Fatecnológica Jahu

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

MACHINE LEARNING COMO ESTRATÉGIA DE APOIO À GESTÃO ACADÊMICA: UM ESTUDO SOBRE PREDIÇÃO DE EVASÃO NO ENSINO SUPERIOR

MACHINE LEARNING AS A STRATEGY TO SUPPORT ACADEMIC MANAGEMENT: A STUDY ON DROPOUT PREDICTION IN HIGHER EDUCATION

Rodrigo Minutti Recchia
Mestre, Fatec Jahu, rodrigo.recchia@cps.sp.gov.br

Simone Cristina Mussio
Doutora, Fatec Jahu, simone.mussio@cps.sp.gov.br

DOI: https://doi.org/10.54628/issn2763-5600.v20.1.2026.347

RESUMO

A evasão escolar no ensino superior constitui um desafio recorrente para instituições educacionais, impactando indicadores de permanência, conclusão e eficiência acadêmica. Este artigo apresenta uma revisão bibliográfica sobre a aplicação de técnicas de *Machine Learning* na previsão da evasão em cursos de graduação, com ênfase em estudos que utilizam Mineração de Dados Educacionais. São discutidos trabalhos nacionais e internacionais que empregam algoritmos de classificação, como Árvores de Decisão, *Random Forest*, *Support Vector Machine e XGBoost*, destacando seus desempenhos, atributos mais relevantes e contribuições para a gestão acadêmica. Como referência ilustrativa, analisa-se um estudo de caso amplamente citado na literatura (Lanes; Alcântara, 2018), realizado na Universidade Federal do Rio Grande (FURG), que evidencia o potencial dessas técnicas na identificação precoce de estudantes em risco de evasão. Conclui-se que a aplicação de métodos de aprendizado de máquina representa uma estratégia promissora para subsidiar sistemas de apoio à decisão e políticas institucionais de permanência estudantil.

Palavras-chave: Mineração de Dados Educacionais. Evasão Escolar. *Machine Learning*. Algoritmos de Classificação. Ensino Superior.

ABSTRACT

*Student dropout in higher education constitutes a recurring challenge for educational institutions, affecting indicators of student retention, completion, and academic efficiency. This article presents a literature review on the application of Machine Learning techniques for predicting dropout in undergraduate programs, with emphasis on studies that employ Educational Data Mining. National and international works that use classification algorithms—such as Decision Trees, Random Forest, Support Vector Machine, and XGBoost—are discussed, highlighting their performance, the most relevant attributes, and their contributions to academic management. As an illustrative reference, a widely cited case study conducted at the Federal University of Rio Grande (*FURG*) (Lanes & Alcântara, 2018) is analyzed, demonstrating the potential of these techniques for the early identification of students at risk of dropout. It is concluded that the application of machine learning methods represents a promising strategy to support decision-making systems and institutional student retention policies.*

Keywords: *Educational Data Mining. Student Dropout. Machine Learning. Classification Algorithms. Higher Education.*

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

1 INTRODUÇÃO

A evasão escolar no ensino superior brasileiro tem sido amplamente discutida na literatura acadêmica em função de seus impactos sociais, econômicos e institucionais. Estudos como os de Silva Filho *et al.* (2007) evidenciam que elevados índices de abandono comprometem a eficiência do sistema educacional, reduzem as taxas de conclusão dos cursos e geram desperdício significativo de recursos públicos, além de afetarem negativamente as trajetórias acadêmicas e profissionais dos estudantes. Nesse sentido, a evasão não deve ser compreendida apenas como um fenômeno individual, mas como um problema sistêmico, relacionado a fatores acadêmicos, institucionais, socioeconômicos e pessoais.

Diante desse cenário, pesquisadores têm recorrido a abordagens baseadas em Mineração de Dados Educacionais (*Educational Data Mining – EDM*) e técnicas de *Machine Learning* como estratégias para compreender padrões associados à evasão e subsidiar a tomada de decisão institucional. A crescente digitalização dos processos acadêmicos tem possibilitado a coleta e o armazenamento de grandes volumes de dados educacionais, criando um ambiente propício para a aplicação de métodos computacionais capazes de identificar tendências, relações e fatores de risco não evidentes por meio de análises tradicionais. Trabalhos como os de Baker e Yacef (2009), Manhães *et al*. (2011) e Digiampietri *et al.* (2016) demonstram que modelos preditivos podem identificar estudantes em situação de risco com níveis relevantes de acurácia, contribuindo para o desenvolvimento de sistemas de alerta precoce e para a formulação de políticas de permanência mais eficazes.

Nesse contexto, observa-se um crescente interesse da literatura em comparar diferentes algoritmos de classificação, tais como Árvores de Decisão, *Random Forest*, *Support Vector Machine* e métodos baseados em *boosting*, com o objetivo de avaliar seu desempenho preditivo, grau de interpretabilidade e aplicabilidade no contexto da gestão acadêmica. Além da acurácia dos modelos, os estudos enfatizam a relevância de atributos relacionados ao desempenho acadêmico e à integração institucional, reforçando perspectivas teóricas consolidadas sobre a permanência estudantil.

Assim, o presente artigo não se propõe a realizar uma investigação empírica própria, mas sim sistematizar e discutir criticamente a literatura existente sobre a aplicação de técnicas de *Machine Learning* na previsão da evasão no ensino superior. Busca-se destacar os principais algoritmos utilizados, as métricas de avaliação mais recorrentes e os dados mais relevantes reportados nos estudos analisados. Como elemento ilustrativo e de apoio à discussão, é

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

analisado um estudo de caso amplamente referenciado na literatura, conduzido por pesquisadores da Universidade Federal do Rio Grande (FURG) (Lanes; Alcântara, 2018), o qual exemplifica o potencial dessas técnicas no contexto brasileiro e contribui para a compreensão de seus limites e possibilidades de aplicação**.**

2. FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica deste estudo ancora-se nos aportes da Mineração de Dados Educacionais e do Aprendizado de Máquina, com ênfase em tarefas de classificação aplicadas à previsão da evasão no ensino superior. Parte-se do pressuposto de que a crescente digitalização dos processos acadêmicos tem ampliado a disponibilidade de dados institucionais, possibilitando análises preditivas capazes de subsidiar a gestão educacional e a formulação de políticas de permanência estudantil.

Nesse contexto, são discutidos inicialmente os fundamentos da *Educational Data Mining*, destacando sua contribuição para a compreensão de fenômenos educacionais complexos, como a evasão. Em seguida, apresentam-se os princípios do aprendizado de máquina supervisionado, com foco em algoritmos de classificação amplamente utilizados na literatura. Posteriormente, são descritos e analisados algoritmos específicos (Árvores de Decisão (J48), *Random Forest*, *Support Vector Machine* (SVM) e *XGBoost*), enfatizando suas características, potencial preditivo e limitações, especialmente no que se refere à interpretabilidade dos modelos em contextos educacionais. Por fim, são abordadas as principais métricas de avaliação empregadas em estudos de evasão, considerando suas implicações práticas e analíticas.

2.1 *Educational Data Mining* e Algoritmos de Classificação

A Mineração de Dados Educacionais (*Educational Data Mining – EDM*) configura-se como um campo interdisciplinar que integra conceitos da Ciência da Computação, Estatística, Psicologia Educacional e Ciências da Educação, com o objetivo de analisar dados provenientes de ambientes educacionais e gerar conhecimento útil para a compreensão dos processos de ensino e aprendizagem. Romero e Ventura (2010) definem a EDM como a aplicação de métodos computacionais para a descoberta de padrões em grandes volumes de dados educacionais, permitindo identificar fatores associados ao desempenho acadêmico, à permanência e à evasão estudantil.

A literatura aponta que o crescimento dos sistemas acadêmicos informatizados possibilitou o armazenamento de dados relacionados ao perfil socioeconômico dos estudantes,

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

histórico escolar, desempenho acadêmico e participação institucional, criando condições favoráveis para análises preditivas mais sofisticadas (Baker; Yacef, 2009). Nesse contexto, a evasão no ensino superior tem se destacado como um dos principais fenômenos investigados pela Mineração de Dados Educacionais (EDM), devido ao seu impacto direto nos indicadores institucionais e nas políticas públicas educacionais.

Nesse cenário, pesquisas recentes têm evidenciado o avanço das investigações voltadas à predição da evasão no ensino superior por meio de técnicas de mineração de dados e aprendizado de máquina. Nascimento et al. (2022), em uma revisão sistemática sobre modelos preditivos aplicados à evasão acadêmica, destacam que diferentes abordagens computacionais vêm sendo utilizadas para identificar padrões associados ao risco de abandono estudantil. Os autores ressaltam, entretanto, que, embora esses modelos apresentem resultados promissores na identificação precoce de estudantes em risco, permanecem desafios relacionados à seleção de variáveis, à capacidade de generalização para diferentes contextos institucionais e, principalmente, à transformação das previsões em estratégias efetivas de acompanhamento e permanência estudantil. Dessa forma, os modelos preditivos devem ser compreendidos não apenas como ferramentas de classificação, mas como recursos de apoio à tomada de decisão acadêmica.

Estudos como os de Manhães *et al.* (2011) e Digiampietri *et al.* (2016) demonstram que técnicas de mineração de dados podem ser empregadas para identificar estudantes em risco de evasão ainda nos estágios iniciais do curso, permitindo a implementação de estratégias preventivas. De forma recorrente, a literatura evidencia que variáveis relacionadas ao desempenho acadêmico e à integração institucional tendem a apresentar maior poder explicativo do que atributos exclusivamente demográficos ou pré-ingresso.

2.2 Aprendizado de Máquina e Tarefas de Classificação

O aprendizado de máquina (*Machine Learning*) constitui um subcampo da Inteligência Artificial voltado ao desenvolvimento de algoritmos capazes de aprender padrões a partir de dados e realizar previsões ou tomadas de decisão sem programação explícita para cada situação (Mitchell, 1997). No âmbito da EDM, os algoritmos de aprendizado supervisionado são amplamente utilizados em tarefas de classificação, nas quais o objetivo consiste em atribuir rótulos previamente definidos – como “evadido” ou “concluinte” – a instâncias com base em seus atributos.

A literatura aponta que o processo de classificação envolve, de modo geral, duas etapas principais: a fase de treinamento, em que o modelo aprende padrões a partir de dados rotulados,

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

e a fase de teste ou validação, na qual se avalia sua capacidade de generalização para novos dados (Witten; Frank; Hall, 2011). No contexto educacional, a escolha do algoritmo não depende apenas da acurácia obtida, mas também de aspectos como interpretabilidade, custo computacional e aplicabilidade prática no ambiente institucional.

Diversos estudos analisados destacam a relevância de comparar diferentes algoritmos de classificação para a previsão da evasão, buscando identificar aqueles que oferecem melhor equilíbrio entre desempenho preditivo e transparência na interpretação dos resultados (Brito *et al*., 2015; Digiampietri *et al*., 2016).

2.3 Árvores de Decisão (J48)

As Árvores de Decisão figuram entre os algoritmos mais utilizados na literatura sobre previsão de evasão no ensino superior. O algoritmo J48, implementação do C4.5 proposto por Quinlan (1993), constrói modelos hierárquicos nos quais cada nó interno representa um teste sobre um atributo, e cada folha corresponde a uma classe de saída.

A literatura destaca que o J48 utiliza medidas baseadas em entropia e ganho de informação para selecionar os atributos mais relevantes em cada divisão da árvore. Uma de suas principais vantagens consiste na geração de regras explícitas e facilmente interpretáveis, característica considerada particularmente relevante no contexto educacional, pois permite que gestores acadêmicos compreendam os fatores associados à evasão sem necessidade de conhecimento técnico aprofundado em ciência de dados (Witten; Frank; Hall, 2011).

Estudos baseados em Árvores de Decisão, como o algoritmo J48, são frequentemente destacados na literatura por aliarem bom desempenho preditivo e elevada interpretabilidade, o que favorece sua adoção em sistemas de apoio à decisão no ensino superior (Quinlan, 1993; Witten; Frank; Hall, 2011).

2.4 *Random Forest*

O *Random Forest* é um método de aprendizado por conjunto (*ensemble learning*) proposto por Breiman (2001), que combina múltiplas Árvores de Decisão construídas a partir de amostras aleatórias do conjunto de dados. Cada árvore é treinada de forma independente, utilizando subconjuntos distintos de instâncias e atributos, e a predição final é obtida por meio de votação majoritária.

A literatura aponta que essa abordagem reduz a variância do modelo e minimiza problemas de sobreajuste *(overfitting*), frequentemente associados a árvores de decisão individuais. Em estudos sobre evasão escolar, o *Random Forest* tem sido destacado por

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

apresentar desempenho preditivo robusto, especialmente em bases de dados com grande número de atributos (Brito *et al*., 2015).

Entretanto, apesar do bom desempenho, diversos autores ressaltam que a interpretabilidade do *Random Forest* é inferior à das Árvores de Decisão simples, o que pode limitar sua aplicabilidade direta em contextos nos quais a compreensão dos fatores de risco é tão importante quanto a predição em si.

2.5 *Support Vector Machine* (SVM)

As Máquinas de Vetores de Suporte (*Support Vector Machine* – SVM), desenvolvidas por Vapnik (2000), constituem algoritmos baseados na maximização da margem entre classes, buscando encontrar o hiperplano que melhor separa os dados em um espaço de características. Para problemas não lineares, as SVM utilizam funções kernel, permitindo a projeção dos dados em espaços de maior dimensionalidade.

A literatura aponta que as SVM são particularmente eficazes em problemas com fronteiras de decisão complexas e conjuntos de dados de alta dimensionalidade. No contexto da evasão escolar, diversos estudos relatam desempenho competitivo desse algoritmo, embora ressaltem sua menor transparência interpretativa quando comparada a modelos baseados em regras explícitas (Manhães *et al*., 2011).

Assim, embora as SVM apresentem potencial preditivo relevante, sua utilização em ambientes educacionais deve considerar a necessidade de explicabilidade dos modelos para fins de gestão acadêmica e formulação de políticas institucionais.

2.6 *XGBoost* e Métodos de *Boosting*

Os métodos de *boosting* representam uma classe de algoritmos que constroem modelos de forma sequencial, em que cada novo modelo busca corrigir os erros cometidos pelos anteriores. O *XGBoost*, proposto por Chen e Guestrin (2016), destaca-se na literatura por sua eficiência computacional, capacidade de paralelização e mecanismos de regularização que reduzem o risco de sobreajuste.

A literatura aponta que o *XGBoost* tem sido amplamente utilizado em competições de ciência de dados e aplicações reais, apresentando elevado desempenho preditivo em diferentes domínios. Em estudos sobre evasão no ensino superior, esse algoritmo tem demonstrado resultados promissores, especialmente em bases de dados maiores e mais complexas (Chen; Guestrin, 2016).

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

Contudo, assim como ocorre com outros métodos de *ensemble*, sua menor interpretabilidade em comparação a Árvores de Decisão simples é frequentemente apontada como uma limitação no contexto educacional.

2.7 Métricas de Avaliação em Estudos de Evasão

A literatura indica que a avaliação de modelos preditivos de evasão envolve o uso de métricas como acurácia, precisão, *recall* e *F1-score*, além da análise da matriz de confusão. Essas métricas permitem avaliar não apenas a proporção geral de acertos, mas também a capacidade do modelo de identificar corretamente estudantes em risco de evasão, aspecto fundamental para a implementação de ações preventivas (Tan; Steinbach; Kumar, 2005).

É recorrente observar nos estudos que erros do tipo falso negativo – estudantes evadidos classificados como concluintes – são considerados mais críticos do que falsos positivos, uma vez que podem impedir a realização de intervenções antecipadas. Dessa forma, a literatura enfatiza a necessidade de interpretar os resultados dos modelos de forma cuidadosa, considerando suas implicações práticas e éticas.

3. PROCEDIMENTOS METODOLÓGICOS

O presente trabalho caracteriza-se como uma pesquisa bibliográfica, de abordagem qualitativa e caráter exploratório-descritivo, voltada à análise de estudos que abordam a aplicação de técnicas de *Machine Learning* e Mineração de Dados Educacionais na previsão da evasão no ensino superior. Foram considerados artigos científicos, dissertações e trabalhos publicados em periódicos e anais de eventos nacionais e internacionais.

Para a composição do *corpus* de análise, foram selecionados estudos publicados entre 2007 e 2024, priorizando pesquisas desenvolvidas no contexto brasileiro, sem desconsiderar contribuições internacionais relevantes para a compreensão do fenômeno da evasão e das técnicas utilizadas em sua predição. A seleção considerou a relevância temática e a pertinência dos estudos aos objetivos da pesquisa, especialmente aqueles que empregaram algoritmos de classificação aplicados à previsão da evasão acadêmica.

A análise dos estudos foi orientada por critérios previamente definidos, considerando o tipo de algoritmo empregado, os conjuntos de dados utilizados, as métricas de avaliação reportadas, os atributos considerados relevantes para a predição, as vantagens e limitações dos modelos e sua aplicabilidade no contexto da gestão acadêmica. Sempre que disponíveis, foram consideradas métricas quantitativas como acurácia, precisão, *recall* e *F1-score*, buscando

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

compreender o desempenho dos modelos sem estabelecer comparações diretas entre resultados obtidos em bases de dados e contextos institucionais distintos.

Por se tratar de uma revisão bibliográfica de caráter narrativo, a análise foi conduzida de forma interpretativa, privilegiando a articulação crítica entre os estudos selecionados. Dessa forma, os resultados não foram tratados como uma comparação estatística ou sistemática entre algoritmos, mas como uma análise das características, potencialidades e limitações das técnicas empregadas na previsão da evasão no ensino superior. Nesse conjunto, destaca-se o estudo de caso desenvolvido na Universidade Federal do Rio Grande (FURG) (Lanes; Alcântara, 2018), selecionado por sua pertinência à discussão comparativa dos algoritmos de classificação aplicados à evasão acadêmica no contexto brasileiro.

4. RESULTADOS E DISCUSSÃO

A análise da literatura revisada evidencia um consenso quanto ao potencial das técnicas de Mineração de Dados Educacionais e Aprendizado de Máquina para a previsão da evasão no ensino superior. De forma recorrente, os estudos analisados relatam desempenhos preditivos satisfatórios na aplicação de algoritmos de classificação, especialmente quando utilizados dados acadêmicos institucionais consolidados, como histórico de desempenho, participação em programas institucionais e indicadores de integração acadêmica.

Entre os algoritmos mais frequentemente investigados, destacam-se as Árvores de Decisão, o *Random Forest*, as Máquinas de Vetores de Suporte (SVM) e métodos baseados em *boosting*, como o *XGBoost*. A literatura aponta que, embora algoritmos de *ensemble* e métodos mais sofisticados frequentemente apresentem ganhos marginais em termos de acurácia, modelos baseados em Árvores de Decisão, como o J48, continuam sendo amplamente adotados em estudos educacionais devido à sua elevada interpretabilidade e facilidade de aplicação em contextos institucionais (Manhães *et al*., 2011; Brito *et al.*, 2015; Digiampietri *et al*., 2016).

Aprofundando essa questão, a preferência pelo J48 em contextos educacionais não se justifica apenas pelo desempenho preditivo, mas sobretudo pela possibilidade de extrair regras interpretáveis que gestores acadêmicos possam compreender e aplicar sem conhecimento técnico especializado. O *XGBoost*, embora apresente acurácia superior em determinados cenários, opera como um modelo de caixa-preta, dificultando a identificação dos fatores que levam à predição de evasão. Em um ambiente institucional no qual a legitimidade das decisões depende da compreensão dos critérios adotados, essa distinção é determinante: saber que estudantes com coeficiente de rendimento abaixo de determinado limiar no primeiro semestre apresentam maior risco de evasão é uma informação acionável, enquanto uma probabilidade

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

gerada por um ensemble de centenas de árvores não o é da mesma forma. Berens *et al*. (2019) corroboram esse argumento ao demonstrar que sistemas de alerta precoce baseados em modelos interpretáveis apresentam maior adesão por parte de equipes pedagógicas e gestores, precisamente porque os critérios de classificação são compreensíveis e auditáveis.

As implicações dessas escolhas para a formulação de políticas institucionais são igualmente relevantes. Hussain *et al.* (2019) destacam que a adoção de modelos preditivos na gestão acadêmica requer não apenas acurácia, mas também mecanismos que permitam traduzir os resultados em ações concretas – como programas de tutoria direcionada, suporte financeiro ou acompanhamento psicopedagógico. Essa perspectiva reforça a necessidade de que a escolha do algoritmo considere, além das métricas de desempenho, critérios de interpretabilidade e adequação ao contexto institucional de uso.

A Tabela 1 apresenta uma síntese comparativa dos principais algoritmos de classificação identificados na literatura revisada, considerando sua interpretabilidade, desempenho preditivo, limitações e aplicabilidade no contexto da gestão acadêmica. Embora os estudos analisados apresentem métricas como acurácia, precisão, *recall* e *F1-score*, as diferenças entre bases de dados e contextos institucionais limitam a comparação direta dos resultados. Dessa forma, a tabela organiza qualitativamente as principais características, potencialidades e limitações dos modelos aplicados à previsão da evasão no ensino superior.

Tabela 1 – Síntese comparativa dos algoritmos discutidos na literatura revisada

Algoritmo	**Interpretabilidade**	**Desempenho Preditivo**	**Limitação Principal**	**Adoção em Gestão**
J48 (Árvore de Decisão)	Alta	Bom a Muito Bom	Sobreajuste em bases pequenas	Alta – regras legíveis
Random Forest	Baixa a Média	Muito Bom	Difícil interpretação	Moderada
SVM	Baixa	Bom a Muito Bom	Black-box; custo computacional	Baixa
XGBoost	Baixa	Excelente	Requer técnicas extras de explicabilidade	Baixa a Moderada

Fonte: Elaborado pelos autores com base nos estudos revisados.

Como elemento ilustrativo, o estudo conduzido na Universidade Federal do Rio Grande (FURG) (Lanes; Alcântara, 2018), amplamente referenciado na literatura nacional, evidencia a relação entre desempenho preditivo e interpretabilidade dos modelos de *Machine Learning* aplicados à evasão acadêmica. O estudo de Lanes e Alcântara (2018) demonstra a aplicabilidade do algoritmo J48 na identificação de estudantes em risco de evasão na FURG, com acurácia de

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

90,7%, destacando-se pela possibilidade de geração de regras interpretáveis para gestores acadêmicos. Essa vantagem interpretativa é corroborada pela literatura mais ampla, que aponta desempenho competitivo do J48 mesmo quando comparado a modelos como *Random Forest*, SVM e métodos de *boosting* (Manhães *et al.*, 2011; Brito *et al.*, 2015; Digiampietri *et al.*, 2016). Essa característica representa um aspecto relevante para a adoção prática de modelos preditivos no ambiente educacional, uma vez que permite compreender os fatores associados ao risco de evasão e orientar estratégias institucionais de acompanhamento e permanência estudantil.

A análise das regras geradas pelo modelo baseado em Árvores de Decisão no estudo de Lanes e Alcântara (2018) evidencia padrões recorrentes também identificados em outras pesquisas. Variáveis relacionadas ao desempenho acadêmico, como coeficiente de rendimento, emergem como os principais discriminadores entre estudantes evadidos e concluintes, reforçando questões consolidadas na literatura sobre a importância da integração acadêmica para a permanência estudantil (Tinto, 1993). Outrossim, atributos associados ao envolvimento institucional, como participação em programas de bolsas ou atividades extracurriculares, mostram-se relevantes na redução do risco de evasão, corroborando modelos teóricos baseados no envolvimento estudantil (Astin, 1984).

Outro aspecto amplamente discutido nos estudos analisados refere-se à relevância relativa de diferentes tipos de atributos. De modo geral, características pré-ingresso, como tipo de escola de origem, média no exame de ingresso ou estado de procedência, tendem a apresentar menor poder explicativo quando comparadas a indicadores de desempenho e integração durante o curso. Esses dados reforçam a perspectiva de que a evasão deve ser compreendida como um fenômeno dinâmico, construído ao longo da trajetória acadêmica, e não apenas como resultado de condições iniciais do estudante.

No que se refere à avaliação dos modelos, a literatura enfatiza a importância de métricas que vão além da acurácia global, destacando a relevância de indicadores como recall e análise da matriz de confusão. Em particular, observa-se que erros do tipo falso negativo – estudantes em risco classificados como concluintes – são considerados mais críticos, pois podem inviabilizar intervenções preventivas. Essa discussão evidencia a necessidade de interpretar os resultados dos modelos à luz de suas implicações práticas e éticas, especialmente quando aplicados a políticas institucionais de permanência.

Embora os modelos de Machine Learning apresentem potencial para previsão da evasão, ressalta-se que sua efetividade depende da qualidade dos dados utilizados, da definição adequada das variáveis preditoras e da interpretação contextualizada dos resultados. Assim, a contribuição da Inteligência Artificial não deve ser compreendida apenas pela capacidade de

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

prever eventos, mas também pela possibilidade de apoiar decisões institucionais fundamentadas.

A análise crítica dos estudos revisados também permite identificar lacunas e pontos de tensão na literatura. A maioria dos trabalhos analisados concentra-se em instituições específicas, o que limita a generalização dos modelos para outros contextos. Além disso, variáveis comportamentais e psicossociais – como engajamento, satisfação com o curso e condições de saúde mental – são raramente incorporadas aos modelos, embora sua influência sobre a permanência estudantil seja amplamente reconhecida. Zhao *et al*. (2023) apontam que modelos que incorporam esse tipo de variável tendem a apresentar maior robustez preditiva e maior relevância prática, sugerindo um caminho promissor para pesquisas futuras. Há também divergências entre estudos quanto ao peso relativo de fatores socioeconômicos e acadêmicos na predição da evasão, o que reforça a necessidade de modelos contextualizados e adaptados a cada realidade institucional.

Do ponto de vista teórico, os resultados dialogam com modelos clássicos de permanência estudantil, mas também evidenciam limites de sua aplicação quando considerados isoladamente. A predominância de indicadores de desempenho acadêmico entre os principais preditores da evasão corrobora a relevância da integração acadêmica discutida por Tinto (1993). Entretanto, Santos Junior e Real (2017), em revisão da produção nacional sobre evasão, destacam o caráter multidimensional do fenômeno, evidenciando a interação entre fatores institucionais, curriculares e socioeconômicos. Nesse sentido, modelos preditivos baseados predominantemente em dados administrativos podem representar apenas parte da complexidade associada à evasão, reforçando a necessidade de abordagens teóricas e computacionais capazes de integrar diferentes dimensões do fenômeno.

Diante disso, pesquisas futuras podem avançar da identificação de quais algoritmos apresentam melhor desempenho para a investigação de como esses modelos podem ser implementados, interpretados e incorporados aos processos efetivos de acompanhamento acadêmico. Esse deslocamento do “o que” para o “como” pode contribuir para aproximar o desenvolvimento dos modelos preditivos de sua aplicação prática na gestão da permanência estudantil.

No plano prático, os resultados da revisão indicam que a incorporação de modelos preditivos à gestão acadêmica demanda mais do que a escolha de um algoritmo com boa acurácia. É necessário definir protocolos institucionais de uso – como limiares de intervenção, fluxos de acompanhamento e responsabilidades das equipes envolvidas – além de assegurar que os resultados dos modelos sejam interpretados de forma crítica e contextualizada, considerando

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

as limitações técnicas e as implicações éticas do uso de sistemas preditivos no ambiente educacional.

5. CONSIDERAÇÕES FINAIS

Este artigo teve como objetivo sistematizar e discutir criticamente a literatura sobre a aplicação de técnicas de Mineração de Dados Educacionais e Aprendizado de Máquina na previsão da evasão no ensino superior. A análise dos estudos revisados evidencia que algoritmos de classificação têm apresentado resultados consistentes na identificação de estudantes em risco, contribuindo para o desenvolvimento de sistemas de alerta precoce e para o apoio à tomada de decisão institucional.

Os trabalhos analisados indicam que, embora algoritmos mais complexos, como *Random Forest* e *XGBoost*, frequentemente apresentem elevado desempenho preditivo, modelos baseados em Árvores de Decisão, como o J48, mantêm relevância significativa no contexto educacional devido à sua interpretabilidade. A possibilidade de extrair regras claras e compreensíveis emerge como um fator decisivo para a adoção desses modelos por gestores acadêmicos, reforçando a ideia de que a escolha do algoritmo deve considerar não apenas a acurácia, mas também a transparência e a aplicabilidade prática.

De forma consistente, a literatura aponta que o desempenho acadêmico e a integração institucional figuram entre os principais fatores associados à evasão, enquanto atributos exclusivamente demográficos ou pré-ingresso apresentam menor capacidade explicativa. Esses dados dialogam com modelos teóricos consolidados sobre permanência estudantil, destacando a evasão como um fenômeno multifatorial e dinâmico, construído ao longo da experiência universitária.

Como limitações dos estudos analisados, são frequentemente mencionados aspectos como o desbalanceamento das classes, a restrição a dados administrativos, a análise concentrada em uma única instituição e a ausência de variáveis comportamentais ou psicossociais. Tais limitações indicam caminhos promissores para pesquisas futuras, incluindo o uso de técnicas de balanceamento de dados, a incorporação de atributos temporais e comportamentais, a validação externa dos modelos e o aprofundamento das discussões éticas relacionadas ao uso de sistemas preditivos no contexto educacional.

Conclui-se que a aplicação de técnicas de *Machine Learning* na previsão da evasão no ensino superior constitui uma abordagem promissora, desde que orientada por critérios de interpretabilidade, responsabilidade ética e alinhamento com políticas institucionais de

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

permanência. Espera-se que esta sistematização contribua para o avanço das pesquisas na área e subsidie gestores e pesquisadores no desenvolvimento de estratégias mais eficazes para a redução da evasão e a promoção do sucesso acadêmico. Como desdobramento desta pesquisa, sugere-se a realização de estudos aplicados em diferentes instituições de ensino superior, incorporando variáveis acadêmicas, socioeconômicas e psicossociais, além da validação comparativa dos modelos em bases reais.

REFERÊNCIAS

ASTIN, A. W. Student involvement: A developmental theory for higher education. **Journal of College Student Personnel**, v. 25, n. 4, p. 297-308, 1984

BAKER, R. S.; YACEF, K. **The state of educational data mining in 2009:** A review and future visions. Journal of Educational Data Mining, v. 1, n. 1, p. 3-17, 2009.

BERENS, J.; SCHNEIDER, K.; GÖRTZ, S.; OSTER, S.; BURGHOFF, J. **Early Detection of Students at Risk – Predicting Student Dropouts**. Administrative Sciences, v. 9, n. 1, p. 1-20, 2019.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

BRITO, D. S. et al. Predição de Reprovação de Alunos de Educação a Distância Utilizando Mineração de Dados. In: **Simpósio Brasileiro de Informática na Educação**, 26., 2015, Maceió. Anais... Porto Alegre: SBC, 2015. p. 1118-1127.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: **ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 22., 2016, San Francisco. Proceedings... New York: ACM, 2016. p. 785-794.

DIGIAMPIETRI, L. A. et al. Mineração de Dados para Identificação de Alunos com Alto Risco de Evasão: Um Estudo de Caso. **Revista Brasileira de Informática na Educação**, v. 24, n. 1, p. 147-159, 2016.

HUSSAIN, S. et al. **Using Machine Learning to Predict Student Difficulties from Learning Session Data**. Artificial Intelligence Review, v. 52, n. 1, p. 381-407, 2019.

LANES, M. A.; ALCÂNTARA, C. S. Predição de Alunos com Risco de Evasão: estudo de caso usando mineração de dados. In: **Simpósio Brasileiro de Informática na Educação**, 29., 2018, Fortaleza. Anais... Porto Alegre: SBC, 2018. p. 1921-1925.

MANHÃES, L. M. B. et al. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. In: **Simpósio Brasileiro de Informática na Educação**, 22., 2011, Aracaju. Anais... Porto Alegre: SBC, 2011. p. 150-159.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997.

NASCIMENTO, J. C. S. et al. **Predição de Evasão no Ensino Superior:** Uma Revisão Sistemática. Revista Brasileira de Informática na Educação, v. 30, p. 1-25, 2022.

ISSN 1980-8526 (print)
ISSN 2763-5600 (online)

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Mateo: Morgan Kaufmann, 1993.

ROMERO, C.; VENTURA, S. Educational Data Mining: A Review of the State of the Art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C**, v. 40, n. 6, p. 601-618, 2010.

SANTOS JUNIOR, G.; REAL, L. M. C. **A evasão na educação superior:** o estado da arte das pesquisas no Brasil a partir de 1990. Avaliação: Revista da Avaliação da Educação Superior, v. 22, n. 2, p. 365-384, 2017.

SILVA FILHO, R. L. L. et al. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, v. 37, n. 132, p. 641-659, 2007.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Boston: Addison-Wesley, 2005.

TINTO, V. **Leaving College: Rethinking the Causes and Cures of Student Attrition**. 2nd ed. Chicago: University of Chicago Press, 1993.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. 2nd ed. New York: Springer, 2000.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining:** Practical Machine Learning Tools and Techniques. 3rd ed. Burlington: Morgan Kaufmann, 2011.

ZHAO, Y. et al. **Prediction and Analysis of College Student Dropout Using Machine Learning**. Journal of Educational Computing Research, v. 61, n. 4, p. 840-864, 2023.