WIDaT 2025
VIII WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA
MARÍLIA • SP

Uma abordagem preditiva da evasão na educação a distância por meio de técnicas de mineração de dados educacionais

A predictive approach to dropout in distance education through educational data mining

Weslley Rodrigues Sepúlvida[1]
http://lattes.cnpq.br/4152945701822073
https://orcid.org/0009-0002-6132-3017

Edilson Ferneda[2]
http://lattes.cnpq.br/2531761427648020
https://orcid.org/0000-0003-4164-5828

Fernando Antônio Hello[3]
http://lattes.cnpq.br/2518246712240428
https://orcid.org/0000-0003-1421-386X

Joyce Siqueira[4]
http://lattes.cnpq.br/7006325340635761
https://orcid.org/0000-0003-0246-8443

Djennifer Hipólito Portugal Amâncio Kneip[5]
http://lattes.cnpq.br/2963176937865068
https://orcid.org/0009-0007-5564-6940

Resumo

A Educação a Distância tem se consolidado como a principal escolha dos estudantes no Brasil, impulsionada por um crescimento expressivo na última década. No entanto, esse avanço trouxe o desafio da evasão estudantil, que compromete o desenvolvimento educacional e social, afeta investimentos públicos e privados e limita o

[1] Mestre, Universidade Católica de Brasília (UCB), Brasília, DF, Brasil. Docente, Faculdade de Tecnologia e Inovação Senac-DF, Brasília, DF, Brasil. weslleyrsepulvida@gmail.com.

[2] Doutor, LIRMM-Université Montpellier II, França. Docente, Universidade Católica de Brasília (UCB), Brasília, DF, Brasil. eferneda@p.ucb.br.

[3] Doutor, Universidade de Campinas (Unicamp), Campinas, SP, Brasil. Empresa Brasileira de Pesquisa Agropecuária, Brasília, DF, Brasil. hello@unicamp.br.

[4] Doutora, Universidade de Brasília (UnB), Brasília, DF, Brasil. Docente, Universidade Católica de Brasília (UCB), Brasília, DF, Brasil. joycitta@gmail.com.

[5] Graduanda, Universidade Católica de Brasília (UCB), Brasília, DF, Brasil. Assistente de Tecnologias Educacionais, Universidade Católica de Brasília (UCB), Brasília, DF, Brasil. kneipdjennifer@outlook.com.

crescimento individual e coletivo dos estudantes. Nesse contexto, a Mineração de Dados Educacionais (*Educational Data Mining - EDM*) surge como uma estratégia eficaz para desenvolver modelos preditivos que auxiliem na redução da evasão em cursos de EAD. Este estudo analisa o comportamento de estudantes evadidos em cursos de graduação de uma universidade do Centro-Oeste, utilizando dados do Ambiente Virtual de Aprendizagem (AVA) e do Sistema de Gestão Acadêmica (SGA) para apoiar decisões preventivas. O processo de análise seguiu as metodologias Knowledge Discovery in Database (*KDD*) e CRISP-DM, estruturado em quatro etapas: revisão da literatura, análise dos dados do AVA e SGA, aplicação da ferramenta Weka com cinco algoritmos classificadores (Bayes Net, SMO, J48, Naive Bayes e Multilayer Perceptron) e a proposição de ações gerenciais para mitigar a evasão. Os resultados demonstram que a análise dos dados nas primeiras quatro semanas do curso permite prever a evasão com precisão significativa, oferecendo um recurso estratégico para a gestão acadêmica e possibilitando intervenções preventivas focadas na permanência estudantil.

Palavras-chave: educação a distância; *Knowledge Discovery in Database (KDD)*; mineração de dados; *Educational Data Mining (*EDM); predição de evasão.

Abstract

Distance Education (EAD) has become the primary choice for students in Brazil, driven by significant growth over the past decade. However, this expansion has brought the challenge of student dropout, which undermines educational and social development, impacts public and private investments, and limits both individual and collective growth. In this context, Educational Data Mining (EDM) emerges as an effective strategy for developing predictive models that help reduce dropout rates in distance learning programs. This study analyzes the behavior of students who have dropped out of undergraduate courses at a university in the Midwest region, using data from the Virtual Learning Environment (AVA) and the Academic Management System (SGA) to support preventive decision-making. The analysis followed the Knowledge Discovery in Database (KDD) and CRISP-DM methodologies, structured into four stages: a literature review, data analysis from the AVA and SGA, application of the Weka tool with five classification algorithms (Bayes Net, SMO, J48, Naive Bayes, and Multilayer Perceptron), and the proposal of management actions to mitigate dropout. The results show that analyzing data within the first four weeks of the course enables the prediction of dropout with significant accuracy, providing a strategic resource for academic management and enabling preventive interventions focused on student retention.

Keywords: distance education; Knowledge Discovery in Database (KDD); data mining; Educational Data Mining (EDM); dropout prediction.

1 INTRODUÇÃO

Com o crescimento contínuo da Educação a Distância (EAD), diversas pesquisas têm sido conduzidas para compreender e reduzir a evasão nessa modalidade. Contextualizando o cenário, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), com base nos resultados do Censo da Educação Superior 2021, destaca que o Ensino a Distância cresceu 474% em uma década, mas especificamente, entre os anos 2011 e 2021 (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022). O Censo EAD.BR 2022 (Associação Brasileira de Educação à Distância, 2024) indicou que 73% dos alunos ingressantes no ano de 2022 optaram pela EAD. Assim, segundo o Censo EAD.BR 2022 “A sociedade optou pela EAD” (Associação Brasileira de Educação à Distância, 2022, p. 109).

A partir de números tão expressivos em relação ao ingresso de estudantes na modalidade, o Censo EAD.BR 2022 também apresenta a difícil realidade referente à alta evasão dos estudantes, cujo índice se concentra na faixa de até 40% (Associação Brasileira de Educação à Distância, 2024).

Os dados evidenciam que a evasão é um dos principais desafios enfrentados pelas Instituições de Ensino Superior (IES) no Brasil. O Censo EAD.BR 2022 (Associação Brasileira de Educação à Distância, 2024) destaca que o controle da evasão impacta tanto o desenvolvimento educacional quanto o social, afetando investimentos públicos e privados e limitando o crescimento individual e coletivo dos estudantes.

Ainda no Censo EAD.BR 2022 (Associação Brasileira de Educação à Distância, 2024), há fatores conhecidos que contribuem para a evasão no ensino superior, entre eles, as dificuldades financeiras, a falta de adaptação aos modelos pedagógicos e os desafios no uso de ferramentas tecnológicas. O documento ressalta que "O controle da evasão no ensino superior deve ser compreendido como um elemento crucial para o desenvolvimento educacional e social" (Associação Brasileira de Educação à Distância, 2024, p. 89). A permanência dos estudantes está diretamente relacionada à equidade e à inclusão (Associação Brasileira de Educação à Distância, 2024, p. 90).

Retomando a frase inicial deste texto: diversas pesquisas vêm sendo realizadas a fim de mitigar esse fenômeno. Neste contexto, trataremos especificamente de soluções computacionais capazes de auxiliar os gestores na tomada de decisões sobre o tema. Discutiremos uma linha de pesquisa que utiliza Técnicas de Extração de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Databases* – KDD) para extrair do Ambiente Virtual de Aprendizagem (AVA) e do Sistema de Gestão Acadêmica (SGA), conhecimento sobre o perfil e o comportamento dos estudantes evadidos.

Fayyad, Piatetsky-Shapiro e Smyth (1996) definem o KDD como um processo não trivial de extração de informações implícitas, previamente desconhecidas e potencialmente úteis a partir dos dados armazenados em um banco de dados. Esse processo envolve a execução de cinco etapas: a seleção, com a escolha dos dados relevantes para a análise; o pré-processamento, com a limpeza e preparação dos dados; a transformação, com a conversão dos dados para formatos adequados às técnicas de mineração; a mineração de dados, com a

aplicação de algoritmos para identificar padrões significativos; e a interpretação e avaliação, com a análise e validação dos padrões descobertos. Trata-se de um processo iterativo e interativo, onde todas as etapas estão conectadas e envolvem tarefas e decisões a serem realizadas pelos usuários.

Neste universo em que a Ciência da Computação se mostra como área dominante, é importante destacarmos a importante interdisciplinaridade com a Ciência da Informação (CI). Gomes, Pimenta e Schneider (2019) realizaram estudo sobre a mineração de dados na pesquisa em CI, no qual, relatam os inúmeros estudos de como os algoritmos têm impactado os aspectos sociais, econômicos e culturais na sociedade e de como o estudo crítico sobre a influência algorítmica e como as novas tecnologias de informação impactam a sociedade são uma das principais vocações do campo da CI.

Técnicas de mineração de dados são cada vez mais usadas em ciências sociais e humanidades. Gomes, Pimenta e Schneider (2019) destacam a classificação como um elemento comum na mineração de dados, cujo objetivo é identificar padrões que unem ou separam os dados analisados. Cada processo de mineração resulta em novos dados ou atributos para os conjuntos analisados, e as classificações emergentes, especialmente as complexas, necessitam de sistemas para organizar os resultados.

As pesquisas sobre evasão incluíram técnicas de KDD, utilizando informações geradas e mantidas nas bases de dados educacionais, a fim de identificar o comportamento dos estudantes com tendência a evadir, emitir alertas e/ou fornecer informações importantes para que medidas preventivas possam ser tomadas.

O *Education Data Mining* (EDM) é considerado um sinônimo de KDD quando aplicado a bases educacionais, embora o KDD tenha um escopo mais amplo. O EDM tem sido usado para identificar padrões de evasão e medir a motivação estudantil, auxiliando na tomada de decisões para reduzir a evasão. Suas atividades incluem análises preditivas e descritivas, mas nesta pesquisa, o foco é preditivo.

Dessa forma, considerando que: (i) a evasão de estudantes é apontada como o maior obstáculo enfrentado pelos cursos da modalidade EAD; (ii) partindo do pressuposto de que é possível identificar os estudantes com tendência a evadir de forma preditiva, por meio de técnicas do KDD; e (iii) a importância de gerar informações que forneçam insumo ao processo

de tomada de decisão, o objetivo da pesquisa é à construção de um modelo preditivo de evasão de estudantes de Educação a Distância (EAD), utilizando técnicas de KDD aplicadas a dados gerados por um Sistema de Gestão Acadêmica (SGA) e pelo Ambiente Virtual de Aprendizagem (AVA).

A compreensão deste comportamento, obtida por meio da interpretação dos dados dos alunos, fornecerá subsídios para a implementação de ações de mitigação da evasão, contribuindo para a manutenção de índices satisfatórios de permanência dos estudantes.

Assim, em posse de um modelo de predição de evasão, os gestores poderão receber, em tempo hábil, informações relevantes para acompanhar e agir preventivamente no combate a evasão, podendo colocar-se à disposição dos estudantes com alguma dificuldade.

Diante do cenário desafiador da evasão, este estudo reforça a relevância do uso de técnicas de Mineração de Dados Educacionais (EDM) como ferramenta estratégica para apoiar a tomada de decisões em instituições de ensino superior. Assim, os próximos passos desta pesquisa envolvem a apresentação detalhada da metodologia empregada e dos resultados obtidos com a aplicação dos algoritmos classificadores.

2 PROCEDIMENTOS METODOLÓGICOS

Trata-se de uma pesquisa de natureza aplicada, com abordagem mista, combinando métodos quantitativos e qualitativos, e com objetivos exploratórios. A investigação possui um meio experimental, utilizando a técnica de revisão bibliográfica e o método de pesquisa de estudo de caso.

O estudo de caso foi realizado em uma Instituição de Ensino Superior do Centro-Oeste, com foco nos alunos dos cursos de Tecnologia da Informação. A análise abrange dados de dois períodos letivos, totalizando 110 estudantes que utilizam o Ambiente Virtual de Aprendizagem (AVA) e o Sistema de Gestão Acadêmica (SGA). O estudo envolveu a seleção de dados e interações dos estudantes no primeiro semestre, comparando com o status desses mesmos alunos no segundo semestre. A partir dessa seleção, é possível identificar a classe de 'evadidos', que é o foco deste trabalho, e analisar os comportamentos dos estudantes por meio das técnicas de *Knowledge Discovery from Data* (KDD).

Para a investigação da base de dados na busca do comportamento do estudante propenso a evadir, foi utilizado a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM) que possui um conjunto de seis fases (compreensão do negócio, compreensão dos dados, preparação de dados, modelagem, avaliação e implantação) que devem ser cumpridas para cobrir todo o ciclo de vida de um projeto de *Data Mining* (Hotz, 2024; Piatetsky, 2014). O CRISP-DM pode ser considerado uma metodologia aplicada dentro do ciclo KDD, focando principalmente nas etapas de mineração, avaliação e implantação.

No contexto do processo de KDD, foi aplicada a técnica de mineração de dados por classificação, que tem como objetivo construir um modelo capaz de mapear os descritores de um objeto a um conjunto predefinido de classes, possibilitando a predição da classe de novos exemplos.

Os algoritmos utilizados para essa técnica estão implementados e disponíveis em diversos pacotes de análise de dados. Para esta pesquisa, foi utilizada a ferramenta WEKA, um software livre e gratuito que oferece uma ampla gama de recursos estatísticos para a avaliação e comparação de desempenho de algoritmos de mineração de dados (WEKA WIKI, 2024).

Na ferramenta WEKA foram avaliados os seguintes classificadores: Bayes Net, *Sequential Minimal Optimization* (SMO), J48, Naive Bayes e MultilayerPerceptron, cada um representando abordagens distintas de algoritmos de classificação. O Bayes Net utiliza redes bayesianas, enquanto o Naive Bayes aplica o teorema de Bayes. O J48 constrói árvores de decisão. O SMO é um otimizador para máquinas de vetores de suporte (SVM). Por fim, o MultilayerPerceptron representa uma rede neural artificial com múltiplas camadas.

Esses algoritmos foram aplicados a três cenários distintos: um conjunto de dados não variantes no tempo (*dataset*-1), que se originou do SGA; um conjunto de dados variantes no tempo (*dataset*-2), que se originou no AVA (dados extraídos 30 dias após o início das aulas); e a fusão desses dois conjuntos. As estruturas desses *datasets* são apresentadas, respectivamente, nos Quadros 1 e 2.

Quadro 1 - Estrutura do dataset-1, do SGA, com dados não variantes no tempo

Nome	Descrição
Tipo de ingresso	Vestibular / PROUNI / Transferência interna / Transferência externa / Ingresso como graduados
Tipo de matrícula	Calouro / Reabertura / Renovação / Retorno

Situação no período letivo	Cancelado / Desistente / Desligado / Formado / Matriculado / Trancado / Transferência interna
Período letivo do estudante	*numeral*
Quantidade de períodos cursados	*numeral*
Polo de educação a distância	***Polos no Brasil:*** Belém / Belo Horizonte / Brasília / Coronel Fabriciano / Fortaleza / Manaus / Porto Alegre / Recife / Rio de Janeiro / São José dos Campos / São Paulo / Vitória ***Polos do exterior:*** Angola / Estados Unidos / Japão
Sexo	M – Masculino / F – Feminino
Estados	***Brasil:*** AM / CE / DF / MG / PA / PE / PI / RJ / RS / SC / SP ***Exterior:*** Luanda/Dondo (AO) / Boston (EUA) / Tókio/Nagoia (JP)
Idade	Abaixo de 20 / 20 a 25 / 26 a 30 / 31 a 35 / 36 a 40 / 41 a 45 / Acima de 45
Estado civil	Casado / Solteiro / Outros / Não informado
Situação final	Evadido / Matriculado

Fonte: Elaboração própria.

Quadro 2 - Estrutura do dataset-2, do AVA, com dados variantes no tempo

Nome	Descrição
Quantidade de acessos à lista de participantes	*numeral*
Quantidade de acessos ao curso	*numeral*
Quantidade de acessos ao recurso perfil	*numeral*
Quantidade de acessos ao quadro de notas	*numeral*
Quantidade de participações nos fóruns	*numeral*
Quantidade de acessos aos fóruns	*numeral*
Quantidade de acessos aos exercícios	*numeral*
Quantidade de acessos ao plano de ensino	*numeral*
Quantidade de acessos aos conteúdos	*numeral*
Classe	Evadido / Matriculado

Fonte: Elaboração própria.

Conforme os Quadros 1 e 2, os dados tratados incluem apenas informações relevantes para a análise educacional, tais como características acadêmicas e comportamentais dos estudantes no AVA e no SGA, garantindo o anonimato dos participantes. Para o treinamento e a avaliação dos algoritmos, os dados são categorizados exclusivamente com o objetivo de identificar padrões relacionados à evasão ou permanência no curso. Dessa forma, nenhuma informação pessoal é exposta, respeitando os princípios da Lei Geral de Proteção de Dados (LGPD).

Após a importação dos dados para uma planilha eletrônica, eles foram submetidos ao WEKA separadamente e, posteriormente, os dados foram unificados. A extração das fontes distintas foi necessária para junção dos dados não variantes no tempo (SGA) e os dados variantes no tempo (AVA).

Os classificadores são aplicados utilizando o método *Cross Validation*, que garante

que cada instância do dado será utilizada pelo menos uma vez para treinamento e testes. Em cada uma dessas vezes, um dos grupos serve para testes e os outros são utilizados para treinamento, garantindo que todos os grupos sejam utilizados uma vez para testes e k-1 vezes para treinamento.

A comparação entre os resultados dos semestres é feita por meio de três métricas de classificação:

a) Matriz de Confusão, que relaciona classificações corretas com as predições realizadas para as classes "evadido" e "matriculado", permitindo identificar os verdadeiros positivos e negativos, assim com os falsos positivos e negativos;
b) precisão, calculada a partir da razão entre previsões verdadeiras positivas e o total de previsões positivas;
c) acurácia, que representa o percentual de classificações corretas em relação ao total de predições.

Portanto, diante do processo metodológico estruturado e das análises realizadas, este estudo avança na proposição de soluções baseadas em técnicas de Mineração de Dados Educacionais (EDM) para apoiar a gestão acadêmica. A etapa seguinte deste estudo concentra-se na apresentação dos resultados obtidos com a aplicação dos algoritmos de classificação.

3 RESULTADOS E DISCUSSÕES

A partir dos *datasets* foram realizados seis experimentos, conforme mostrado no Quadro 3.

Quadro 3 - Experimentos da pesquisa

Datasets	Experimento	Status dos Estudantes				
		Matriculado	Formado	Cancelado	Desistente	Trancado
1	1	X		X	X	
	2	X	X	X	X	
2	3	X		X	X	
	4	X	X	X	X	
1 e 2	5	X		X	X	
	6	X	X	X	X	

Fonte: Elaboração própria.

Do SGA foram extraídos inicialmente dez atributos, conforme Quadro 1. Contudo, após submissão dos atributos à função do WEKA que avalia a relevância ou não dos atributos, eles foram reduzidos para três atributos: tipo de matrícula, período letivo do estudante e estado.

Do AVA foram extraídos inicialmente nove atributos, conforme Quadro 2. Entretanto, após a submissão da relevância dos atributos via WEKA, foram utilizados cinco atributos: quantidade de acessos ao curso, quantidade de acesso aos fóruns, quantidade de acesso ao plano de ensino, quantidade de acesso aos conteúdos, quantidade de acesso à lista de participantes.

A partir desses dados, foram realizados seis experimentos com cinco classificadores em cada experimento, totalizando 30 testes, sendo que os melhores resultados de cada experimento são mostrados na Tabela 1.

Tabela 1 - Melhores resultados obtidos nos seis experimentos

Experimento	**1**	**2**	**3**	**4**	**5**	**6**
Classificador	BayesNet/ SMO	Multilayer perception	Naive Bayes	Naive Bayes	Naive Bayes	Naive Bayes
Estudantes formados	Não	Sim	Não	Sim	Não	Sim
Dados	Não variantes no tempo	Não variantes no tempo	Variantes no tempo	Variantes no tempo	Não variantes e variantes no tempo	Não variantes e variantes no tempo
Acurácia	79,7872%	83,6364%	58,5106%	57,2727%	60,6383%	61,8182%
Classificação correta dos evadidos	4 de 21	8 de 21	17 de 21	17 de 21	16 de 21	15 de 21
Precisão	0,1904	0,3809	0,8095	0,8095	0,7619	0,7142

Fonte: Elaboração própria.

Nos experimentos, foram consideradas três métricas de qualidade: matriz de confusão, precisão e acurácia. A análise dos melhores resultados revela a predominância de classificadores bayesianos, especialmente Bayes Net e Naive Bayes. Dos seis classificadores mais eficazes, cinco pertencem a essa categoria, sendo que Naive Bayes aparece quatro vezes e Bayes Net, uma vez.

Outra observação é em relação à inclusão dos estudantes formados no experimento 2, usando os dados “não variantes no tempo”. Tal inclusão apresentou 50% de melhora na quantidade de acertos dos estudantes propensos a evadir, passando de 4

acertos em 21 prováveis no primeiro experimento para 8 acertos em 21 prováveis no segundo experimento. Apesar da melhora, os experimentos mostram a dificuldade na predição da evasão utilizando somente os dados “não variantes no tempo”.

Em relação aos experimentos utilizando dados “variantes no tempo”, os resultados apresentados foram superiores aos experimentos com dados “não variantes no tempo”, sendo que a inclusão dos estudantes formados nos experimentos não apresentou ganho significativo na quantidade de acertos dos estudantes propensos a evadir.

Em relação à acurácia dos resultados, a maior obtida não representou o melhor desempenho para a pesquisa. A maior acurácia foi alcançada no experimento 2, chegando a 83,64%; entretanto, a quantidade de acertos dos estudantes propensos a evadir foi de apenas 8 em um total de 21. Esse resultado demonstra uma limitação da métrica de acurácia nesse contexto, pois ela considera o percentual geral de classificações corretas, sem diferenciar a importância relativa de cada classe. Como a evasão representa uma parcela menor do total de estudantes, um modelo que classifica corretamente a maioria dos alunos como matriculados pode apresentar alta acurácia, mas sem oferecer valor preditivo significativo para a tomada de decisão.

Já nos experimentos 3 e 4, apesar da menor acurácia (58,51% e 57,27%, respectivamente), ambos classificaram corretamente 17 dos 21 estudantes propensos a evadir. Nesse caso, a precisão na identificação do grupo-alvo — os alunos com maior risco de evasão — é mais relevante do que a acurácia geral, pois permite que a instituição intervenha de forma preventiva para reduzir as taxas de abandono. Portanto, um modelo com maior sensibilidade para identificar estudantes em risco, mesmo com menor acurácia geral, oferece maior utilidade para a gestão acadêmica.

Em relação à quantidade de acertos dos estudantes propensos a evadir, dois experimentos se destacaram: os experimentos 3 e 4. Em comum, ambos utilizaram dados “variantes no tempo”, extraídos do AVA, e o classificador Naive Bayes. Os experimentos classificaram corretamente 17 dos 21 estudantes propensos a evadir, com a precisão, métrica de qualidade utilizada para aferir os resultados obtidos, de 0,81%.

Com a união dos dados “não variantes no tempo” e “variantes no tempo”,

esperava-se que os experimentos 5 e 6 apresentassem melhora tanto na classificação dos estudantes propensos a evadir quanto na precisão. Contudo, houve uma discreta piora em relação aos experimentos que utilizaram somente os dados “variantes no tempo”. Com a junção dos dados, foram classificados corretamente 16 dos 21 estudantes propensos a evadir, alcançando uma precisão de 76%.

Com os resultados obtidos, observa-se que foi possível predizer os estudantes propensos a evadir, utilizando cinco atributos variantes no tempo, extraídos do AVA, com 30 dias de início das aulas, obtendo uma precisão de 81%.

Os resultados apresentados devem subsidiar os gestores em tomadas de decisão, a partir da elaboração de modelos para tomada de decisão. Esse modelos devem prever ações de acompanhamento dos estudantes durante o semestre letivo, visando: (i) colocar-se à disposição dos estudantes com alguma dificuldade, (ii) solicitar maior empenho dos estudantes que não acessam com frequência o AVA, (iii) lembrar dos prazos das atividades e encontros presenciais avaliativos, (iv) estimular a participação dos estudantes nas atividades acadêmicas, e /ou (vii) envio de informações incentivadoras estimulando a motivação e vínculo do estudante, (viii) provocar os professores para que estimulem os estudantes a participarem das atividades propostas nas disciplinas, (ix) premiar professores com menores índices de evasão.

4 CONSIDERAÇÕES FINAIS

Os resultados deste estudo indicam que a aplicação do processo de KDD nos dados do SGA e do AVA é eficaz na predição de estudantes em risco de evasão. Foram testados algoritmos de classificação mais simples, como Naive Bayes e Bayes Net, que demonstraram boa capacidade preditiva, especialmente ao utilizar dados variantes no tempo. A principal vantagem desses modelos é a sua simplicidade, o que facilita a implementação por parte das universidades, mesmo sem infraestrutura avançada ou equipes especializadas em aprendizado de máquina. Isso permite que instituições de ensino superior adotem estratégias preditivas sem grandes investimentos em tecnologia.

Além disso, a análise dos dados permitiu identificar padrões de comportamento associados à evasão, possibilitando intervenções mais direcionadas por parte da gestão

acadêmica. Isso reforça a importância do uso de técnicas de Mineração de Dados Educacionais (EDM) para o desenvolvimento de estratégias preventivas, contribuindo para a retenção estudantil e para a melhoria da qualidade da EAD.

É importante destacar que este estudo utiliza uma amostra composta por 110 alunos de cursos da área de TI. Os Censos da EaD de diferentes anos apontam que a taxa de evasão no ensino superior varia consideravelmente entre os cursos, o que pode influenciar os resultados de previsões relacionadas à evasão estudantil. Por outro lado, compreendemos que os dados extraídos dos AVA e do SGA possuem natureza genérica, o que pode favorecer sua aplicabilidade a diferentes áreas do conhecimento.

Como proposta para estudos futuros, recomenda-se ampliar a base de dados incluindo um número maior de estudantes, com outros cursos ou áreas e abarcando mais semestres letivos, a fim de aumentar a robustez do modelo e avaliar a possível diferença entre as áreas. Além disso, a utilização de algoritmos mais sofisticados, como Random Forest e redes neurais profundas, pode proporcionar maior precisão na predição da evasão, permitindo a identificação de padrões mais complexos e dinâmicos. Esses modelos avançados podem aprimorar ainda mais a capacidade de análise de dados educacionais, possibilitando intervenções mais eficazes e personalizadas por parte das instituições de ensino.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE EDUCAÇÃO A DISTÂNCIA. **Censo EAD.BR 2022**: relatório analítico da aprendizagem a distância no Brasil 2022. Curitiba: InterSaberes, 2024. Disponível em: https://abed.org.br/arquivos/CENSO%20EAD%202022%20PORTUGUES+INGLES.pdf. Acesso em: 24 fev. 2025.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI Magazine**, [S. l.], v. 17, n. 3, p. 37, 1996. DOI: https://doi.org/10.1609/aimag.v17i3.1230. Disponível em: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230. Acesso em: 18 maio 2025.

GOMES, Josir Cardoso; PIMENTA, Ricardo Medeiros; SCHNEIDER, Marco. Mineração de dados na pesquisa em ciência da informação: desafios e oportunidades. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 20., 2019, Florianópolis. **Anais**

[...]. Florianópolis: UFSC, 2019. Disponível em: https://conferencias.ufsc.br/index.php/enancib/2019/paper/download/546/517. Acesso em: 24 fev. 2025.

HOTZ, Nick. **Data science process alliance**: what is CRISP-DM. 2024. Disponível em: https://www.datascience-pm.com/crisp-dm-2/. Acesso em: 24 fev. 2025.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). **Censo da educação superior: ensino a distância cresce 474% em uma década.** 2022. Disponível em: https://www.gov.br/inep/pt-br/assuntos/noticias/censo-da-educacao-superior/ensino-a-distancia-cresce-474-em-uma-decada. Acesso em: 24 fev. 2025.

PIATETSKY, Gregory. **CRISP-DM, still the top methodology for analytics, data mining, or data science projects.** 28 out. 2014. Disponível em: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html. Acesso em: 24 fev. 2025.

WEKA WIKI. **Home**. 2024. Disponível em: https://waikato.github.io/weka-wiki/. Acesso em: 24 fev. 2025.

4.0 Internacional

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:
SEPÚLVIDA, Weslley Rodrigues; FERNEDA, Edilson; HELLO, Fernando Antônio; SIQUEIRA, Joyce; KNEIP, Djennifer Hipólito Portugal Amâncio. Uma abordagem preditiva da evasão na educação a distância por meio de técnicas de mineração de dados educacionais. *In*: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: https:/doi.org/10.22477/viii.widat.220.