Revista Pleiade – Centro Universitário Descomplica UniAmérica
ISSN 1980-8666 (Impresso) – ISSN 2674-8231 (Eletrônico) – Classificação Capes B3 (2017-2020)

Desempenho Acadêmico na Previsão da Evasão no Ensino Superior: Comparação entre Modelos Bagging, Boosting e Ensemble de Votação Majoritária

Academic Performance in Predicting Dropout Rates in Higher Education: A Comparison Between Bagging, Boosting, and Majority Voting Ensemble Models

Ruminiki Pavei Schmoeller, Vanessa Demarchi Peron e Walter Mesquita Filho

1. Mestre em Tecnologias Computacionais para o Agronegócio (UTFPR). Docente do curso de Bacharelado em Engenharia de Software do Centro Universitário Descomplica UniAmérica. ORCID: https://orcid.org/0009-0006-5046-4390 2. Mestre em Tecnologias, Gestão e Sustentabilidade (Unioeste). ORCID: https://orcid.org/0000-0002-1516-7693 3. Mestre em Entomologia (UNESP) e Doutor em Entomologia (USP). ORCID: https://orcid.org/0000-0003-4695-9471

ruminikis@gmail.com

Palavras-chave
Ensino Superior
Evasão Escolar
LightGBM
Machine Learning
Random Forest
XGBoosting

Keywords
Higher Education
School Dropout
LightGBM
Machine Learning
Random Forest
XGBoosting

Resumo:
A pesquisa comparou modelos de machine learning para predição da evasão no ensino superior, utilizando atributos de desempenho acadêmico, demográficos e socioeconômicos. Quatro modelos foram avaliados (Random Forest, XGBoost, LightGBM e Ensemble). Os resultados da validação cruzada indicaram a superioridade dos algoritmos baseados em Gradient Boosting. O LightGBM apresentou o melhor desempenho geral, com acurácia de 91,05% e AUC-ROC de 0,9694. A análise da matriz de confusão do LightGBM revelou Taxa de Falsos Negativos de 9,82% e Taxa de Falsos Positivos de 7,27%. O estudo dos erros revelou que os falsos negativos, estudantes que cancelaram e não foram identificados, abandonaram o curso apesar do desempenho acadêmico semelhante ao dos ativos, sugerindo evasão motivada por fatores não acadêmicos e não capturados pelo modelo. A importância dos atributos, medida pelo SHAP, estabeleceu que a eficiência acadêmica, principalmente média de aprovações por período e taxa de carga horária de sucesso no último semestre são os preditores dominantes. Concluiu-se: modelos de boosting são eficazes para sistemas de alerta precoce e baixo desempenho é forte indicador de risco. A limitação em captar fatores não acadêmicos aponta para a necessidade de incorporar dados mais abrangentes para previsão de maior impacto nas estratégias de retenção.

Abstract:
This research compared machine learning models for predicting dropout rates in higher education, using academic performance, demographic, and socioeconomic attributes. Four models were evaluated (Random Forest, XGBoost, LightGBM, and Ensemble). Cross-validation results indicated the superiority of Gradient Boosting-based algorithms. LightGBM showed the best overall performance, with an accuracy of 91.05% and an AUC-ROC of 0.9694. Analysis of the confusion matrix of LightGBM revealed a False Negative Rate of 9.82% and a False Positive Rate of 7.27%. The error study revealed that false negatives—students who dropped out and were not identified—abandoned their courses despite academic performance similar to that of active students, suggesting dropout motivated by non-academic factors not captured by the model. The importance of attributes, as measured by SHAP, established that academic efficiency, mainly average pass rates per semester and the rate of successful coursework in the last semester, are the dominant predictors. It was concluded that boosting models are effective for early warning systems, and low performance is a strong indicator of risk. The limitation in capturing non-academic factors points to the need to incorporate more comprehensive data for predicting a greater impact on retention strategies.

Artigo recebido em: 15.10.2025.
Aprovado para publicação em: 07.11.2025.

INTRODUÇÃO

A evasão no ensino superior é um fenômeno persistente e multifatorial, com impactos profundos nas instituições de educação superior, nas trajetórias individuais dos estudantes e nas políticas públicas educacionais (Behr, 2020). Embora a decisão de abandonar um curso resulte de uma complexa interação de fatores pessoais e institucionais (Minhoto, Smaili e Arantes, 2023), o desempenho acadêmico é um indicativo da adaptação do estudante, com influência inclusive sobre sua motivação para continuar o curso (Fior et al, 2024; Goren; Galve-González, 2024; Apumayta, 2024; Cabezas, 2022).

A busca por compreender quais características levam o estudante a abandonar o curso tem sido tema de diversos estudos. As primeiras teorias começaram a ser desenvolvidas na década de 1970 a partir dos trabalhos de Spady (1970), Tinto (1975), Bean (1980) e Pascarella e Terenzini (1980), que destacaram fatores como integração social, expectativas e satisfação dos estudantes, aspectos acadêmicos e organizacionais. Pesquisas recentes apontam para causas semelhantes, reforçando dificuldades financeiras, baixo desempenho acadêmico, falta de motivação e engajamento, não identificação e falta de perspectiva com o curso, circunstâncias pessoais e fatores institucionais (Schuhardt et al., 2024).

A compreensão das causas da evasão ganha ainda mais relevância diante do contexto de expansão do ensino superior no Brasil. No período de 2010 a 2024 houve um aumento de 51% na oferta de vagas em Instituições Federais de Ensino, chegando a 586.246 vagas. Apesar disso, a taxa de ingresso não acompanhou o crescimento, mantendo-se no mesmo patamar, com uma média de 339 mil ingressos anuais (INEP, 2025).

A ociosidade inicial, decorrente do descompasso entre a oferta de vagas e a taxa de ingresso, é agravada pela evasão, que chega a 30% nos dois primeiros anos e a 60% após o período regular do curso, com apenas 40% de estudantes diplomados (INEP, 2025). Como comparação, nos países da Organização para a Cooperação e Desenvolvimento Econômico, a taxa de diplomados é de 39% no tempo apropriado do curso e sobe para 68% após o tempo regular (OECD, 2022).

Apesar do avanço teórico, o paradoxo da expansão do acesso desacompanhada da ocupação das vagas e a dificuldade de retenção, reforça a necessidade de novos estudos. A análise do desempenho acadêmico se mostra relevante para identificar o risco de abandono. A disponibilidade de modelos preditivos permite às instituições adotar uma postura preventiva e concentrar esforços no sucesso do estudante desde o início de sua trajetória, por meio da implementação de mecanismos de alertas em seus sistemas informatizados (USP, 2025).

Tem-se, portanto, um problema prático relevante: a dificuldade das instituições de implementar intervenções preventivas eficazes, direcionadas e em tempo hábil. A ausência de ferramentas preditivas acuradas, com foco em dados obtidos desde o início da jornada acadêmica, limita a capacidade de identificar proativamente os discentes com maior probabilidade de abandonar os estudos, resultando em estratégias de retenção frequentemente reativas e de menor impacto.

Neste contexto, a presente pesquisa busca responder à seguinte questão: o desempenho acadêmico, medido a partir dos semestres iniciais, pode ser um indicador do risco de evasão estudantil? O objetivo é identificar um conjunto de atributos relevantes para a predição da evasão e desenvolver modelos de aprendizado supervisionado capazes de antecipar o risco de abandono de matrícula. Será conduzida uma etapa de seleção e engenharia de atributos a fim de otimizar a generalização e interpretabilidade dos modelos. Ao integrar técnicas de ciência de dados com a literatura educacional, espera-se contribuir para a formulação de estratégias mais eficazes de combate à evasão e de promoção da permanência estudantil.

ESTADO DA ARTE / TRABALHOS RELACIONADOS

A evasão universitária é reconhecida como um fenômeno multifacetado e complexo, raramente dependente de um fator isolado, mas sim do resultado da inter-relação de múltiplos determinantes (Behr, 2020; Galve-González, 2024). Com alguma variação, os estudos têm aplicado majoritariamente modelos de aprendizado supervisionado, avaliados atributos demográficos, socioeconômicos e de desempenho acadêmico (Alvarado-Uribe et al., 2022; Viloria et al., 2019; Alvarez, Callejas e Griol, 2020; Aulck et al., 2017; Niyogisubizo et al., 2022; Flores et al., 2022; Gonçalves et al., 2018; Perez et al., 2018; Sara et al., 2015). Alguns, no entanto, avançaram para estudar atributos familiares, como ocupação e escolaridade dos pais (Pedroza et al., 2019).

De La Cruz-Campos et al. (2023), Lorenzo-Quiles et al. (2023), Behr et al. (2020) e Da Silva (2021), identificaram que características pessoais e sociais (90,4%), acadêmicas (76,1%), demográficas (57,1%) e socioeconômicas (42,8%) são as mais comumente analisadas na busca de identificação de grupos de risco para a evasão.

Galve-González (2024), aponta que problemas de desempenho acadêmico e dificuldades de aprendizado estão diretamente relacionados a condições inadequadas de estudo e a uma gestão de tempo deficiente, que se referem a aspectos socioeconômicos e comportamentais do estudante. Niyogisubizo et al. (2022), reforçam que a dificuldade financeira se configura como fator crítico para o desempenho e a continuidade dos estudos.

Demetriou e Schmitz-Sciborski (2011), destacaram a relevância da integração social, motivação, forças pessoais e otimismo. Variáveis como satisfação com a escolha do curso, engajamento acadêmico e a aplicação de estratégias de autorregulação da aprendizagem são consideradas fundamentais. Por sua vez, o desinteresse pelo conteúdo curricular tem sido apontado como um fator chave na previsão da evasão (Galve-González, 2024).

Dada a complexidade do fenômeno da evasão, diversos estudos têm focado em abordagens quantitativas para investigar suas causas e possíveis intervenções (Flores et al., 2022; Apumayta, 2024). Alvarado-Uribe et al. (2022) observaram que a maior parte da literatura utiliza algoritmos tradicionais de aprendizado de máquina, como regressão logística, kNN e modelos baseados em árvores de decisão. Viloria et al. (2019), Flores et al. (2022) e Tete et al. (2022), observam que o método Random Forest é comumente o mais utilizado na predição da evasão, seguido de redes bayesianas e Regressão Logística.

Estudos mais recentes apontam para a tendência crescente no uso de modelos de conjunto (*ensemble*) e técnicas de otimização, visando superar desafios persistentes como o desequilíbrio de classes — situação em que a maioria dos estudantes não abandona o curso. Niyogisubizo et al. (2022) propuseram um modelo de conjunto empilhado (*stacked ensemble*) de duas camadas, combinando Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB) e Redes Neurais de Alimentação Direta (FNN). Esse método superou os modelos isolados, alcançando uma acurácia de teste de 92,18% e uma AUC de 0,983.

Novos estudos, conduzidos por Villar e Andrade (2024), reafirmam a superioridade dos algoritmos de *boosting*, especialmente quando combinados com a otimização de hiperparâmetros. Os achados indicaram que LightGBM e CatBoost, após a otimização, apresentaram desempenho superior aos algoritmos convencionais.

Apesar de ainda ser um desafio complexo, a previsão da evasão vem apresentando avanços com o emprego de técnicas modernas de aprendizado de máquina. A relevância dos diferentes estudos reside em iden-

tificar e confirmar variáveis importantes para a previsão da evasão, levando em conta o seu treinamento em diferentes contextos e conjuntos de dados.

MATERIAIS E MÉTODOS

Para a realização desta pesquisa, foram utilizados dados anonimizados de estudantes de uma Universidade Federal do Brasil, relacionados ao desempenho acadêmico, características demográficas e socioeconômicas. As atividades foram organizadas segundo a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM), desenvolvidas em ciclos iterativos para (1) entendimento do negócio; (2) entendimento dos dados; (3) preparação dos dados; (4) modelagem; e (5) avaliação (Wirth e Hipp, 2000).

Na etapa de entendimento do negócio e entendimento dos dados foram compreendidas as questões relacionadas à evasão escolar e como elas são tratadas na Universidade. Foram avaliadas normas referentes aos processos de matrícula e registro acadêmico do discente, considerando os procedimentos relativos aos cancelamentos e trancamentos. A amostra selecionada teve 14.800 registros e 26 atributos, cuja relação é apresentada no Quadro 1.

Quadro 1: Relação dos atributos do conjunto de dados

Campo	
idade_ingresso_curso	acompanhamentos_pedagogicos_por_periodo
forma_ingresso	carga_horaria_matriculada_por_periodo
carga_horaria_integralizada_por_periodo	carga_horaria_total_primeiro_semestre
turno	taxa_cobertura_auxilio_alimentacao
tempo_ensino_medio_e_ingresso	taxa_cobertura_auxilio_transporte
total_faltas_primeiro_semestre	taxa_engajamento_em_projetos
total_faltas_ultimo_semestre	carga_horaria_sucesso_por_periodo
media_notas	carga_horaria_insucesso_por_periodo
media_percentual_frequencia	carga_horaria_sucesso_primeiro_semestre
media_numero_faltas	carga_horaria_insucesso_primeiro_semestre
aprovacoes_por_periodo	carga_horaria_sucesso_ultimo_semestre
reprovacoes_por_periodo	carga_horaria_insucesso_ultimo_semestre
trancamentos_por_periodo	situacao

Na fase de preparação dos dados, foram removidos registros de estudantes que não tiveram pelo menos um semestre consolidado e desconsiderados os estudantes que tiveram a matrícula cancelada por transferência para outra instituição; decisão judicial; falecimento; e desistência antes do início do curso (Andifes, 1996; Coimbra, Silva e Costa, 2021).

Atributos somatórios e de contagem que continham valores faltantes foram atualizados para zero. As demais variáveis foram proporcionalizadas em relação ao total de períodos cursados. Uma etapa de engenharia de atributos (*feature engineering*) foi realizada. Nesse processo foram combinados atributos quantitativos de forma a gerar taxas associadas ao desempenho acadêmico.

A situação do estudante foi categorizada em duas classes: ativos e cancelados. Segundo as normas da Universidade, foram classificados como cancelados os estudantes que estavam com o vínculo inativo e aqueles que atingiram o limite de trancamentos e não estavam matriculados em nenhuma disciplina.

Uma etapa de seleção de características foi realizada, com objetivo de encontrar um número reduzido de atributos, suficiente para uma boa previsão da variável resposta e que favoreça a definição de um modelo mais compacto e de melhor interpretabilidade (Hall, 1999; Becher et al., 2000; Genuer et al., 2010; Li et al., 2017). Neste trabalho foi utilizado o método de avaliação da informação mútua (Peng et al., 2005), uma medida da dependência entre duas variáveis aleatórias, usada para quantificar a quantidade de informação compartilhada entre um atributo (*feature*) e a variável alvo (*target*). Ao final da etapa de preparação, o conjunto consolidado teve 9.430 observações com 33 atributos. Desses, 4.950 identificados como **ativos** e 4.480 identificados como **cancelados**.

Na etapa de modelagem foram treinados modelos Random Forest (Breiman, 2001), XGBoosting (Chen e Guestrin, 2016), LightGBM (Ke et al. 2017) e em seguida um modelo *ensemble* por votação majoritária que combina a classificação gerada pelos três modelos. A escolha dos modelos está alinhada aos principais estudos na área que reforçam o uso de modelos de árvores, e a tendência recente para o uso de modelos de boosting.

Foi estabelecida uma *pipeline* com etapas de pré-processamento dos dados com a codificação de variáveis categóricas usando a técnica de *one hot encoding* e padronização de variáveis numéricas pela transformação dos dados em *z-score* para que cada característica se assemelhe a uma distribuição normal padrão (Cabello-Solorzano, 2023). Os conjuntos foram divididos aleatoriamente em treinamento (com 80% das observações) e teste (com 20% das observações). Em todos os modelos foi aplicada a estratégia de validação cruzada com dez dobras (*10 k-fold cross-validation*).

Os modelos foram avaliados de forma comparativa por meio das métricas de acurácia, precisão, revocação, F1-Score e área sob a curva ROC, que são adequadas para problemas de classificação (Fawcett, 2006; Fávero, 2017). Além das métricas de desempenho, o resultado dos modelos foi comparado pelo teste McNemar (McNemar, 1947). O teste é usado para determinar se existe uma diferença estatisticamente significativa entre as taxas de erro de dois modelos quando eles são aplicados no mesmo conjunto de dados. Diferente de outros testes, o teste de McNemar foca especificamente nas instâncias em que os modelos discordam.

A importância dos atributos e interpretabilidade do modelo foi avaliada pelo método SHAP (*SHapley Additive exPlanations*). O SHAP calcula a importância dos atributos, quantificando o impacto que cada atributo tem na predição final do modelo. Fornece uma forma consistente e aditiva de decompor a predição de uma instância em contribuições individuais dos seus atributos. Permite a interpretabilidade de uma instância específica e a importância geral dos atributos para o modelo (Lundberg e Lee, 2017).

Resultados e Discussão

Modelos de Classificação

A análise dos resultados da validação cruzada revela um desempenho muito próximo entre os modelos, com superioridade para modelos baseados em *Gradient Boosting* (LightGBM, XGBoost). O modelo *ensemble* também apresentou desempenho semelhante (Tabela 1).

A Tabela 2, exibe os resultados da análise comparativa dos modelos Random Forest, LightGBM, XGBoost e *ensemble*, empregando o teste de McNemar para avaliar suas performances. O modelo Random Forest apresentou um desempenho estatisticamente inferior aos demais modelos. Não houve diferença significativa no desempenho entre LightGBM, XGBoost e o *Ensemble*, mostrando que os erros que esses modelos cometem são muito semelhantes.

Tabela 1: Desempenho dos modelos com o método de validação cruzada

Modelo	Acurácia	Precisão	Recall	F1-Score	AUC-ROC
LightGBM	0,9105 ± 0,0073	0,9132 ± 0,0102	0,8970 ± 0,0067	0,9050 ± 0,0075	0,9694 ± 0,0035
XGBoost	0,9093 ± 0,0115	0,9115 ± 0,0169	0,8965 ± 0,0095	0,9039 ± 0,0116	0,9690 ± 0,0043
Ensemble	0,9099 ± 0,0078	0,9132 ± 0,0131	0,8956 ± 0,0099	0,9042 ± 0,0080	0,9676 ± 0,0044
Random Forest	0,8900 ± 0,0094	0,8916 ± 0,0177	0,8753 ± 0,0059	0,8832 ± 0,0089	0,9545 ± 0,0060

Tabela 2: Comparação do desempenho dos modelos pelo teste McNemar

Modelo 1	Modelo 2	Estatística McNemar	Valor-p	Significante (p<0,05)	Melhor Modelo
Random Forest	LightGBM	18,0	0.0000	Sim	LightGBM
Random Forest	XGBoost	25,0	0.0003	Sim	XGBoost
Random Forest	Ensemble	14,0	0.0000	Sim	Ensemble
LightGBM	XGBoost	16,0	0.0725	Não	N/A
LightGBM	Ensemble	9,0	0.1221	Não	N/A
XGBoost	Ensemble	11,0	0.5572	Não	N/A

A taxa de revocação de 89,70% sugere que o LightGBM tem maior capacidade de identificar os alunos que estão em risco real de abandonar o curso - que é o principal objetivo de um sistema de alerta precoce. Ao mesmo tempo, demonstrou maior precisão e acurácia. Quando o modelo classifica um aluno como risco de evasão, ele está correto em 91,32% das vezes, minimizando o risco de falsos positivos, e consequentemente, que os recursos de intervenção sejam desperdiçados em alunos que não precisam.

Em relação à área sob a curva ROC, os modelos Light GBM, XGBoost e Ensemble apresentam curvas bastante próximas, com vantagem para o LightGBM que apresentou AUC de 96,94%, sugerindo que o modelo tem alto poder de discriminação entre as classes positivas e negativas. Os modelos divergiram pouco, em apenas 5,6% dos casos. Isso reforça a compreensão do teste McNemar, ao apresentar que as divergências entre os modelos de boosting e ensemble não foram significativas.

Análise da matriz de confusão

A partir do resultado do desempenho dos modelos, o LightGBM foi escolhido para análise. Observa-se bom desempenho na classificação do conjunto de testes, com acurácia de 90,91%. Dos 1.886 casos avaliados, 1.726 foram corretamente classificados, com 918 verdadeiros negativos (alunos ativos) e 808 verdadeiros positivos (previsão de cancelamento), conforme pode ser visto na Tabela 3.

A taxa de falso positivo (erro Tipo I) é 7,27%, ou seja, cerca de 7% dos alunos que continuariam ativos foram classificados como em risco de evasão. A taxa de erro Tipo II, que corresponde aos falsos negativos, é maior, atingindo aproximadamente 9,82%, com 88 alunos cancelados classificados como ativos. Esse tipo de erro é mais crítico nesse contexto, pois representa alunos em risco de evasão não identificados pelo modelo, comprometendo a eficiência das ações de retenção. Falsos positivos podem levar a gastos desnecessários com alunos que não precisam de ajuda, enquanto falsos negativos deixam alunos em risco sem atendimento.

A análise dos erros, especificamente dos Falsos Positivos (FP) e Falsos Negativos (FN), revela os perfis atípicos que o modelo não conseguiu capturar (Tabela 4 e Tabela 5). Quanto aos falsos positivos, observa-se que o perfil de desempenho deste grupo é inferior ao do aluno ativo e semelhante ao de um aluno que cancelou a matrícula. Esse grupo apresentou dificuldade acadêmica e de engajamento, no entanto, se manteve ativo. A sua manutenção pode estar relacionada ao apoio pedagógico recebido, que foi superior.

Tabela 3: Análise da matriz de confusão

Métrica	Valor	Descrição / Interpretação
Total de Instâncias	1886	Número total de alunos na amostra de teste.
Verdadeiros Negativos (TN)	918	Alunos ATIVOS que o modelo previu corretamente como ATIVOS.
Falsos Positivos (FP)	72	Alunos ATIVOS que o modelo previu erroneamente como CANCELADOS.
Falsos Negativos (FN)	88	Alunos CANCELADOS que o modelo previu erroneamente como ATIVOS.
Verdadeiros Positivos (TP)	808	Alunos CANCELADOS que o modelo previu corretamente como CANCELADOS.
Taxa de Erro Tipo I (Falso Alarme)	7,27%	Indica que 7,27% dos alunos ativos foram classificados de forma errada como evadidos.
Taxa de Erro Tipo II (Falso Negativo)	9,82%	Indica que 9,82% dos alunos evadidos não foram identificados pelo modelo.

Tabela 4: Análise dos falsos positivos (alunos ativos classificados como cancelados)

Característica	Padrão Encontrado (Falso Positivo)	Comparação com Ativos	Comparação com Cancelados
Taxa de Sucesso (Último Semestre)	27%	44%	27%
Média de Notas	5,47	7,12	5,23
Reprovações por Período	1,44	0,51	1,46
Frequência Média	83%	93%	83%
Evolução do Desempenho	-0,48	-0,05	-0,45
Engajamento Extracurricular	0,018	0,05	0,01
Carga Horária Matriculada	271	207	284
Carga Horária Integralizada	166	322	160
Acompanhamento Pedagógico	0,52	0,37	0,40
Cobertura de Auxílios	18%	37%	17%

Quanto aos falsos negativos (Tabela 5), são alunos com desempenho intermediário que apesar do elevado número de acompanhamentos pedagógicos, decidiram abandonar o curso. A opção pelo abandono, nesse caso, pode ter sido influenciada por fatores externos, não captados pelo conjunto de dados.

Tabela 5: Análise dos falsos negativos (alunos cancelados classificados como ativos)

Característica	Padrão Encontrado (Falso Negativo)	Comparação com Ativos	Comparação com Cancelados
Taxa de Sucesso (Último Semestre)	66%	81%	27%
Média de Notas	7,12	7,66	5,23
Reprovações por Período	0,83	0,51	1,46
Frequência Média	92%	93%	83%
Evolução do Desempenho	-0,19	-0,05	-0,45
Engajamento Extracurricular	0,04	0,05	0,01
Carga Horária Matriculada	263	207	284
Carga Horária Integralizada	262	322	160
Acompanhamento Pedagógico	0,57	0,37	0,40
Cobertura de Auxílios	26%	37%	17%

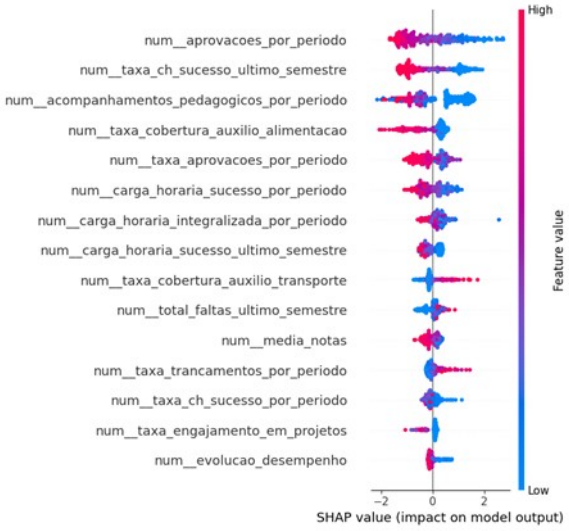
Pleiade, *19*(49): 42-55, Out.-Dez., 2025
DOI: 10.32915/pleiade.v19i49.1195

A análise dos Falsos Positivos e Falsos Negativos revela a complexidade da predição de evasão, mostrando que o modelo, embora eficaz, apresenta desafios importantes.

Análise da Importância dos Atributos

A importância dos atributos, obtida pela média do valor absoluto de SHAP, apresenta o impacto individual de cada variável na previsão final do modelo LightGBM. A análise fornece uma compreensão granular de como cada atributo se comporta e se associa ao risco de evasão (valores SHAP positivos) ou à permanência (valores SHAP negativos), ilustrando a intensidade e a direção da influência de cada característica (Gráfico 1).

Gráfico 1: Importância dos atributos pelo método SHAP



No Gráfico 1, é possível observar que as variáveis de desempenho acadêmico contribuem positivamente para a classificação e portanto, são um bom indicador da separação entre as classes, com o menor desempenho acadêmico associado a estudantes com risco de evasão.

A análise revelou que 11 das 15 variáveis mais importantes para predizer o risco de evasão se relacionam à eficiência acadêmica, três ao suporte institucional e duas ao comportamento do estudante. No nível superior (SHAP > 0,5), o modelo prioriza o desempenho acadêmico e a busca por suporte institucional como preditores dominantes (Tabela 6).

A quantidade de aprovações no curso e a taxa de carga horária de sucesso no último semestre são os principais preditores do sucesso acadêmico, a incapacidade de progredir leva à evasão. O acompanhamento pedagógico, com impacto superior a métricas de desempenho, surge como forte indicador, possivelmente agindo como *proxy* para alunos de alto risco. A vulnerabilidade financeira, indicada pela taxa de cobertura de auxílio alimentação, também é um forte indicador socioeconômico de retenção.

No segundo nível de relevância (SHAP entre 0,25 e 0,5), o modelo incorpora variáveis de desempenho acadêmico, como taxa de aprovação, carga horária de sucesso por período e carga horária integralizada, que indicam a progressão do estudante e sua capacidade de concluir disciplinas. A carga horária de sucesso no último semestre reforça a importância dos dados mais recentes para a capacidade preditiva do modelo.

Tabela 6: Nível de importância dos atributos

Variável	SHAP Value	Contexto
aprovacoes_por_periodo	0,950	Eficiência Acadêmica
taxa_ch_sucesso_ultimo_semestre	0,912	Eficiência Acadêmica
acompanhamentos_pedagogicos_por_periodo	0,846	Suporte Institucional
taxa_cobertura_auxilio_alimentacao	0,509	Suporte Institucional
taxa_aprovacoes_por_periodo	0,432	Eficiência Acadêmica
carga_horaria_sucesso_por_periodo	0,423	Eficiência Acadêmica
carga_horaria_integralizada_por_periodo	0,274	Eficiência Acadêmica
carga_horaria_sucesso_ultimo_semestre	0,264	Eficiência Acadêmica
taxa_cobertura_auxilio_transporte	0,258	Suporte Institucional
total_faltas_ultimo_semestre	0,191	Comportamental
media_notas	0,188	Eficiência Acadêmica
taxa_trancamentos_por_periodo	0,175	Eficiência Acadêmica
taxa_ch_sucesso_por_periodo	0,160	Eficiência Acadêmica
taxa_engajamento_em_projetos	0,154	Comportamental
evolucao_desempenho	0,134	Eficiência Acadêmica

O terceiro nível (SHAP < 0,25) inclui sinais menos influentes, como notas e comportamento. A média das notas tem menos peso que as taxas de aprovação, indicando que o modelo prioriza a aprovação sobre a qualidade da nota. Faltas e trancamentos sinalizam desengajamento como fator de risco, embora menor que o fracasso acadêmico direto. Atributos como engajamento em projetos extracurriculares (fator protetivo, possivelmente por vínculos mais sólidos) e a evolução do desempenho (diferença entre o primeiro e o último período) refinam a previsão, reforçando a evasão estar mais ligada à reprovação do que ao desempenho médio.

O teste t de Welch (Tabela 7), confirmou a significância das diferenças entre as médias dos atributos mais relevantes.

Tabela 7: Comparação da média dos atributos de estudantes identificados como ativos e cancelados

Variável	Estatística T	p-valor	Média Ativos	Média Cancelados
aprovacoes_por_periodo	64,98	0,00E+00*	4,79	2,44
taxa_ch_sucesso_ultimo_semestre	85,76	0,00E+00*	0,84	0,27
acompanhamentos_pedagogicos_por_periodo	-2,94	3,25E-03*	0,38	0,41
taxa_cobertura_auxilio_alimentacao	27,32	5,26E-158*	0,37	0,18
taxa_aprovacoes_por_periodo	76,92	0,00E+00*	0,82	0,49
carga_horaria_sucesso_por_periodo	25,13	6,72E-135*	72,61	67,93
carga_horaria_integralizada_por_periodo	70,54	0,00E+00*	322,51	160,15
carga_horaria_sucesso_ultimo_semestre	59,72	0,00E+00*	314,71	99,90
taxa_cobertura_auxilio_transporte	6,29	3,37E-10*	0,18	0,14
total_faltas_ultimo_semestre	-38,14	3,39E-290*	18,71	46,26
media_notas	66,77	0,00E+00*	7,66	5,23
taxa_trancamentos_por_periodo	-23,44	6,02E-117*	0,09	0,15
taxa_ch_sucesso_por_periodo	39,52	0,00E+00*	0,45	0,29
taxa_engajamento_em_projetos	29,02	1,20E-175*	0,05	0,01
evolucao_desempenho	53,03	0,00E+00*	-0,05	-0,45

Nota: *Diferença significativa ao nível de significância de 1% (p < 0,01).

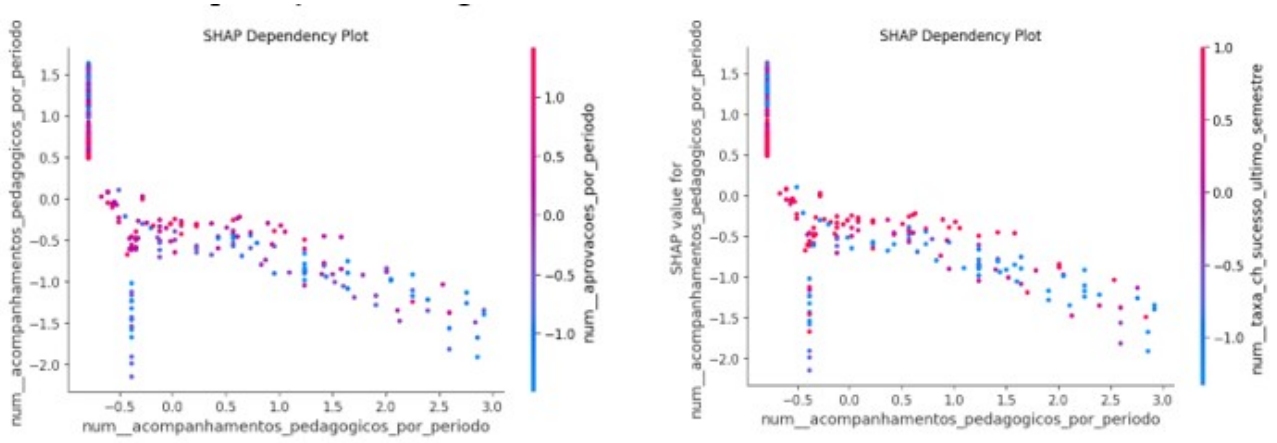
Pleiade, *19*(49): 42-55, Out.-Dez., 2025
DOI: 10.32915/pleiade.v19i49.1195

A análise das médias revela diferenças estatisticamente significativas entre os estudantes ativos e cancelados em todas as variáveis. O grupo cancelado apresentou médias significativamente mais altas em indicadores de insucesso, como total_faltas_ultimo_semestre e taxa_trancamentos_por_periodo, reforçando a hipótese de que o baixo desempenho, a frequência insuficiente e a maior taxa de trancamentos são sintomas importantes para o abandono.

As variáveis socioeconômicas, taxa de cobertura de auxílio alimentação e taxa de cobertura de auxílio transporte apresentaram menor cobertura no grupo cancelado. Já a quantidade de acompanhamentos pedagógicos foi semelhante. Por um lado, reforça a importância do apoio institucional para manter estudantes ativos, e por outro, revela limitações na capacidade de retenção.

O aprofundamento sobre os acompanhamentos pedagógicos, revela que a variável atua como um *proxy* para estudantes em risco de evasão (Gráfico 2). A relação linear indica que mais acompanhamentos se associam a pior desempenho (menor sucesso em carga horária e menos aprovações). Contudo, a presença de pontos azuis em valores mínimos de acompanhamento sugere que alguns estudantes em risco não estão sendo identificados pelas ações de suporte institucional.

Gráfico 2. Gráfico de dependência de acompanhamentos pedagógicos em relação às aprovações e a carga horária de sucesso no último semestre



Os resultados confirmam a relação entre desempenho acadêmico e evasão, a exemplo de Fior et al. (2022), que mostraram que evadidos tinham notas significativamente mais baixas (M = 4,31) que os matriculados (M = 6,41). Um melhor desempenho acadêmico está ligado a um menor risco de abandono, evidenciando o impacto do rendimento na decisão dos estudantes.

Considerando o desempenho uma variável dinâmica e preditora, os autores sugerem que as intervenções institucionais foquem no fortalecimento da autoeficácia estudantil. Ações diretas para aumentar as notas, como monitorias e reestruturação curricular, podem ser eficazes na melhoria da permanência, a exemplo dos acompanhamentos pedagógicos evidenciados neste trabalho.

Também o estudo de Vaarma e Li (2024), reforça que o desempenho acadêmico é o fator preditivo mais poderoso. Para os autores, "créditos acumulados" e "número de reprovações" são as variáveis mais importantes. Neste trabalho, se assemelham a baixa carga horária integralizada (créditos acumulados) e baixo número de aprovações, que caracterizam o grupo de estudantes cancelados.

Goren et al. (2024), corroboram com o ponto de vista e afirmam que o desempenho acadêmico é o dado mais preditivo de abandono. Embora a qualidade da predição (AUC) melhore com dados de exames parciais

e ao final do semestre, os autores apontam um paradoxo temporal: esses indicadores, apesar de melhores, são tardios, limitando o tempo para intervenções eficazes. Isso gera um *trade-off* entre capacidade preditiva e intervenção precoce. Essa limitação esteve presente neste trabalho, que considerou apenas estudantes com, no mínimo, um semestre concluído.

O baixo desempenho acadêmico se configura, portanto, como um sintoma claro e consistente de risco para evasão. A principal contribuição dos estudos (Fior et al., 2022; Goren et al., 2024) está em evidenciar que, embora o desempenho seja o melhor preditor do abandono, ele é condicionado por fatores psicológicos, como a autoeficácia, e decisões pessoais, como a escolha do curso.

Considerações Finais

A análise realizada neste estudo confirmou a capacidade preditiva associada a variáveis de desempenho acadêmico e demonstrou que estudantes que possuem melhor desempenho têm menor risco para a evasão. Os modelos desenvolvidos apresentaram alta capacidade preditiva, com destaque para os algoritmos de *boosting*, em especial ao LightGBM, com acurácia acima de 90% e área sob a curva ROC próxima de 97%. O modelo *ensemble*, que combina a predição dos modelos, reforçou a semelhança entre eles. A partir da análise dos erros e da importância das variáveis pelo método SHAP, foi possível traçar um perfil entre os grupos ativos e cancelados.

Academicamente, o aluno em risco de evasão, demonstra baixo rendimento, com menor taxa de aprovação, menor capacidade de cumprimento da carga horária matriculada, notas mais baixas. Comportamentalmente apresenta alto absenteísmo, matrícula excessiva em disciplinas e baixo engajamento extracurricular. Institucionalmente possui menor cobertura de auxílios estudantis, sugerindo por um lado a importância do auxílio para a manutenção dos estudantes ativos, e por outro, a necessidade de aprofundamento institucional para verificar se a decisão pelo abandono foi influenciada pela falta de assistência. Os acompanhamentos pedagógicos emergiram como variável importante para o modelo, no entanto, as médias próximas entre estudantes ativos e cancelados indica a necessidade de aprofundamento para verificar a efetividade das ações ou as limitações institucionais, quando a decisão pelo abandono não pode ser revertida.

Esses achados fornecem à instituição um roteiro para possíveis ações de retenção que levem em conta a intervenção precoce baseada no desempenho desde o primeiro semestre. A identificação precoce de quedas no rendimento permite intervenções oportunas, evitando que o baixo desempenho se torne um fator irreversível. O acompanhamento pelo setor pedagógico se mostra essencial também para a compreensão dos diversos sinais que o estudante possa emitir em relação a sua decisão futura.

Referências

ALVARADO-URIBE, J. et al. Student Dataset from Tecnologico de Monterrey in Mexico to Predict Dropout in Higher Education. **Data**, v. 7, n. 9, p. 119, set. 2022. Disponível em: https://doi.org/10.3390/data7090119. Acesso em 10 mar. 2024.

ALVAREZ, N. L.; CALLEJAS, Z.; GRIOL, D. Predicting Computer Engineering Students' Dropout In Cuban Higher Education With Pre-Enrollment and Early Performance Data. **Journal of Technology and Science Education**, v. 10, n. 2, p. 241–258, 2020. Disponível em: https://doi.org/10.3926/jotse.922. Acesso em 05 mar. 2025.

ANDIFES. **Diplomação, retenção e evasão nos cursos de graduação em IES públicas:** Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras. Brasília, DF: [s. n.], 1996. Disponível em:

http://www.andifes.org.br/wp-content/files_flutter/Diplomacao_Retencao_Evasao_Graduacao_em_IES_Publicas-1996.pdf. Acesso em 29 mar. 2024.

APUMAYTA, Raul Quincho; CAYLLAHUA, Javier Carrillo; PARI, Abraham Ccencho et al. University dropout: a systematic review of the main determinant factors (2020–2024). **F1000Research**, v. 13, p. 942, 2024. Disponível em: https://doi.org/10.12688/f1000research.154263.2. Acesso em 08 out. 2025.

AULCK, L.; VELAGAPUDI, N.; BLUMENSTOCK, J.; WEST, J. Predicting student dropout in higher education. 2017. Disponível em: http://arxiv.org/abs/1606.06364. Acesso em 15 mar. 2024.

BEAN, John P. Dropouts and turnover: the synthesis and test of a causal model of student attrition. **Research in Higher Education**, v. 12, p. 155–187, 1980.

BEHR, A.; GIESE, M.; TEGUIM KAMDJO, H. D.; THEUNE, K. Dropping out of university: a literature review. **Review of Education**, v. 8, n. 2, p. 614–652, 2020. Disponível em: https://doi.org/10.1002/rev3.3202. Acesso em 06 out. 2023.

BOEHMKE, Bradley. Hands-on machine learning with R. Disponível em: https://bradleyboehmke.github.io/HOML/index.html. Acesso em: 14 out. 2023.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CABELLO-SOLORZANO, K.; ORTIGOSA DE ARAUJO, I.; PEÑA, M.; CORREIA, L. J.; TALLÓN-BALLESTEROS, A. Impact of data normalization on the accuracy of machine learning algorithms: a comparative analysis. **18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023)**. Disponível em: https://doi.org/10.1007/978-3-031-42536-3_33. Acesso em 10 out. 2024.

CHEN, T.; GUESTRIN, C. XGBoost: a scalable tree boosting system. 2016. Disponível em: https://doi.org/10.1145/2939672.2939785. Acesso em 22 set. 2024.

COIMBRA, C. L.; SILVA, L. B.; COSTA, N. C. Evasion in higher education: definitions and trajectories. **Educação e Pesquisa**, v. 47, p. 1–18, 2021. Disponível em: https://doi.org/10.1590/S1678-4634202147228764. Acesso em 25 ago. 2024.

DE LA CRUZ-CAMPOS, J.; VICTORIA-MALDONADO, J.; MARTÍNEZ-DOMINGO, J.; CAMPOS-SOTO, M. Causes of academic dropout in higher education in Andalusia and proposals for its prevention at university: a systematic review. **Frontiers in Education**, v. 8, 2023. Disponível em: https://doi.org/10.3389/feduc.2023.1130952. Acesso em 17 ago. 2024.

FÁVERO, L. P. Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. São Paulo: **GEN LTC**, 2017.

FAWCETT, Tom. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006.

FIORE, Camila Alves et al. Impacto da autoeficácia e do rendimento acadêmico no abandono de estudantes do ensino superior. **Psicologia Escolar e Educacional**, v. 26, e235218, Campinas, SP, 2022. Disponível em: https://doi.org/10.1590/2175-35392022235218. Acesso em 10 out. 2025.

FLORES, Vaneza; HERAS, Stella; JULIAN, Vicente. Comparison of predictive models with balanced classes using the SMOTE method for the forecast of student dropout in higher education. **Electronics**, v. 11, n. 3, p. 457, 2022. Disponível em: https://doi.org/10.3390/electronics11030457. Acesso em 6 set. 2024.

GALVE-GONZÁLEZ, C.; BERNARDO, A. B.; CASTRO-LÓPEZ, A. Understanding the dynamics of college transitions between courses: uncertainty associated with the decision to drop out studies among first and second year students. **European Journal of Psychology of Education**, v. 39, n. 2, p. 959–978, 2024. Disponível em: https://doi.org/10.1007/s10212-023-00732-2. Acesso em 8 set. 2024.

GENUER, Robin; POGGI, Jean-Michel; TULEAU-MALOT, Christine. Variable selection using random forests. **Pattern Recognition Letters**, v. 31, n. 14, p. 2225–2236, 2010.

GONÇALVES, T. C.; SILVA, J. C. da; CORTES, O. A. C. Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do Instituto Federal do Maranhão. **Revista Brasileira de Computação Aplicada**, v. 10, n. 3, p. 11–20, 2018. Disponível em: https://doi.org/10.5335/rbca.v10i3.8427. Acesso em 20 set. 2024.

GORE, Or; COHEN, Liron; RUBINSTEIN, Amir. Early prediction of student dropout in higher education using machine learning models. In: PAASSEN, B.; EPP, C. D. (eds.). **Proceedings of the 17th International Conference on Edu-**

cational Data Mining. Atlanta: International Educational Data Mining Society, 2024. p. 349–359. Disponível em: https://doi.org/10.5281/zenodo.12729834. Acesso em 6 out. 2024.

HENRIQUEZ CABEZAS, N.; VARGAS ESCOBAR, D. Predictive models of academic achievement and dropout of first year students of a Chilean public university. **Revista de Estudios y Experiencias en Educación**, v. 21, n. 45, p. 299–316, 2022.

HSU, Hui-Huang et al. Feature selection via correlation coefficient clustering. **Journal of Software**, v. 5, n. 12, p. 1371–1377, 2010.

INEP. Relatório do 5º ciclo de monitoramento das metas do Plano Nacional de Educação – 2024. 2. ed. Brasília, DF: **INEP**, 2024. 625 p. ISBN 978-65-5801-074-6 (impresso); 978-65-5801-071-5 (on-line).

INEP. Censo da Educação Superior, 2025. Disponível em: https://app.powerbi.com/view?r=eyJrIjoiMGJiMmNiNTAtOTY1OC00ZjUzLTg2OGUtMjAzYzNiYTA5YjliIiwidCI6IjI2ZjczODk3LWM4YWMtNGIxZS05NzhmLWVhNGMwNzc0MzRiZiJ9&pageName=ReportSection4036c90b8a27b5f58f54. Acesso em: 5 set. 2025.

KE, G. et al. LightGBM: a highly efficient gradient boosting decision tree. **Advances in Neural Information Processing Systems**, 2017.

LI, J. et al. Feature selection: a data perspective. **ACM Computing Surveys**, v. 50, n. 6, 2017. Disponível em: https://doi.org/10.1145/3136625. Acesso em 10 fev. 2024.

LORENZO-QUILES, O.; GALDÓN-LÓPEZ, S.; LENDÍNEZ-TURÓN, A. Dropout at university: variables involved on it. **Frontiers in Education**, v. 8, 2023. Disponível em: https://doi.org/10.3389/feduc.2023.1159864. Acesso em 15 mar. 2024.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Advances in Neural Information Processing Systems**, v. 30, 2017.

MCNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. **Psychometrika**, v. 12, n. 2, p. 153–157, 1947.

MINHOTO, Maria Angélica; SMAILI, Soraya; ARANTES, Pedro. 2,3 milhões abandonaram curso superior em 2021. **Folha de São Paulo**, 23 fev. 2023. Disponível em: https://www1.folha.uol.com.br/blogs/sou-ciencia/2023/02/23-milhoes-abandonaram-curso-superior-em-2021.shtml. Acesso em: 15 fev. 2024.

NIYOGISUBIZO, Jovial et al. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization. **Computers and Education: Artificial Intelligence**, v. 3, p. 100066, 2022. Disponível em: https://doi.org/10.1016/j.caeai.2022.100066. Acesso em 15 set. 2024.

NOGUEIRA, C. M. M.; NONATO, B. F.; RIBEIRO, G. M.; FLONTINO, S. R. D. Promessas e limites: o Sisu e sua implementação na Universidade Federal de Minas Gerais. **Educação em Revista**, v. 33, 2017. Disponível em: https://doi.org/10.1590/0102-4698161036. Acesso em 15 out. 2024.

OECD. Education at a Glance 2022: OECD Indicators. Paris: **OECD Publishing**, 2022. Disponível em: https://doi.org/10.1787/3197152b-en. Acesso em 25 set. 2025.

PASCARELLA, Ernest T.; TERENSINI, Patrick T. Predicting freshman persistence and voluntary dropout decisions from a theoretical model. **The Journal of Higher Education**, v. 51, n. 1, p. 60–75, 1980.

PENG, H.; LONG, F.; DING, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 27, n. 8, p. 1226–1238, 2005.

PEREZ, B.; CASTELLANOS, C.; CORREAL, D. Applying data mining techniques to predict student dropout: a case study. 2018.

SARA, Nicolae-Bogdan; HALLAND, Rasmus; IGEL, Christian; ALSTRUP, Stephen. High-school dropout prediction using machine learning: a Danish large-scale study. In: **Proceedings. ESANN 2015: 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning**, p. 319–324, 2015.

SCHUHARDT, Oscar Luiz et al. A evasão no ensino superior brasileiro na percepção dos alunos evadidos: motivos e fatores apontados nos estudos entre os anos de 2014 e 2023. **SIMPPA – Simpósio de Pós-Graduação e Pesquisa em Administração**, 4., 2024 p. 1–13. Disponível em: https://ppa.uem.br/iv-simppa2024x/anais. Acesso em: 25 nov. 2024.

SPADY, William G. Dropouts from higher education: an interdisciplinary review and synthesis. **Interchange**, v. 1, n. 1, p. 64–85, 1970.

TETE, M. et al. Predictive models for higher education dropout: a systematic literature review. **Education Policy Analysis Archives**, v. 30, p. 149, 2022.

TINTO, Vincent. Dropout from higher education: a theoretical synthesis of recent research. **Review of Educational Research**, v. 45, n. 1, p. 89–125, 1975.

USP. USP desenvolve ferramenta para monitorar probabilidade de aluno não concluir o curso. **Jornal da USP**, São Paulo, 16 out. 2025. Disponível em: https://jornal.usp.br/institucional/usp-desenvolve-ferramenta-para-monitorar-probabilidade-de-aluno-nao-concluir-o-curso/. Acesso em: 2 nov. 2025.

VAARMA, Matti; LI, Hongxiu. Predicting student dropouts with machine learning: an empirical study in Finnish higher education. **Technology in Society**, v. 76, p. 102474, 2024. Disponível em: https://doi.org/10.1016/j.techsoc.2024.102474. Acesso em: 15 mar. 2025.

VILORIA, Amelec; LEZAMA, Omar Bonerge Pineda; VARELA, Noel. Bayesian classifier applied to higher educa**tion dropout. Procedia Computer Science, v. 160, p. 573–577, 2019.**

VILLAR, A.; DE ANDRADE, C. R. V. Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. **Discover Artificial Intelligence**, v. 4, n. 1, 2024. Disponível em: https://doi.org/10.1007/s44163-023-00079-z. Acesso em: 16 mar. 2025.

WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: towards a standard process model for data mining. **Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining**, 2000. p. 29–39.

YANG, S.; BERDINE, G. The receiver operating characteristic (ROC) curve. **The Southwest Respiratory and Critical Care Chronicles**, v. 5, n. 19, p. 34, 2017. Disponível em: https://doi.org/10.12746/swrccc.v5i19.391. Acesso em: 15 ago. 2025.