

Evasão no Ensino Superior brasileiro: Mapeamento Estruturado da Literatura em *ML*, *EDM* e *LA*, e *Benchmark*

Eduardo Fernandes Bueno^{1,2}, Decio Wey Berti Junior¹, e Jacques Duílio Brancher²

¹Assessoria de Tecnologia de Informação (ATI)

Universidade Estadual de Londrina (UEL)

Rodovia Celso Garcia Cid, PR-445, Km 380 – Campus Universitário

CEP 86057-970 – Londrina – PR – Brasil

²Departamento de Computação

Universidade Estadual de Londrina (UEL)

Rodovia Celso Garcia Cid, PR-445, Km 380 – Campus Universitário

CEP 86057-970 – Londrina – PR – Brasil

bueno.fernandes.eduardo@uel.br, decio@uel.br, jacques@uel.br

Abstract. *Dropout in Brazilian Higher Education is a relevant problem with official rates of 60%. This article presents a Scoping Review aligned with PRISMA-ScR, covering 88 Brazilian studies published between 2018 and 2026 classified according to their predominant orientation in ML, EDM, or LA. Of these, 78 studies used their own institutional bases, classification targets and non-temporal validations; only 10 explicitly used INEP/MEC data, 49 used Random Forest and 14 reported artifacts. The review identifies key methodological gaps. In the temporal benchmark using Random Forest and INEP's official database, there is a prediction of the evasion rate in $t + 1$ from variables observed in t . The Random Forest achieved regression performance in temporal holdout with $R^2 = 0,3337$, $MAE = 0,1778$, $RMSE = 0,2331$, and $MSE = 0,0543$, with reproducible synthesis and benchmark for dropout prediction.*

Keywords: *Higher Education Dropout; Machine Learning; Educational Data Mining; Learning Analytics; Random Forest.*

Resumo. *A evasão no Ensino Superior brasileiro é um problema relevante com taxas oficiais de 60%. Este artigo apresenta uma Scoping Review alinhada ao PRISMA-ScR, abrangendo 88 estudos brasileiros publicados entre 2018 e 2026, classificados segundo sua orientação predominante em ML, EDM ou LA. 78 estudos adotaram bases institucionais próprias, alvos classificatórios e validações não temporais e 10 estudos basearam-se em dados do INEP. 49 estudos aplicaram Random Forest (RF) e 14 reportaram artefatos. A revisão identificou lacunas metodológicas. No benchmark temporal com RF da base de dados oficiais do INEP, há predição da taxa de evasão em $t + 1$ a partir de variáveis observadas em t . O RF obteve desempenho regressivo no holdout temporal com $R^2 = 0,3337$, $MAE = 0,1778$, $RMSE = 0,2331$ e $MSE = 0,0543$, com síntese reprodutível e benchmark para predição de evasão.*

Palavras-chave: *Evasão no Ensino Superior; Aprendizado de Máquina; Mineração de Dados Educacionais; Learning Analytics; Random Forest.*

1. Introdução

A evasão constitui um problema crítico no Ensino Superior brasileiro. Segundo dados do MEC/INEP, os "indicadores de trajetória dos estudantes em cursos de licenciatura para a coorte de ingressantes de 2015", no grupo "BRASIL (Todos os cursos)", apresentam 60% de desistência acumulada; no mesmo conjunto de dados observa-se que a "Taxa de Evasão do Sistema de Educação Superior de 2010 a 2024" aumentou de 11,3% a 17,5% [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira 2025].

Esse fenômeno afeta estudantes, instituições, políticas públicas e a alocação de recursos educacionais. Na rede pública, compromete a eficiência do investimento em educação; no setor privado, afeta sustentabilidade econômico-financeira, ocupação de vagas e planejamento institucional.

Há necessidade de métodos capazes de monitorar trajetórias acadêmicas, identificar padrões de risco e apoiar decisões institucionais de permanência estudantil, com uso de *Machine Learning* (ML), *Educational Data Mining* (EDM) e *Learning Analytics* (LA).

As contribuições deste artigo são: i) uma revisão e mapeamento estruturado de 88 estudos brasileiros sobre evasão/permanência no Ensino Superior, classificados segundo sua orientação predominante em ML, EDM ou LA; ii) a estratificação da literatura entre estudos com e sem uso explícito de dados oficiais INEP/MEC, e de dados próprios institucionais; iii) uma síntese quantitativa de algoritmos, validações, métricas, tipos de alvo e artefatos computacionais; e iv) um *benchmark* temporal com *RF*, formulado como resposta experimental às lacunas identificadas na revisão, especialmente com uso de dados INEP, validação temporal, redução do risco de *data leakage* e comparação com *baselines*.

O *benchmark* não constitui um estudo independente ou concorrente à revisão, mas uma instância experimental mínima, auditável e reproduzível das lacunas metodológicas mapeadas. Seu objetivo é demonstrar como uma formulação temporal com dados oficiais do INEP/MEC pode servir de referência comparável para pesquisas futuras, sem substituir estudos institucionais baseados em dados individuais ou contextos locais.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica; a Seção 3 detalha a metodologia da revisão; a Seção 4 discute o *benchmark experimental*, a Seção 5 apresenta as ameaças à validade; a Seção 6 apresenta a conclusão; e a Seção 7 indica trabalhos futuros.

2. Fundamentação Teórica

O uso de *Machine Learning* (ML) em modelagem computacional de padrões em dados acadêmicos, permite estimar trajetórias, desempenho e situações de risco de evasão estudantil. Essa perspectiva dialoga com a definição clássica de aprendizado de máquina como campo voltado à modelagem de aprendizagem e à melhoria de desempenho em tarefas específicas [Carbonell et al. 1983].

Educational Data Mining ou Mineração de Dados Educacionais aplica métodos computacionais a dados educacionais para identificar padrões, perfis e relações úteis à compreensão de fenômenos como evasão, permanência e desempenho acadêmico [Baker et al. 2011].

Já a *Learning Analytics* enfatiza medição, coleta, análise e comunicação de dados sobre estudantes e seus contextos, com o objetivo de compreender e otimizar processos de

aprendizagem [Long and Siemens 2011]. Na prática, a EDM tende a enfatizar descoberta de padrões de modelagem e a LA aproxima esses resultados de relatórios, visualizações e apoio à decisão institucional [Jones 2019].

Neste trabalho, portanto, ML, EDM e LA não são tratados como metodologias equivalentes nem como campos de fronteiras mutuamente exclusivas. Para fins de síntese, cada estudo recebeu uma orientação predominante, definida pelo enquadramento adotado pelos autores e pela coerência entre seu objetivo, sua tarefa analítica e o uso declarado dos resultados [Carbonell et al. 1983, Baker et al. 2011, Long and Siemens 2011]

O *Random Forest* destaca-se por combinar múltiplas árvores de decisão treinadas com componentes aleatórios, agregando suas previsões para aumentar robustez e reduzir instabilidade do modelo [Breiman]. Seu desempenho depende do equilíbrio entre a força preditiva das árvores individuais e a baixa correlação entre elas, o que explica a sua recorrência em tarefas preditivas de evasão.

Data leakage ou *vazamento de dados* ocorre quando informações indisponíveis no momento real da predição são introduzidas, direta ou indiretamente, durante o treinamento, na seleção de atributos, no pré-processamento ou na avaliação do modelo. Em problemas educacionais, esse risco pode ocorrer quando variáveis contemporâneas ou posteriores ao desfecho são usadas como preditores ou quando o particionamento dos dados permite que informações de períodos futuros influenciem o treinamento. Nesses casos, o desempenho estimado tende a ser artificialmente otimista, pois, o modelo passa a se beneficiar de informação que não estaria disponível em um cenário prospectivo de uso [Kaufman et al. 2012, Kuhn and Johnson 2013].

Portanto, estratégias de validação temporal são particularmente relevantes em séries educacionais ordenadas no tempo. Particionamentos aleatórios podem misturar registros de anos distintos entre treino e teste; logo, a validação temporal preserva a ordem cronológica dos dados, treinando modelos com observações do passado e avaliando-os em períodos posteriores. Essa separação busca aproximar a avaliação experimental de uma situação real de predição futura, reduzindo o risco de vazamento temporal [Bergmeir and Benítez 2012, Roberts et al. 2017].

3. Metodologia

Este estudo caracteriza-se como *Scoping Review*, alinhado ao PRISMA-ScR e com elementos de *Systematic Mapping*, voltado ao mapeamento de evidências, padrões metodológicos e lacunas de pesquisa [Kitchenham et al. 2022].

Para garantir rastreabilidade e replicabilidade, o protocolo definiu previamente bases de busca, *strings*, período de publicação, tipos documentais elegíveis, critérios de inclusão e exclusão, processo de triagem no *Rayyan* e campos padronizados.

3.1. Objetivo, Questões da Pesquisa (QP1–QP4) e PICOC

O objetivo da revisão literária é identificar e sintetizar estudos primários sobre Evasão, Permanência, Retenção ou Conclusão no Ensino Superior brasileiro. Para a síntese quantitativa, cada estudo foi classificado segundo sua orientação predominante, com a distinção entre modelagem computacional (ML), mineração de dados educacionais (EDM) e uso analítico de evidências para compreensão e apoio à decisão (LA).

A revisão adota estratificação analítica entre estudos sem uso explícito de dados INEP/MEC, classificados como `UNIVERSO_A`; e estudos com uso explícito desses dados, classificados como `UNIVERSO_B`. Estudos que utilizaram *Random Forest* foram identificados para comparação com o *benchmark* experimental.

As questões de pesquisa são: **QP1**: Como os estudos sobre predição ou monitoramento de evasão se distribuem segundo sua orientação predominante em ML, EDM ou LA no Ensino Superior brasileiro; e como se distribuem entre estudos com ou sem uso explícito de dados INEP/MEC? **QP2**: Quais estratégias de validação e métricas são reportadas? **QP3**: Quais tipos de sistemas, *dashboards* e interfaces são propostos? **QP4**: Quais tipos de alvo são empregados nos estudos, tais como taxa de evasão, risco de evasão, evasão binária, permanência ou retenção?

Foi utilizado o PICOC da seguinte forma: **População**: estudos sobre evasão no Ensino Superior brasileiro com uso de dados educacionais. **Intervenção**: aplicação de ML/EDM/LA e/ou construção de sistemas para predição ou monitoramento. O uso de RF será estratificado para análise e *benchmark*. **Comparação**: quando houver algoritmos, compará-los entre forma de validação, variáveis e/ou abordagens com ou sem sistema. **Outcome (resultados)**: desempenho, evidências de utilidade, presença de artefato, reprodutibilidade e implicações para decisão. **Contexto**: Brasil, Ensino Superior/Graduação, com estratificação de uso ou não de dados INEP/MEC.

3.2. Estratégia de Busca e Bases

A estratégia de busca primária envolveu bases de dados e bibliotecas digitais internacionais a saber: *Scopus*, *Web of Science* (WoS), *IEEE Xplore*, *ACM Digital Library*, *ScienceDirect*, *SpringerLink* e *ERIC*, além do *Google Scholar* como buscador complementar. Para cobertura do cenário nacional, tendo o Simpósio Brasileiro de Informática na Educação (SBIE) e a Revista Brasileira de Informática na Educação (RBIE) fez-se consultas à biblioteca digital *SOL/SBC*, à rede *SciELO* e ao Portal de Periódicos *CAPES*.

A estratégia foi organizada em cinco *strings* complementares: **S1**, voltada à busca principal em inglês; **S2**, com âncora em Brasil e INEP/MEC; **S3**, voltada à produção nacional em português; **S4**, orientada à identificação de artefatos, *dashboards* e apoio à decisão; e, **S5**, específica para estudos com *Random Forest*.

A *string* fundamental que serviu de base para S1–S5 é a seguinte:

("dropout"OR attrition OR retention OR "dropout prediction"OR "dropout risk") AND ("higher education"OR undergraduate OR university OR HEI OR "higher education institution") AND (Brazil OR Brazilian) AND ("machine learning"OR "educational data mining"OR "learning analytics"OR "data mining"OR EDM)

O período de 2018–2026 justifica-se pela necessidade de capturar o estado da arte em técnicas de ML/EDM/LA, além de alinhar-se à vigência da *Lei Geral de Proteção de Dados no Brasil* (LGPD), que tem a anonimização de dados como um princípio norteador. As publicações elegíveis foram artigos completos de periódicos e *conference papers*; excluindo-se patentes, capítulos de livros, monografias (TCCs), dissertações, teses e *short papers*. Os campos priorizados compreenderam título, resumo e palavras-chave, conforme disponibilidade de cada base de dados.

3.3. Processo de Seleção, Critérios de Inclusão e Exclusão

O processo de seleção foi conduzido em três fases no *Rayyan*: **Triagem** por título e resumo no *screening*; **Full-text screening**, com leitura também da introdução, conclusão e texto completo quando necessário; e inclusão no *Data Extraction*.

As bases que mais se destacaram foram a *ERIC* 1.138; *Google Scholar* 1.095; e *SOL/SBC* 1.048. As bases com menor número de registros foram a *ScienceDirect* 4 e *SciELO* 7, conforme Tabela 1.

De um total de 3.827 registros, que analisados com *Rayyan*, retirando duplicados restaram 2.350. Em uma segunda triagem com *Screening*, e reavaliação com *Full-text screening*, restaram 88 estudos para avaliação de texto completo e documentação da decisão metodológica. A síntese do fluxo de registros no *Rayyan* é apresentada na Tabela 2.

Tabela 1. Estudos Importados

Fonte	Quantidade
ERIC	1.138
Google Scholar	1.095
SOL/SBC	1.048
Web of Science	169
SpringerLink	148
IEEE Xplore	82
Scopus	60
ACM Digital Library	45
CAPES	31
SciELO	7
ScienceDirect	4
Total	3.827

Tabela 2. Fluxo de registros no Rayyan

Etapas	Quantidade
Referências importadas no <i>Rayyan</i>	3.827
Sinalização de duplicidade	2.738
Registros removidos por duplicidade	1.477
Referências para triagem	2.350
Excluídos no <i>screening</i>	2.260
Incluídos no <i>screening</i>	90
Registros reavaliados no <i>FTS</i>	29
Total avaliados no <i>Full-text screening</i>	119
Excluídos após texto completo	31
Incluídos para extração de dados	88

De acordo com o protocolo, os **Critérios de Inclusão (CIs)** são os descritos na Tabela 3 e os **Critérios de Exclusão (CEs)** na Tabela 4, respectivamente a seguir:

Tabela 3. Critérios de Inclusão

Código	Descrição
CI1	Estudos com foco em evasão, permanência, retenção ou risco de evasão no Ensino Superior brasileiro, em nível de graduação, publicados a partir de 2018.
CI2	Estudos aplicados ao contexto brasileiro de Ensino Superior, com menção explícita ao Brasil ou uso de dados, validação ou instituição situada no país.
CI3	Estudos que aplicam ML, EDM, LA ou método computacional equivalente para predição, monitoramento, classificação, regressão, detecção de padrões ou tarefa analítica relacionada à evasão/permanência.
CI4	Artigos completos, publicados em periódicos ou conferências, com texto completo acessível.
CI5	Estudos publicados em português ou inglês.
CI6	Estudos com ou sem uso explícito de dados INEP/MEC, desde que aderentes ao recorte Ensino Superior brasileiro + ML/EDM/LA + evasão/permanência.
CI7	Estudos com ou sem uso de <i>Random Forest</i> ; o uso de RF foi tratado como recorte analítico específico para comparação com o <i>benchmark</i> , não como critério obrigatório de inclusão.

Tabela 4. Critérios de Exclusão

Código	Descrição
CE1	Excluir estudos que não tratam de evasão, permanência, retenção, abandono ou risco de evasão no Ensino Superior brasileiro.
CE2	Excluir estudos que não são do contexto brasileiro de Ensino Superior (HE-BR), incluindo estudos exclusivamente internacionais ou de outros níveis educacionais, como educação básica, técnica, K-12 ou <i>secondary education</i> .
CE3	Excluir estudos que não aplicam <i>Machine Learning</i> (ML), <i>Educational Data Mining</i> (EDM), <i>Learning Analytics</i> (LA) ou método analítico computacional equivalente, incluindo textos apenas teóricos, ensaísticos, normativos ou descritivos sem modelagem ou analítica computacional.
CE4	Excluir resumos, pôsteres, editoriais, slides, prefácios, <i>short papers</i> , <i>extended abstracts</i> , <i>book chapters</i> , <i>patents</i> , TCCs, dissertações e teses.
CE5	Excluir duplicatas, mantendo-se a versão mais completa para extração.
CE6	Excluir estudos com texto completo indisponível.
CE7	A ausência de INEP/MEC explícito não constitui motivo de exclusão. Estudos sem INEP/MEC explícito, mas aderentes ao recorte HE-BR + ML/EDM/LA + evasão/permanência, permanecem elegíveis no UNIVERSO_A. Apenas os estudos com INEP/MEC explícito são adicionalmente estratificados no UNIVERSO_B.
CE8	A ausência de <i>Random Forest</i> (RF) não constitui motivo de exclusão.

3.4. Extração de Dados e Síntese dos Resultados

A extração analítica dos dados foi estruturada em planilha eletrônica própria, com campos padronizados e síntese quantitativa, e para evitar dupla contagem, cada estudo recebeu uma única orientação predominante: ML, EDM ou LA. A atribuição considerou a coerência entre o objetivo do estudo, a tarefa analítica executada e a forma de utilização dos resultados. Essa codificação operacional não implica equivalência conceitual entre ML, EDM e LA, nem ausência de intersecções entre esses campos.

A síntese dos resultados tem três níveis complementares: 1) **Descritivo**, com distribuição dos estudos por ano, algoritmos, validação e métricas; 2) **Temático**, com categorias de apoio à tomada de decisão, apenas quando presente algum artefato computacional; 3) **Estratificado**, com comparação entre Universos A/B, e estudos com RF.

A síntese segue o encadeamento: **Problema** → **Dados** → **Alvo** → **Variáveis** → **Modelo** → **Validação** → **Métricas** → **Artefato** → **Evidência de apoio à decisão**.

3.4.1. Caracterização Geral dos Estudos

A distribuição temporal mostra concentração de estudos entre 2020–2025, sugerindo crescimento recente das aplicações de ML/EDM/LA em evasão e permanência no Ensino Superior e possível interesse institucional por monitoramento acadêmico.

A predominância de publicações em português indica forte presença da literatura nacional, inclusive vinculados à Informática/Computação e Educação. Estudos em inglês 27 (30,7%) evidenciam que a produção brasileira busca diálogo com a comunidade internacional de LA e EDM.

3.5. Resultados e Estratificação Analítica

A maior parte dos estudos foi classificada no UNIVERSO_A, sem uso explícito de dados INEP/MEC e com uso exclusivo de bases institucionais próprias, totalizando 78 trabalhos (88,6%), dos quais 42 utilizam RF. Os demais 10 estudos (11,4%) utilizaram explicitamente dados INEP/MEC, compondo o UNIVERSO_B; nesse grupo, 7 utilizaram RF. Essa diferença quantitativa constitui achado relevante: embora INEP/MEC seja a principal fonte pública nacional sobre Ensino Superior, sua exploração explícita em estudos de evasão com ML/EDM/LA ainda aparece de forma minoritária.

O uso de *Random Forest* foi identificado em 49 estudos (55,7%), sendo 7 pertencentes ao subconjunto mais diretamente comparável ao *benchmark* empírico, isto é, estudos do UNIVERSO_B com uso explícito de dados INEP/MEC e *Random Forest*. A Tabela 5 sintetiza a composição geral do *corpus* final, incluindo universo analítico, idioma, uso de *Random Forest* e presença de artefatos computacionais.

Tabela 5. Caracterização geral do *corpus* final da revisão

Indicador	n	%
Total de estudos	88	100,0
Universo A – sem INEP/MEC explícito	78	88,6
Universo B – com INEP/MEC explícito	10	11,4
Estudos com <i>Random Forest</i>	49	55,7
Subgrupo INEP/MEC + RF	7	8,0
Com artefato computacional	14	15,9
Sem artefato computacional	74	84,1
Publicações em português	61	69,3
Publicações em inglês	27	30,7

Embora o UNIVERSO_B não represente o maior volume da literatura, constitui o núcleo de maior comparabilidade com este trabalho, por envolver dados oficiais, cobertura nacional e potencial de replicação interinstitucional.

3.6. Orientações Predominantes em ML / EDM / LA

Tabela 6. Principais grupos de algoritmos

Grupo de algoritmo	n	%
<i>Random Forest</i>	49	55,7
Decision Tree/J48/CART	41	46,6
Regression/SVR/Ridge/Linear	29	33,0
Logistic Regression	25	28,4
SVM/SVC	23	26,1
KNN	21	23,9
Boosting/AdaBoost/GBM/CatBoost	21	23,9
Neural Network/MLP	19	21,6
Naive Bayes	17	19,3
XGBoost	13	14,8
Clustering/K-means	5	5,7
LightGBM	4	4,5
Association Rules/Apriori	3	3,4
Explainability/XAI	3	3,4

Tabela 7. Estratégias de validação identificadas nos estudos

Tipo de validação	n	%
<i>Cross-validation</i>	28	31,8
<i>Holdout</i>	19	21,6
<i>Holdout e cross-validation</i>	14	15,9
Não clara	13	14,8
Temporal	7	8,0
<i>Holdout, cross-validation, Monte Carlo</i>	1	1,1
Qualitativa por grupo focal	1	1,1
Distância de <i>cluster</i>	1	1,1
Experimental e análise exploratória	1	1,1
Sem validação empírica	1	1,1
Método do cotovelo e análise exploratória	1	1,1
Holdout não especificado	1	1,1

Em relação à QP1, 57 estudos (64,8%) foram classificados como predominantemente orientados a *Machine Learning*; 27 trabalhos (30,7%) em *Educational Data Mining*; e 4 (4,5%) em *Learning Analytics*. Esses valores correspondem a uma codificação analítica exclusiva para fins de síntese e não significam ausência de sobreposição entre os campos. Nos estudos que empregavam mais de uma denominação, prevaleceu a orientação mais coerente com o objetivo central, a tarefa analítica e o uso declarado dos resultados. A distribuição dos principais grupos de algoritmos e das estratégias de validação identificadas é apresentada nas Tabelas 6 e 7.

As Tabelas 8 e 9 indicam que a literatura brasileira analisada tem orientação predominantemente algorítmica e preditiva, com forte presença de modelos supervisionados e métricas classificatórias. A predominância de alvos do tipo risco de evasão (68,2%) sugere que muitos estudos formulam o problema como classificação de estudantes ou registro em grupos de risco, e não como estimação temporal de taxas futuras. Essa escolha é compreensível em bases institucionais, nas quais há maior proximidade com dados acadêmicos individuais; mas, dificulta a comparação direta com estudos baseados em dados oficiais agregados.

Tabela 8. Métricas reportadas

Grupo de métrica	n	%
Acurácia/Accuracy	61	69,3
Precisão/Precision	47	53,4
Recall/Sensibilidade	45	51,1
F1/F-score	42	47,7
Matriz de confusão	26	29,5
AUC/ROC	22	25,0
Lift/Confiança/Suporte	8	9,1
Kappa	7	8,0
MAE	4	4,5
MSE	3	3,4
MCC	2	2,3
RMSE	1	1,1
R^2 /Pseudo- R^2	1	1,1

Tabela 9. Tipos de alvo empregados nos estudos

Tipo de alvo	n	%
Risco de evasão	60	68,2
Permanência/retenção	9	10,2
Evasão binária	8	9,1
Taxa de evasão	6	6,8
Conclusão	1	1,1
Evasão/permanência	1	1,1
Conclusão, tempo ideal, retenção e risco de evasão	1	1,1
Evasão/reprovação	1	1,1
Previsão de evasão	1	1,1

Outro ponto crítico é a baixa frequência de validação temporal. Embora *cross-validation* e *holdout* sejam estratégias úteis, sua aplicação sem separação cronológica pode superestimar o desempenho quando registros de anos próximos ou futuros influenciam o treinamento. Nesse sentido, a predominância de validações não temporais reforça a contribuição do *benchmark* proposto, que avalia a predição da evasão futura em $t + 1$ a partir de variáveis observadas em t .

3.7. Artefatos Computacionais e Apoio à Decisão

A presença de artefatos computacionais foi analisada como evidência de conversão dos modelos preditivos em instrumentos práticos de apoio à gestão acadêmica, como *dashboards*, sistemas *web*, APIs etc. Dos 88 estudos, apenas 14 (15,9%) reportam algum tipo de artefato. A Tabela 10 sintetiza a existência, o tipo de artefato e a categoria de apoio à decisão identificada. Esses resultados sugerem que a maior parte da literatura permanece concentrada na etapa experimental de modelagem, sem avançar para mecanismos de operacionalização, visualização ou uso recorrente por gestores.

Tabela 10. Artefatos computacionais e categorias de apoio à decisão

Existência de artefato			Tipo de artefato			Categoria de apoio à decisão		
Não	74	84,1	Não se aplica	74	84,1	Não se aplica	74	84,1
Sim	14	15,9	Dashboard	3	3,4	Priorização de intervenção	5	5,7
			Sistema web	3	3,4	Monitoramento	3	3,4
			API	2	2,3	Exploratório/visualização	2	2,3
			Protótipo	2	2,3	Diagnóstico analítico de gestão	1	1,1
			Dashboard SABIA	1	1,1	Apoio e explicabilidade	1	1,1
			Arquitetura ML/XAI	1	1,1	Planejamento acadêmico	1	1,1
			Plataforma LANSE	1	1,1	Alerta precoce	1	1,1
			Relatório digital	1	1,1			

3.8. Discussão dos Achados da Revisão

A baixa presença de artefatos (15,9%) pode decorrer de diferentes fatores: escopo limitado dos estudos, ausência de equipes multidisciplinares para transformar modelos em sistemas, dificuldades de integração com bases acadêmicas institucionais, preocupações com privacidade e governança de dados ou ainda a predominância de avaliações centradas em desempenho algorítmico. Do ponto de vista da informática na Educação, essa lacuna é relevante porque modelos de predição só se tornam efetivamente úteis quando conectados a processos de interpretação, priorização de intervenção e monitoramento institucional.

A leitura conjunta de ML, EDM e LA deve ser interpretada como uma estratégia de mapeamento transversal da produção nacional e não como fusão conceitual entre campos com objetivos distintos. Para permitir a comparação, foram aplicadas dimensões comuns aos estudos –fonte dos dados, definição do alvo, variáveis, modelos, validação, métricas e artefatos– sem desconsiderar que ML enfatiza métodos de modelagem, EDM privilegia a descoberta de padrões em contextos educacionais e LA aproxima a análise de processos de interpretação, comunicação e apoio à decisão.

A partir dos achados da revisão, o *benchmark* experimental foi definido como uma etapa de operacionalização metodológica e não como um segundo estudo desconectado. A revisão mostrou três lacunas recorrentes no *corpus*: uso minoritário de dados oficiais INEP/MEC, baixa frequência de validação temporal e pouca explicitação de *baselines* comparativos. O *benchmark*, então, foi construído para materializar essas lacunas em uma formulação experimental controlada: predizer evasão futura em $t + 1$ a partir de variáveis observadas em t , usando dados oficiais agregados, separação temporal e comparação com modelos. Assim, o experimento não busca substituir os estudos institucionais analisados, mas, oferecer uma referência mínima para pesquisas futuras baseadas em dados oficiais.

4. Benchmark Experimental: Dados, Modelagem e Validação

O *benchmark* foi desenhado para avaliar a capacidade preditiva em cenário temporal, evitando particionamentos aleatórios que poderiam misturar observações futuras no treino. Variáveis observadas no ano-base t foram utilizadas para estimar a `taxa_evasao_alvo` no ano-alvo $t + 1$, com separação temporal entre treino e teste, comparação com *baselines* explícitos e exclusão de variáveis contemporâneas diretamente derivadas da evasão.

A coleta e organização dos microdados do Censo da Educação Superior utilizaram `requests`, `BeautifulSoup` e `PyArrow`; a preparação e modelagem utilizaram

pandas, NumPy, scikit-learn e joblib, com RandomForestRegressor e LinearRegression. A visualização foi apoiada por Matplotlib, Seaborn, Plotly e Streamlit. A reprodutibilidade foi verificada por testes com pytest e configuração prevista de padronização com Ruff.

As variáveis da entrada foram: `ingressantes_t`, `concluintes_t`, `matriculados_t`, `vagas_totais_t` e `inscritos_totais_t`. Variáveis contemporâneas diretamente derivadas da evasão, como `taxa_evasao` e `taxa_conclusao`, não foram utilizadas como atributos preditivos no modelo principal (RF).

A base preditiva principal foi o arquivo `dados_modelagem_temporal_sbie.csv`, com 400.195 registros antes da remoção de valores ausentes e 310.757 registros completos após critério de completude das variáveis temporais.

A avaliação utilizou validação temporal com treino em `ano_base <= 2018` e teste em `ano_base > 2018`. Após o *dropna*, o conjunto de treino compreendeu 199.429 registros dos anos-base de 2009–2018, enquanto o conjunto de teste compreendeu 111.328 registros dos anos-base de 2019, 2021, 2022 e 2023, correspondendo aos anos-alvo 2020, 2022, 2023 e 2024.

O *Random Forest* foi comparado com três *baselines*: média histórica do treino, persistência temporal e Regressão Linear. No *holdout* temporal, o RF obteve melhor desempenho regressivo entre os modelos avaliados, com $R^2 = 0,3337$, $MAE = 0,1778$, $RMSE = 0,2331$ e $MSE = 0,0543$. Essa superioridade, contudo, deve ser interpretada no eixo regressivo, pois, a leitura binária alta/baixa evasão foi utilizada como análise auxiliar e apresentou comportamento menos uniforme entre as métricas classificatórias. Assim, a comparação principal do *benchmark* considera métricas regressivas, enquanto acurácia, precisão, recall e F1 são interpretadas apenas como evidência complementar sobre a separação entre cenários de maior e menor evasão, como demonstra a Tabela 11.

Tabela 11. Comparação entre baselines e Random Forest no holdout temporal

Modelo	R^2	MAE	RMSE	MSE	Acc.	Prec.	Recall	F1
Média histórica	-0,0049	0,2429	0,2862	0,0819	0,7167	0,7167	1,0000	0,8350
Persistência	-0,0559	0,1967	0,2934	0,0861	0,7432	0,8373	0,7965	0,8164
Regressão Linear	0,1595	0,2172	0,2617	0,0685	0,7379	0,7388	0,9811	0,8429
Random Forest	0,3337	0,1778	0,2331	0,0543	0,7404	0,8095	0,8342	0,8216

Experimentos complementares reforçaram a consistência da formulação temporal. Na ablação, executada com configuração própria de *Random Forest*, a remoção de `concluintes_t` reduziu o desempenho de $R^2 = 0,4075$ para $R^2 = 0,2683$, indicando valor preditivo relevante dessa variável. Mas, seu uso exige explicitar a hipótese operacional de que a predição ocorre após o fechamento anual do ano-base t . Caso o cenário de uso envolva decisões durante o ano corrente, `concluintes_t` deve ser removida ou substituída por variável parcial disponível naquele momento, conforme mostra Tabela 12.

Tabela 12. Experimento de ablação sem `concluintes_t`

Modelo	R^2	MAE	RMSE	F1
<i>Random Forest</i> completo	0,4075	0,1697	0,2198	0,8452
<i>Random Forest</i> sem <code>concluintes_t</code>	0,2683	0,1951	0,2442	0,8356

Também foi executada validação temporal do tipo *rolling/expanding*. Nesse experimento complementar, o *Random Forest* teve obtido R^2 médio de 0,4372, MAE médio de 0,1639 e F1 médio de 0,8335, enquanto a Regressão Linear obteve R^2 médio de 0,1993, MAE médio de 0,2122 e F1 médio de 0,8260. Esses resultados não substituem o *holdout* temporal principal, mas, funcionam como evidência adicional de estabilidade temporal.

Uma limitação observada no diagnóstico de completude foi a ausência de `ano_base=2020` após a remoção de linhas incompletas. Antes do critério de completude, havia 28.507 registros com `ano_base=2020`; entretanto, nenhum permaneceu após o *dropna*, pois, todas essas observações apresentavam ausência simultânea nas cinco variáveis preditoras obrigatórias em *t*: `ingressantes_t`, `concluintes_t`, `matriculados_t`, `vagas_totais_t`, e `inscritos_totais_t`. O alvo `taxa_evasao_alvo` estava disponível, mas, não havia observações completas das variáveis explicativas correspondentes.

Assim, o ano-alvo 2020 permanece no teste a partir de `ano_base=2019`, mas, não houve avaliação com `ano_base=2020` e, conseqüentemente, não houve predição para `ano_alvo=2021`. Optou-se por não imputar ou interpolar esse ano-base, preservando apenas observações completas. Essa decisão evita introduzir valores sintéticos em um *benchmark* temporal sensível à ordem cronológica dos dados, especialmente em um intervalo potencialmente atípico como 2020–2021. A imputação temporal foi, portanto, tratada como possibilidade para análise de sensibilidade futura.

Permanecem limitações: ausência de variáveis socioeconômicas e exógenas, uso de dados agregados em vez de dados individuais, dependência do fechamento anual para uso de `concluintes_t`, desbalanceamento de classes na leitura binária, perda do ano-base 2020 após critério de completude e desempenho regressivo moderado.

5. Ameaças à Validade

Validade de construção: A ameaça decorre da variação conceitual entre os estudos primários sobre permanência, retenção, conclusão e evasão, bem como de diferenças de granularidade e formas de mensuração. Para mitigação, foram definidos critérios explícitos de inclusão e exclusão, blocos conceituais de triagem e estratificação entre estudos com e sem uso explícito de dados INEP; e formulação temporal do alvo e exclusão de variáveis contemporâneas diretamente derivadas da evasão no *Benchmark*.

Validade interna: A validade interna pode ser afetada por erros de classificação durante a triagem, extração e categorização dos estudos, especialmente na identificação de artefatos computacionais, tipos de alvo e estratégias de validação. No experimento, a ameaça interna decorre do critério de completude, que remove registros incompletos e afeta a cobertura temporal. Essa ameaça é particularmente relevante para `ano_base=2020`, que não apresentou observações completas nas variáveis preditoras obrigatórias e não compôs o conjunto final de modelagem.

Validade externa: A generalização dos resultados deve ser interpretada com cautela. A revisão está restrita a estudos sobre Ensino Superior brasileiro, publicados entre 2018 e 2026, dependente das bases consultadas e das *strings* utilizadas. No experimento, os resultados refletem dados agregados oficiais do INEP/MEC e podem não representar adequadamente cenários institucionais específicos.

Validade de conclusão: pode ser limitada pela predominância de síntese descritiva, heterogeneidade dos estudos primários e ausência de metanálise estatística. As frequências reportadas indicam padrões observados no *corpus*, mas, não implicam superioridade causal de algoritmos, métricas ou estratégias de validação. No experimento, o desempenho do *RF* deve ser interpretado como evidência de um *benchmark* temporal e não como prova definitiva de capacidade preditiva geral ou de causalidade sobre a evasão.

6. Conclusão

A revisão estruturada da literatura brasileira sobre evasão, permanência, retenção e conclusão no Ensino Superior com identificação da predominância se é ML, EDM ou LA, combinada a um *benchmark* experimental temporal com dados oficiais do INEP/MEC, contribui na integração entre três dimensões ainda pouco articuladas na literatura nacional: mapeamento sistemático, estratificação por fonte de dados institucionais próprias ou oficiais e avaliação experimental com formulação temporal orientada à redução da evasão.

Os resultados indicam que a produção brasileira sobre evasão no Ensino Superior é expressiva, mas, apresenta lacunas metodológicas e operacionais. A concentração em bases institucionais próprias mostra a importância de estudos locais, mas, limita a comparabilidade entre trabalhos e a construção de evidências nacionais. A predominância de alvos classificatórios e validações não temporais sugere que parte da literatura prioriza desempenho preditivo imediato, sem avaliar a estabilidade dos modelos em cenários futuros. Além disso, a baixa presença de artefatos computacionais de apoio à tomada de decisão revela uma distância entre experimentos de modelagem e sua incorporação em processos institucionais de monitoramento, priorização de intervenção e permanência estudantil.

O *benchmark* experimental complementa a revisão ao oferecer uma formulação reprodutível baseada em dados oficiais, variáveis observadas no ano-base t e alvo futuro em $t + 1$. O *Random Forest* obteve o melhor desempenho regressivo no *holdout* temporal, mas a análise binária auxiliar não indicou superioridade uniforme em todas as métricas classificatórias. Esse resultado reforça a interpretação do experimento como referência metodológica auditável, e não como solução definitiva para previsão de evasão, evitando conclusões superestimadas a partir de dados agregados.

7. Trabalhos Futuros

Pretende-se aprofundar a análise qualitativa dos estudos do UNIVERSO_B, especialmente aqueles que utilizam dados INEP/MEC e *Random Forest*, comparando definição alvo, granularidade dos dados, variáveis, validação, métricas e reprodutibilidade. Expandir o benchmark com novos algoritmos, validações temporais e formulações do problema, incluindo classificação de risco, regressão da taxa de evasão e recortes por área do conhecimento, modalidade de ensino, categoria administrativa e região geográfica. A incorporação de variáveis socioeconômicas, institucionais, territoriais e temporais poderá ampliar a robustez analítica, desde que acompanhada de cuidados com vies, vazamento de dados, interpretabilidade e governança.

Por fim, recomenda-se evoluir o experimento para artefato computacional completo de apoio à decisão, com *pipeline* reprodutível, rastreabilidade de experimentos, documentação de variáveis, monitoramento de métricas e interface analítica.

Disponibilidade de Artefatos

Os artefatos associados a este trabalho estão disponíveis em repositório público no GitHub: <https://github.com/EddieFerb/sbie2026-evasao-ensino-superior-artefatos>. O repositório inclui a planilha de extração dos 88 estudos primários, documentos de protocolo, condução e relato da revisão, exportações e evidências do Rayyan, tabelas consolidadas do *benchmark* temporal, capturas da aplicação AppEvasão, documentação de disponibilidade dos dados, manifesto dos artefatos, arquivo de citação, licença e checksums.

A aplicação pública AppEvasão, desenvolvida em Streamlit para visualização das evidências do *benchmark* temporal, está disponível em: <https://appevasaosbie.streamlit.app>. A aplicação tem finalidade acadêmica, demonstrativa e metodológica, não devendo ser interpretada como sistema definitivo de decisão institucional nem como predição individual de evasão.

Os microdados brutos oficiais do INEP/MEC não são redistribuídos no repositório e devem ser obtidos diretamente nas fontes públicas oficiais. Os microdados utilizados no Benchmark estão disponíveis em Censo da Educação Superior: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>.

Os *scripts* e componentes internos de implementação não são disponibilizados nesta versão por integrarem artefato institucional em desenvolvimento. Contudo, os procedimentos experimentais, variáveis, critérios de filtragem, divisão temporal, métricas e resultados consolidados são descritos no artigo e nos materiais suplementares, favorecendo transparência, auditabilidade e reprodutibilidade metodológica.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio de bolsa DS, Código de Financiamento 001. Foi desenvolvido junto ao Programa de Pós-Graduação em Ciência da Computação no Mestrado em Computação da Universidade Estadual de Londrina (UEL).

Declaração sobre uso de Inteligência Artificial

Os autores declaram que *Rayyan*, *ChatGPT*, *Codex* e *Zotero* foram utilizadas como apoio à revisão textual, organização e verificação de clareza do manuscrito. Todo o conteúdo científico, decisões metodológicas, análise dos dados, interpretação dos resultados, referências e versão final do artigo foram revisados, validados e assumidos integralmente pelos autores.

Cuidados Éticos

Este estudo utiliza revisão de literatura e dados oficiais agregados e públicos do INEP/MEC, sem coleta direta com seres humanos, intervenção com participantes ou tratamento de dados pessoais identificáveis. Assim, não houve acesso a informações individualizadas de estudantes nem possibilidade de identificação direta de sujeitos. O uso dos dados foi restrito a fins acadêmicos, com análise agregada e observância aos princípios de minimização, finalidade e transparência no tratamento de dados educacionais.

Referências

- Baker, R., Isotani, S., and Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de informática na educação*, 19(02):03.
- Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.
- Breiman, L. Random forests. 45(1):5–32.
- Carbonell, J. G., Michalski, R. S., and Mitchell, T. M. (1983). *An Overview of Machine Learning*, pages 3–23. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2025). Apresentação do censo da educação superior 2024: divulgação dos resultados. Apresentação institucional. Divulgação realizada em 17 set. 2025. Disponível em: https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2024/apresentacao_censo_da_educacao_superior_2024.pdf. Acesso em: 20 abr. 2026.
- Jones, K. M. (2019). Learning analytics and higher education: a proposed model for establishing informed consent mechanisms to promote student privacy and autonomy. *International Journal of Educational Technology in Higher Education*, 16(1):24.
- Kaufman, S., Rosset, S., and Perlich, C. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4):15:1–15:21.
- Kitchenham, B., Madeyski, L., and Budgen, D. (2022). Segress: Software engineering guidelines for reporting secondary studies. *IEEE Transactions on Software Engineering*, 49(3):1273–1298.
- Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer, New York.
- Long, P. and Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review (Online)*.
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillerá-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.