

Aplicação de algoritmos de *Machine Learning* na predição da evasão em cursos de exatas: uma revisão sistemática

Débora Pompilio¹

¹Universidade Federal do Paraná, Curitiba, Brasil
(deboraleite@ufpr.br)

Resumo: A evasão em cursos de exatas é um desafio relevante no ensino superior. Este trabalho apresenta uma revisão sistemática com base em 10 artigos sobre o uso de *Machine Learning* na predição desse fenômeno. Observa-se a predominância de algoritmos como *Random Forest*, árvores de decisão e redes neurais, além do uso de variáveis acadêmicas. Os resultados indicam bom desempenho preditivo e potencial para apoiar ações de retenção.

Palavras-chave: evasão estudantil; *machine learning*; cursos de exatas; predição; ensino superior

INTRODUÇÃO

A formação de profissionais nas áreas de Ciências Exatas, Tecnologias, Engenharias e Matemática (STEM) constitui um pilar fundamental para o desenvolvimento tecnológico de um país (Reis, 2012). No entanto, o ensino superior nessas áreas enfrenta um desafio persistente e complexo: as elevadas taxas de evasão estudantil. Este fenômeno não apenas compromete a trajetória individual dos discentes, mas também gera prejuízos institucionais consideráveis e agrava o déficit de mão de obra qualificada em setores estratégicos da sociedade (Silva et al., 2020).

A literatura acadêmica aponta que a desistência em cursos de exatas, como Engenharia, Computação, Matemática e Economia, apresenta contornos específicos que a distinguem de outras áreas do conhecimento (Silva et al., 2025). Frequentemente, o abandono está concentrado nos períodos iniciais,

muitas vezes associado ao desempenho em disciplinas do ciclo básico, que exigem maior domínio de conteúdos fundamentais (Godoy, 2017). No entanto, a evasão envolve diferentes fatores, incluindo aspectos acadêmicos, socioeconômicos, vocacionais e institucionais (Adachi, 2017). A dificuldade em identificar precocemente os estudantes em situação de vulnerabilidade impede que as instituições de ensino implementem estratégias de apoio e retenção de forma oportuna.

Nesse cenário, a gestão educacional tem buscado transitar de uma postura reativa para uma abordagem preventiva. A disponibilidade crescente de dados nos sistemas de informações acadêmicas abriu caminho para o uso de modelos computacionais capazes de identificar padrões de comportamento e prever o risco de abandono (Tete, 2022). Diferente das análises estatísticas tradicionais, que muitas vezes oferecem uma visão retrospectiva do problema, as técnicas de aprendizado de máquina permitem o processamento de grandes volumes de variáveis para antecipar a probabilidade de desligamento de um aluno antes que o evento ocorra (Primão, 2022).

Embora a aplicação de algoritmos preditivos na educação tenha se expandido nos últimos anos, a dispersão dos estudos e a diversidade de metodologias aplicadas dificultam uma compreensão consolidada sobre quais abordagens são mais eficazes no contexto específico das ciências exatas (Nascimento, 2022). A particularidade desses cursos, marcada por currículos densos e sequenciais, exige uma análise criteriosa de como as ferramentas de inteligência artificial estão sendo empregadas para mitigar a evasão.

Diante desse contexto, o presente trabalho propõe uma Revisão Sistemática de Literatura (RSL) focada no emprego de algoritmos de aprendizado de máquina para a previsão de evasão em cursos de exatas. O objetivo é mapear o estado da arte dessa interseção temática, identificando os algoritmos mais recorrentes, as variáveis de entrada mais influentes e as métricas de avaliação dos modelos mais utilizadas.

MATERIAL E MÉTODOS

A metodologia desta revisão de literatura foi estruturada seguindo as diretrizes do protocolo de Briner e Denyer (2012). De acordo com esse referencial, a pesquisa é organizada em cinco etapas sucessivas: o delineamento da pesquisa, o mapeamento dos estudos nas bases de dados, a triagem qualitativa, a consolidação dos achados e, por fim, a apresentação dos resultados e evidências extraídas.

A seguir, são detalhados os elementos fundamentais que compõem este percurso metodológico, incluindo as questões de pesquisa estabelecidas, os repositórios digitais consultados e a estratégia de busca aplicada. Além disso, são apresentados os critérios de inclusão e exclusão utilizados para filtrar a literatura e a composição final do portfólio de estudos selecionados para análise.

Questões de Pesquisa

O objetivo do estudo é responder as seguintes questões de pesquisa:

QP1: Quais são os algoritmos e técnicas de Machine Learning predominantes na literatura para a predição de evasão especificamente em cursos de exatas?

QP2: Quais variáveis e atributos (acadêmicos, comportamentais ou socioeconômicos) são identificados como os preditores de maior impacto para o risco de abandono em cursos de exatas?

QP3: Quais são as métricas de avaliação de desempenho utilizadas para validar os modelos e quais os níveis de eficácia (acurácia, precisão, etc.) alcançados pelos estudos no contexto de exatas?

Base de dados e estratégia de busca

Inicialmente, foram conduzidas buscas exploratórias nas bases de dados Scopus e Science Direct, com o objetivo de avaliar e ajustar os termos e operadores a serem empregados no estudo. Como estratégia de busca, definiu-se a consulta nos campos de título, resumo e palavras-chave dos artigos, utilizando a seguinte combinação: ("*student dropout*" OR "*student attrition*" OR "*academic dropout*") AND ("*machine learning*" OR "*data mining*" OR "*predictive*")

model") AND ("higher education" OR university). O recorte temporal adotado foi de 2019 a 2025.

Tabela 1. Resultados iniciais da busca por base de dados

Base de dados	Quantidade de artigos encontrados
ScienceDirect	77
Scopus	105
Total Geral	182

Após essa etapa, foi realizado um refinamento dos resultados, considerando apenas estudos relacionados a cursos de exatas. Na sequência, foram aplicados critérios de exclusão com o objetivo de manter apenas trabalhos alinhados ao escopo da pesquisa, envolvendo Machine Learning, evasão e cursos de exatas. A Tabela 2 apresenta a sistematização do processo de filtragem e seleção dos 10 artigos finais.

Tabela 2. Sistematização do processo de filtragem e seleção

Critério de exclusão	Quantidade de artigos
Número inicial de artigos	182
Artigos duplicados entre as bases	1
Exclusão por recorte temporal	4
Exclusão por tipo de documento	12
Fora do tema "Exatas/STEM"	47
Fora do alvo "Evasão"	67
Contexto universitário geral (sem recorte específico de exatas)	41
Total de artigos excluídos	172

Número de artigos final

10

RESULTADOS E DISCUSSÃO

Nesta seção, são respondidas e discutidas as questões de pesquisa apresentadas na seção anterior.

QP1: Quais são os algoritmos e técnicas de *Machine Learning* predominantes na literatura para a predição de evasão especificamente em cursos de exatas?

A análise detalhada dos dez estudos que compõem esta revisão revela um cenário tecnológico diversificado, onde a predição de evasão em cursos de exatas transita de modelos estatísticos consolidados para arquiteturas de Inteligência Artificial altamente complexas. Ao responder à primeira questão de pesquisa (QP1), observa-se que as técnicas podem ser categorizadas desde abordagens clássicas e lineares até modelos híbridos de fronteira.

No patamar das técnicas clássicas, a Regressão Logística (RL) permanece como uma ferramenta fundamental e recorrente. Orozco-Rodríguez et al. (2025) utilizam a Regressão Logística para modelar a probabilidade de abandono em uma escola de engenharia, ressaltando sua utilidade na interpretação direta de variáveis demográficas e de ingresso.

Em uma linha similar, Vega-Rebolledo et al. (2025) aplicam a Regressão Logística Multinomial como um modelo comparativo robusto para classificar trajetórias de estudantes de Engenharia de *Software*, alcançando níveis de eficácia que servem de base para algoritmos mais complexos. É importante notar que a RL também é empregada de forma estratégica como um Meta-Classificador em arquiteturas de fusão, onde sua função é consolidar as predições geradas por outros modelos mais pesados, conforme demonstrado por Kannan et al. (2024).

Avançando para os modelos baseados em árvores e aprendizado de conjunto (*ensembles*), encontramos o grupo de algoritmos mais frequente e com melhor desempenho em dados educacionais. As Árvores de Decisão são aplicadas por

Nuankaew (2019) para gerar regras de decisão claras, agindo como um nó raiz que permite aos gestores entenderem os pontos de ruptura acadêmica. No entanto, o *Random Forest* e o XGBoost (*Extreme Gradient Boosting*) emergem como os grandes destaques da literatura atual. Deleña et al. (2025) realizaram um estudo comparativo exaustivo entre dez algoritmos e concluíram que o XGBoost superou todos os demais em acurácia e precisão no contexto da educação em tecnologia, sendo particularmente eficaz em lidar com o ruído e a complexidade dos registros históricos.

Reforçando essa tendência, Osunbunmi et al. (2025) apontam o *Random Forest* como o modelo mais equilibrado e preciso para prever a persistência de alunos de engenharia ao longo de uma década de dados, superando as limitações de generalização de modelos mais simples. Kuzilek et al. (2021) também validam o uso do RF, destacando sua estabilidade ao analisar comportamentos sequenciais em exames de disciplinas críticas como Matemática e Física.

Além disso, os estudos mais recentes propõem a hibridização de paradigmas de aprendizado. Christou et al. (2023) introduzem um algoritmo inovador chamado FSC4RBF, que combina Evolução Gramatical com Redes Neurais de Função de Base Radial (RBF). Esta técnica utiliza computação evolutiva para construir e selecionar características automaticamente antes do processamento pela rede neural, resultando em uma capacidade de generalização superior para prever a duração dos estudos em departamentos de Física e Química.

Outra abordagem de alta robustez é apresentada por Khan et al. (2025), que desenvolveram uma arquitetura híbrida de empilhamento (*stacking*) que funde o poder de extração de padrões das Redes Neurais Convolucionais (CNN) com a solidez do *Random Forest*, utilizando novamente o XGBoost como meta-aprendiz para refinar as decisões finais.

Em síntese, embora algoritmos como *Support Vector Machines* (SVM) e *K-Nearest Neighbors* (KNN) apareçam como modelos de base em quase todos os experimentos comparativos (Kuzilek et al., 2021; Osunbunmi et al., 2025; Deleña et al., 2025), a técnica predominantemente utilizada e recomendada para o contexto de exatas é o XGBoost, seguido de perto pelo *Random Forest*. A

preferência por esses modelos de *ensemble* justifica-se pela sua superioridade técnica em identificar padrões não lineares em grandes volumes de dados institucionais, enquanto as novas pesquisas apontam para a fusão desses algoritmos com técnicas de *Deep Learning* e métodos evolutivos como o caminho para predições ainda mais precisas e confiáveis.

QP2: Quais variáveis e atributos (acadêmicos, comportamentais ou socioeconômicos) são identificados como os preditores de maior impacto para o risco de abandono em cursos de exatas?

No âmbito dos preditores acadêmicos, a literatura destaca o desempenho inicial como o fator determinante. Osunbunmi et al. (2025) identificam que o GPA (média geral de notas) dos dois primeiros anos e a pontuação em SAT Math são os preditores mais influentes para a persistência em engenharia, ressaltando ainda o impacto significativo de notas em disciplinas críticas como Física, Química e Cálculo I. Complementando essa visão, Vega-Rebolledo et al. (2025) apontam que, em Engenharia de *Software*, o progresso de créditos e o índice GPA atual superam até mesmo a nota do exame de admissão na predição de trajetórias de sucesso. Na mesma linha, Kannan et al. (2024) reforçam que as notas obtidas nos dois primeiros semestres e o desempenho em níveis intermediários de educação são as variáveis mais utilizadas para prever o abandono. Nuankaew (2019) traz uma especificidade técnica ao isolar seis disciplinas curriculares como fatores associados à evasão, sendo os Fundamentos de Tecnologia da Informação e a Introdução à Programação as matérias com maior peso na decisão do estudante de Computação. Por fim, Christou et al. (2023) observam que notas baixas combinadas com um aumento na duração prevista dos estudos são sinais claros de risco iminente nos departamentos de Física, Matemática e Química.

A dimensão comportamental surge como um refinamento crucial para a precisão dos modelos. Kuzilek et al. (2021) inovam ao identificar que o comportamento durante os exames (a ordem e a frequência com que o aluno se submete às provas de disciplinas como Matemática I e Física I) manifesta suas habilidades de gestão de tempo, sendo que padrões de "inatividade" em semanas

consecutivas são preditores fortíssimos de falha no primeiro ano de Engenharia Mecânica. Já Khan et al. (2025) destacam o número de tentativas anteriores em disciplinas e o volume de créditos cursados como variáveis comportamentais que segregam alunos de alto desempenho daqueles em risco de retirada institucional.

Quanto aos preditores socioeconômicos e demográficos, os estudos revelam tendências marcantes para o nicho de exatas. Orozco-Rodríguez et al. (2025) definem o perfil de maior risco em uma escola de engenharia como sendo o de estudantes do sexo masculino, oriundos de escolas secundárias privadas de ensino geral e que residem muito longe da universidade. O fator gênero também é citado por Coelho et al. (2025), que utilizam análise visual para demonstrar que mulheres residentes em dormitórios universitários apresentam taxas de saída significativamente diferentes de seus pares masculinos, embora o impacto global do gênero varie conforme o suporte institucional. Além disso, Deleña et al. (2025) inserem a renda parental e a escolaridade dos pais como variáveis de contexto que, somadas à idade e ao estado civil, ajudam a compor um modelo de predição de retenção mais robusto em cursos de TI. Em suma, os achados indicam que o preditor de maior impacto é o desempenho em matemática e ciências básicas nos primeiros semestres, mas que fatores como o tipo de escola secundária (Orozco-Rodríguez et al., 2025), a idade de ingresso (Vega-Rebolledo et al., 2025) e a eficácia da gestão do tempo no período de exames (Kuzilek et al., 2021) são as variáveis que permitem às instituições de exatas personalizar suas intervenções.

QP3: Quais são as métricas de avaliação de desempenho utilizadas para validar os modelos e quais os níveis de predição (acurácia, precisão, etc.) alcançados pelos estudos no contexto de exatas?

De acordo com a literatura revisada, as métricas de avaliação de desempenho predominantemente utilizadas são a acurácia, precisão, *recall* (sensibilidade) e *F1-score*, as quais aparecem de forma consistente nos trabalhos de Reymark D. Deleña et al. (2025), Sunawar Khan et al. (2025), K. Rajesh Kannan et al. (2024) e Ibukun Osunbunmi et al. (2025). Além destas, a área sob a curva ROC (AUC-

ROC) é empregada para validar a capacidade discriminatória dos modelos em estudos como os de Isabel Vega-Rebolledo et al. (2025), Sunawar Khan et al. (2025) e K. Rajesh Kannan et al. (2024). Pesquisas que utilizam abordagens de regressão ou modelos probabilísticos, como as de Vasileios Christou et al. (2023) e Reymark D. Deleña et al. (2025), incorporam ainda o Erro Quadrático Médio (MSE) e o *Log Loss* para quantificar a confiabilidade das estimativas e a margem de erro das previsões realizadas.

Em termos de níveis de eficácia alcançados, os estudos demonstram que as técnicas de *Machine Learning* atingem bom desempenho preditivo no contexto de exatas, especialmente quando aplicados modelos de alta performance. Reymark D. Deleña et al. (2025) reportaram que o algoritmo XGBoost alcançou uma acurácia de 93,79% e um F1-score de 93,88%, superando outros nove modelos no contexto de TI. De forma similar, Isabel Vega-Rebolledo et al. (2025) obtiveram 90% de acurácia com a Regressão Multinomial em Engenharia de Software, além de uma AUC de 0,97 para o modelo de análise de sobrevivência, o que indica uma capacidade excepcional do sistema em distinguir corretamente entre estudantes que concluem o curso e aqueles que evadem. Na área de Computação Aplicada aos Negócios, Pratya Nuankaew (2019) validou um modelo baseado em Árvore de Decisão com 87,21% de acurácia, mantendo uma precisão de 80% especificamente para a classe de evasão.

O uso de arquiteturas híbridas também se destaca como uma tendência para alcançar melhor desempenho preditivo. O modelo de hibridização proposto por Sunawar Khan et al. (2025), que funde Redes Neurais Convolucionais com *Random Forest* e XGBoost, atingiu 88% de acurácia e uma AUC-ROC de 0,90. Já o algoritmo FSC4RBF, testado por Vasileios Christou et al. (2023) em departamentos de Física e Matemática, obteve 90,90% de acerto, registrando o menor erro quadrático médio (MSE de 0,19) entre todos os modelos comparados. Complementando essa frente tecnológica, o modelo de fusão híbrida (stacking) de K. Rajesh Kannan et al. (2024) também demonstrou alta performance, alcançando uma AUC de 0,91 para a predição de abandono institucional.

A literatura ressalta ainda que a sensibilidade (*recall*) é uma métrica tão crucial quanto a acurácia, pois minimiza o risco de deixar alunos vulneráveis sem assistência. Ibukun Osunbunmi et al. (2025) alcançaram um *recall* de 89,65% com *Gradient Boosting*, garantindo a detecção da vasta maioria dos alunos em risco de não persistir na engenharia. Por sua vez, Jakub Kuzilek et al. (2021) demonstraram que a introdução de dados sequenciais de comportamento em exames aumentou o *F-measure* para a classe de alunos que falham em aproximadamente 0,3, tornando a detecção muito mais estável. Em síntese, os níveis de eficácia observados variam entre 77% e 93% de acurácia, com o uso consistente de validação cruzada por autores como Reymark D. Deleña et al. (2025) e Sunawar Khan et al. (2025) assegurando que os modelos propostos para cursos de exatas tenham alto poder preditivo.

CONCLUSÃO

Esta revisão sistemática consolidou a aplicação de Machine Learning para a predição de evasão em cursos de exatas. Em resposta à QP1, os resultados indicam que algoritmos de *Ensemble*, especificamente *XGBoost* e *Random Forest*, são as técnicas predominantes devido ao alto desempenho em bases de dados educacionais desbalanceadas.

No que se refere à QP2, os estudos convergem que o desempenho acadêmico inicial, representado pelo GPA e notas em disciplinas de base como Cálculo, Física e Química, é o indicador de maior impacto para a permanência do aluno. Variáveis sociodemográficas também se mostraram essenciais para a calibração dos modelos, com destaque para o gênero e a idade de ingresso, que permitem às instituições mapear perfis específicos de vulnerabilidade.

Quanto à QP3, as técnicas analisadas atingiram elevados níveis de eficácia, com acurácias variando majoritariamente entre 77% e 94%. A utilização sistemática de métricas como *F1-score* e *AUC-ROC* garante que os modelos sejam robustos o suficiente para sustentar Sistemas de Alerta Precoce operacionais. Conclui-se que a integração desses algoritmos ao cotidiano universitário possibilita a

realização de intervenções pedagógicas precoces, antes que o abandono se concretize, restando para pesquisas futuras a investigação de dados comportamentais em tempo real para o refinamento dessas previsões.

REFERÊNCIAS

- ADACHI, Ana Amélia Chaves Teixeira. Evasão de estudantes de cursos de graduação da USP: ingressantes nos anos de 2002, 2003 e 2004. 2017. Tese (Doutorado) – Universidade de São Paulo, São Paulo, 2017.
- BRINER, Rob; DENYER, David. Systematic review and evidence synthesis as a practice and scholarship tool. 2012.
- CHRISTOU, V. et al. Performance and early drop prediction for higher education students using machine learning. *Expert Systems with Applications*, v. 225, p. 120079, 2023.
- COELHO, D.; PAPENHAUSEN, E.; MUELLER, K. Evolutionary design of a visual analytics interface to study predictive patterns in high dimensional data. *Visual Informatics*, 2025.
- SILVA, Fernanda Cristina; DE OLIVEIRA CABRAL, Thiago Luiz; PACHECO, Andressa Sasaki Vasques. Evasão ou permanência? Modelos preditivos para a gestão do ensino superior. *Education Policy Analysis Archives*, v. 28, p. 149, 2020.
- DELEÑA, R. D. et al. Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors. *Systems and Soft Computing*, v. 7, p. 200352, 2025.
- GODOY, Elenilton Vieira; DE ALMEIDA, Eustáquio. A evasão nos cursos de engenharia e a sua relação com a matemática: uma análise a partir do COBENGE. *Educação Matemática Debate*, v. 1, n. 3, p. 339–361, 2017.
- KANNAN, K. R.; ABARNA, K. T. M.; VAIRACHILAI, S. Enhancing student academic performance forecasting in technical education: a hybrid fusion

method. SSRG International Journal of Electronics and Communication Engineering, v. 11, n. 12, p. 146–153, 2024.

KHAN, S. et al. Predictive analytics in education: enhancing student achievement through machine learning. Social Sciences & Humanities Open, v. 12, p. 101824, 2025.

KUZILEK, J.; ZDRAHAL, Z.; FUGLIK, V. Student success prediction using student exam behaviour. Future Generation Computer Systems, v. 125, p. 661–671, 2021.

NASCIMENTO, F. F. Machine learning aplicados à predição de estudantes em risco de evasão estudantil em diferentes cursos do ensino superior. 2022. Dissertação (Mestrado em Computação Aplicada) – Universidade Federal Rural do Semi-Árido, Mossoró, 2022.

NUANKAEW, P. Dropout situation of business computer students, University of Phayao. International Journal of Emerging Technologies in Learning, v. 14, n. 19, p. 117–131, 2019.

OROZCO-RODRÍGUEZ, C. et al. Dropout rate model analysis at an engineering school. Education Sciences, v. 15, n. 3, p. 287, 2025.

OSUNBUNMI, I. et al. Artificial intelligence in engineering education research: using machine learning models to predict undergraduate engineering students' persistence to graduation. Journal of Engineering Education, v. 114, n. 4, 2025.

PRIMÃO, Aline Pacheco et al. Uso de algoritmos de machine learning para prever a evasão escolar no ensino superior: um estudo no Instituto Federal de Santa Catarina. 2022.

REIS, Vivian W.; CUNHA, Paulo J. M.; SPRITZER, I. M. P. A. Evasão no ensino superior de engenharia no Brasil: um estudo de caso no CEFET/RJ. In: CONGRESSO BRASILEIRO DE ENSINO DE ENGENHARIA (COBENGE), 2012, Belém. Anais [...]. Belém, 2012.

SILVA, Jardel J. O. et al. O paradoxo dos cursos de ciência, tecnologia, engenharia e matemática (STEM): uma abordagem com base na teoria da escolha racional. ARACÊ, v. 7, n. 8, p. e7387, 2025.

TETE, Marcelo Ferreira et al. Aplicação de métodos preditivos em evasão no ensino superior: uma revisão sistemática da literatura. Education Policy Analysis Archives, v. 30, p. 149, 2022.

VEGA-REBOLLEDO, I. et al. Applying survival analysis and explainable artificial intelligence to understand academic success in software engineering students. Array, v. 28, p. 100540, 2025.