade Federal de Pelotas, Pelotas, 2021.

COUTO, Diego da Costa do. **Mineração de dados educacionais aplicada à busca de perfis de alunos em casos de evasão ou retenção: uma abordagem através de Redes Bayesianas**. 2017. 89 f. Dissertação (Mestrado) – Mestrado em Engenharia Elétrica, Universidade Federal do Pará, Belém, 2017.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI Magazine,** Palo Alto, v. 17, n. 3, p. 37-54, mar. 1996. Disponível em: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230. Acesso em: 01 jun. 2025.

FIDA, Sanam. **Student performance prediction by using Cluster Analysis**. 2020. 68 f. Dissertação (Mestrado) – Mestrado em Ciência da Computação, Capital University Of Science & Technology, Islamabad, 2020.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining: um guia prático**. 4. ed. Rio de Janeiro: Elsevier, 2005.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data mining: concepts and techniques**. 3. ed. Waltham: Morgan Kaufmann Publishers, 2012.

KUNCHALA, Vikas. **Predicting undergraduate student dropout using Artificial Intelligence, Big Data and Machine Learning**. 2021. 64 f. Dissertação (Mestrado) - University Of Georgia, Athens, Geórgia (EUA), 2021.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica**. 7. ed. São Paulo: Atlas, 2010.

NUNES, Hélder Antero Amaral. **Mineração de dados socioeconômicos e educacionais de discentes para predição de evasão e retenção escolar**. 2023. 95 f. Dissertação (Mestrado) – Mestrado em Tecnologia Educacional, Universidade Federal do Ceará, Fortaleza, 2023.

O'NEILL, Kelly. **Predicting first year retention for undergraduate educational opportunity fund students**. 2024. 121 f. Dissertação (Mestrado) – Mestrado em Matemática Aplicada, Ramapo College Of New Jersey, Mahwah, 2024.

RAMOS, Ana Carolina; GONÇALVES JUNIOR, Oswaldo. Abandono e evasão escolar sob a ótica dos sujeitos envolvidos. **Educação e Pesquisa**, São Paulo, v. 50, p. e268037, abr. 2024. DOI: https://doi.org/10.1590/S1678-4634202450268037. Disponível em: https://revistas.usp.br/ep/article/view/224106. Acesso em: 31 maio 2025.

SILVA, Alex Sandro Siqueira da. **Mineração de dados aplicada à predição da evasão escolar no ensino médio**. 2021. 144 f. Tese (Doutorado) – Doutorado em Engenharia Elétrica, Universidade Estadual Paulista, Ilha Solteira, 2021.

SILVA, Maria Onelia Santos. Evasão escolar: desafios e perspectivas da educação no Brasil. *In*: KOCHHANN, Andrea (Org.). **Rumo ao futuro da Educação: tendências e desafios.** Campina Grande: Licuri, 2024. p. 239-251. Disponível em: https://editoralicuri.com.br/index.php/ojs/article/view/486. Acesso em: 01 jun. 2025.

SINGER, Cody Gene. **Educational data mining: an application of a predictive model of online student enrollment decisions**. 2023. 169 f. Tese (Doutorado) – Doutorado em Educação, Arizona State University, Tempe, 2023.

SOUZA, Alex Marques de. **Machine learning e a evasão escolar: análise preditiva no suporte à tomada de decisão.** 2020. 134 f. Dissertação (Mestrado) – Mestrado em Sistemas de Informação e Gestão do Conhecimento, Universidade FUMEC, Belo Horizonte, 2020.

TAMADA, Mariela Mizota. **Predição de evasão de cursos técnicos em EaD através de técnicas de aprendizado de máquina em duas etapas**. 2022. 155 f. Tese (Doutorado) – Doutorado em Informática, Universidade Federal do Amazonas, Manaus, 2022.