



Predição da Evasão no Ensino Superior: Uma Revisão Sistemática sobre Algoritmos e Variáveis do Aprendizado de Máquina (2020-2025)

Amanda Moreira Rapold Souza¹, Andressa Urbano de Sousa Darzé¹,
Davi Beserra Caldas¹, Gabriel Antonio Torres Diogo¹,
Vitor Santos da Mota¹, Cleber Brito Santos¹

¹Escola de Tecnologias
Universidade Católica do Salvador (UCSAL)
Av. Prof. Pinto de Aguiar, 2589 Pituaçu, CEP: 41740-090
Salvador / BA, Brasil

{amandamoreira.souza, andressa.darze, davi.caldas

gabriel.diogo, vitorsantos.mota}@edu.ucsal.br

cleber.brito@pro.ucsal.br

Resumo. Este estudo apresenta uma revisão sistemática da literatura sobre técnicas de aprendizado de máquina aplicadas à predição da evasão estudantil no ensino superior, considerando publicações de 2020 a 2025. A revisão, conduzida segundo o protocolo de Kitchenham, resultou na identificação de nove estudos primários, dos quais apenas três disponibilizaram dados ou detalhamento metodológico suficiente para permitir reprodução experimental.

A síntese qualitativa indica a predominância de modelos supervisionados, especialmente árvores de decisão, florestas aleatórias, métodos de ensemble e redes neurais, apoiados por variáveis acadêmicas, socioeconômicas, demográficas e comportamentais.

A comparação entre resultados originais e reproduzidos evidenciou diferenças relevantes na reprodutibilidade, com maior estabilidade em modelos de ensemble e maior sensibilidade metodológica em redes neurais. Nesse contexto, persistem desafios significativos, como a escassez de dados abertos, a insuficiência de transparência metodológica e a heterogeneidade na seleção de métricas.

Os achados buscam identificar o estado da arte e apontam possíveis direções para avanços futuros, especialmente quanto à melhoria da documentação experimental e ao uso ampliado de variáveis comportamentais.

1. Introdução

A evasão de estudantes no ensino superior configura-se como um dos principais desafios da educação contemporânea, sendo um fenômeno complexo e multifatorial influenciado

por fatores acadêmicos, socioeconômicos e pessoais. Esse problema afeta diretamente a trajetória formativa dos alunos e a sustentabilidade das instituições de ensino superior (IES) [Moreno and Támara 2024]. Além disso, estatísticas educacionais evidenciam a gravidade e a persistência do fenômeno, motivando um interesse contínuo da comunidade científica em investigar suas causas e aprimorar modelos de predição [González-Morales et al. 2025].

Paralelamente, a crescente digitalização dos processos educacionais tem possibilitado o registro contínuo de dados acadêmicos, como interações em plataformas virtuais [Marcolino et al. 2025], avaliações e trajetórias discentes, constituindo uma fonte valiosa de informações sobre o comportamento estudantil e abrindo novas oportunidades para compreender padrões de aprendizagem e subsidiar decisões pedagógicas baseadas em evidências [Aljawarneh and Lara 2021]. A análise desses dados consolidou-se como um campo de pesquisa dedicado a aplicar métodos de ciência de dados na educação [Romero and Ventura 2020]. Tais análises permitem identificar fatores que influenciam o desempenho acadêmico, prever comportamentos de risco, como a evasão, e orientar políticas institucionais voltadas à melhoria da aprendizagem e da permanência discente. Nesse contexto, as técnicas de Aprendizado de Máquina (*Machine Learning, ML*) destacam-se pela capacidade de aprender padrões a partir dos dados e realizar predições de forma automática [Brown and et al. 2020]. Essa abordagem tornou-se central nos estudos recentes que buscam antecipar eventos de evasão no ensino superior [Niyogisubizo et al. 2022].

Apesar do crescimento das pesquisas sobre predição de evasão, persistem lacunas na sistematização das diferentes abordagens analíticas empregadas [Ngulube and Ncube 2025]. Os estudos existentes divergem quanto aos algoritmos de ML utilizados, aos conjuntos de variáveis considerados e às métricas de avaliação adotadas, o que dificulta comparações diretas e compromete a reprodutibilidade dos resultados [Glandorf et al. 2024]. Essa diversidade metodológica reforça a importância de uma revisão sistemática capaz de mapear o estado da arte e identificar padrões e lacunas no conhecimento atual sobre predição de evasão.

Em síntese, o objetivo geral deste trabalho é realizar uma revisão sistemática da literatura sobre a aplicação de técnicas de aprendizado de máquina na predição da evasão no ensino superior, consolidando o conhecimento produzido entre 2020 e 2025 acerca desse tema.

2. Fundamentação Teórica

A fundamentação teórica tem como finalidade apresentar os conceitos essenciais que sustentam a análise desenvolvida neste estudo. Inicialmente, discute-se o aprendizado de máquina, abordando seus princípios gerais e sua aplicação na construção de modelos capazes de identificar padrões e realizar predições a partir de dados. São destacados, em particular, o paradigma do aprendizado supervisionado e os principais instrumentos de avaliação de desempenho.

Além disso, a seção contempla os fundamentos metodológicos relacionados à revisão sistemática da literatura empregada no trabalho. Assim, a articulação entre os conceitos teóricos de aprendizado de máquina e os procedimentos da revisão sistemática fornecem o suporte necessário para a compreensão dos resultados e para a análise crítica das evidências apresentadas.

2.1. Aprendizado de Máquina

O aprendizado de máquina corresponde a um conjunto de métodos computacionais que permitem que modelos aprendam padrões a partir de dados, produzindo previsões ou decisões sem depender de regras explicitamente programadas. De acordo com Badillo et al. 2020, um algoritmo de ML ajusta seus parâmetros internos com base em exemplos observados, de modo a melhorar progressivamente seu desempenho em uma tarefa específica.

2.1.1. Aprendizado supervisionado

O aprendizado supervisionado é o paradigma predominante nas pesquisas que aplicam técnicas de ML à análise educacional [Romero and Ventura 2020]. Ele se baseia em bases históricas que já contêm exemplos rotulados, isto é, registros acompanhados do desfecho verdadeiro. No caso da predição de evasão, esses rótulos correspondem, em geral, a indicadores de permanência ou desligamento do estudante. O modelo aprende a relação entre os atributos observados e o resultado conhecido, ajustando seus parâmetros para, posteriormente, estimar o desfecho de novos casos de forma sistemática [Aljawarneh and Lara 2021].

Por essa razão, os estudos incluídos nesta revisão, assim como as reproduções experimentais realizadas, empregam majoritariamente técnicas de aprendizado supervisionado, formulando o problema como uma tarefa de classificação. O objetivo é aprender uma função capaz de generalizar esse mapeamento, ou seja, realizar previsões confiáveis para novos casos. Esse paradigma inclui problemas de classificação, quando o rótulo é categórico, e de regressão, quando o rótulo é numérico.

Classificadores supervisionados, como árvores de decisão (*Decision Tree, DT*), florestas aleatórias (*Random Forest, RF*), redes neurais artificiais (*Artificial Neural Networks, ANN*) ou métodos de conjunto (*Ensemble Methods*), são amplamente utilizados em áreas como saúde, finanças, educação e sistemas de recomendação [Gutierrez-Pachas et al. 2023].

2.1.2. Matriz de confusão

A avaliação de modelos de classificação costuma partir da chamada matriz de confusão, que organiza os resultados das previsões em quatro categorias [Badillo et al. 2020]:

- **TP (*True Positive*) – Verdadeiros Positivos:** previsão correta da classe positiva;
- **FP (*False Positive*) – Falsos Positivos:** previsão incorreta da classe positiva;
- **TN (*True Negative*) – Verdadeiros Negativos:** previsão correta da classe negativa;
- **FN (*False Negative*) – Falsos Negativos:** previsão incorreta da classe negativa.

A matriz de confusão possibilita derivar métricas específicas de desempenho que capturam diferentes aspectos do comportamento do modelo.

2.1.3. Métricas de desempenho

Acurácia (ACC)

Proporção total de acertos do modelo.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

É intuitiva, porém pode ser pouco informativa quando as classes estão desbalanceadas.

Precisão (P)

Indica a proporção de previsões positivas que estavam corretas.

$$P = \frac{TP}{TP + FP}$$

Útil quando o objetivo é reduzir falsos positivos.

Sensibilidade (S) ou Taxa de Verdadeiros Positivos (*True Positive Rate, TPR*)

Proporção de casos positivos que o modelo conseguiu identificar.

$$S = \frac{TP}{TP + FN}$$

Importante quando a falha em detectar a classe positiva tem alto custo.

Especificidade ou Taxa de Verdadeiros Negativos (*True Negative Rate, TNR*)

Proporção de casos negativos corretamente identificados.

$$TNR = \frac{TN}{TN + FP}$$

Complementa a sensibilidade ao avaliar o desempenho sobre a classe negativa.

F1-score

Média harmônica entre precisão e sensibilidade [Saito and Rehmsmeier 2015].

$$F1 = 2 \times \frac{P \times S}{P + S}$$

Útil quando há compensações relevantes entre falsos positivos e falsos negativos.

Área Sob a Curva (*Area Under The Curve, AUC*)

A AUC corresponde à área sob a curva ROC (*Receiver Operating Characteristic*) e fornece uma medida global da capacidade discriminativa de um modelo classificatório. Valores próximos de 1 indicam maior habilidade do modelo em diferenciar corretamente entre as classes, enquanto valores próximos de 0,5 sugerem desempenho equivalente ao acaso [Choi et al. 2020]. É uma métrica robusta, pois considera todos os limiares possíveis de decisão.

Síntese

De forma geral, o aprendizado supervisionado fornece ferramentas para estimar resultados a partir de conjuntos de dados observacionais, permitindo capturar padrões complexos e realizar previsões de maneira sistemática. A compreensão da matriz de confusão e de métricas como acurácia, precisão, sensibilidade, especificidade, *F1-score* e

AUC é essencial para interpretar adequadamente o desempenho dos modelos discutidos neste artigo, independentemente do contexto em que são aplicados.

2.2. Revisão Sistemática da Literatura

A revisão sistemática da literatura (RSL) é um método científico estruturado para identificar, analisar e sintetizar evidências sobre um tema de pesquisa. Diferentemente das revisões tradicionais, baseadas na seleção livre de fontes e na interpretação subjetiva do pesquisador, a RSL segue um protocolo definido que assegura rastreabilidade, reprodutibilidade e transparência em etapas como busca, triagem, análise e síntese [Dos Santos et al. 2020, Page and et al. 2021]. Essa abordagem confere à RSL o status de estudo secundário, pois foca na consolidação crítica e sistemática de estudos primários já existentes em vez de gerar novos dados empíricos [Cajal et al. 2020].

A RSL distingue-se de outras modalidades de revisão, como a revisão integrativa (mais ampla e descritiva) e a revisão narrativa (interpretativa, porém menos sistemática), ao trabalhar com perguntas de pesquisa explícitas, critérios de seleção replicáveis e síntese transparente das evidências [Brignardello-Petersen et al. 2025, Kitchenham 2004]. Em domínios como ciência de dados, essa sistematização é crucial para separar achados robustos de resultados circunstanciais, avaliando o conhecimento acumulado de forma estruturada e evitando interpretações isoladas. Desse modo, a RSL promove maior consistência teórica e metodológica entre estudos [Kitchenham et al. 2015]. Como resultado, a RSL aumenta a confiabilidade das conclusões científicas e subsidia decisões baseadas em evidências, tanto na pesquisa acadêmica quanto em práticas profissionais orientadas por dados [Cajal et al. 2020].

2.2.1. Princípios e Etapas do Processo Sistemático

Conduzir uma RSL envolve três macroetapas principais: planejamento, execução e relato dos resultados. Essas fases, propostas Kitchenham and Charters 2007, constituem a espinha dorsal da metodologia de revisão sistemática em computação:

- **Planejamento:** formulação das perguntas de pesquisa, definição do escopo, dos critérios de inclusão/exclusão e das estratégias de busca. Essa etapa assegura que a revisão responda a perguntas bem definidas e replicáveis.
- **Execução:** aplicação prática do protocolo. Envolve realizar buscas estruturadas, triagem dos estudos (leitura de títulos, resumos e textos completos) e extração dos dados relevantes. A rastreabilidade das decisões nessa fase garante a transparência do processo.
- **Relato dos resultados:** síntese das evidências encontradas, apresentando as principais tendências, lacunas e padrões identificados nos estudos analisados. Essa etapa exige clareza na comunicação científica e discussão crítica dos achados e limitações.

Essas três fases configuram mais do que procedimentos operacionais, pois expressam um compromisso epistemológico com a produção de conhecimento rigoroso e verificável. Cada decisão metodológica, desde a delimitação do conjunto de estudos examinados até a estratégia adotada para a síntese dos resultados, influencia de maneira direta

a validade interna da revisão, entendida como a consistência lógica e a precisão do processo analítico, assim como a validade externa, relativa ao grau em que as conclusões podem ser estendidas a outros contextos investigativos.

2.2.2. Controle de Viés e Validade das Evidências

Um fundamento central da RSL é a mitigação de vieses, entendidos como fatores que podem distorcer a interpretação das evidências [Kitchenham 2004]. O viés pode ocorrer em várias etapas – na formulação da pergunta, na seleção das fontes, na extração dos dados ou na interpretação dos resultados. Protocolos consolidados, como o de Kitchenham [Kitchenham and Charters 2007] e o PRISMA [Page and et al. 2021], orientam o registro de todas as decisões metodológicas e a justificativa para inclusão ou exclusão de estudos, garantindo transparência e rastreabilidade em todo o processo.

O controle rigoroso de vieses reforça a validade interna da revisão (coerência entre as etapas) e a validade externa (capacidade de replicar os resultados em outros contextos). Esse cuidado metodológico fortalece o rigor científico e a confiança nas conclusões da RSL, permitindo seu uso confiável em pesquisas futuras e aplicações práticas baseadas em evidências [Kitchenham et al. 2015].

2.2.3. Protocolo Kitchenham

A opção por adotar o protocolo de Kitchenham 2004 nesta pesquisa deve-se à sua ampla aceitação e consolidação como referência metodológica nas áreas de Computação e Engenharia de Software. Derivado de diretrizes da área médica e adaptado ao contexto computacional, o protocolo Kitchenham tornou-se um padrão reconhecido de rigor, transparência e reprodutibilidade em revisões sistemáticas [Kitchenham and Charters 2007].

Esse protocolo orienta de forma clara todas as etapas da revisão, desde a formulação das perguntas de pesquisa até a busca, seleção, extração e síntese dos estudos. Ele oferece critérios objetivos de inclusão e exclusão, bem como procedimentos sistemáticos para avaliar a qualidade dos estudos, assegurando um processo imparcial e replicável. Sua utilização minimiza vieses individuais do pesquisador, garante rastreabilidade nas decisões metodológicas e facilita a comparabilidade dos resultados com revisões anteriores, reforçando a credibilidade e a consistência dos achados.

Além disso, estudos apontam que o protocolo Kitchenham permanece entre os mais citados e replicados na área, servindo de base para outros métodos e boas práticas [Pereira da Silva 2009]. Sua adoção reflete, portanto, a busca por elevado rigor metodológico, transparência no processo de seleção e confiabilidade na síntese das evidências, mostrando-se a escolha mais adequada aos objetivos desta pesquisa. As etapas conduzidas segundo o protocolo Kitchenham forneceram, assim, a base metodológica para a análise e síntese dos resultados apresentados na seção seguinte.

3. Revisão Sistemática da Literatura

Essa seção apresenta a condução da revisão sistemática da literatura, seguindo o protocolo proposto por Kitchenham 2004. A estrutura metodológica abrange as etapas recomenda-

das para RSL: definição das questões de pesquisa, elaboração e aplicação da *string* de busca, estabelecimento de critérios de inclusão e exclusão, processo de triagem e seleção dos estudos, extração/organização de dados e justificativa metodológica. Essas etapas asseguram rastreabilidade, transparência e reprodutibilidade do procedimento, em conformidade com os fundamentos conceituais discutidos previamente. A seguir, cada etapa do planejamento e execução da revisão é detalhada.

3.1. Questões de Pesquisa

A definição das questões de pesquisa é o ponto de partida da RSL, pois orienta todas as decisões subsequentes, da estratégia de busca à análise dos resultados. Segundo Diego Dermeval e Ig Bittencourt e Leonardo Marques 2020, as perguntas devem estar diretamente alinhadas ao objetivo geral do estudo, garantindo foco e coerência metodológica ao processo.

No contexto deste trabalho, formularam-se quatro questões de pesquisa visando identificar, organizar e compreender as principais técnicas de aprendizado de máquina empregadas na predição da evasão de estudantes no ensino superior. Pretende-se, assim, mapear o estado da arte do tema, reunindo evidências sobre métodos, variáveis e métricas utilizadas, bem como apontando lacunas e desafios na literatura recente. As quatro perguntas definidas foram:

- **QP1:** Quais técnicas analíticas e modelos de aprendizado de máquina têm sido empregados para prever a evasão de estudantes em instituições de ensino superior?
- **QP2:** Quais tipos de variáveis são mais utilizadas nos estudos de predição de evasão discente?
- **QP3:** Quais métricas e estratégias de avaliação são adotadas para medir o desempenho dos modelos preditivos de evasão?
- **QP4:** Quais desafios e lacunas metodológicas persistem nas pesquisas recentes sobre evasão estudantil baseadas em análise de dados?

Essas questões foram elaboradas com base em estudos prévios de predição de evasão no ensino superior, de forma a sintetizar o conhecimento existente e compreender as principais abordagens e tendências atuais. A revisão sistemática, portanto, busca oferecer uma visão abrangente e crítica do tema, servindo como base teórica para trabalhos futuros, como pesquisas de pós-graduação ou mestrado focadas em aprimorar estratégias de monitoramento e prevenção da evasão discente.

3.2. String de Busca

A *string* de busca é fundamental para determinar a abrangência, precisão e relevância dos estudos recuperados na revisão sistemática. Como destacam Kitchenham and Charters 2007, uma *string* de busca é uma estrutura lógica projetada para interagir com bases de dados acadêmicas, responsável por retornar estudos potencialmente relevantes ao tema. A seleção criteriosa e refinamento dos termos de busca impactam diretamente a qualidade e completude da revisão, garantindo rigor metodológico e reprodutibilidade.

Para construir uma *string* de busca robusta e alinhada aos objetivos do estudo, utilizaram-se operadores lógicos *AND* e *OR* de forma adequada. Conforme Rosen 2019, o operador *AND* restringe os resultados aos documentos que contenham todos os termos

especificados simultaneamente, enquanto o operador *OR* amplia o escopo ao recuperar documentos que contenham ao menos um dos termos associados. O uso equilibrado desses operadores evita tanto perder estudos relevantes, quanto excesso de resultados irrelevantes, alcançando um meio-termo entre abrangência e especificidade.

Segundo Diego Dermeval e Ig Bittencourt e Leonardo Marques 2020, a construção da *string* envolve agrupar termos semelhantes em blocos conectados internamente pelo operador *OR*, e então associar esses blocos entre si pelo operador *AND*. Essa estrutura garante alinhamento aos temas das questões de pesquisa e ao escopo da revisão, tornando a busca ampla o suficiente para captar estudos relevantes, porém precisa o bastante para excluir trabalhos fora do tema.

Em conformidade com Kitchenham and Charters 2007, também adotou-se a abordagem *PICO* – acrônimo para População (*Population, P*), Intervenção (*Intervention, I*), Comparação (*Comparison, C*), Resultado (*Outcome, O*) – adaptada ao contexto da evasão estudantil para estruturar a *string* de busca. A partir dessa estrutura, foram definidos quatro blocos semânticos:

- **População:** termos referentes às instituições de ensino superior — "*higher education*" *OR* "*university*" *OR* "*college*"
- **Intervenção:** termos associados a técnicas de aprendizado de máquina e modelagem preditiva — "*machine learning*" *OR* "*predictive model*" *OR* "*predictive modelling*"
- **Comparação:** termos relativos à evasão e à permanência discente — "*student dropout*" *OR* "*student retention*"
- **Resultado:** termos que delimitam o foco em estudantes de graduação — "*undergraduate student*" *OR* "*bachelor student*" *OR* "*college student*"

Esses blocos foram interligados pelo operador *AND*, assegurando que cada estudo recuperado abordasse simultaneamente educação superior, preditivas de aprendizado de máquina e evasão discente. A *string* de busca foi refinada iterativamente, testando diferentes combinações de termos até maximizar a recuperação de estudos relevantes e minimizar resultados irrelevantes, em consonância com as recomendações de Kitchenham and Charters 2007, resultando na seguinte formulação final:

String final: ("*higher education*" *OR* "*university*" *OR* "*college*") *AND* ("*machine learning*" *OR* "*predictive model*" *OR* "*predictive modelling*") *AND* ("*student dropout*" *OR* "*student retention*") *AND* ("*undergraduate student*" *OR* "*bachelor student*" *OR* "*college student*")

3.3. Critérios de Inclusão e Exclusão

Definidos os termos de busca e obtido um conjunto inicial de estudos, estabeleceram-se critérios de inclusão e exclusão para filtrar os documentos. Essa etapa é crucial para garantir consistência, transparência e qualidade metodológica na seleção dos estudos. Os critérios foram estabelecidos já na fase de planejamento da revisão, assegurando, dessa forma, que apenas trabalhos relevantes e metodologicamente adequados integrassem a análise. Kitchenham and Charters 2007 sugerem considerar fatores como idioma, veículo de publicação, autores, contexto do estudo, participantes, desenho metodológico, método de amostragem e data de publicação ao formular os filtros de seleção. Ao adaptar essas

diretrizes ao escopo desta pesquisa, foram definidos os seguintes critérios:

Cr terios de Inclus o (CI):

- **CI1:** Publica es entre 2020 e 2025 para um recorte temporal recente, assim, garantindo atualidade da revis o.
- **CI2:** Estudos focados em modelos de aprendizado de m quina aplicados   preven o da evas o no ensino superior, mantendo a coer ncia com o tema central.
- **CI3:** Trabalhos que apresentam resultados sobre evas o, incluindo descri o de m tricas de desempenho dos modelos para evidenciar as aplica es pr ticas.
- **CI4:** Estudos prim rios que apresentem dados originais produzidos pelos autores, permitindo avaliar evid ncias emp ricas diretas.
- **CI5:** Disponibilidade do texto completo do estudo, permitindo an lise integral de metodologia, resultados e limita es.

Cr terios de Exclus o (CE):

- **CE1:** Optou-se por remover artigos n o redigidos em ingl s, idioma predominante das publica es de maior impacto em ci ncia de dados e ML. A inclus o de outras l nguas exigiria ampliar o escopo para m ltiplos idiomas, aumentando a complexidade e comprometendo a consist ncia metodol gica.
- **CE2:** Documentos duplicados ou trabalhos muito similares. No caso de estudos similares do mesmo autor, reteve-se apenas a vers o mais recente para evitar contagem dupla de evid ncias.
- **CE3:** Trabalhos que n o s o estudos prim rios, por exemplo, revis es sistem ticas, mapeamentos, livros, disserta es ou teses, com o intuito de prevenir duplicidade de informa es e manter comparabilidade dos resultados.
- **CE4:** Estudos publicados em ve culos de baixa credibilidade ou fora da  rea, priorizando artigos revisados por pares em peri dicos/confer ncias reconhecidos, a fim de garantir rigor cient fico e relev ncia na  rea de engenharia de *software*/educa o.
- **CE5:** Estudos inacess veis ou sem texto completo dispon vel, pois sem o conte do integral n o   poss vel avaliar metodologias e resultados de forma adequada, comprometendo a validade da revis o.

3.4. Sele o dos Estudos

O processo de sele o dos estudos seguiu as diretrizes de Kitchenham 2004 e Kitchenham and Charters 2007, privilegiando a transpar ncia, rastreabilidade e reprodutibilidade. Ap s definir a *string* de busca e os crit rios, procedeu-se   busca autom tica na base Google Acad mico (executada em 20/09/2025), obtendo-se um total inicial de 5.770 registros. As refer ncias foram exportadas e organizadas com aux lio de um gerenciador bibliogr fico, o Mendeley. A remo o de duplicatas foi realizada manualmente (via planilhas e cruzamento de metadados), pois n o foi poss vel automatiz la completamente. Em seguida, a sele o passou por cinco filtros sucessivos:

- **Filtro 1 – Ano e idioma:** Aplicou-se o recorte temporal (2020–2025) e lingu stico (ingl s) definido pela revis o. Resultaram 3.080 documentos ap s excluir trabalhos fora desse intervalo e idioma.

- **Filtro 2 – Duplicatas:** Removeram-se registros duplicados ou múltiplos resultados do mesmo estudo, reduzindo o total para 2.772 documentos únicos.
- **Filtro 3 – Título e palavras-chave:** Realizou-se uma triagem inicial pelos títulos e palavras-chave para excluir publicações claramente fora do escopo; por exemplo, estudos sobre outros níveis de ensino ou sem foco em modelagem preditiva de evasão. Após esse filtro temático, restaram 223 artigos potencialmente relevantes.
- **Filtro 4 – Resumo e conclusões:** Procedeu-se à leitura dos resumos e das conclusões de cada um dos 223 artigos, verificando pertinência temática, tipo de metodologia e alinhamento com as questões de pesquisa. Essa análise mais refinada levou à exclusão de 171 trabalhos, mantendo 52 estudos candidatos.
- **Filtro 5 – Leitura completa:** Na etapa final, os 52 estudos restantes foram lidos na íntegra. Nessa fase, avaliou-se a qualidade metodológica e a credibilidade de cada publicação por meio do sistema de pontuação descrito na Seção 3.4.1. Após a aplicação desse processo de avaliação, 9 estudos primários atenderam aos requisitos e foram selecionados para compor a amostra final da revisão sistemática.

Todas as etapas de identificação, triagem e elegibilidade foram rigorosamente documentadas em planilha eletrônica, registrando o número de artigos identificados, excluídos e selecionados em cada fase, conforme recomendado por Kitchenham 2004. Esse registro permite rastrear cada decisão de inclusão/exclusão e reforça a transparência do processo. Os resultados do processo de seleção estão sumarizados na Tabela 1 que lista os 9 artigos incluídos na amostra final juntamente com suas principais informações.

Com a aplicação bem-sucedida do protocolo de revisão sistemática e a definição do conjunto final de estudos, estabeleceu-se a base metodológica necessária para a síntese dos resultados, apresentada nas seções seguintes.

Tabela 1: Síntese dos estudos selecionados.

ID	Título original	Autor(es)	Fonte	Ano	Objetivo
A01	<i>Predictive analytics study to determine undergraduate students at risk of dropout</i>	Gonzalez-Nucamendi A.; Noguez J.; Neri L.; Robledo-Rella V.; García-Castelán	Frontiers	2023	Detectar estudantes de graduação em risco de evasão usando variáveis quantitativas.
A02	<i>Predicting Student Dropout Through Text and Media Content Analysis of VKontakte Profiles</i>	Sergei S. Goncharov; Dmitry I. Ignatov; Anastasia Yu. Chernyshova	IEEE Xplore	2025	Prever evasão a partir da análise de conteúdo de redes sociais.
A03	<i>Deep Learning Model to Predict Students Retention Using BLSTM and CRF</i>	D. Uliyan; A. S. Aljaloud; A. Alkhalil; H. S. A. Amer; M. A. E. A. Mohamed; A. F. M. Alogali	IEEE Xplore	2021	Prever retenção usando deep learning (BLSTM + CRF).
A04	<i>Predicting Student Dropout Using Machine Learning Algorithms</i>	Suleyman Alpasan Sulak; Isik Nigmet Koçlu	Imines	2024	Aplicar algoritmos de Machine Learning para prever evasão escolar.
A05	<i>Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors through Machine Learning and Survival Analysis Methods in the Latin American Context</i>	Daniel A. Gutierrez-Pachas; Germain Garcia-Zambia; Ernesto Cuadra-Vargas; Guillermo Camara-Chavez; Erick Gomez-Nieto	MDPI	2023	Suportar decisões sobre evasão analisando fatores sociais e acadêmicos.
A06	<i>Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics</i>	Sandra C. Metz; Christina S. Bukow; Heinrich Peters; Christine Deacon; Alice Dinu; Clemens Stachl	Nature	2022	Prever retenção estudantil usando engajamento em apps e dados sociodemográficos.
A07	<i>Statistical and machine learning models for predicting university dropout and scholarship impact</i>	Stephan Romero; Xinyue Liao	Plos One	2023	Predizer evasão e impacto de bolsas usando ML e métodos estatísticos.

Continua na próxima página

ID	Título original	Autor(es)	Fonte	Ano	Objetivo
A08	<i>Decision trees for predicting dropout in Engineering Course students in Brazil</i>	Mariano A.; Ferreira A.; Santos M.	Elsevier	2022	Classificar estudantes de engenharia quanto ao risco de evasão.
A09	<i>Early Risk Detection of Bachelor's Student Withdrawal or Long-Term Retention</i>	Caicedo-Castro I. et al.	IARIA Congress	2022	Detectar precocemente o risco de deserção.

3.4.1. Avaliação dos Estudos

Nesta subseção detalha-se como os estudos primários foram avaliados para a revisão sistemática. Seguindo recomendações metodológicas clássicas, utilizou-se um *checklist* estruturado com 6 critérios de qualidade (QA1–QA6). Cada estudo recebeu de 0 a 1 ponto por critério; notas mais altas indicam maior rigor metodológico e clareza. Foi estabelecido o critério de inclusão: pontuação total $\geq 4,0$. A seguir, apresentam-se os critérios adotados (Tabela 2) e a forma de pontuação.

Tabela 2. Critérios de avaliação de qualidade (QA) utilizados na revisão

Fonte: Elaborado pelos autores (2025).

Código	Descrição do Critério
QA1	Clareza dos objetivos: o estudo apresenta objetivos ou hipóteses claramente definidos e alinhados ao tema da revisão?
QA2	Detalhamento metodológico e replicabilidade: a metodologia (pré-processamento, particionamento, configuração dos modelos) é descrita com detalhe suficiente para permitir replicação?
QA3	Qualidade e descrição das fontes de dados: a origem, estrutura, tamanho e eventuais tratamentos aplicados aos dados são apresentados de maneira clara e completa?
QA4	Avaliação e métricas: o estudo utiliza métricas adequadas (ex.: Acurácia, F1, AUC) e descreve o protocolo de avaliação empregado (<i>cross-validation</i> , <i>holdout</i> , etc.)?
QA5	Discussão de limitações e vieses: o estudo explicita suas limitações metodológicas, potenciais vieses ou ameaças à validade?
QA6	Relevância e contribuição: o estudo contribui de forma direta para responder às questões de pesquisa desta revisão (QP1–QP4)?

Pontuação Final e Seleção dos Artigos

Cada estudo selecionado pelo filtro 5 foi avaliado segundo os critérios QA1–QA6. Entre os 52 estudos analisados, as pontuações totais variaram de 0,0 a 6,0. Para a seleção final, foram considerados apenas os estudos com pontuação maior ou igual a 4,0, resultando em 9 trabalhos.

Esse ponto de corte foi definido buscando um equilíbrio entre rigor acadêmico e as características pragmáticas da literatura sobre evasão estudantil, garantindo a inclusão de estudos com clareza de objetivos e apresentação adequada dos resultados [Kitchenham and Charters 2007]. Ao mesmo tempo, evitou-se um critério excessivamente restritivo, de

modo a não excluir pesquisas potencialmente relevantes. Ressalta-se que esse limiar pode ser ajustado em revisões futuras, conforme a necessidade de maior rigor ou flexibilidade.

Com isso, a pontuação dos estudos, de acordo com os critérios de qualidade, pode ser observada na Tabela 3.

Tabela 3. Pontuação dos estudos selecionados segundo os critérios de qualidade

Fonte: Elaborado pelos autores (2025).

Artigo	QA1	QA2	QA3	QA4	QA5	QA6	Total
A01	1.0	1.0	1.0	1.0	0.5	1.0	5.5
A02	1.0	1.0	1.0	1.0	1.0	1.0	6.0
A03	1.0	1.0	1.0	1.0	0.5	1.0	5.5
A04	1.0	1.0	1.0	1.0	0.5	1.0	5.5
A05	1.0	0.5	0.5	0.5	0.5	1.0	4.0
A06	1.0	0.5	0.5	0.5	0.5	1.0	4.0
A07	1.0	1.0	1.0	1.0	1.0	1.0	6.0
A08	1.0	0.5	1.0	0.5	0.5	1.0	4.5
A09	1.0	1.0	1.0	1.0	0.5	1.0	5.5

Legenda: 1 = Satisfaz totalmente; 0.5 = Parcialmente; 0 = Não satisfaz.

3.5. Síntese Qualitativa e Discussão Sobre as Questões de Pesquisa

1. **QP1:** Quais técnicas analíticas e modelos têm sido empregados para prever a evasão de estudantes em instituições de ensino superior?

De modo geral, os estudos analisados indicam que a predição da evasão estudantil no ensino superior tem se apoiado predominantemente em métodos de aprendizado supervisionado, com destaque para modelos baseados em árvores de decisão e suas variações, como florestas aleatórias e técnicas de impulsionamento de gradiente, amplamente exploradas em pesquisas como A01, A02 e A04. Além desses, observa-se o uso frequente de regressão logística, máquinas de vetor de suporte e redes neurais artificiais, empregadas tanto como modelos principais quanto como referências comparativas em experimentos de predição, como ilustrado em A03 e A07. Em menor proporção, alguns estudos incorporam métodos de regularização voltados à seleção de variáveis, bem como modelos de análise de sobrevivência utilizados para estimar o risco e o tempo até a evasão, abordagem presente sobretudo em A05. No conjunto, evidencia-se uma tendência de consolidação dos modelos supervisionados e das abordagens de conjunto como principais estratégias analíticas na literatura recente sobre predição da evasão estudantil.

2. **QP2:** Quais tipos de variáveis são mais utilizadas nos estudos sobre predição de evasão?

Os estudos analisados evidenciam que as variáveis acadêmicas constituem o grupo mais recorrente nos modelos de predição da evasão, englobando indicadores de desempenho em disciplinas, notas em exames de admissão, coeficiente

de rendimento, histórico de reprovações e tempo de permanência no curso, conforme observado em pesquisas como A01, A03 e A04. Em segundo lugar, aparecem variáveis socioeconômicas, como situação financeira, concessão de bolsas e financiamentos estudantis e escolaridade ou ocupação dos pais, que refletem o impacto das condições de suporte econômico sobre a continuidade dos estudos, elementos presentes em A04 e A07. As variáveis demográficas, como idade, sexo e estado civil, são amplamente utilizadas como fatores de controle ou de complementação às dimensões anteriores. Observam-se ainda variáveis institucionais e comportamentais, relacionadas à forma de ingresso, ao turno de estudo, ao nível de engajamento acadêmico e às interações estabelecidas pelo estudante no ambiente universitário, dimensões evidenciadas em trabalhos como A02 e A06. Em conjunto, os resultados indicam que os modelos de predição da evasão tendem a combinar indicadores de desempenho e características individuais com informações institucionais e econômicas, buscando representar de maneira abrangente os fatores que influenciam a permanência estudantil.

3. **QP3:** Quais métricas e estratégias de avaliação são adotadas para medir o desempenho dos modelos de predição?

Os estudos analisados evidenciam o uso predominante de métricas derivadas da matriz de confusão como principal referência de avaliação, com destaque para acurácia, precisão, sensibilidade e pontuação F1, que mensuram a capacidade dos modelos em distinguir corretamente estudantes evadidos e não evadidos, prática comum em trabalhos como A01 e A04. Os modelos são treinados e testados a partir da divisão dos dados em conjuntos de treinamento e teste, procedimento que permite avaliar o desempenho em observações não utilizadas na etapa de ajuste. Observa-se também a adoção abrangente de técnicas de validação cruzada, especialmente o método *K-Fold*, aplicado com diferentes números de partições em estudos como A05, A07 e A09, com o propósito de reduzir o viés e aumentar a confiabilidade e a generalização dos resultados. Em síntese, prevalecem métricas clássicas de classificação e procedimentos de validação cruzada, o que reforça a busca por rigor e consistência metodológica nas pesquisas sobre predição da evasão estudantil.

4. **QP4:** Quais desafios e lacunas metodológicas ainda persistem nas pesquisas recentes sobre evasão estudantil baseadas em análise de dados?

Os estudos analisados revelam que, embora haja avanços significativos na aplicação de técnicas analíticas à predição da evasão estudantil, persistem desafios e limitações metodológicas que comprometem a generalização e a reprodutibilidade dos resultados. Entre as principais dificuldades destacam-se a escassez e a fragmentação dos dados disponíveis, evidenciadas em estudos como A02, que apontam limitações na obtenção de informações completas sobre os estudantes. A limitação amostral constitui outro obstáculo recorrente, uma vez que diversas pesquisas utilizam bases reduzidas ou restritas a uma única instituição,

como observado em A08 e A09, o que dificulta a extrapolação dos resultados para outros contextos. Além disso, nota-se a presença de viés institucional e regional, especialmente em pesquisas baseadas em dados provenientes de uma única fonte ou de um recorte geográfico restrito, como ocorre em A06 e A07. Outra lacuna importante diz respeito à ausência de variáveis não observadas, como aspectos relacionados à motivação, saúde mental e suporte institucional, apontada de forma explícita em A07. Também é relatada a dificuldade de padronização de variáveis e critérios de coleta, o que limita a comparabilidade entre estudos, conforme mencionado em A01. Em conjunto, essas limitações apontam para a necessidade de bases de dados mais amplas e integradas, bem como de modelos mais generalizáveis e transparentes, capazes de incorporar múltiplas dimensões do fenômeno da evasão e de reduzir o impacto de vieses contextuais e amostrais.

3.6. Considerações Finais

A revisão sistemática desenvolvida estruturou todas as etapas necessárias para mapear o estado da arte sobre técnicas de aprendizado de máquina aplicadas à predição de evasão no ensino superior. A síntese qualitativa das evidências mostrou que os modelos supervisionados, especialmente árvores de decisão, florestas aleatórias, métodos de conjunto e técnicas de redes neurais artificiais, constituem o núcleo das abordagens contemporâneas. Tais modelos são sustentados predominantemente por variáveis acadêmicas, indicadores de desempenho, fatores socioeconômicos e características institucionais.

Os achados evidenciam não apenas o predomínio de determinadas técnicas, mas também lacunas metodológicas persistentes, como limitações amostrais, dificuldades na padronização de variáveis e insuficiência de dados comportamentais. Esses aspectos apontam oportunidades concretas para o avanço da área, especialmente por meio da incorporação de fontes de dados mais diversificadas, do desenvolvimento de modelos híbridos e da ampliação da transparência metodológica.

4. Síntese Quantitativa e Discussão Comparativa dos Resultados

Esta seção apresenta a síntese quantitativa dos estudos selecionados na revisão sistemática e faz uma comparação entre os resultados originais e aqueles obtidos por meio da nossa reprodução experimental. Embora nove estudos tenham sido identificados como relevantes no período analisado, apenas três disponibilizaram informações metodológicas suficientemente detalhadas ou bases de dados acessíveis para permitir uma replicação rigorosa. Esses três trabalhos foram, portanto, escolhidos como base para a análise quantitativa.

A partir dessa seleção, a seção examina o contexto de cada estudo, confronta as métricas de desempenho reportadas pelos autores com os resultados reproduzidos e discute padrões observados, divergências e limitações decorrentes tanto das características das bases de dados quanto do nível de transparência metodológica dos estudos originais.

4.1. Contexto dos estudos selecionados

Os trabalhos selecionados adotam configurações distintas de variáveis e algoritmos, o que possibilita uma comparação consistente entre abordagens preditivas. Apresentam-se, a seguir, os elementos necessários para contextualizar cada estudo antes da análise quantitativa.

- **Estudo A** – *Predicting Student Dropout Using Machine Learning Algorithms* [SULAK and KOKLU 2024]
Este estudo utilizou um conjunto de dados público contendo informações de 4.424 estudantes de uma instituição de ensino superior portuguesa. A base contempla 37 variáveis acadêmicas e demográficas. O problema é modelado como uma tarefa de classificação multiclasse, com três categorias de saída: evadido, matriculado e graduado. Foram testados três algoritmos tradicionais de aprendizado supervisionado: DT, RF e ANN de camada única.
- **Estudo B** – *Statistical and Machine Learning Models for Predicting University Dropout and Scholarship Impact* [Romero and Liao 2025]
Também baseado na mesma base pública do Estudo A, com 4.424 registros, este estudo aborda a evasão como um problema de classificação binária (evasão versus permanência). Foram comparados cinco modelos preditivos: regressão penalizada via Lasso, RF, *XGBoost*, Modelo Aditivo Generalizado (*Generalized Additive Model, GAM*) e uma ANN de uma camada.
- **Estudo C** – *Using Machine Learning to Predict Student Retention from Socio-Demographic Characteristics and App-Based Engagement Metrics* [Matz et al. 2023]
O terceiro estudo analisou a retenção de estudantes após o primeiro semestre em quatro instituições de ensino superior norte-americanas, totalizando aproximadamente 50.095 registros. Foram consideradas tanto variáveis institucionais, como desempenho acadêmico e perfil socioeconômico, quanto métricas de engajamento extraídas de um aplicativo estudantil, incluindo participação em eventos e interações em rede social. Os autores compararam o desempenho dos algoritmos *Elastic Net (EN)* e RF em três diferentes conjuntos de atributos: todas as variáveis combinadas, apenas variáveis de engajamento e apenas variáveis institucionais. Os modelos foram avaliados por meio de validação cruzada aninhada, com registro de métricas como AUC, TPR, TNR e tempo de treinamento.

4.2. Resultados da reprodução e comparação com os estudos originais

A reprodução experimental foi conduzida utilizando os mesmos conjuntos de dados empregados nos estudos, de modo a assegurar condições comparáveis para a avaliação dos modelos. Para cada trabalho selecionado, foram implementados os algoritmos descritos pelos autores, respeitando o máximo possível as configurações relatadas e aplicando procedimentos padronizados de validação para obtenção das métricas de desempenho. As tabelas apresentadas nas subseções seguintes sintetizam os valores originais e os resultados reproduzidos, constituindo a base para a análise quantitativa e comparativa dos três estudos. Além disso, a Tabela 4.2 mostra a relação dos pacotes utilizados e suas versões na reprodução dos modelos implementados nos artigos selecionados.

Tabela 4. Relação das dependências e suas versões.

Fonte: Elaborado pelos autores (2025).

Artigo 1	Artigo 2	Artigo 3
Python 3.11.5	Python 3.11.5	R 3.0.0
Pandas 2.0.3	Pandas 2.0.3	Tidyverse 2.0.0
Numpy 1.24.4	Numpy 1.24.4	Mlr 2.19.3
Scikit-Learn 1.3.2	Scikit-Learn 1.3.2	MlrCPO 0.3.8
	Matplotlib 3.7.2	ParallelMap 1.5.1
	Seaborn 0.12.2	Skimr 2.2.1
	XGBoost 1.7.6	GGplot2 4.0.0
	Pygam 0.10.1	Psych 2.5.6

Estudo A

O estudo A foi reproduzido utilizando o conjunto de dados público disponibilizado pela *UCI Machine Learning Repository*. A Tabela 5 apresenta a comparação entre as me-

Tabela 5. Estudo A - Comparação de Métricas: Original vs. Reprodução

Fonte: Elaborado pelos autores (2025), com base em SULAK and KOKLU 2024.

Modelo	Original				Reprodução			
	ACC (%)	P (%)	S (%)	F1 (%)	ACC (%)	P (%)	S (%)	F1 (%)
DT	70,1	70,0	70,1	70,0	67,8	61,4	61,8	61,5
RF	75,5	73,9	75,5	74,2	71,8	64,6	57,1	52,1
ANN	77,3	76,0	77,3	76,2	56,8	47,2	41,5	37,0

tricas originais e os valores obtidos na reprodução para os três modelos analisados: DT, RF e ANN. A comparação entre os resultados originais e reproduzidos revela uma perda de desempenho em todos os modelos avaliados, ainda que em magnitudes distintas. DT apresenta a menor discrepância entre os dois cenários, indicando que sua reprodução foi relativamente estável e pouco sensível a variações metodológicas. RF, embora mantenha desempenho superior ao da DT no estudo original, demonstra uma queda mais acentuada na reprodução, especialmente nas métricas associadas à identificação da classe de interesse, sugerindo influência de diferenças no pré-processamento ou na configuração do modelo. Já ANN apresenta a maior divergência entre os resultados originais e reproduzidos, indicando forte dependência de detalhes da arquitetura e dos hiperparâmetros, cuja ausência no estudo original dificulta substancialmente a replicação fiel.

De forma geral, observa-se que modelos mais complexos exibem maior sensibilidade a pequenas variações no processo de reprodução, reforçando a necessidade de documentação metodológica completa e precisa para assegurar reprodutibilidade em estudos de aprendizado de máquina.

Estudo B

O Estudo B utilizou o mesmo conjunto de dados público empregado no Estudo A, disponibilizado pela *UCI Machine Learning Repository*, o que permitiu aplicar, em

condições controladas, os mesmos modelos descritos pelos autores. A Tabela 6 apresenta a comparação entre as métricas originais e aquelas obtidas na reprodução, abrangendo os modelos Lasso, RF, *XGBoost* e uma ANN de camada única. O modelo GAM, embora incluído na publicação original, permaneceu apenas como referência, em razão de limitações computacionais que inviabilizaram sua implementação nesta etapa. A seguir, analisam-se a aderência entre os resultados e as variações observadas entre os modelos avaliados.

Tabela 6. Estudo B - Comparação de Métricas: Original vs. Reprodução

Fonte: Elaborado pelos autores (2025), com base em Romero and Liao 2025.

Modelo	Original					Reprodução				
	ACC	AUC	P	S	F1	ACC	AUC	P	S	F1
Lasso	0,875	0,919	0,873	0,798	0,834	0,934	0,970	0,930	0,898	0,914
RF	0,879	0,935	0,939	0,872	0,904	0,919	0,968	0,931	0,856	0,892
XGB	0,882	0,924	0,947	0,871	0,907	0,916	0,967	0,911	0,870	0,890
GAM	0,868	0,919	0,930	0,863	0,895	–	–	–	–	–
ANN	0,862	0,912	0,932	0,854	0,891	0,930	0,965	0,927	0,891	0,908

O conjunto de dados do Estudo B apresenta desbalanceamento moderado entre as classes, com aproximadamente 32% de evasão e 68% de permanência. Diante dessa característica, a avaliação do desempenho dos modelos buscou evitar interpretações enviesadas das métricas. Em cenários de classes desbalanceadas, a acurácia isolada pode se mostrar artificialmente elevada, não refletindo adequadamente a capacidade de discriminação da classe minoritária. Assim, a análise privilegiou métricas mais informativas, como o *F1-score* e a Área Sob a Curva da curva ROC (AUC), que fornecem uma avaliação mais robusta do desempenho preditivo.

A comparação entre os resultados originais e aqueles obtidos na reprodução evidencia elevada consistência entre as métricas reportadas. De modo geral, os valores reproduzidos apresentaram acurácia, AUC e F1 ligeiramente superiores aos do estudo original, com destaque para os modelos Lasso e ANN, cujos incrementos de AUC ultrapassam 0,05. Essas diferenças podem ser atribuídas a variações no particionamento dos dados e à adoção de procedimentos mais abrangentes de ajuste de hiperparâmetros, aliados à validação cruzada estratificada.

Os modelos baseados em árvores, como *Random Forest* e *XGBoost*, mantiveram desempenho elevado e próximo aos resultados originais, reafirmando sua robustez e estabilidade mesmo em contextos de desbalanceamento. Esse comportamento é consistente com o uso de técnicas de ensemble e regularização, além da presença de variáveis preditivas fortemente associadas ao risco de evasão, como indicadores acadêmicos iniciais e características socioeconômicas.

Em síntese, a proximidade entre os valores obtidos confirma a reprodutibilidade metodológica do Estudo B e sustenta suas conclusões quanto ao bom desempenho de modelos de ensemble e técnicas regularizadas na predição de evasão estudantil, mesmo diante de classes desbalanceadas.

Estudo C

No Estudo C, os autores disponibilizaram tanto o conjunto de dados quanto o código utilizado na implementação dos modelos, o que permitiu uma reprodução metodologicamente mais fiel das etapas originais. Os modelos foram avaliados em diferentes cenários definidos pela combinação de variáveis institucionais e métricas de engajamento estudantil, possibilitando examinar o impacto incremental dessas informações no desempenho preditivo. A Tabela 7 apresenta a comparação entre as métricas originais e os valores reproduzidos para os modelos *RF* e *EN* nos dois cenários considerados.

Tabela 7. Estudo C - Comparação de Métricas: Original vs. Reprodução

Fonte: Elaborado pelos autores (2025), com base em Matz et al. 2023.

Modelo	Cenário	Original				Reprodução			
		F1	AUC	TPR	TNR	F1	AUC	TPR	TNR
RF	1	0.61	0,84	0.59	0.90	0,60	0,81	0,58	0,89
EN	1	0.44	0.65	0.71	0.51	0,47	0,65	0,75	0,52
RF	2	0.50	0.71	0.66	0.66	0,47	0,67	0,67	0,60
EN	2	0.46	0.73	0.58	0.71	0,46	0,71	0,62	0,69

Cenário 1 - Variáveis de engajamento; Cenário 2 - Variáveis institucionais;

Nota: Resultados referentes a universidade 1.

A comparação entre os resultados originais e reproduzidos do Estudo C indica que os desempenhos obtidos seguem o mesmo comportamento geral descrito pelos autores. Em ambos os cenários, o modelo *RF* manteve desempenho superior ao *EN*, sobretudo na métrica F1, refletindo maior capacidade de identificar corretamente estudantes retidos e não retidos. A reprodução apresentou pequenas variações em relação aos valores originais, com reduções discretas de AUC e F1 no *RF* e ligeiros aumentos de sensibilidade no *EN*, possivelmente decorrentes de diferenças no ajuste interno dos modelos e no processo de validação. Apesar dessas oscilações, o padrão comparativo entre os modelos permaneceu consistente: o desempenho melhora quando variáveis de engajamento são incorporadas, e o *RF* continua a apresentar resultados mais robustos. Esses achados reforçam a conclusão do estudo de que indicadores comportamentais contribuem significativamente para a melhoria das previsões de retenção estudantil.

4.3. Limitações da síntese quantitativa

Esta síntese possui limitações que afetam a generalização dos resultados:

- **Reprodução parcial** – Apenas um dos estudos (C) disponibilizou dados e código completos. Nos outros casos, a reprodução baseou-se em descrições textuais, o que pode explicar divergências de desempenho.
- **Limitação computacional na reprodução experimental** - Durante o processo de reprodução dos modelos preditivos, enfrentamos restrições de poder computacional que limitaram a execução completa de algumas etapas experimentais. Especificamente no Estudo C, foi necessário reduzir o número de combinações de hiperparâmetros avaliadas na validação cruzada interna, o que pode ter afetado a otimização final dos modelos. Além disso, no Estudo C, optamos por realizar os

experimentos apenas com os dados correspondentes à Universidade 1, devido ao custo computacional elevado associado à simulação e treinamento dos modelos para as demais instituições. Essa escolha restringe a generalização dos achados e impede uma avaliação completa da portabilidade dos modelos entre contextos institucionais distintos. Tais limitações reforçam a importância de se considerar recursos computacionais adequados em estudos de replicação e *benchmarking* de modelos de aprendizado de máquina.

- **Heterogeneidade de cenários** – Os estudos variam quanto à definição do problema (evadido, graduado ou matriculado), tamanho da base, balanceamento de classes e seleção de variáveis. Portanto, não é metodologicamente adequado comparar diretamente métricas entre estudos ou estabelecer um *ranking* único de modelos.
- **Foco restrito a três estudos** – A revisão identificou mais seis trabalhos relevantes (2020–2025) com abordagens de aprendizado profundo ou séries temporais. Contudo, esses estudos não foram reproduzidos devido à indisponibilidade dos conjuntos de dados e à ausência dos códigos dos modelos aplicados ou de descrição metodológica suficientemente detalhada para permitir sua replicação. Assim, tendências recentes como redes neurais profundas ou métodos de detecção precoce não estão incluídas na síntese quantitativa.

4.4. Considerações finais

A comparação entre os resultados originais e reproduzidos nos três estudos evidencia padrões consistentes relacionados à robustez dos algoritmos, ao papel dos tipos de variáveis empregadas e ao impacto da transparência metodológica na reprodutibilidade.

Em primeiro lugar, observa-se que modelos baseados em árvores, especialmente RF e *XGBoost*, apresentam desempenho estável nos três estudos. No Estudo B, esses modelos mantiveram AUC acima de 0,93 e valores de F1 próximos de 0,90 tanto nos resultados originais quanto na reprodução. No Estudo C, o RF superou sistematicamente o EN em todos os cenários, destacando sua capacidade de capturar interações complexas entre variáveis institucionais e comportamentais. No Estudo A, embora o desempenho absoluto tenha sido inferior ao dos demais estudos, a DT e a RF apresentaram quedas moderadas quando comparadas à ANN, sugerindo maior resiliência a variações metodológicas.

Em segundo lugar, a análise comparativa demonstra que a disponibilidade de dados e código desempenha papel central na reprodutibilidade dos resultados. No Estudo B, cuja base de dados pública e metodologia estavam claramente documentadas, as métricas reproduzidas foram altamente próximas das originais, inclusive com ganhos pontuais atribuíveis a diferenças na validação cruzada e no ajuste de hiperparâmetros. De modo semelhante, o Estudo C forneceu tanto os dados quanto o código experimental, permitindo uma reprodução estruturada e fiel dos cenários originais; pequenas variações observadas podem ser atribuídas a detalhes específicos da execução e à adaptação necessária para o processamento dos diferentes cenários. Em contraste, o Estudo A ofereceu documentação limitada sobre o pré-processamento e sobre a arquitetura da ANN, resultando em divergências substanciais — como a queda de F1 de 76% para 37% — e confirmando que a falta de transparência metodológica compromete a replicabilidade, sobretudo em modelos mais sensíveis.

Outro aspecto transversal diz respeito ao papel das variáveis comportamentais. No

Estudo C, a inclusão de métricas de engajamento estudantil elevou sensivelmente o desempenho dos modelos, evidenciando que variáveis institucionais, embora relevantes, não são suficientes para capturar padrões associados à permanência. Os ganhos observados na AUC e no F1 reforçam que indicadores de interação real com o ambiente acadêmico ampliam a capacidade preditiva e devem ser incorporados em estudos futuros sempre que possível.

Por fim, constata-se que a reprodutibilidade varia substancialmente entre os estudos, refletindo graus distintos de detalhamento metodológico. Quando modelos, hiperparâmetros, estratégias de validação e pré-processamentos foram claramente especificados (o Estudo B, parcialmente, e o Estudo C), os resultados reproduzidos foram consistentes e próximos aos originais. Nos casos em que tais informações estavam incompletas ou ausentes (Estudo A), o processo de reprodução enfrentou limitações relevantes, sobretudo em modelos de maior complexidade.

No conjunto, as evidências apontam para três tendências gerais: (i) modelos de ensemble mantêm desempenho robusto em diferentes contextos; (ii) variáveis comportamentais ampliam significativamente o poder preditivo; e (iii) a transparência metodológica permanece como condição indispensável para assegurar reprodutibilidade e comparabilidade entre estudos.

5. Conclusões e Trabalhos Futuros

De modo geral, a análise dos estudos evidencia que a predição da evasão estudantil tem se apoiado em um conjunto amplo de variáveis e em procedimentos metodológicos que buscam garantir rigor e consistência. Os modelos frequentemente combinam indicadores acadêmicos, características socioeconômicas e demográficas, além de informações institucionais e comportamentais, compondo uma representação multifatorial dos determinantes da permanência estudantil. Paralelamente, a predominância de métricas baseadas na matriz de confusão e a utilização sistemática de estratégias de validação cruzada demonstram um esforço contínuo para assegurar confiabilidade dos modelos desenvolvidos.

Apesar desses avanços, persistem lacunas significativas que apontam direções importantes para estudos futuros. A disponibilização de dados anonimizados e de *scripts* completos de análise surge como uma medida essencial para fortalecer a transparência e facilitar a reprodutibilidade por parte da comunidade científica. Ademais, investigações conduzidas em diferentes contextos institucionais permitiriam avaliar com maior precisão as limitações e o alcance dos modelos, ampliando sua aplicabilidade. Outro ponto promissor refere-se à expansão do uso de variáveis comportamentais, especialmente aquelas que capturam o engajamento em ambientes digitais, as quais têm demonstrado impacto relevante na melhoria do desempenho preditivo.

Por fim, a RSL apresentada neste trabalho estabelece uma base que pode auxiliar uma possível análise quantitativa, contribuindo para a consolidação estatística dos achados e para o amadurecimento do campo de pesquisa sobre predição de evasão no ensino superior.

Referências

- [Aljawarneh and Lara 2021] Aljawarneh, S. and Lara, J. A. (2021). Data science for analyzing and improving educational processes. *Journal of Computing in Higher Education*,

33:545–550.

- [Badillo et al. 2020] Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., and Zhang, J. D. (2020). An introduction to machine learning. *Clinical Pharmacology & Therapeutics*, 107:871–885.
- [Brignardello-Petersen et al. 2025] Brignardello-Petersen, R., Santesso, N., and Guyatt, G. H. (2025). Systematic reviews of the literature: an introduction to current methods. *American Journal of Epidemiology*, 194(2):536–542.
- [Brown and et al. 2020] Brown, T. and et al. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- [Cajal et al. 2020] Cajal, B., Jim’enez, R., Gervilla, E., and Monta no, J. J. (2020). Doing a systematic review in health sciences. *Clinical and Health*, 31(2):77–83.
- [Choi et al. 2020] Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., and Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science & Technology*, 9:14–14.
- [Diego Dermeval e Ig Bittencourt e Leonardo Marques 2020] Diego Dermeval e Ig Bittencourt e Leonardo Marques, Lúcia Dellagnelo, S. I. (2020). Editorial da edição especial mapeamento sistemático e revisão sistemática da literatura sobre o uso de tecnologias educacionais na educação básica. *Revista Brasileira de Informática na Educação*, 28(0).
- [Dos Santos et al. 2020] Dos Santos, M. B. F., Agostini, B. A., Bassani, R., Pereira, G. K. R., and Sarkis-Onofre, R. (2020). Protocol registration improves reporting quality of systematic reviews in dentistry. *BMC medical research methodology*, 20(1):57.
- [Glandorf et al. 2024] Glandorf, D., Lee, H. R., Orona, G. A., Pumptow, M., Yu, R., and Fischer, C. (2024). Temporal and between-group variability in college dropout prediction. *ACM International Conference Proceeding Series*, pages 486–497.
- [González-Morales et al. 2025] González-Morales, M. O., López-Aguilar, D., Álvarez Pérez, P. R., and Toledo-Delgado, P. A. (2025). Dropping out of higher education: Analysis of variables that characterise students who interrupt their studies. *Acta Psychologica*, 252:104669.
- [Gutierrez-Pachas et al. 2023] Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G., and Gomez-Nieto, E. (2023). Supporting decision-making process on higher education dropout by analyzing academic, socioeconomic, and equity factors through machine learning and survival analysis methods in the latin american context. *Education Sciences 2023, Vol. 13, Page 154*, 13:154.
- [Kitchenham 2004] Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical report, Department of Computer Science University of Durham.
- [Kitchenham and Charters 2007] Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report. Technical Report, Version 2.3.

- [Kitchenham et al. 2015] Kitchenham, B. A., Budgen, D., and Brereton, P. (2015). *Evidence-based software engineering and systematic reviews*. CRC press.
- [Marcolino et al. 2025] Marcolino, M. R., Porto, T. R., Primo, T. T., Targino, R., Ramos, V., Queiroga, E. M., Munoz, R., and Cechinel, C. (2025). Student dropout prediction through machine learning optimization: insights from moodle log data. *Scientific Reports*, 15.
- [Matz et al. 2023] Matz, S. C., Bukow, C. S., Peters, H., et al. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, 13:5705.
- [Moreno and Támara 2024] Moreno, S. P. B. and Támara, L. G. (2024). Complexities of student dropout in higher education: a multidimensional analysis. *Frontiers in Education*, 9:1461650.
- [Ngulube and Ncube 2025] Ngulube, P. and Ncube, M. M. (2025). Predicting academic success and identifying at-risk students: A systematic review of data analytics and machine learning approaches in higher education institutions. *Theory and Practice*, 2025:117–134.
- [Niyogisubizo et al. 2022] Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., and Nshimyumukiza, P. C. (2022). Predicting student’s dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, 3:100066.
- [Page and et al. 2021] Page, M. J. and et al. (2021). Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372.
- [Pereira da Silva 2009] Pereira da Silva, É. R. (2009). *Métodos para Revisão e Mapeamento Sistemático da Literatura (Methods for Systematic Literature Reviews and Systematic Mapping Studies)*. PhD thesis, Universidade Federal do Rio de Janeiro.
- [Romero and Ventura 2020] Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3):e1355.
- [Romero and Liao 2025] Romero, S. and Liao, X. (2025). Statistical and machine learning models for predicting university dropout and scholarship impact. *PLOS ONE*, 20(6):1–19.
- [Rosen 2019] Rosen, K. H. (2019). *Discrete Mathematics and Its Applications*. McGraw-Hill Education, New York, 8th edition.
- [Saito and Rehmsmeier 2015] Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432.
- [SULAK and KOKLU 2024] SULAK, S. A. and KOKLU, N. (2024). Predicting student dropout using machine learning algorithms. *Intelligent Methods In Engineering Sciences*, 3(3):91–98.