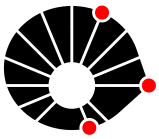

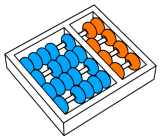

Predição de Evasão Escolar no Ensino Superior por meio de Técnicas de Inteligência Artificial

Lilian Fontan de Oliveira
Marcelo da Silva Reis

Relatório Técnico - IC-PFG-25-58
Projeto Final de Graduação
2025 - Dezembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Predição de Evasão Escolar no Ensino Superior por meio de Técnicas de Inteligência Artificial

Lilian Fontan de Oliveira[2,3]
Marcelo da Silva Reis[1,2,3]

[1] Hub de Inteligência Artificial e Arquiteturas Cognitivas (H.IAAC);
[2] Laboratório de Inteligência Artificial (recod.ai);
[3] Instituto de Computação, Universidade Estadual de Campinas (UNICAMP).
lilian.fontan@gmail.com; msreis@unicamp.br

Campinas, 19 de dezembro de 2025

Resumo

Este trabalho apresenta uma revisão sistemática da literatura dos últimos cinco anos sobre métodos de predição de evasão no ensino superior baseados em técnicas de inteligência artificial. Foram analisados doze estudos que empregaram metodologias variadas — desde algoritmos clássicos de machine learning até abordagens mais sofisticadas de deep learning — revelando diferentes graus de desempenho conforme o contexto institucional e os dados disponíveis. A maioria dos estudos utilizou estruturas de dados semelhantes, compostas principalmente por informações acadêmicas, socioeconômicas e demográficas. Nesse cenário, algoritmos baseados em árvores de decisão e técnicas de ensemble destacaram-se por oferecer um bom equilíbrio entre desempenho, interpretabilidade e complexidade computacional, apresentando resultados consistentes para esse tipo de aplicação. Como trabalho futuro, propõe-se investigar o impacto das políticas institucionais de permanência estudantil — como auxílios financeiros e moradia — sobre a probabilidade de evasão, integrando modelos preditivos com análises que permitam estimar o efeito potencial dessas intervenções.

Sumário

1 Introdução **3**
1.1 Organização Deste Trabalho 5

2 Metodologia **5**
2.1 Processo de Triagem 6

3 Resultados da Revisão Sistemática **6**
3.1 Modelos Clássicos de Machine Learning 7
3.2 Construção e Integração de Bases de Dados Educacionais 11
3.3 Modelos com Deep Learning ou Híbridos 12
3.4 Investigação do Momento Ideal para Realizar a Previsão de Risco 13

4 Discussão **15**
4.1 Perspectivas de Trabalhos Futuros 15

5 Conclusão **17**

Agradecimentos **18**

Referências **19**

1 Introdução

A evasão no ensino superior constitui um problema complexo e relevante, pois gera prejuízos não apenas para o estudante, mas também para a universidade e para a sociedade. Apesar de não haver um consenso absoluto sobre sua definição, a evasão pode ser entendida como a saída definitiva do estudante do curso sem sua conclusão.

Segundo Freitas [1], a evasão pode assumir diferentes formas, tais como o não retorno do aluno às atividades após um período, a desistência formal do curso, a transferência interna para outro curso dentro da mesma instituição, a exclusão por normas institucionais e a transferência para o mesmo curso em outra instituição. Tais distinções são fundamentais, pois delimitam o escopo de análise — diferenciando, por exemplo, a evasão do curso da evasão institucional.

Diversos fatores podem levar um estudante a evadir; entre os mais recorrentes encontram-se:

- dificuldades pessoais e de adaptação à vida universitária;
- falta de motivação ou baixa identificação com o curso escolhido;
- desempenho insuficiente e desafios acadêmicos persistentes;
- limitações econômicas que dificultam a permanência;
- questões familiares ou sociais que impactam a continuidade dos estudos.

Embora esses fatores sejam amplamente reconhecidos na literatura, dados institucionais da Unicamp evidenciam que muitos deles também aparecem no contexto local. A Figura 1, extraída do Relatório de Avaliação Institucional da Unicamp [2], apresenta os principais motivos informados por estudantes ativos que não renovaram a matrícula no primeiro semestre de 2022.

Figura 1: Motivos Apresentados por Estudantes Para Estarem sem Matrícula no 1° semestre de 2022. Fonte: Relatório Final da Avaliação Institucional UNICAMP 2019 - 2023 [2].

No contexto da Universidade Estadual de Campinas (UNICAMP) — reconhecida em 2024 pelo Times Higher Education (THE) como a segunda melhor universidade da América Latina — essa questão assume um peso ainda maior. Por ser uma instituição pública financiada majoritariamente por recursos do Estado, a saída prematura de um estudante implica a perda parcial do investimento público destinado à sua formação.

Em 2024, a taxa de evasão na Unicamp foi de 6,53%, o que corresponde a 1.295 estudantes que deixaram seus cursos apenas nesse ano. Embora o percentual possa parecer reduzido à primeira vista, o número absoluto evidencia a magnitude do problema. A Figura 2 mostra a evolução percentual da evasão na universidade ao longo dos últimos anos.

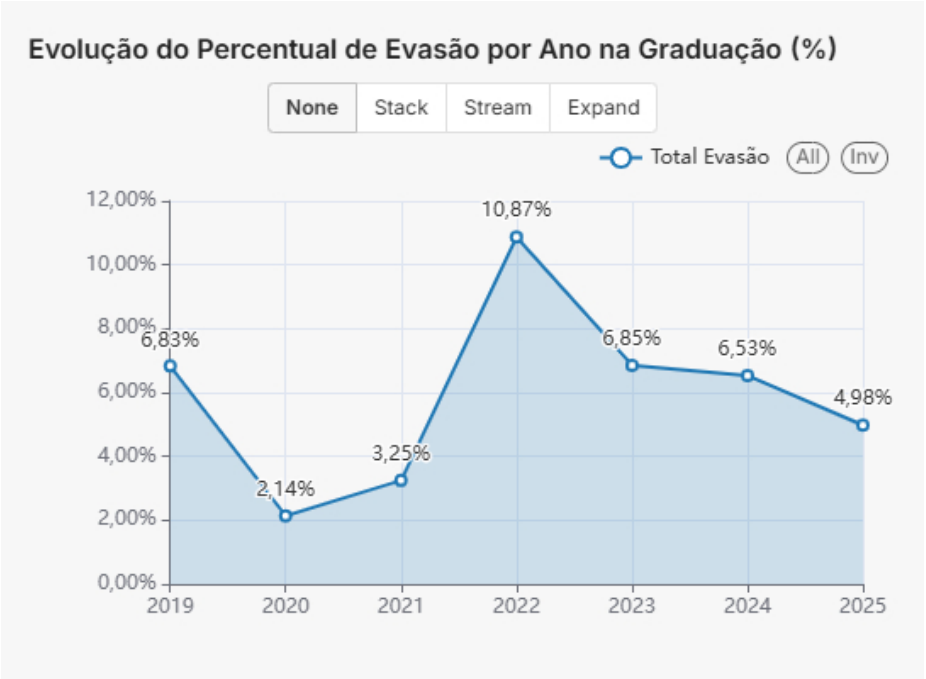

Figura 2: Evolução do percentual de evasão por ano na Graduação da Unicamp. Fonte: Geplanes – Indicadores Estratégicos [3].

Quando analisada por área do conhecimento, a evasão se mostra ainda mais significativa. Na área de Exatas, por exemplo, a taxa registrada em 2024 alcançou 10,08% nos cursos integrais e 11,83% nos cursos noturnos, revelando um impacto proporcionalmente maior nesse conjunto de cursos. A Figura 3 apresenta a evolução dessas proporções.

Diante desse cenário, a Unicamp tem buscado compreender e enfrentar o fenômeno da evasão. A identificação precoce de estudantes com risco de abandonar o curso constitui uma ferramenta essencial, pois permite não apenas ações interventivas em nível individual — atuando diretamente com cada aluno —, mas também uma análise ampla do problema. Essa perspectiva macro possibilita o desenvolvimento de políticas institucionais e públicas mais eficazes, capazes de reduzir a evasão de forma estrutural.

A proposta deste trabalho é realizar uma revisão sistemática da literatura composta por

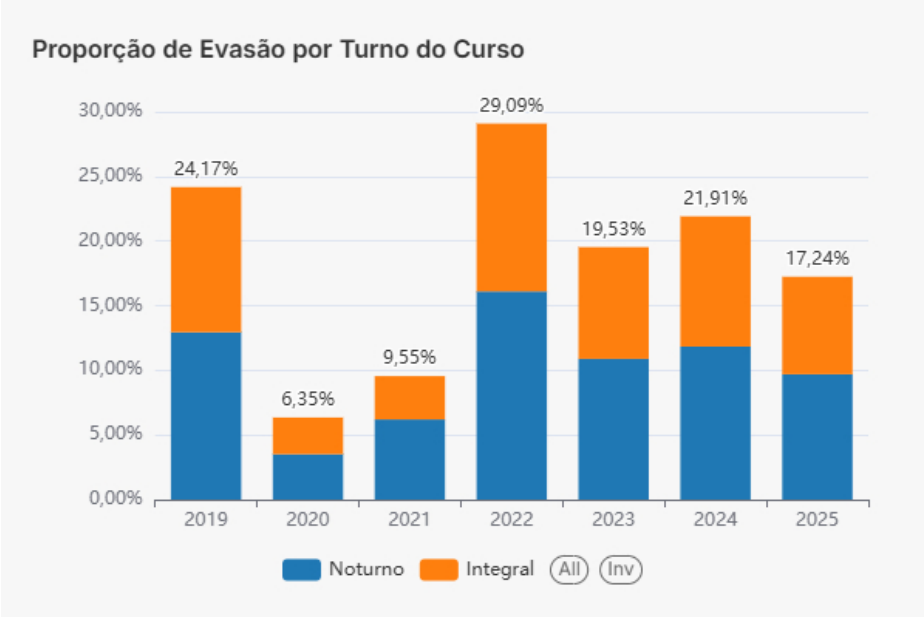

Figura 3: Proporção de evasão nos cursos da área de Ciências Exatas da Unicamp, separada por turno. A parte azul de cada barra representa a porcentagem de alunos dos cursos noturnos que evadiram, enquanto a parte laranja corresponde aos cursos integrais. O valor indicado acima de cada barra representa a soma das duas proporções. Fonte: Geplanes – Indicadores Estratégicos [3].

estudos de evasão alinhados ao contexto da Unicamp. Isso implica selecionar pesquisas que abordem a previsão de evasão em instituições de ensino superior com cursos presenciais. Além disso, como a Unicamp atualmente não dispõe de mecanismos para coletar dados comportamentais de estudantes em plataformas digitais, estudos baseados nesse tipo de informação não foram considerados.

1.1 Organização Deste Trabalho

A estrutura deste PFG está distribuída em quatro seções principais. A Seção 2 descreve os procedimentos metodológicos adotados para a condução da revisão sistemática e os critérios utilizados na seleção dos estudos. A Seção 3 apresenta a síntese e a avaliação dos artigos identificados, com foco em seus principais resultados. A Seção 4 analisa criticamente esses achados à luz do contexto da Unicamp e apresenta possíveis direções para trabalhos futuros. Por fim, a Seção 5 apresenta as considerações finais do estudo.

2 Metodologia

Para identificar estudos recentes sobre predição de evasão e desempenho acadêmico por meio de técnicas de inteligência artificial, foi realizada uma busca sistemática na base de dados

Scopus, considerada uma das principais plataformas indexadoras de literatura científica internacional.

A busca foi efetuada no dia 08 de novembro de 2025 utilizando a seguinte query, construída para combinar termos relacionados a inteligência artificial, educação e evasão estudantil:

```
    ("artificial intelligence" OR "machine learning")
AND ("educational data mining" OR "learning analytics")
AND ("student dropout" OR "student retention"
     OR "student persistence" OR "academic attrition")
AND ("academic performance" OR "student performance"
     OR "learning outcomes" OR "academic achievement"
     OR "grades" OR "student success")
```

Foram considerados estudos publicados entre 2020 e 2025, redigidos em português ou inglês, e classificados na Scopus como Article, Conference Paper ou Review.

2.1 Processo de Triagem

Como o objetivo deste trabalho é identificar metodologias aplicáveis ao contexto da Unicamp, foram excluídos estudos que analisavam ensino remoto, híbrido ou totalmente online, bem como aqueles que dependiam de dados comportamentais extraídos de Learning Management Systems (LMS), como Moodle ou Blackboard — tipos de dados que não são atualmente coletados pela instituição. Além disso, foram considerados apenas estudos conduzidos no ensino superior presencial.

Após a filtragem inicial, os títulos e resumos dos artigos retornados foram avaliados, e aqueles potencialmente relevantes foram lidos integralmente.

3 Resultados da Revisão Sistemática

Nesta seção, apresentaremos e discutiremos trabalhos que selecionamos da busca sistemática. Essa busca identificou doze estudos publicados entre 2020 e 2025 que abordaram a previsão de evasão ou desempenho acadêmico no ensino superior por meio de técnicas de inteligência artificial. Esses trabalhos podem ser organizados em quatro categorias principais:

- estudos baseados em algoritmos clássicos de Machine Learning - Seção 3.1;
- pesquisas focadas na construção e integração de bases de dados educacionais - Seção 3.2;
- abordagens que empregam deep learning ou modelos híbridos - Seção 3.3; e
- estudos que investigam o momento ideal para realizar a previsão de risco - Seção 3.4.

Nas subseções seguintes, apresentamos os principais achados de cada grupo.

3.1 Modelos Clássicos de Machine Learning

Uma quantidade considerável dos estudos que encontramos utilizou algoritmos clássicos de Machine Leaning incluindo modelos lineares, métodos de vizinhança, SVM e, com especial frequência, técnicas baseadas em árvores de decisão e ensembles — como Random Forest, Gradient Boosting e XGBoost. Esses modelos apareceram com grande frequência porque se comportam bem com dados tabulares, apresentam interpretabilidade relativamente elevada e boa performance mesmo com a questão inerente da evasão de desbalanceamento entre as classes

Em 2020, Ribeiro e Canedo trataram da previsão de evasão da Universidade de Brasília (UnB) [4]. Em alguns cursos, a UnB apresentava uma taxa alarmante de mais de 50% de evasão. Os autores coletaram dados acadêmicos e sociais de estudantes de graduação que concluíram ou evadiram entre 2006 e 2018, utilizando-os para treinar modelos Generalized Linear Model (GLM), Gradient Boosting Machine (GBM), Support Vector Machines (SVM) e Random Forest (RF). A partir da comparação das confusion tables e das métricas de validação, concluíram que o modelo GBM apresentou o melhor desempenho, com AUC próximo de 86% e menor taxa de erro entre as classes. Além disso, os autores identificaram os principais fatores associados à evasão, destacando que variáveis de desempenho acadêmico e fatores sociais foram as mais relevantes.

Uma limitação importante do estudo é a ausência de uma análise temporal da evasão: todos os alunos evadidos foram tratados como um único grupo, sem considerar em que momento da trajetória a evasão ocorreu. Outra limitação é a falta de uma discussão sobre o desbalanceamento das classes. Embora alguns cursos apresentem evasão superior a 50%, isso não implica que o conjunto de dados consolidado mantenha essa proporção. É provável que o dataset utilizado seja desbalanceado, e a ausência de qualquer tratamento ou análise desse aspecto pode ter impactado as métricas preditivas reportadas.

Assim como Ribeiro e Canedo, outros autores também investigaram algoritmos clássicos de Machine Learning aplicados à evasão. No artigo de Gonzalez *et al.* [5], os autores realizaram um estudo considerando 14.495 estudantes de uma universidade privada do México, cuja taxa de evasão é de 8,5%. O objetivo central do trabalho foi identificar os fatores mais relevantes associados à evasão, de modo que a instituição pudesse intervir de forma eficiente junto aos alunos com maior risco.

O estudo avaliou oito classificadores distintos e concluiu que o Random Forest apresentou o desempenho mais interessante entre eles, pois ofereceu um bom equilíbrio entre a precisão na classe de evasão e a precisão na classe de retenção, além de manter uma alta acurácia geral. Outro ponto decisivo para sua escolha foi a capacidade do Random Forest de fornecer uma análise clara da importância das variáveis, algo essencial para identificar fatores críticos de risco.

Um aspecto importante do artigo é a forma como os autores lidam com o desbalanceamento dos dados — um fenômeno inerente ao problema de evasão no ensino superior. Para isso, adotaram a abordagem conhecida como Probability Threshold Approach, que consiste em ajustar o limiar de probabilidade utilizado para classificar as instâncias. Em vez de utilizar o valor convencional de 50%, o limiar é deslocado para favorecer a detecção da classe minoritária (evasão), equilibrando melhor a sensibilidade entre as duas classes. Esse

método se destaca por dispensar técnicas como undersampling e oversampling, evitando a geração ou remoção artificial de dados.

As análises permitiram concluir que as variáveis mais relevantes para prever evasão foram: (a) a média obtida no primeiro período do primeiro semestre, (b) a média do nível anterior (Previous Level Average, PNA), e (c) o desempenho na avaliação geral de matemática. Além disso, utilizando um threshold diferente de 50%, os autores alcançaram uma precisão de 81% para retenção, 51% para evasão e uma precisão global de 78,5%.

Embora o artigo apresente uma contribuição relevante, suas conclusões devem ser interpretadas à luz de algumas limitações importantes. A primeira delas é que o conjunto de dados inclui apenas estudantes do primeiro ano, não contemplando estudantes veteranos, cuja dinâmica de evasão pode ser distinta. A segunda limitação é que o estudo utilizou apenas variáveis quantitativas, deixando de fora fatores qualitativos como características socioeconômicas ou aspectos emocionais. Além disso, diversos registros foram removidos do dataset devido à ausência de valores completos, o que pode introduzir viés de seleção.

Assim como Gonzalez *et al.*, Martinez Neda *et al.* [6] também exploram modelos preditivos clássicos, mas com foco na identificação precoce de estudantes com risco de entrar em academic probation na UCI. Os autores investigam a previsão dos estudantes de Ciência da Computação da Universidade da Califórnia, Irvine (UCI), com maior risco de entrar em academic probation — situação em que o aluno apresenta desempenho insuficiente e pode até ser desligado do curso — ainda no primeiro ano da graduação.

Para isso, foram avaliados diversos classificadores de Machine Learning, incluindo Logistic Regression, Linear SVM, Random Forest, Naive Bayes, AdaBoost, Gradient Boost, XGBoost, k-NN e Decision Trees. O objetivo central era maximizar o recall, isto é, identificar o maior número possível de estudantes que realmente entrarão em probation, mesmo que isso aumentasse o número de falsos positivos.

Entre todos os modelos testados, os melhores resultados foram obtidos pelo Linear SVM (recall de 0,838) e pela Logistic Regression (recall de 0,812). Embora alguns modelos apresentassem acurácia ligeiramente maior, esse indicador não era adequado para o estudo devido ao forte desbalanceamento do conjunto de dados (apenas 24% dos estudantes entram em probation). Assim, priorizar recall era mais apropriado para o objetivo de identificar alunos em risco.

O estudo também revelou dois fatores especialmente importantes para a previsão: (i) o estudante não ter realizado o exame AP CS A — o que indica menor experiência prévia com programação — e (ii) variáveis demográficas, que apresentaram impacto maior que as notas e demais indicadores acadêmicos prévios. Os autores argumentam que esses padrões refletem desigualdades estruturais de acesso à educação, e não características individuais dos estudantes, reforçando a necessidade de intervenções institucionais mais direcionadas.

Embora esse estudo apresente contribuições importantes, é importante levar em consideração que o conjunto de dados analisado é restrito a estudantes de Ciência da Computação, o que pode prejudicar a generalização dos resultados. Além disso, como a maior parte das variáveis refere-se ao período anterior ao ingresso, o modelo não captura elementos importantes do desempenho ao longo do primeiro ano, o que limita seu potencial para intervenções realmente adaptativas durante o curso.

Diferentemente do estudo anterior, focado na previsão precoce baseada em características

prévias ao ingresso, Kumar *et al.* [7] adotam uma abordagem mais tradicional, comparando o Random Forest a outros métodos clássicos de machine learning e buscando maximizar o desempenho global do modelo. O artigo investiga a evasão acadêmica utilizando o algoritmo Random Forest, com otimização por meio do processo de Grid Search. O estudo emprega um conjunto de dados composto por 5.018 estudantes distribuídos em 12 cursos presenciais, obtidos de diferentes campi da instituição analisada. Na comparação com outros métodos clássicos de machine learning — Decision Tree, KNN, Gradient Boosting, Naive Bayes e SVM — o modelo Random Forest apresentou o melhor desempenho, alcançando acurácia de 0,955 e F1-score de 0,955. Um resultado particularmente relevante é a taxa de falso negativo próxima de zero na validação, o que indica que praticamente nenhum aluno em risco de evasão deixaria de ser identificado pelo modelo.

Entretanto, os autores não especificam quais hiperparâmetros foram incluídos no processo de Grid Search, o que limita a compreensão detalhada da otimização realizada e compromete parcialmente a reprodutibilidade do estudo.

Ainda dentro das abordagens baseadas em modelos clássicos de Machine Learning, Bussaman *et al.* [8] ampliam a análise utilizando métodos de Machine Learning para prever os riscos de evasão dos estudantes da Faculty of Science and Technology da Rajabhat Maha Sarakham University, na Tailândia. O estudo abrange 2.361 estudantes distribuídos em cinco cursos entre os anos de 2010 e 2022. Os autores buscaram não apenas identificar estudantes com risco de evasão, mas também aqueles que não se formariam no tempo mínimo. Para isso, os estudantes foram classificados em quatro categorias: On schedule (formados dentro do prazo), Not on schedule (formados com atraso), Dropped out (evadidos sem conclusão) e Resigned (desligamento formal do curso).

Seis algoritmos supervisionados foram avaliados: Decision Tree, Naive Bayes, Neural Networks, Gradient Boosting, Random Forest e Majority Voting. O melhor desempenho foi obtido pelo modelo de Majority Voting, que alcançou acurácia de 88.14%. Em relação ao desempenho por classe, destacaram-se os valores de F1-score: 94.07% para On schedule, 85.50% para Dropped out e 66.06% para Resigned. No entanto, o desempenho para a classe Not on schedule foi consideravelmente inferior, com F1-score de apenas 36.30%, revelando grande dificuldade em identificar estudantes que se formam com atraso.

Apesar dessa limitação, a proposta de distinção entre atraso, evasão e conclusão regular é bastante relevante. Um estudo dentro dessa perspectiva, aprimorando essa análise poderia ser muito benéfico nos contextos das universidades públicas não apenas para tentar evitar a evasão mas também para evitar o atraso dos alunos, na tentativa de otimizar os recursos públicos.

Finalmente, encerrando os estudos de baseados em algoritmos clássiscos de Machine Learning, o artigo de Kuntintara *et al.* [9] apresenta um estudo sobre a evasão dos estudantes de graduação da Bangkok University. Foram considerados dados demográficos, acadêmicos, financeiros e comportamentais (incluindo histórico de uso de Wi-Fi, obtido por meio de registro de MAC address) dos ingressantes de 2021. Para a tarefa preditiva, os autores avaliaram sete algoritmos: Logistic Regression, KNN, SVM, Decision Tree, Random Forest, Gradient Boosting e XGBoost. A principal métrica utilizada na comparação foi a área sob a curva Precision–Recall (AUC-PR), apropriada para bases desbalanceadas. Com base nesses resultados, realizou-se um ajuste de hiperparâmetros via Grid Search, e o XGBoost

apresentou o melhor desempenho (AUC-PR = 88.44% e F1 = 81.79%). Após essa etapa, os autores aplicaram técnicas de balanceamento, combinando oversampling e undersampling. Finalmente, o modelo resultante foi utilizado para prever a evasão dos estudantes do ano subsequente, alcançando uma acurácia de 84.96

No estudo, o balanceamento dos dados foi realizado apenas após a etapa de comparação dos algoritmos e após o ajuste de hiperparâmetros via Grid Search. Essa decisão metodológica pode ter impactado a seleção do modelo, pois a comparação inicial foi realizada em um cenário desbalanceado, favorecendo algoritmos mais robustos a esse problema, como o XGBoost, e prejudicando modelos mais sensíveis, como SVM e Logistic Regression. Além disso, o Grid Search foi executado sobre dados desbalanceados, enquanto o modelo final foi treinado em dados balanceados, o que pode gerar inconsistências entre o processo de otimização e o cenário efetivo utilizado na predição.

O artigo de Vemulapalli *et al.* [10] investiga modelos híbridos de Machine Learning para prever o desempenho acadêmico de estudantes do ensino superior. Inicialmente, os autores comparam quatro modelos individuais — LSTM (Long Short-Term Memory), XGBoost (Extreme Gradient Boosting), RNN (Recurrent Neural Network) e CNN (Convolutional Neural Network). Entre eles, os melhores desempenhos foram obtidos pelo LSTM, com acurácia de 90,81%, e pelo XGBoost, com acurácia de 93,76%. A partir desses resultados, os autores propõem um modelo híbrido que combina as características sequenciais capturadas pela LSTM com o poder de classificação do XGBoost. Esse modelo integrado apresenta desempenho superior aos demais, alcançando acurácia de 96,54

Um ponto crítico do estudo é que a descrição do conjunto de dados é excessivamente superficial. O artigo não informa o tamanho da amostra, a instituição de origem, o contexto educacional, nem se os dados utilizados são reais ou simulados. A ausência desses elementos compromete a reprodutibilidade e dificulta uma avaliação rigorosa da validade externa dos resultados.

Já no artigo de Alejandra *et al.* [11] tem como objetivo identificar as melhores condições de aplicação de técnicas de machine learning para a previsão de evasão escolar, concentrando-se em dois aspectos principais: o balanceamento das classes e a otimização de hiperparâmetros para aprimorar o desempenho dos classificadores. O estudo utilizou um dataset com 4424 instâncias e 34 atributos, normalizado por meio do RobustScaler, e aplicou três técnicas de reamostragem: ADASYN, que realiza oversampling adaptativo ao gerar mais instâncias sintéticas em regiões de difícil aprendizado; SVM-SMOTE, uma variante do SMOTE voltada para identificar instâncias de fronteira; e SMOTE + ENN, um método híbrido que combina oversampling com a limpeza de ruídos via Edited Nearest Neighbors.

Três modelos foram avaliados: Random Forest (RF), Support Vector Machine (SVM) e XGBoost, sendo cada um testado tanto com hiperparâmetros padrão quanto com otimização via Bayesian Optimization. Os resultados indicam de forma consistente que a otimização de hiperparâmetros melhora todas as métricas analisadas, independentemente da técnica de balanceamento ou do classificador utilizado. Entre os modelos, o XGBoost apresentou o melhor desempenho, alcançando um F1-score de 93,39%, valor consideravelmente superior ao observado na literatura relacionada.

A abordagem de otimização de hiperparâmetros proposta pelos autores se destaca como um diferencial relevante, especialmente porque a grande maioria dos estudos semelhan-

tes atinge F1-scores próximos de 75%, enquanto a otimização elevou o desempenho para patamares acima de 90%. A continuidade dessa linha de pesquisa — explorando outras estratégias de busca e ajuste fino — tem potencial para gerar resultados ainda mais expressivos e consolidar práticas mais robustas para a previsão de evasão.

Os estudos analisados mostraram uma gama bem diversa de modelos clássicos de Machine Learning, mostrando que eles ainda apresentam um papel central na previsão de evasão e desempenho acadêmico, apresentando bons resultados bastante interessantes. Entretanto, existem desafios que foram considerados principalmente na confecção dos datasets para esses estudos. Na próxima seção, discutimos estudos que se dedicam justamente a esse desafio, concentrando-se na construção e integração de datasets educacionais voltados à previsão de evasão.

3.2 Construção e Integração de Bases de Dados Educacionais

Uma parte importante da literatura não se restringe à aplicação de modelos preditivos, mas volta-se à própria construção das bases de dados que viabilizam essas análises. Nesse contexto, destaca-se o trabalho de Realinho *et al.* [12], que apresenta o desenvolvimento de um dataset abrangente voltado à previsão de evasão e desempenho estudantil utilizando técnicas de Machine Learning.

Para a construção desse dataset, os autores integraram dados provenientes de diversas fontes institucionais, estruturando-os em quatro grupos principais de atributos: demográficos, socioeconômicos, macroeconômicos e acadêmicos. O conjunto possui três variáveis-alvo — Dropout (evasão), Enrolled (ainda matriculado após o período padrão) e Graduate (Graduado) — permitindo múltiplos cenários de modelagem preditiva. O processo de preparação dos dados incluiu etapas de limpeza, deduplicação, seleção de atributos relevantes e balanceamento das classes, garantindo que variáveis irrelevantes ou desbalanceamentos estruturais não prejudicassem o desempenho dos modelos. O resultado final foi um dataset com 4.424 estudantes e 35 atributos, abrangendo cursos de diferentes áreas. Embora o dataset seja útil para estudos comparativos sobre métodos de previsão de evasão, é importante reconhecer que ele reflete a realidade de uma universidade portuguesa, onde fatores institucionais, econômicos e culturais — como o pagamento de mensalidades — diferem significativamente do contexto da Unicamp e de outras universidades públicas brasileiras. Por esse motivo, o dataset não pode ser aplicado diretamente para treinar modelos voltados à nossa realidade. Ainda assim, a metodologia de construção, especialmente o rigor no tratamento, integração e balanceamento dos dados, representa uma referência valiosa e poderia orientar iniciativas semelhantes no contexto brasileiro. Nesse sentido, seria altamente relevante que universidades públicas disponham de um dataset estruturado com esse nível de qualidade para pesquisas sobre evasão e desempenho estudantil.

Esse tipo de integração de dados desempenha um papel essencial no desenvolvimento de modelos preditivos robustos, pois define os limites e o potencial das análises posteriores. A partir desse ponto, a literatura também revela estudos que, em vez de focar apenas na construção das bases de dados, buscam explorar arquiteturas mais complexas — incluindo métodos de deep learning e combinações híbridas — para aprimorar a capacidade de previsão.

3.3 Modelos com Deep Learning ou Híbridos

Além dos estudos que focam na construção de bases de dados e no uso de algoritmos clássicos, a literatura recente também explora abordagens mais sofisticadas, baseadas em deep learning e em arquiteturas híbridas. Esses modelos buscam capturar padrões complexos e relações não lineares nos dados, especialmente em cenários de grande escala ou alta dimensionalidade.

Dentro deste contexto, o artigo de [13] apresenta um estudo baseado em mais de dez anos de dados acadêmicos reais da Louisiana State University (LSU), com o objetivo de prever a probabilidade de conclusão do curso por parte dos estudantes. Um dos pontos muito relevantes deste trabalho é o extenso pré-processamento dos dados, no qual é feito um processo de limpeza e utiliza-se estratégias de imputação dependo da natureza da variável como, por exemplo, conversão de CEPs para coordenadas geográficas. Esse tratamento tão detalhado minimiza distorções nos modelos primitivos.

A partir desse dataset, os autores propõem um framework baseado em Convolutional Autoencoders (CAE) com seis camadas convolucionais, utilizando-o como mecanismo de redução e extração de atributos. O objetivo é transformar os 197 atributos originais em representações latentes mais compactas — no caso, 141 atributos — preservando relações não lineares e padrões complexos associados ao desempenho e trajetória acadêmica dos estudantes. Essas representações são então utilizadas como entrada para o modelo preditivo final, um Random Forest.

O estudo compara o desempenho do Random Forest utilizando os atributos originais com o desempenho obtido ao utilizar os atributos extraídos pelo CAE. O Random Forest treinado com as variáveis originais apresentou o melhor resultado, alcançando uma AUC de aproximadamente 0,90. Já o modelo alimentado pelos atributos latentes do CAE obteve uma AUC levemente inferior (cerca de 0,87), mas com ganhos relevantes em termos de eficiência computacional, sendo mais leve e mais rápido de treinar. Assim, apesar da pequena perda de desempenho, o CAE demonstrou capacidade de preservar padrões essenciais e reduzir substancialmente a dimensionalidade.

Esse estudo apresenta uma abordagem inovadora que evidencia o potencial de técnicas avançadas de redução de dimensionalidade em cenários educacionais de grande escala, mas poderia ir além, explorando também o uso dessas representações em modelos preditivos mais sofisticados ou híbridos, e não apenas como etapa de pré-processamento.

Já o artigo de Vemulapalli *et al.* [10] avança no uso de técnicas mais complexas ao empregar modelos híbridos de Machine Learning para prever o desempenho acadêmico de estudantes do ensino superior. Inicialmente, os autores comparam quatro modelos individuais — LSTM (Long Short-Term Memory), XGBoost (Extreme Gradient Boosting), RNN (Recurrent Neural Network) e CNN (Convolutional Neural Network). Entre eles, os melhores desempenhos foram obtidos pelo LSTM, com acurácia de 90,81%, e pelo XGBoost, com acurácia de 93,76%. A partir dessa análise, os autores propõem um modelo híbrido que combina as capacidades de extração de padrões temporais da LSTM com o poder de classificação do XGBoost. Esse modelo integrado apresenta desempenho superior aos demais, alcançando acurácia de 96,54%.

Um ponto crítico do estudo é que a descrição do conjunto de dados é excessivamente

superficial. O artigo não informa o tamanho da amostra, a instituição de origem, o contexto educacional, nem se os dados utilizados são reais ou simulados. A ausência desses elementos compromete a reprodutibilidade e dificulta uma avaliação rigorosa da validade externa dos resultados.

Seguindo ainda dentro das abordagens que empregam modelos mais sofisticados, o artigo de Zanellati *et al.* [14] apresenta um estudo que busca conciliar alto desempenho preditivo com boa capacidade de explicação dos modelos, aspecto particularmente relevante no contexto de evasão acadêmica. Para isso, os autores comparam dois modelos — Random Forest (RF) e Feature Tokenizer Transformer (FTT) — utilizando uma base de dados composta por 44.875 estudantes de uma universidade italiana, abrangendo os anos de 2018 a 2021 e diversos cursos. Foram construídos cinco datasets temporais correspondentes a 0, 3, 6, 9 e 12 meses após a matrícula.

No que diz respeito ao desempenho preditivo, o FTT superou o RF em praticamente todas as métricas, o que é consistente com a natureza dos modelos: enquanto o RF tende a capturar padrões mais simples e robustos, o FTT é capaz de modelar relações complexas entre variáveis, como ocorre nesse conjunto de dados. O melhor desempenho foi obtido pelo FTT aos 12 meses, com acurácia de 0,87 e sensibilidade de 0,81.

Quanto à explicabilidade, os autores também observaram resultados superiores para o FTT. Em particular, as técnicas aplicadas apontaram de forma consistente que o número de créditos acumulados (ECTS) é a variável mais influente, sendo que estudantes com ECTS baixos apresentam maior risco de evasão.

Apesar da solidez metodológica, uma limitação importante do estudo é que o melhor desempenho ocorre aos 12 meses, quando boa parte das evasões já se concretizou — inclusive segundo os próprios autores. Nesse momento, a possibilidade de intervenção institucional é reduzida, o que compromete a utilidade prática do modelo para ações preventivas. Além disso, o FTT apresenta custo computacional significativamente mais elevado, o que restringe sua adoção em cenários em que intervenções mais precoces são desejáveis.

Os estudo mostrados evidenciam que abordagens baseadas em deep learning e modelos híbridos têm potencial para ampliar a capacidade preditiva em cenários educacionais, especialmente quando lidam com bases extensas ou relações não lineares. Além dessas questões, um aspecto cada vez mais relevante é o momento em que a previsão é realizada. Por essa razão, diversos estudos recentes têm voltado sua atenção não apenas ao como prever, mas ao quando prever — tema discutido na próxima subseção.

3.4 Investigação do Momento Ideal para Realizar a Previsão de Risco

Além da escolha do modelo preditivo, um aspecto crucial destacado na literatura recente diz respeito ao momento em que a previsão é realizada. Mesmo modelos altamente precisos perdem utilidade prática se identificam o risco apenas quando a evasão já está próxima ou consumada. Nesse contexto, o estudo de Monica Martins *et al.* [15] busca justamente determinar qual é o momento mais adequado do primeiro ano acadêmico para prever desempenho e risco de evasão.

O artigo de Monica Martins *et al.*[15] tem como objetivo identificar qual é o melhor momento do primeiro ano acadêmico para realizar previsões de desempenho e risco de

evasão. Para isso, os autores dividem os dados em três conjuntos: o primeiro (S0), contendo apenas as informações disponíveis logo após a matrícula; o segundo (S1), contendo também o desempenho acadêmico do primeiro semestre; e o terceiro (S2), que adiciona as notas do segundo semestre. É importante notar que, se um estudante desiste entre o final do primeiro e o segundo semestre, seus registros aparecem em S1, mas não em S2, o que reduz ainda mais a representatividade da classe "Dropout" neste último conjunto. Em todos os datasets, os estudantes são classificados em três categorias: "Graduate" (conclusão no tempo previsto), "Enrolled" (permanência após o tempo regular) e "Dropout" (evasão).

Para cada uma dessas versões do dataset, os autores testaram cinco modelos distintos: SMOTE + Random Forest, SVMSMOTE + Random Forest, RUSBoost, Balanced Random Forest (BRF) e Easy Ensemble (EE). Os três primeiros abordam o desbalanceamento diretamente no nível dos dados, por meio da inserção de exemplos sintéticos ou remoção de exemplos majoritários, enquanto os dois últimos tratam o desbalanceamento no próprio algoritmo, utilizando subamostragem interna da classe majoritária.

A avaliação foi conduzida utilizando o F1-score global, o F1-score por classe e a balanced accuracy. O F1-score corresponde à média harmônica entre precisão e recall, o que é especialmente relevante em cenários desbalanceados. O melhor desempenho obtido pelos autores ocorreu com o conjunto S1, atingindo um F1-score máximo de 0,745. A conclusão apresentada é que, embora S2 contenha maior quantidade de dados — o que intuitivamente poderia sugerir um melhor desempenho — o desbalanceamento acentuado prejudicou a performance dos modelos. Além disso, os melhores resultados foram obtidos com modelos baseados em Random Forest. Os autores destacam ainda que, ao final do primeiro semestre, as variáveis acadêmicas tornam-se substancialmente mais relevantes que variáveis demográficas ou socioeconômicas para fins preditivos.

Este estudo apresenta uma abordagem diferenciada ao realizar experimentos não apenas ao final do primeiro ano, mas também ao final do primeiro semestre, demonstrando inclusive resultados superiores nesta fase intermediária. Isso significa que é possível identificar estudantes em risco de forma mais precisa e, sobretudo, mais precoce, permitindo intervenções institucionais em um momento em que ainda há maior chance de impactar positivamente sua trajetória acadêmica. Nesse sentido, a Unicamp poderia se beneficiar significativamente de uma investigação semelhante, já que também dispõe das notas do primeiro semestre como primeiro indicador concreto de desempenho dos estudantes.

Os estudos analisados evidenciam que a previsão de evasão e desempenho acadêmico tem evoluído de modo substancial, abrangendo desde modelos clássicos de Machine Learning até arquiteturas híbridas e abordagens baseadas em deep learning. Cada uma dessas linhas de pesquisa oferece contribuições importantes para lidar com um problema tão complexo quanto relevante nas instituições de ensino superior.

Na seção de Discussão 4, examinamos essas contribuições e limitações de forma integrada, identificando convergências, lacunas e oportunidades para pesquisas futuras — especialmente à luz do contexto das universidades públicas brasileiras.

4 Discussão

Os estudos analisados revelam uma expressiva diversidade de métodos aplicados à previsão de evasão no ensino superior. A Tabela 1 mostra que não existe um consenso sobre um único modelo superior; ao contrário, diferentes técnicas alcançam melhor desempenho dependendo do contexto institucional, da natureza das variáveis disponíveis e das escolhas metodológicas de cada pesquisa. Observa-se que a literatura combina desde modelos clássicos de *machine learning* até abordagens mais recentes de *deep learning* e modelos híbridos, refletindo a complexidade multifatorial do fenômeno da evasão e a necessidade de abordagens flexíveis capazes de capturar padrões variados.

Um aspecto relevante é a forte presença de algoritmos baseados em árvores de decisão e ensembles, como Random Forest, Gradient Boosting e XGBoost. A recorrente eficácia desses modelos está associada à sua capacidade de lidar com dados tabulares, heterogêneos e frequentemente desbalanceados — características típicas de bases educacionais. Esses modelos também oferecem interpretabilidade razoável e robustez estatística, o que os torna candidatos naturais em aplicações institucionais.

A Tabela 2 evidencia que não há apenas diversidade de modelos, mas também de *métricas* de avaliação. Enquanto alguns trabalhos priorizam a acurácia global, outros dão maior ênfase a métricas mais sensíveis ao desbalanceamento, como recall, F1-score ou AUC-PR. O conjunto de resultados reforça que a escolha da métrica é tão importante quanto a escolha do modelo, pois define qual tipo de erro é mais tolerável em cada contexto institucional.

Além dos modelos baseados em árvores, alguns estudos exploraram técnicas mais sofisticadas, como o Feature Tokenizer Transformer (FTT), autoencoders convolucionais (CAE) e arquiteturas híbridas que combinam redes neurais com modelos tradicionais — como no caso de Vemulapalli *et al.* [10], cujo modelo LSTM + XGBoost alcançou acurácia de 96,54%. Tais abordagens tendem a apresentar desempenhos superiores quando há grande volume de dados, alta complexidade relacional e possibilidade de exploração de estruturas temporais. Contudo, também requerem maior custo computacional e maior capacidade de infraestrutura, fatores que precisam ser ponderados pelas instituições.

Dessa forma, embora não haja um modelo universalmente superior, emergem padrões consistentes: métodos baseados em árvores são altamente competitivos em cenários gerais, enquanto redes profundas e modelos híbridos se destacam quando há disponibilidade de dados extensos e necessidade de capturar relações intrincadas. Para instituições que buscam sistemas de alerta precoce — como seria o caso da Unicamp — os estudos sugerem que modelos de complexidade moderada, treinados com dados do primeiro semestre, tendem a oferecer o melhor equilíbrio entre desempenho e aplicabilidade prática.

4.1 Perspectivas de Trabalhos Futuros

Os resultados levantados nesta revisão mostram que é possível prever a evasão com boa precisão e também identificar os fatores que mais contribuem para esse risco. Um caminho natural para pesquisas futuras consiste em avançar do diagnóstico para a avaliação de impacto de políticas institucionais, isto é, estimar como diferentes intervenções poderiam

Tabela 1: Algoritmos de Machine Learning utilizados nos estudos.

Estudo	**Algoritmos testados**	**Melhor desempenho**
Ribeiro & Canedo (2020) [4]	GLM, GBM, SVM, Random Forest	GBM
Shoorangiz *et al.* (2025) [13]	Convolutional Autoencoder (CAE) e Random Forest	Convolutional Autoencoder (CAE) e Random Forest
Martinez Neda *et al.* (2023) [6]	Logistic Regression, Linear SVM, Random Forest, Naive Bayes, AdaBoost, Gradient Boost, XGBoost, k-NN, Decision Trees	Linear SVM
Gonzalez *et al.* (2023) [5]	Oito classificadores distintos	Random Forest
Martins *et al.* (2023) [15]	SMOTE + RF, SVM-SMOTE + RF, RUSBoost, Balanced Random Forest, Easy Ensemble	Balanced Random Forest (S1)
Cuevas Chávez *et al.* (2024) [11]	Random Forest, SVM, XGBoost	XGBoost
Zanellati *et al.* (2024) [14]	Random Forest, Feature Tokenizer Transformer (FTT)	FTT
Kumar *et al.* (2024) [7]	Random Forest, Decision Tree, KNN, Gradient Boosting, Naive Bayes, SVM	Random Forest
Bussaman *et al.* (2024) [8]	Decision Tree, Naive Bayes, Neural Networks, Gradient Boosting, Random Forest, Majority Voting	Majority Voting
Kuntintara *et al.* (2023) [9]	Logistic Regression, KNN, SVM, Decision Tree, Random Forest, Gradient Boosting, XGBoost	XGBoost
Vemulapalli *et al.* (2025) [10]	LSTM, RNN, CNN, XGBoost; modelo híbrido LSTM + XGBoost	Modelo híbrido LSTM + XGBoost

Tabela 2: Métricas de desempenho dos melhores modelos por estudo.

Estudo	**Métrica**	**Valor**	**Observação**
Ribeiro & Canedo (2020) [4]	Acurácia	86%	GBM apresentou melhor desempenho geral
Shoorangiz *et al.* (2025) [13]	AUC	85%	Apesar da inclusão do CAE ter reduzido em desempenho, ganhou em custo computacional
Martinez Neda *et al.* (2023) [6]	Recall	0,838 (SVM)	Base altamente desbalanceada; recall era fundamental
Gonzalez *et al.* (2023) [5]	Acurácia	78,5%	Probability Threshold Approach para lidar com desbalanceamento
Martins *et al.* (2023) [15]	F1-score	0,745 (S1)	S1 superou S2 devido ao desbalanceamento intenso
Cuevas Chávez *et al.* (2024) [11]	F1-score	93,39%	XGBoost otimizado via Bayesian Optimization
Zanellati *et al.* (2024) [14]	Acurácia	0,87 (12 meses, FTT)	Alta precisão, mas previsão tardia
Kumar *et al.* (2024) [7]	Acurácia / F1	0,955	Falso negativo praticamente zero
Bussaman *et al.* (2024) [8]	Acurácia	88,14%	Melhor desempenho via Majority Voting
Kuntintara *et al.* (2023) [9]	AUC-PR	88,44%	XGBoost foi o melhor modelo
Vemulapalli *et al.* (2025) [10]	Acurácia	96,54%	Modelo híbrido LSTM + XGBoost superou os demais

alterar a probabilidade de evasão prevista pelos modelos.

Esse tipo de análise é particularmente relevante no contexto da Unicamp. A Figura 4 ilustra a evolução da evasão geral da universidade entre 2019 e 2025, comparada à evasão entre estudantes bolsistas. Observa-se que, apesar das variações significativas na evasão total — chegando a 10,87% em 2022 — a evasão entre bolsistas permanece consistentemente muito baixa, sempre inferior a 1%. Embora tais dados não permitam inferir causalidade direta, eles sugerem que as políticas de permanência podem desempenhar um papel importante na retenção estudantil.

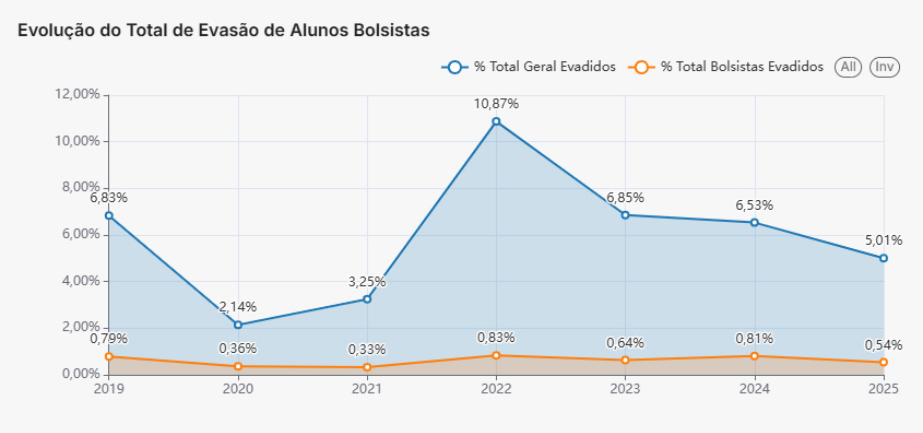

Figura 4: Evasão na Unicamp de Alunos Bolsistas. Fonte: Geplanes – Indicadores Estratégicos [3].

Nesse sentido, um trabalho futuro promissor seria desenvolver um modelo capaz de simular o impacto de diferentes intervenções sobre a trajetória dos estudantes identificados como de alto risco. Por exemplo, se um modelo aponta que um aluno tende a evadir por motivos financeiros, seria possível estimar a probabilidade de permanência caso diferentes tipos de auxílio fossem oferecidos, como bolsa alimentação, bolsa transporte ou moradia. Isso permitiria responder questões como: qual intervenção é mais eficaz para cada perfil de estudante?, até que ponto o apoio institucional reduz o risco previsto?, qual o retorno esperado por unidade de investimento em permanência estudantil?

Por fim, a construção de um estudo de caso aplicado à Unicamp, combinando predição de risco, explicabilidade e avaliação de políticas institucionais, poderia orientar decisões estratégicas voltadas à redução tanto da evasão quanto da retenção prolongada. Se bem-sucedida, essa metodologia poderia ser estendida a outras universidades públicas brasileiras, que enfrentam desafios semelhantes de financiamento, vulnerabilidade estudantil e permanência no ensino superior.

5 Conclusão

Esse estudo realizou uma revisão sistemática de trabalhos recentes sobre predição de evasão no ensino superior utilizando técnicas de inteligência artificial. Consideraram-se estudos

publicados entre 2020 e 2025, abrangendo abordagens que vão desde algoritmos clássicos de *machine learning* até métodos mais sofisticados baseados em *deep learning*. Os resultados mostram que a evasão tem sido analisada sob múltiplas perspectivas, refletindo tanto a diversidade dos contextos institucionais quanto a variedade de dados disponíveis — predominantemente informações acadêmicas, socioeconômicas e demográficas.

A partir dessa análise, observa-se que não existe um consenso universal sobre o melhor método de predição de evasão. O desempenho dos modelos depende fortemente das características da base de dados e das especificidades de cada instituição. Ainda assim, é possível identificar um padrão recorrente: algoritmos baseados em árvores de decisão e técnicas de *ensemble* apresentam um equilíbrio particularmente interessante entre complexidade moderada, interpretabilidade e robustez estatística, o que os torna adequados para instituições que trabalham principalmente com dados tabulares. Assim, métodos como Random Forest, Gradient Boosting e XGBoost constituem soluções eficazes e amplamente aplicáveis no contexto institucional.

Como trabalhos futuros, sugerimos avançar do diagnóstico para a avaliação de impacto, investigando como diferentes políticas institucionais de permanência estudantil podem influenciar a probabilidade de evasão prevista pelos modelos. Essa abordagem permitiria não apenas identificar estudantes em risco, mas também estimar a eficácia potencial de intervenções específicas, oferecendo subsídios valiosos para decisões estratégicas no âmbito das universidades públicas.

Agradecimentos

Agradeço ao meu orientador, Prof. Marcelo da Silva Reis, pela orientação, disponibilidade e contribuições ao longo do desenvolvimento deste trabalho.

Agradeço ao Instituto de Computação da Universidade Estadual de Campinas pelo suporte acadêmico e pela formação oferecida ao longo da graduação.

Agradeço à equipe do Escritório de Dados e Apoio à Transformação (EDAT) pelo suporte e apoio ao longo deste ano.

Por fim, agradeço à minha família e às pessoas que estiveram ao meu lado durante essa trajetória, pelo apoio e incentivo constantes.

Referências

[1] Rafael Scarassatti Freitas. A ocorrência da evasão do ensino superior: uma análise das diferentes formas de mensurar. Dissertação de mestrado, Universidade Estadual de Campinas (Unicamp), Faculdade de Educação, Campinas, SP, 2016. Orientação de Elizabeth Nogueira Gomes da Silva Mercuri.

[2] Universidade Estadual de Campinas. Relatório final da avaliação institucional unicamp 2019-2023. `https://cgu.unicamp.br/wp-content/uploads/sites/14/2024/11/Relatorio-Final-da-Avaliacao-Institucional-Ciclo-2019-a-2023.pdf`, 2025. Acesso em: 09 dez. 2025.

[3] Universidade Estadual de Campinas. Geplanes – indicadores estratégicos. `https://geplanes.unicamp.br/indicadores/`, 2025. Acesso em: 08 dez. 2025.

[4] Renato Carauta Ribeiro and Edna Dias Canedo. Using data mining techniques to perform school dropout prediction: A case study. In *17th International Conference on Information Technology – New Generations (ITNG 2020)*, volume 1134 of *Advances in Intelligent Systems and Computing*, pages 211–221. Springer Nature Switzerland AG, 2020.

[5] Andres Gonzalez-Nucamendi, Julieta Noguez, Luis Neri, Víctor Robledo-Rella, and Rosa María Guadalupe García-Castelán. Predictive analytics study to determine undergraduate students at risk of dropout. *Frontiers in Education*, 8:1–14, 2023.

[6] Barbara Martinez Neda, Max Wang, Amanjeet Singh, Sergio Gago-Masague, and Jennifer Wong-Ma. Staying ahead of the curve: Early prediction of academic probation among first-year cs students. In *2023 3rd International Conference on Applied Artificial Intelligence (ICAPAI)*, pages 1–7, Irvine, CA, USA, 2023. IEEE.

[7] Devendra Kumar, Arnav Kothiyal, Raj Kumar, Hemantha C., and Ramya Maranan. Random forest approach optimized by the grid search process for predicting the dropout students. In *2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET)*. IEEE, 2024.

[8] Sittichai Bussaman, Patchara Nasa-Ngium, Wongpanya S. Nuankaew, Thapanapong Sararat, and Pratya Nuankaew. Ensemble learning approaches to strategically shaping learner achievement in thailand higher education. *Education and Information Technologies*, 2024.

[9] Wichukorn Kuntintara, Piya Warabuntaweesuk, and Sountaree Rattapasakorn. Student dropout prediction using machine learning. *Applied Sciences*, 13(4):2247, 2023.

[10] Saritha Vemulapalli, Poonam Bhagat, Somendra Shukla, Nityangini Jhala, D. Jayanthi, and Dileep Pulugu. Predicting student performance and academic success in higher education using a hybrid xgboost–lstm model. In *Proceedings of the 8th International Conference on Computing Methodologies and Communication (ICCMC 2025)*, pages 1501–1506. IEEE, 2025.

[11] P. Alejandra Cuevas-Chávez, Samuel Narciso, Eduardo Sánchez-Jiménez, Itzel Celerino Pérez, Yasmín Hernández, and Javier Ortiz-Hernandez. School dropout prediction with class balancing and hyperparameter configuration. In *MICAI 2023 Workshops*, volume 14502 of *Lecture Notes in Artificial Intelligence*, pages 12–20. Springer Nature Switzerland, 2024.

[12] Valentim Realinho, Jorge Machado, Luís Baptista, and Mónica V. Martins. Predicting student dropout and academic success. *Data*, 7(11):146, 2022.

[13] Mohammad Erfan Shoorangiz and Michal Brylinski. Harnessing large-scale university registrar data for predictive insights: A data-driven approach to forecasting undergraduate student success with convolutional autoencoders. *Machine Learning and Knowledge Extraction*, 7(3):80, 2025.

[14] Andrea Zanellati, Stefano Pio Zingaro, and Maurizio Gabbrielli. Balancing performance and explainability in academic dropout prediction. *IEEE Transactions on Learning Technologies*, 17:2086–2099, 2024.

[15] Mónica V. Martins, Luís Baptista, Jorge Machado, and Valentim Realinho. Multi-class phased prediction of academic performance and dropout in higher education. *Applied Sciences*, 13(8):4702, 2023.