ISSN: 2358-8829

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

Isabella Lie Oshima [1]
Maria das Graças J. M. Tomazela [2]

RESUMO

A evasão escolar é um problema social persistente que afeta o desenvolvimento da educação no Brasil. A falta de conhecimento aprofundado sobre suas causas limita a eficácia das políticas públicas e das decisões escolares para enfrentá-la. 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 é utilizar técnicas de mineração de dados para identificar padrões e analisar as variáveis associadas à evasão escolar, auxiliando professores e gestores a compreender melhor os fatores que a motivam. Para isso, foi realizada uma pesquisa experimental com abordagem qualitativa, que envolveu a seleção de variáveis capazes de influenciar a evasão escolar, como características sociais e situação de trabalho, além da definição de meios para observar os efeitos dessas variáveis. Os dados escolares foram coletados da PNAD Contínua de 2021 a 2024, filtrando pela faixa etária de 14 a 18 anos. Em seguida, foi realizada a limpeza e transformação desses dados; a *clusterização* dos dados por meio da biblioteca Scikit-learn e, por fim, a análise dos *clusters*. Os resultados gerados mostraram um forte vínculo com o ambiente escolar nos primeiros anos da juventude. Por outro lado, o grupo formado predominantemente por jovens pardos de 18 anos, do sexo masculino e das regiões Nordeste, Sudeste e Sul, apresentou a menor taxa de frequência escolar e a maior taxa de inserção no mercado de trabalho. Além disso, foi identificado um grupo composto por jovens de 16 a 18 anos, majoritariamente da região Nordeste e de cor/raça parda, com uma taxa relevante de evasão, mas pouca inserção no mercado de trabalho, reforçando a complexidade do abandono escolar.

Palavras-chave: Evasão escolar, Mineração de dados, *Clusterização*, Inteligência artificial.

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

[1] Graduanda do Curso de Análise e Desenvolvimento de Sistemas da Faculdade de Tecnologia de Indaiatuba – FATEC-ID, isabella.oshima@fatec.sp.gov.br;
[2] Doutora pelo Curso de Engenharia de Produção da Universidade Metodista de Piracicaba - UNIMEP, graca.tomazela@fatec.sp.gov.br.

ISSN: 2358-8829

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. A questão norteadora deste trabalho foi: "As técnicas de mineração de dados são capazes de identificar fatores significativos que influenciam a evasão escolar, revelando correlações entre diversas variáveis?". O objetivo dessa pesquisa foi utilizar essas técnicas para identificar padrões e analisar os principais fatores associados à evasão escolar.

2 Metodologia

Esse trabalho caracteriza-se como uma pesquisa de abordagem qualitativa, buscando compreender profundamente comportamentos sociais; de natureza aplicada, elaborada para a solução de problemas práticos (Gil, 2002); com objetivo explicativo, ao investigar os fatores que contribuem para a ocorrência dos fenômenos; e com procedimentos experimentais, ao analisar as variáveis que podem influenciar o objeto de estudo definido.

Para o desenvolvimento deste trabalho foi utilizado o processo de KDD, realizando as etapas de pré-processamento, mineração de dados e pós-processamento, conforme apresentado nas próximas seções. Os resultados alcançados foram documentados, e a pesquisa foi concluída.

2.1 Pré-processamento

Na etapa de pré-processamento, os dados foram obtidos da PNAD (Pesquisa Nacional por Amostra de Domicílios) Contínua de 2021 a 2024, inicialmente em formato TXT e posteriormente convertidos para o formato CSV, com o uso da biblioteca Pandas. Os dados foram filtrados pela faixa etária de 14 a 18 anos, e foram extraídos atributos relacionados à

ISSN: 2358-8829

situação escolar, às características sociais e à situação de trabalho.

As bases de dados foram consolidá-las em um único arquivo, adicionando uma nova coluna denominada ‘Ano’. Com base na coluna UF, também foi criada uma nova coluna correspondente à região. Removeram-se os *outliers* (94 registros), identificados somente no atributo “Horas trabalhadas por semana”, e os registros inconsistentes (39 registros), que possuíam o nível de instrução “Superior Completo”, mesmo após o filtro de idade de até 18 anos. Os valores que representavam a opção “Não aplicável” foram padronizados com o valor 0. A variável ‘Horas trabalhadas por semana’ passou por um processo de discretização por frequência. As colunas ‘Ano’ e ‘Idade’ foram ajustadas para que elas ficassem na mesma escala que os outros atributos. Por último, foi aplicado o método de One Hot Encoding para transformar as variáveis categóricas ‘Sexo’, ‘Região’ e ‘Cor/raça’ em colunas binárias, evitando que os algoritmos interpretem essas categorias como hierárquicas.

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 abaixo.

Quadro 1: Atributos da base de dados e suas características

Nome do campo	**Descrição**	**Tipo**
Ano	Ano de referência (2021 a 2024)	Nominal
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)	Nominal
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
Nivel de instrucao	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, Sem Carteira Assinada, etc.)	Ordinal

Fonte: Elaboração própria

ISSN: 2358-8829

2.2 Mineração de Dados

Finalizado o pré-processamento, foi realizada a etapa de mineração de dados. 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 três métricas de avaliação: Silhoeutte Score, índice Davies-Bouldin e índice Calinski-Harabasz (Tabela 1).

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

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. O índice Davies-Bouldin mede a compactação e separação dos *clusters*, sendo que valores menores indicam uma melhor divisão. 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. 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.

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%

ISSN: 2358-8829

Fonte: Elaboração própria

3 Referencial Teórico

3.1 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 (Han; Kamber; Pei, 2012).

Nesse cenário, foi desenvolvido o processo de Descoberta de Conhecimento em Base de Dados, também conhecido como KDD (*Knowledge Discovery in Databases*). O processo de KDD refere-se à extração ou mineração de conhecimento a partir de grandes volumes de dados (Han; Kamber; Pei, 2012). 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, realiza-se a limpeza, integração, seleção e transformação dos dados.

A mineração de dados tem como objetivo a busca efetiva por conhecimentos novos e úteis. A descoberta pode ser dividida em duas atividades: 1) Previsão: aplica o aprendizado supervisionado, trabalha com conjuntos de dados rotulados, assim o pesquisador espera uma saída de dados corretamente rotulada para cada objeto de entrada; 2) Descrição: aplica o aprendizado não supervisionado, as classes não estão definidas e os algoritmos procuram estabelecer relacionamentos entre os dados. 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.

3.2 Clusterização

Clusterização é o processo de agrupar um conjunto de objetos em classes semelhantes, de forma que os objetos de um *cluster* são similares entre si e são diferentes de objetos em outros *clusters*. A *clusterização* pode ser usada para obter entendimento sobre a distribuição de dados, observar as características de cada *cluster* e focar em um conjunto específico de

ISSN: 2358-8829

clusters para análise posterior (Han; Kamber; Pei, 2012).

O *k-means* é um dos algoritmos mais utilizados para *clusterização* e funciona escolhendo aleatoriamente *k* objetos como centros iniciais dos *clusters*; depois, os demais objetos são atribuídos ao cluster mais próximo com base em uma medida de distância. Em seguida, calcula-se a média dos elementos de cada grupo para atualizar os centros, e os objetos são realocados conforme esses novos centros. Esse processo se repete até que não haja mais mudanças nos agrupamentos. O método exige que o valor de *k* seja definido previamente e é sensível a ruídos, pois valores muito diferentes podem afetar os centros dos *clusters*.

3.3 Trabalhos Relacionados

Nesta seção, são apresentados estudos e pesquisas relacionados ao tema deste trabalho.

O trabalho de Fida (2020) teve como objetivo prever quais estudantes apresentariam baixo desempenho, para melhorar a retenção e o desempenho acadêmico. Souza (2020) buscou gerar modelos de predição com algoritmos de classificação que apoiassem os gestores em ações de combate à evasão escolar. Kunchala (2021) propôs a predição da evasão em uma instituição pública de ensino superior nos Estados Unidos, por meio de modelos de classificação. Costa (2021) buscou identificar os atributos que mais afetam a evasão de alunos em Ciência da Computação e Engenharia da UFPel e determinar os melhores algoritmos de classificação para prever alunos em risco de evasão. 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. Tamada (2022) combinou técnicas de agrupamento e classificação para prever o risco de evasão em cursos EaD. Bineid (2022) buscou identificar os alunos em risco de abandono, determinar as variáveis mais influentes e verificar a precisão dos modelos de classificação. 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). 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. 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.

ISSN: 2358-8829

4 Resultados e Discussões

Com os 4 *clusters* gerados, foi feito a análise detalhada de cada um. Após as análises, apresentam-se as discussões dos resultados desse trabalho em relação às pesquisas relacionadas.

O *cluster* 0 é composto por 40483 instâncias. Em relação às características gerais deste *cluster*, 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. Em relação às variáveis relacionadas aos dados escolares, 79,5% dos indivíduos do *cluster* frequentam a escola. O nível de instrução predominante é o “Ensino Médio Incompleto”, representando 65,2% dos casos. 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.

O *cluster* 1 é composto por 28528 instâncias. Em relação às características gerais deste *cluster*, 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. Em relação às variáveis relacionadas aos dados escolares, 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*. Em relação às variáveis associadas à situação de trabalho, o *cluster* apresenta apenas 34 indivíduos que trabalharam em atividades remuneradas, nenhum indivíduo que trabalhou em atividades remuneradas em produtos, 270 indivíduos (0,9%) que ajudaram no trabalho remunerado de terceiros e 4 indivíduos que

ISSN: 2358-8829

estavam afastados de alguma ocupação.

O *cluster* 2 é composto por 17405 instâncias. Em relação às características gerais deste *cluster*, 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. Em relação às variáveis relacionadas aos dados escolares, 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.

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.

O *cluster* 3 é composto por 29934 instâncias. Em relação às características gerais deste *cluster*, 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. Em relação às variáveis relacionadas aos dados escolares, 96,8% dos indivíduos do *cluster* frequentam a escola. O nível de instrução predominante é o “Ensino Fundamental Incompleto”, representando 64.3% dos casos. Em

ISSN: 2358-8829

relação às variáveis associadas à situação de trabalho, o *cluster* apresenta apenas 44 indivíduos que trabalharam em atividades remuneradas, 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.

4.1 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, também 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

ISSN: 2358-8829

que esse padrão de forte vínculo com a escola entre os mais jovens foi mantido ao longo do tempo.

4.2 Discussões

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 trabalho remunerado, região, cor/raça e sexo.

Estudos anteriores também utilizaram abordagens descritivas, mas combinadas a abordagens preditivas. Fida (2020), por exemplo, buscou prever quais estudantes apresentariam baixo desempenho acadêmico, com o objetivo de melhorar a retenção e o rendimento estudantil. Já Tamada (2022) investigou especificamente a evasão no ensino a distância, enquanto O'Neill (2024) focou nos fatores associados ao abandono da faculdade ainda no primeiro ano do curso.

Diversos autores utilizaram exclusivamente técnicas de classificação preditiva. Souza (2020) empregou a metodologia CRISP-DM para prever a evasão escolar, assim como Costa (2021), que investigou a evasão em cursos de Ciência da Computação e Engenharia. Bineid (2022) desenvolveu modelos para identificar alunos em risco de abandono, igualmente utilizando a metodologia CRISP-DM. Silva (2021) abordou a evasão no ensino médio do Instituto Federal de Mato Grosso, também com enfoque preditivo. Kunchala (2021), por sua vez, dedicou-se à evasão no ensino superior.

Além disso, houve pesquisas que ampliaram a discussão sobre métodos preditivos no contexto educacional. Singer (2023) integrou um workshop às técnicas de classificação para prever decisões de matrícula no ensino superior, promovendo debates sobre os resultados obtidos. Já Nunes (2023) desenvolveu um software para auxiliar na previsão de abandono escolar baseado na metodologia CRISP-DM.

CONSIDERAÇÕES FINAIS

Este trabalho aplicou um processo de KDD com o objetivo de identificar as causas de abandono escolar, realizando as etapas de preparação dos dados, definição do número de *clusters*, aplicação da *clusterização* e análise dos *clusters* gerados.

ISSN: 2358-8829

A principal descoberta foi a identificação de um grupo formado majoritariamente por homens pardos de 18 anos, das regiões Nordeste, Sudeste e Sul, com baixa frequência escolar e alta inserção no mercado de trabalho. Esse padrão sugere uma relação direta entre a necessidade de trabalhar e a evasão escolar, evidenciando que muitos jovens abandonaram os estudos devido ao trabalho.

Também foi identificado um grupo com forte vínculo escolar no início da juventude, além de um grupo com evasão escolar relevante sem inserção profissional, sugerindo influência de outros fatores e reforçando a complexidade do fenômeno.

Dessa forma, este trabalho conseguiu caracterizar diferentes perfis de estudantes, demonstrando suas principais características com relação à frequência escolar, inserção no mercado de trabalho e aspectos sociais, e pode ser utilizado como base para tomadas de decisão, a fim de se mitigar os casos de abandono escolar.

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.

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.

ISSN: 2358-8829

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.

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.

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.

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.