Revista Eletrônica de Iniciação Científica em Computação, 2022, 23:1
ISSN: 3085-8461 • doi: 10.5753/reic.2025.5523

Artigo de Pesquisa/Research Paper

Modelos de Inteligência Artificial para Predição de Evasão escolar: Uma revisão sistemática

Artificial Intelligence Models for Predicting School Dropout: A systematic review

Paulo Victor Santos Magalhães [Universidade Federal do Ceará | *paulomagalhaes1206@alu.ufc.br*]
Rubens Abraão da Silva Sousa [Instituto Federal de Ciência e Tecnologia do Ceará |*rubens.sousa@darmlabs.ifce.edu.br*]
Ana Caroline Bento Santos [Instituto Federal de Ciência e Tecnologia do Ceará |*ana.caroline08@aluno.ifce.edu.br*]
Felipe Jose Maia Aguiar [Instituto Federal de Ciência e Tecnologia do Ceará |*felipe.maia@ifce.edu.br*]
Ronaldo Tadeu Pontes Milfont [Instituto Federal de Ciência e Tecnologia do Ceará |*ronaldo.milfont@darmlabs.ifce.edu.br*]

Instituto Federal de Ciência e Tecnologia do Ceará, Rodovia BR 020, km 303, s/nº - Jubaia. Canindé - CE, 67200-000, Brasil.

Resumo. Os Massive Open Online Courses - MOOCs se mostram como plataformas acessíveis para o acesso à educação de forma virtual. No Brasil, o Ministério da Educação (MEC) mostrou que, entre os anos de 2011 a 2021, foi registrado um crescimento de 474% indicando a estratégia que instituições públicas e privadas vêm adotando como política de ensino. Entretanto, existe um grande desafio: as altas taxas de abandono durante o aprendizado por MOOCs. Desta forma, o objetivo deste trabalho é realizar uma revisão sistemática, entre 2018 e 2023, com foco na investigação das técnicas de inteligência artificial (IA) aplicadas à predição da evasão escolar no contexto de MOOCs. Mostrou-se, ao final desse estudo, que os algoritmos Random Forest (RF), Gaussian Naive Bayes (GNB) e Long Short-Term Memory (LSTM) são os mais utilizados na predição. Os datasets observados com grande uso foram o dataset KDD CUP 2015 e o uso de bases próprias. As métricas de avaliação comumente utilizadas nos estudos analisados foram: Precisão, Recall e F1. Durante o estudo, foi possível observar desafios no que tange à limitação e falta de diversidade dos dados; complexidade na medição da eficácia dos modelos e os fatores externos que influenciam o desempenho dos alunos. Foi observada alta adesão de algoritmos como Random Forest e Naive Bayes, além disso, é necessária uma diversificação dos dados, compreender os fatores externos à avaliação como políticas educacionais para análise das previsões.

Abstract. Massive Open Online Courses (MOOCs) have emerged as accessible platforms for virtual education. In Brazil, the Ministry of Education (MEC) reported a 474% increase in MOOC enrollments between 2011 and 2021, indicating a growing trend among public and private institutions. However, high dropout rates remain a significant challenge. This study conducts a systematic review of the literature from 2018 to 2023 to investigate the application of artificial intelligence (AI) techniques for predicting student dropout in MOOCs. The results show that Random Forest (RF), Gaussian Naive Bayes (GNB), and Long Short-Term Memory (LSTM) algorithms are the most commonly used for prediction. The KDD CUP 2015 dataset and custom datasets were frequently employed. Precision, Recall, and F1- score were the primary evaluation metrics. The study identified several challenges, including limited and diverse data, the complexity of measuring model effectiveness, and the influence of external factors on student performance. While Random Forest and Naive Bayes algorithms are popular choices, there is a need for more diverse datasets and a deeper understanding of external factors, such as educational policies, to improve prediction accuracy.

Palavras-chave: Ensino a distância, MOOCs, Evasão escolar, Inteligência artificial, Machine learning, Revisão sistemática

Keywords: Distance learning, MOOCs, School dropout, Artificial intelligence, Machine learning, Systematic review

Recebido/Received: 18 Febrary 2025 • **Aceito/Accepted:** 28 March 2025 • **Publicado/Published:**17 April 2025

1 Introdução

Segundo dados do Ministério da Educação (MEC), o ensino a distância (EaD) teve um aumento de 474% de 2011 a 2021, se tornando cada vez mais popular entre as instituições públicas e privadas [Ministério da Educação, 2022]. Para Valverde-Berrocoso *et al*. [2020], o ensino remoto se caracteriza como o uso de tecnologias para fornecer aprendizagem por meio de ferramentas digitais, permitindo educação sem a presença física. Nesse contexto, surgem as Massive Open Online Courses (MOOCs), plataformas oferecidas de forma online que não exigem pré-requisitos de uma formação acadêmica formal anterior [Jin, 2021]. O crescimento do uso de MOOCs traz consigo o aumento no uso de ferramentas de fácil acesso, que auxiliem com o propósito de melhorar resultados educacionais [Sato *et al*., 2024].

De acordo com Zhang *et al*. [2021], a taxa de abandono em MOOCs é superior a 90%, o que é extremamente alarmante. Desta forma, analisar os dados das plataformas de MOOCs é essencial para entender os fatores que influenciam a retenção e abandono dos cursos, permitindo assim o desenvolvimento de estratégias para mitigar a evasão [de Oliveira *et al*., 2021]. Uma das estratégias que podem ser utilizadas consiste no uso da Inteligência Artificial (IA) para fornecer a adaptação do conteúdo para diferentes necessidades de alunos. Com isso, o uso da IA pode se tornar uma abordagem eficaz para averiguar a persistência, identificar padrões, personalização do aprendizado por meio dos dados obtidos pelas

plataformas educacionais, contribuindo na criação de estratégias que retenham os alunos com risco de abandono [Mrhar *et al*., 2021].

Desta forma, compreender o atual cenário é essencial para identificar algoritmos, métodos, modelos e possibilidades do uso de IA em MOOCs. Diante dessa oportunidade, a pesquisa tem como objetivo realizar uma revisão sistemática dos últimos cinco anos (2018-2023), investigando as técnicas de inteligência artificial aplicadas a MOOCs.

Esta revisão busca identificar, categorizar e descrever as técnicas, métodos, algoritmos e datasets utilizados na literatura no contexto da prevenção e redução da evasão escolar. A capacidade preditiva é essencial, pois possibilita antecipações de intervenções, servindo como base para a síntese dos modelos de Machine Learning (ML). Proporcionando futuras pesquisas e aplicações práticas, inspirando o desenvolvimento de novas abordagens que utilizem ML para aumento da retenção dos alunos em MOOCs e soluções para os desafios na implementação de estratégias de retenção escolar. Dessa forma, busca-se expandir o estado da arte e contribuir com os seguintes pontos:

1. Síntese de modelos e técnicas de ML: possibilita a visualização de forma sistemática de abordagens, técnicas, métodos e algoritmos de inteligência artificial, destacando lacunas, limitações e desafios enfrentados pelos estudos analisados.
2. Mapeamento de datasets: identificação de datasets que podem ser consultados para comparação de pesquisadores e desenvolvedores que têm interesse em utilização de ML em novas possibilidades de pesquisa.
3. Identificação de oportunidade de pesquisa: nossa exploração e sistematização das informações revelam novas lacunas e limitações, possibilitando o desenvolvimento de abordagens, metodologias e percepções significativas para educadores, cientistas da computação e pesquisadores em políticas educacionais focados na previsão da evasão escolar.

Este trabalho está organizado nas seguintes seções: 1) Introdução: contextualização, desafios e motivação para a presente investigação; 2) Trabalhos relacionados: descreve trabalhos semelhantes realizados anteriormente e diferenças do presente estudo; 3) Método: detalha o passo a passo seguido para condução do trabalho; 4) Resultados: apresenta uma descrição dos dados obtidos; 5) Discussões: analisa e argumenta sobre os dados, apresentando uma visão crítica e destacando contribuições; 6) Conclusão: resume as conclusões tidas ao longo do estudo e sugere direções para trabalhos futuros.

2 Trabalhos relacionados

Esta seção apresenta trabalhos relacionados sobre técnicas de inteligência artificial para predição da evasão escolar.

Alalawi *et al*. [2023] realizaram uma revisão sistemática focada em técnicas de ML para a predição de desempenho acadêmico, particularmente na identificação de ações corretivas para estudantes em risco de evasão escolar. O estudo destaca a aplicação de ML, diferenciando-se pela seleção de características específicas para compreender melhor as intervenções educacionais. No entanto, a limitação significativa é o escopo restrito, que se concentra exclusivamente no uso de ML, sem explorar abordagens mais amplas e métodos adicionais que poderiam enriquecer a compreensão da predição de desempenho, como a análise estatística tradicional, modelagem bayesiana ou métodos qualitativos, que trariam contribuições contextuais e comportamentais valiosas.

Silva and Roman [2021], investigaram em sua revisão sistemática a evasão escolar no ensino superior e quais abordagens foram utilizadas para predição, com foco nas técnicas de ML. Embora o estudo tenha abordado questões importantes relacionadas à evasão escolar no ensino superior e o uso de técnicas de aprendizado de máquina (ML) para predição, uma de suas principais limitações foi o foco excessivo no levantamento quantitativo do uso de ML, o que resultou em uma análise superficial dos modelos, técnicas e métodos de predição específicos. A falta de uma discussão mais detalhada sobre as características e eficiência das abordagens preditivas deixa uma lacuna relevante para uma compreensão mais profunda dos métodos aplicados.

MOOCs possuem meios para identificar o desempenho dos estudantes, mas é de interesse compreender como utilizar esses resultados para determinar a precisão das previsões. Moreno-Marcos *et al*. [2019] realizaram uma revisão sistemática sobre as técnicas de ML utilizadas na previsão do desempenho de estudantes em MOOCs. O estudo identificou modelos e características comumente empregados, mas a limitação crucial foi o foco exclusivo em ML, o que restringe a generalização dos resultados e a aplicabilidade em contextos educacionais mais variados.

Balaji *et al*. [2021] buscou as contribuições de modelos de ML para o desempenho acadêmico, analisando conjuntos de dados, métricas e características dos modelos. O foco no contexto educacional específico é uma vantagem, mas a principal limitação foi a falta de consideração de diferentes cenários e métodos de pré-processamento, já que foram analisados apenas artigos relacionados a ML.

Mastour *et al*. [2023] foi direcionada à predição de desempenho em estudantes de ciências médicas, características essenciais, tamanhos de amostras e métricas de avaliação considerando a utilização de ML. Apesar de sua contribuição ao contexto, o estudo é limitado pela falta de generalização, dado o foco em um campo educacional restrito.

Nazir *et al*. [2023] em sua revisão sistemática, exploraram as técnicas de ML aplicadas à visualização do desempenho de estudantes em MOOCs, identificando variáveis e métodos utilizados. No entanto, a limitação reside no escopo reduzido, que não abrange uma análise mais ampla de outras técnicas e métodos de predição, deixando de lado abordagens alternativas de inteligência artificial que poderiam trazer trabalhos valiosos ao estudo.

De outro modo, Gamage *et al*. [2021] realizaram uma revisão sistemática sobre os desafios enfrentados pelos MOOCs no uso de algoritmos de predição de evasão, propondo também análises futuras que considerem aspectos sociais e de aprendizagem colaborativa. Embora o estudo traga contribuições valiosas, nota-se uma limitação devido à ausência de uma exploração mais abrangente de métodos e abordagens.

Desta forma, com base nos estudos anteriores, identificamos a necessidade de um estudo mais abrangente que

aborde os datasets e as lacunas na literatura no período de 2018 a 2023. Enquanto os trabalhos anteriores se concentram em aspectos específicos ou contextos limitados, nossa revisão amplia a visão, explorando áreas inexploradas e proporcionando uma base sólida para futuras pesquisas.

3 Método

Esta seção apresenta o processo metodológico utilizado para chegar aos resultados desta pesquisa, a qual está dividida em: Perguntas de pesquisa, Procedimentos, Estratégia de busca, Critérios de elegibilidade, Avaliação da qualidade e Extração de dados.

3.1 Perguntas de Pesquisa

A pesquisa buscou coletar dados que conseguissem responder a questões-chave relacionadas ao uso de inteligência artificial na predição da evasão escolar, tais como:

- RQ1 Quais os algoritmos de inteligência artificial utilizados na literatura? – Auxilia durante a identificação das soluções mais avançadas e atuais, permitindo que a pesquisa se foque nas práticas mais eficazes e inovadoras;
- RQ2 Quais bases foram utilizadas para os algoritmos de inteligência artificial? – Traz compreensão sobre os métodos utilizados e permite avaliar quais abordagens têm sido mais testadas, validadas e utilizadas nesse campo, para poder identificar oportunidades de melhorias e compreender o que justifica que o campo tenha preferências por uma determinada abordagem em detrimento de outras;
- RQ3 Que métricas foram utilizadas para verificar a acurácia das soluções? – Entender como os estudos estão selecionando e adaptando suas abordagens, levando em conta fatores com disponibilidade de dados, complexidade dos métodos e requisitos específicos do problema de evasão escolar. Cada uma dessas questões contribui para o ganho de um entendimento geral do campo e auxilia na formulação de estratégias mais robustas e eficazes no combate à evasão escolar.
- RQ4 Quais os desafios encontrados na literatura nos últimos 5 anos? – Auxilia na observação do cenário atual a respeito dos problemas que mais ocorrem dentro do contexto da predição da evasão escolar, ajuda a entender o que falta ser feito e o que precisa ser melhorado.

3.2 Procedimentos

Este estudo adotou os passos descritos na Figura 1 conforme orientação do PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [Moher *et al.*, 2009] para fazer a seleção dos artigos a serem analisados. Sua primeira etapa de seleção é a remoção dos duplicados, em seguida a leitura dos títulos, resumos e palavras-chave dos artigos de acordo com os critérios de inclusão.A terceira etapa realiza a leitura dos textos na íntegra.

A busca realizada resultou em 994 artigos, sendo 225 vindos da base de dados da IEEE Digital Library e 769 da ISI Web of Science. Destes, 153 artigos foram eliminados por serem duplicatas. Após a remoção das duplicatas, 753 artigos foram excluídos por fatores relacionados à inadequação aos critérios de exclusão. Restaram 88 artigos para a leitura na íntegra que foram aceitos para o presente estudo.

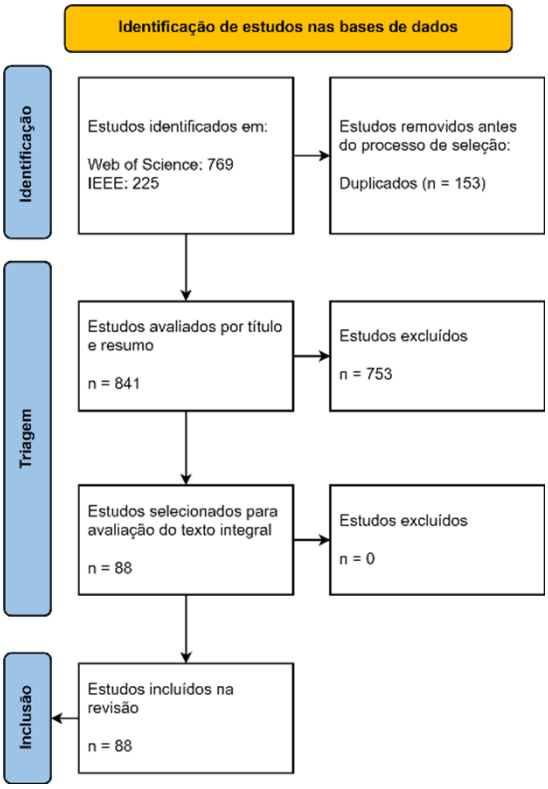

Figura 1. Procedimento de seleção dos estudos.

3.3 Estratégia de busca

A busca dos estudos foi realizada nas bases IEEE Digital Library e Web of Science, utilizando uma lógica baseada em descritores específicos do tema, juntamente com operadores booleanos (AND & OR), com os termos sinônimos isolados por parênteses (). A busca foi feita utilizando os seguintes termos em inglês:

(ML OR DL OR machine learning OR deep learning OR classification OR prediction OR predictive model) AND (approach OR method OR technique OR strategy OR algorithm OR procedure OR model) AND (learning systems OR massive open online courses OR moocs OR mooc OR moodle OR topper OR distance learning program OR teaching and learning OR distance learning OR online course OR online education) AND (forecast OR prediction OR prognosis) AND (school dropout OR student dropout OR dropout prediction OR university dropout OR dropout risk OR high school dropout OR dropout OR retention OR completion OR attrition OR withdrawal)

Essa estratégia leva em consideração o PICOC (Population, Intervention, Comparison, Outcome, and Context). A Tabela 1 mostra a esquematização para chegar aos sinônimos isolados.

3.4 Critérios de elegibilidade

Dois autores realizaram juntos a etapa de seleção dos estudos, buscando minimizar os riscos de viés. Primeiramente, ambos realizaram a seleção com base no título e resumo dos

Tabela 1. Estrutura PICOC aplicada à predição de evasão escolar com machine learning.

PICOC	Descrição	Sinônimos
Population	machine learning aplicado à predição da evasão escolar nos últimos 5 anos	(ML OR DL OR machine learning OR deep learning OR classification OR prediction OR predictive model)
Intervention	Abordagens, técnicas e métodos de machine learning comuns na literatura	(approach OR method OR technique OR strategy OR algorithm OR procedure OR model)
Comparison		
Outcome	Bases de dados utilizadas nos trabalhos de técnicas, abordagens e métodos de machine learning para predição da evasão escolar	(learning systems OR massive open online courses OR moocs OR mooc OR moodle OR topper OR distance learning program OR teaching and learning OR distance learning OR online course OR online education)
Context	Previsão de evasão escolar	(forecast OR prediction OR prognosis) AND (school dropout OR student dropout OR dropout prediction OR university dropout OR dropout risk OR high school dropout OR dropout OR retention OR completion OR attrition OR withdrawal)

artigos, checados por pares para verificar a validade da decisão. Foram tomados os seguintes critérios para inclusão:

- Artigos disponíveis para leitura;
- Artigos escritos em inglês;
- Artigos originais;
- Artigos publicados entre **2019 a 2023**;
- Artigos que apresentem aplicação de técnicas, abordagens e métodos de machine learning em contexto de evasão escolar;
- Artigos que apresentem avaliação em base de dados.

3.5 Avaliação da qualidade

A qualidade de todos os estudos foi avaliada com base em alguns critérios de avaliação, buscando garantir a qualidade estrutural dos artigos a serem analisados. A pontuação máxima foi de 7 para os trabalhos que contemplassem todos os critérios de qualidade definidos pelos autores, o questionário realizado está presente na Tabela 2.

Para cada uma das questões do questionário de qualidade, apenas duas alternativas de pontuação eram permitidas: Sim = 1 ou Não = 0, o que totalizou 7 pontos com todas as questões. A avaliação não utilizou nota de corte por buscar não excluir nenhum estudo relevante para vista de um mapeamento abrangente (UL HAQ, 2008).

3.6 Extração de Dados

Um formulário estruturado de coleta de dados foi utilizado para realizar a extração de dados relevantes para responder às questões de pesquisa. As questões do formulário se destinavam a extrair dados sobre: os algoritmos de machine learning utilizados para a previsão de evasão escolar; as bases de dados que foram utilizadas; qual a taxa de acerto final do algoritmo nas bases de dados; quais métricas os trabalhos utilizaram para avaliar a eficácia dos algoritmos; e quais os desafios e as possibilidades de pesquisa puderam ser identificadas nos artigos. A Tabela 3 mostra o formulário de Extração de Dados.

Tabela 2. Critérios de qualidade da pesquisa.

Critérios de qualidade	Sim	Não
O estudo descreve de forma clara o objetivo da pesquisa?		
O estudo expõe de forma clara as contribuições a partir do trabalho?		
O estudo justifica de forma clara a motivação para o uso do algoritmo/método de machine learning?		
O estudo justifica de forma clara a motivação para o uso do algoritmo/método de machine learning?		
O modelo foi treinado e validado em conjuntos de dados separados?		
O estudo relata claramente os procedimentos adotados?		
Os resultados do estudo foram apresentados de forma clara e concisa?		
As conclusões do estudo foram baseadas nos resultados da pesquisa?		

Tabela 3. Descrição das perguntas relacionadas.

Descrição	Perguntas relacionadas
Quais os algoritmos de machine learning utilizados?	RQ1
Quais bases foram utilizadas?	RQ2
Quais outras métricas foram avaliadas no trabalho (listar resultados)?	RQ3
Desafios e possibilidades de pesquisa?	RQ4

4 Resultados

4.1 Quais são os algoritmos de inteligência artificial utilizados na literatura?

A Figura 2 ilustra a quantidade de vezes que determinadas técnicas foram encontradas nos trabalhos relatados. Pode-se

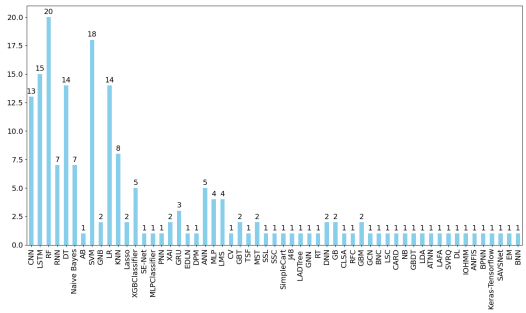

Figura 2. Técnicas mais utilizadas.

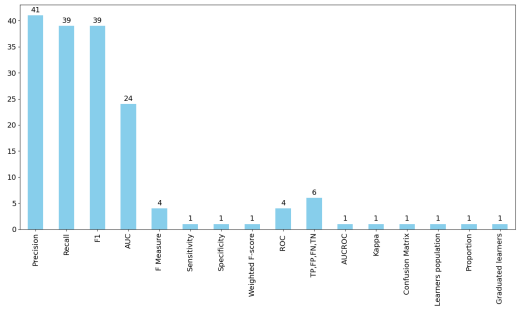

Figura 4. Métricas mais utilizadas.

perceber que a frequência maior é da técnica Random Forest (RF), seguida pelas técnicas Gaussian Naive Bayes (GNB), Long Short-Term Memory (LSTM), K-Nearest Neighbors (KNN), Naive Bayes (NB) e Convolutional Neural Network (CNN).

4.2 Quais bases foram utilizadas para os algoritmos de inteligência artificial?

A quantidade de vezes que determinados datasets foram utilizados nos trabalhos relatados de acordo com a Figura 3. Percebe-se que a maior frequência de utilização é do dataset KDD CUP 2015, seguido pelos datasets Bases Próprias e OULAD. Outros datasets também foram utilizados nos trabalhos, porém em menor quantidade do que os mencionados anteriormente. Exemplos desses datasets são: MOOC Cube, XueTangX, EVG, MITx, HarvardX, Open University, CUP 2015, Unitelma, U.S. Census Bureau, Stanford, CourseRA Course, E-Khool LMS, e Moodle.

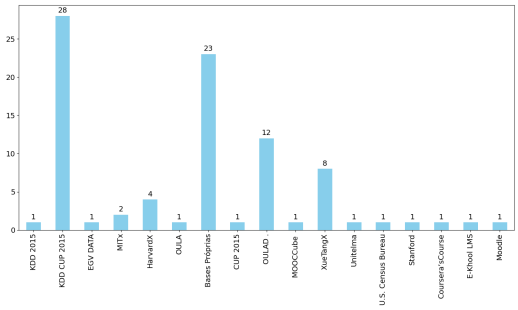

Figura 3. Datasets mais utilizados.

A maior parte dos trabalhos utilizou bases de dados públicas. A base que foi mais utilizada nos artigos foi a KDD CUP 2015, que teve 31,8% de ocorrência. Depois dela, as bases mais utilizadas foram OUDLAD, XueTangX, HarvardX e MITx, com respectivamente 13,6%, 9,1%, 4,5% e 2,3% de ocorrências. O restante das bases mencionadas foi utilizado em poucos trabalhos, com todas as ocorrências equivalendo a 1,1%, sendo essas: KDD 2015, EV.G, Open University, CUP 2015, MOOCCUBE, Unitelma, U.S. Census Bureau, Stanford, Coursera's Course, E-Khool LMS e Moodle.

4.3 Que métricas foram utilizadas para verificar a acurácia das soluções?

A Figura 4 ilustra a quantidade de vezes que determinadas métricas foram utilizadas nos trabalhos relatados. Sendo as mais frequentemente utilizadas Precision, Recall e F1, seguidas por AUC e F Measure.

Outras métricas também foram mencionadas nos trabalhos, mas com menor frequência em comparação com as citadas anteriormente. Exemplos dessas métricas são: Sensitivity, Specificity, Weighted F-Score, ROC, TP/FP/FN/TN, AUCROC, Kappa, Confusion Matrix, Learners population, Proportion, e Graduated learners.

4.4 Quais são os principais desafios relatados?

A Figura 5 mostra que os desafios mais comuns nos trabalhos analisados são a limitação de dados, a restrição de diversidade de dados, a complexidade em medir a eficácia dos modelos e os fatores que influenciam o desempenho dos alunos. Outros desafios, como datasets mais recentes, validação de modelos, novas técnicas de ML e alto custo computacional, aparecem com menor frequência.

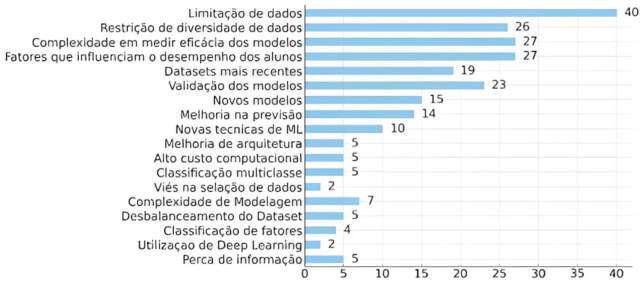

Figura 5. Desafios mais mencionados.

Entre as possibilidades de melhorias mais exploradas foram encontradas: melhoria na previsão, algoritmos mais genéricos e análises interdisciplinares. Além dessas, tivemos outras como: flexibilidade, escalabilidade e personalização dos algoritmos sendo analisadas menos frequentemente.

5 Discussões

Na predição de evasão escolar utilizando algoritmos de ML, questões são abordadas a partir de análises de estudos recentes. A identificação de algoritmos mais usados, bases de dados preferenciais, métricas de avaliação e desafios e possibilidades futuras, estabelecem pontos essenciais para o aprimoramento do uso dessas técnicas.

Observou-se que Random Forest (RF) destacou-se como o mais utilizado. Segundo Herrera *et al*. [2019], o algoritmo RF é especialmente eficaz na manipulação de grandes volumes de dados, uma característica comum em plataformas educacionais. Sua capacidade traz consigo fatores essenciais para datasets com quantidade massiva de dados,

tornando-o auxiliar na seguridade de previsões precisas. O uso de RF atende às necessidades que capturam a complexidade inerente aos dados educacionais, o que explica sua popularidade entre os pesquisadores.

Mostra-se uma concentração em pesquisas utilizando KDD CUP 2015. Assim, proporciona uma base sólida para validações de algoritmos de ML, o que levanta uma preocupação em termos de generalização dos resultados. A dependência excessiva de uma única fonte de dados limita a diversidade dos cenários analisados e, consequentemente, pode restringir a aplicabilidade dos modelos a novos contextos educacionais. Isso sugere uma necessidade urgente de diversificação dos datasets, a fim de garantir a validade externa das conclusões.

As métricas utilizadas para verificar a acurácia das soluções desempenham um papel crucial na compreensão da previsão de alunos com risco de abandono em MOOCs. Quanto maior a precisão das métricas, maior a confiança nas análises dos dados. Assim, o desempenho dos modelos preditivos é fortalecido, assegurando a captação na identificação de padrões.

Quanto aos desafios enfrentados na literatura nos últimos cinco anos, destaca-se a necessidade de dados mais diversificados para melhorar a precisão dos resultados. Como consequência, muitos estudos enfrentam dificuldades em obter conjuntos de dados diversos para treinar os algoritmos de ML. Isso resulta em dificuldades para os algoritmos aprenderem a generalizar para novos cenários, prejudicando a precisão na predição da evasão escolar em plataformas digitais.

Em relação às possibilidades futuras, incluem-se o desenvolvimento de estratégias não só na coleta, mas também na diversificação dos dados. A criação e utilização de dados sintéticos, bem como a formação de parcerias entre instituições para compartilhamento de dados, são caminhos que podem ser explorados. Nessas abordagens, os resultados das predições estariam mais precisos e eficazes na identificação e prevenção de evasão escolar. Com isso, espera-se que os modelos possam fornecer predições mais precisas e contribuir de maneira mais significativa para a identificação e prevenção da evasão escolar em ambientes de ensino remoto.

6 Conclusão

O estudo revisou de forma sistemática as técnicas, abordagens e métodos de ML em relação à predição de evasão escolar em MOOCs. A análise mostrou que os algoritmos Random Forest e Naive são amplamente adotados nos estudos observados, com predominância do dataset KDD CUP 2015. Em contrapartida, a dependência das bases de dados limita a diversificação das fontes para melhorar a generalização dos modelos.

As contribuições acerca da revisão oferecem uma visão ampla sobre as limitações atuais, sendo a necessidade de maior diversidade de dados. Fornecendo base sólida para a construção de modelos mais robustos e generalizáveis. Assim, compreendendo as práticas predominantes e identificando áreas para inovação, facilitando a evolução de técnicas de ML em MOOCs no auxílio à evasão escolar.

Há oportunidades de desenvolver estudos voltados à diversidade de dados que incluem variáveis como contexto econômico, políticas institucionais, fatores culturais e psicológicos. Como também podem guiar o desenvolvimento de algoritmos flexíveis e personalizados, apontando para estratégias utilizadas no cenário atual. Permitindo assim o desenvolvimento de modelos capazes de prever invasão escolar com maior precisão, promovendo uma política educacional personalizada para cada contexto das instituições que utilizam MOOCs. Além disso, é observada a relevância da incorporação dessas técnicas de ML na oferta dos MOOCs como estratégia para aumentar o envolvimento dos alunos e redução do abandono.

Declarações complementares

Financiamento

Esta pesquisa não foi financiada por nenhum órgão governamental e não-governamental.

Contribuições dos autores

Paulo contribuiu para a concepção deste estudo. Rubens e Ana Caroline realizaram a investigação, metodologia, escrita e revisão. Felipe e Ronaldo pela curadoria dos dados, administração de projetos e validação do estudo. Todos os autores leram e aprovaram o manuscrito final.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Disponibilidade de dados e materiais

Os conjuntos de dados (e/ou softwares) gerados e/ou analisados durante o estudo atual poderão ser disponibilizados mediante solicitação.

Referências

Alalawi, K., Athauda, R., and Chiong, R. (2023). Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review. *Engineering Reports*, 5. DOI: 10.1002/eng2.12699.

Balaji, P., Alelyani, S., Qahmash, A., and Mohana, M. (2021). Contributions of Machine Learning Models towards Student Academic Performance Prediction: A Systematic Review. *Applied Sciences*, 11:10007. DOI: 10.3390/app112110007.

de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., and Moreira, F. (2021). How Does Learning Analytics Contribute to Prevent Students' Dropout in Higher Education: A Systematic Literature Review. *Big Data and Cognitive Computing*, 5:64. DOI: 10.3390/bdcc5040064.

Gamage, D., Staubitz, T., and Whiting, M. (2021). Peer assessment in MOOCs: Systematic literature review. *Distance Education*, 42:268–289. DOI: 10.1080/01587919.2021.1911626.

Herrera, V. M., Khoshgoftaar, T. M., Villanustre, F., and Furht, B. (2019). Random forest implementation and optimization for Big Data analytics on LexisNexis's high performance computing cluster platform. *Journal of Big Data*, 6:1–36. DOI: 10.1186/s40537-019-0232-1.

Jin, C. (2021). Dropout prediction model in MOOC based on clickstream data and student sample weight. *Soft Computing*, 25:8971–8988. DOI: 10.1007/s00500-021-05795-1.

Mastour, H., Dehghani, T., Jajroudi, M., Moradi, E., Zarei, M., and Eslami, S. (2023). Prediction of medical sciences students' performance on high-stakes examinations using machine learning models: a protocol for a systematic review. *BMJ Open*, 13:e064956. DOI: 10.1136/bmjopen-2022-064956.

Ministério da Educação (2022). Ensino a distância cresce 474% em uma década. Disponível em: https://www.gov.br/mec/pt-br/assuntos/noticias/2022/ensino-a-distancia-cresce-474-em-uma-decada. Acesso em: 10 abr. 2025.

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6:e1000097. DOI: 10.1371/journal.pmed.1000097.

Moreno-Marcos, P. M., Alario-Hoyos, C., Munoz-Merino, P. J., and Kloos, C. D. (2019). Prediction in MOOCs: A Review and Future Research Directions. *IEEE Transactions on Learning Technologies*, 12:384–401. DOI: 10.1109/TLT.2018.2856808.

Mrhar, K., Benhiba, L., Bourekkache, S., and Abik, M. (2021). A Bayesian CNN-LSTM Model for Sentiment Analysis in Massive Open Online Courses MOOCs. *International Journal of Emerging Technologies in Learning*, 16:216–232. DOI: 10.3991/ijet.v16i23.24457.

Nazir, M., Noraziah, A., Rahmah, M., and Sharma, A. (2023). Examining the potential of machine learning for predicting academic achievement: A systematic review. *Fusion: Practice and Applications*, 13:71–90. DOI: 10.54216/FPA.130207.

Sato, S. N., Moreno, E. C., Rubio-Zarapuz, A., Dalamitros, A. A., Yañez-Sepulveda, R., Tornero-Aguilera, J. F., and Clemente-Suárez, V. J. (2024). Navigating the New Normal: Adapting Online and Distance Learning in the Post-Pandemic Era. *Education Sciences*, 14. DOI: 10.3390/educsci14010019.

Silva, J. J. D. and Roman, N. T. (2021). Predicting Dropout in Higher Education: a Systematic Review. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação (SBIE 2021)*, pages 1107–1117. Sociedade Brasileira de Computação - SBC. DOI: 10.5753/sbie.2021.217437.

Valverde-Berrocoso, J., del Carmen Garrido-Arroyo, M., Burgos-Videla, C., and Morales-Cevallos, M. B. (2020). Trends in Educational Research about e-Learning: A Systematic Literature Review (2009–2018). *Sustainability 2020, Vol. 12, Page 5153*, 12:5153. DOI: 10.3390/SU12125153.

Zhang, J., Gao, M., and Zhang, J. (2021). The learning behaviours of dropouts in MOOCs: A collective attention network perspective. *Computers Education*, 167:104189. DOI: 10.1016/J.COMPEDU.2021.104189.