Análise Comparativa entre Métodos Tradicionais de Aprendizado de Máquina e Aprendizado Ativo na Predição da Evasão Escolar

Felipe Simão H. de Araújo[1], **Luciano de Souza Cabral**[1,2,4], **Rafael Ferreira Mello**[1,3,4]

[1]Centro de Inovação - Centro de Estudos Avançados do Recife (CESAR)
Avenida Cais do Apolo, 77, 50030-220 - Recife-PE - Brasil

[2]Campus Jaboatãos dos Guararapes - Instituto Federal de Pernambuco (IFPE)
Estr. de Bulhões, S/N, Bulhões, Jaboatão dos Guararapes-PE - Brasil

[3]Departamento de Computação - Universidade Federal Rural de Pernambuco (UFRPE)
Rua Dom Manuel de Medeiros, S/N, Dois Irmãos, Recife-PE - Brasil

[4]Núcleo de Excelência em Tecnologias Sociais (NEES)
Universidade Federal de Alagoas (UFAL)
Campus A. C. Simões, Av. Lourival de Melo Mota, Maceió-AL - Brasil

{fsha,lsc6}@cesar.school, rflm@cesar.org.br

Abstract. *This paper presents an automated approach for predicting student dropout in higher education, comparing traditional supervised learning models with active learning strategies. The study addresses two key questions: (1) How can predictive models be built effectively in an automated way? and (2) How can the need for extensive manual labeling be reduced without compromising performance? To this end, two distinct pipelines were developed. The traditional pipeline evaluated supervised algorithms such as XGBoost, LightGBM, and Random Forest, with model selection based on F1-score and hyperparameter optimization using Grid Search, Random Search, BayesSearchCV, and Optuna. The active learning pipeline focused on minimizing labeled data requirements while maintaining competitive performance. Results show that both approaches are effective in anticipating dropout risk, supporting more strategic and data-driven decision-making.*

Resumo. *Este artigo propõe uma abordagem automatizada para a predição da evasão no ensino superior, comparando modelos supervisionados tradicionais com estratégias baseadas em aprendizado ativo. O estudo buscou responder a duas questões centrais: (1) Como construir modelos preditivos eficazes de forma automatizada? e (2) Como minimizar a rotulagem manual sem comprometer o desempenho? Para isso, foram desenvolvidas duas pipelines. A tradicional avaliou algoritmos como XGBoost, LightGBM e Random Forest, com seleção via F1-Score e otimização de hiperparâmetros por Grid Search, Random Search, BayesSearchCV e Optuna. A pipeline com aprendizado ativo priorizou a redução da rotulagem, mantendo resultados competitivos. Os achados indicam que ambas as abordagens são eficazes na antecipação do risco de evasão, oferecendo suporte estratégico à tomada de decisão baseada em dados.*

DOI: 10.5753/wapla.2025.15898

1. Introdução

A evasão escolar é um desafio crítico enfrentado por instituições educacionais em todo o mundo, com impactos significativos na trajetória dos estudantes e no desenvolvimento socioeconômico das nações. No Brasil, o fenômeno afeta de forma mais intensa jovens em situação de vulnerabilidade, ampliando desigualdades sociais e comprometendo a inclusão produtiva [de Araujo et al. 2025].

Diversos fatores contribuem para a evasão, como baixo desempenho acadêmico, desmotivação, dificuldades financeiras, falta de apoio familiar e ausência de políticas públicas eficazes. A identificação precoce de estudantes em risco é, portanto, essencial para possibilitar intervenções direcionadas que promovam a permanência escolar [Bitencourt et al. 2021, Oliveira and Medeiros 2024].

Nos últimos anos, técnicas de aprendizado de máquina têm sido aplicadas com sucesso na previsão de evasão escolar, ao permitir a análise de grandes volumes de dados históricos e a identificação de padrões preditivos [Pimentel et al. 2023, Rimal et al. 2024]. Modelos supervisionados como regressão logística, Random Forest e XGBoost são amplamente utilizados para essa finalidade. No entanto, sua eficácia depende da disponibilidade de dados rotulados — cuja obtenção exige infraestrutura institucional e acompanhamento longitudinal [Silva 2022].

Além disso, dados educacionais reais são frequentemente escassos, desbalanceados e sujeitos a mudanças temporais. Nesses contextos, o aprendizado ativo surge como uma alternativa promissora, ao reduzir a necessidade de rotulagem extensiva por meio da seleção iterativa de exemplos informativos [Settles 2012].

Este trabalho propõe e compara duas pipelines automatizadas para predição da evasão escolar: uma baseada em aprendizado supervisionado tradicional e outra fundamentada em aprendizado ativo. Os resultados indicam que ambas são eficazes, sendo a segunda mais eficiente em cenários com recursos limitados de rotulagem.

2. Trabalhos Relacionados

Esta seção apresenta estudos relacionados à predição de evasão escolar por meio de técnicas de aprendizado de máquina. A seleção dos trabalhos foi realizada de forma exploratória, com foco em pesquisas que abordam: (i) algoritmos supervisionados aplicados à predição de evasão; (ii) métodos de aprendizado ativo em contextos educacionais; e (iii) desafios específicos do domínio, como desbalanceamento de dados e limitações na rotulagem. Embora não se trate de uma revisão sistemática, buscou-se contemplar abordagens representativas e recentes, priorizando publicações com experimentos aplicados a bases de dados públicas ou reais.

No âmbito do aprendizado supervisionado tradicional, Teodoro e Kappel [Teodoro and Kappel 2020] utilizaram cinco algoritmos — Naive Bayes, KNN, Árvores de Decisão, Random Forest e Redes Neurais — em dados públicos do INEP, com o objetivo de prever o risco de evasão no ensino superior público brasileiro. Os melhores resultados foram obtidos com o Random Forest, destacando variáveis como idade, carga horária e participação em atividades extracurriculares como as mais relevantes. Em linha com esses achados, Bitencourt et al. [Bitencourt et al. 2021] também adotaram Random Forest para predição de evasão em bases públicas, enfatizando a importância da seleção

criteriosa de atributos educacionais e corroborando a robustez do algoritmo em cenários supervisionados. De forma complementar, Pimentel et al. [Pimentel et al. 2023] compararam classificadores como SVM, regressão logística e Random Forest, evidenciando o bom desempenho deste último em bases desbalanceadas, especialmente em métricas como o F1-score.

Avançando para abordagens com menor dependência de dados rotulados, Costa et al. [Costa et al. 2020] investigaram o uso de aprendizado ativo em cenários educacionais. A partir de estratégias baseadas em incerteza, os autores demonstraram ser possível reduzir significativamente o número de exemplos rotulados necessários para treinar modelos com desempenho comparável ao de abordagens tradicionais. Um trabalho relacionado é o de Cabral et al. [Cabral 2023], que compararam diretamente aprendizado ativo e supervisionado na predição de desempenho em disciplinas de programação. Utilizando o conjunto de dados CodeBench, observaram que o aprendizado ativo não apenas manteve bons níveis de acurácia com menos dados, como também promoveu maior equidade nos resultados — um fator relevante em contextos educacionais sensíveis à justiça e viés preditivo, ainda que o foco principal não tenha sido evasão.

Complementando as discussões anteriores, Silva [Silva 2022] abordou os desafios operacionais enfrentados em projetos reais de predição de evasão, sobretudo no que diz respeito à rotulagem e à confirmação formal do abandono escolar. O autor destaca a necessidade de infraestrutura institucional para acompanhamento longitudinal e validação dos rótulos, o que torna abordagens como o aprendizado ativo ainda mais relevantes, dado seu foco em otimização do uso de dados rotulados.

Observa-se, portanto, uma lacuna na literatura quanto a comparações diretas e sistemáticas entre pipelines supervisionadas tradicionais e abordagens baseadas em aprendizado ativo no contexto da evasão escolar. Embora estudos anteriores tenham explorado individualmente essas técnicas, são raras as investigações que avaliem, sob condições controladas e com foco em desafios reais como o desbalanceamento de classes e a escassez de rótulos, os impactos relativos dessas estratégias sobre o desempenho preditivo. Essa ausência aponta para uma oportunidade relevante de aprofundamento na área.

3. Metodologia

Este estudo investigou a viabilidade do uso de aprendizado ativo na predição da evasão escolar no ensino superior, com foco na redução da necessidade de dados rotulados sem comprometer o desempenho dos modelos. Para isso, foram desenvolvidas duas pipelines automatizadas: uma tradicional, baseada em aprendizado supervisionado com dados totalmente rotulados, e outra fundamentada em aprendizado ativo com rotulagem progressiva.

3.1. Base de Dados e Pré-processamento

Utilizou-se a base *Higher Education Predictors of Student Retention* (Kaggle)[1], composta por 4.424 registros e 35 variáveis de uma instituição portuguesa. A variável-alvo foi binarizada: evasão (1) versus permanência ou graduação (0). As variáveis independentes abrangem aspectos demográficos, acadêmicos, socioeconômicos e macroeconômicos.

[1] https://encr.pw/QuzoS

O pré-processamento incluiu: (i) inspeção de valores ausentes (nenhum encontrado), (ii) remoção de variáveis altamente correlacionadas ($r > 0{,}85$), (iii) criação de variáveis derivadas (ex.: taxa de aprovação, média global), visando representar padrões de desempenho com menor complexidade. Essas etapas seguiram práticas recomendadas para evitar multicolinearidade e ruído [Dormann et al. 2013, Guyon and Elisseeff 2003].

3.2. Pipeline Tradicional

A pipeline supervisionada foi implementada utilizando a biblioteca *Streamlit*[2] e estruturada para executar automaticamente as etapas de pré-processamento, balanceamento de classes, validação cruzada e treinamento. Utilizou-se *holdout* com 85% dos dados para treino e 15% para teste, além de validação cruzada estratificada ($k = 5$).

Dada a natureza desbalanceada dos dados, foram testadas cinco estratégias de reamostragem: *Oversampling*, *Undersampling*, *SMOTE*, *Tomek Links* e *SMOTE + Tomek*. O conjunto de algoritmos incluiu oito classificadores supervisionados, com otimização de hiperparâmetros via *Grid Search*, *Random Search*, *BayesSearchCV* e *Optuna* [Akiba et al. 2019]. As métricas avaliadas foram Acurácia, Precisão, Recall, F1-Score e Coeficiente Kappa [McHugh 2012].

3.3. Pipeline de Aprendizado Ativo

A pipeline ativa simulou um ambiente com rotulagem limitada. Inicialmente, apenas 20% dos rótulos estavam disponíveis. A cada ciclo, o modelo selecionava exemplos informativos para rotulagem automática, com base em estratégias como *entropy sampling*, *margin sampling*, *random sampling*, *uncertainty + diversity* (via *KMeans*) e *Query by Committee* (QBC) [Ash et al. 2020].

O modelo base foi o *Random Forest*, reavaliado a cada iteração com validação cruzada estratificada e tuning via *Optuna* (TPE) [Nguyen et al. 2022]. As métricas utilizadas foram Acurácia, F1-Score, Precisão, Recall, AUC-ROC e Kappa, com foco na sensibilidade às classes minoritárias [Rodrigues et al. 2019].

4. Experimentos, Resultados e Discussão

Os experimentos foram conduzidos com a mesma base de dados e estratégias de validação padronizadas, permitindo comparação direta entre as pipelines supervisionada tradicional e aprendizado ativo. A avaliação dos modelos foi realizada com base em seis métricas: Acurácia, Precisão, Recall, F1-Score, AUC-ROC e Coeficiente Kappa, com foco específico na classe 1 (estudantes evadidos), por ser a de maior interesse institucional.

Essa ênfase é justificada pela relevância da detecção precoce da evasão, e está alinhada a diretrizes para classificação desbalanceada, que priorizam o desempenho sobre a classe minoritária [Luque et al. 2019]. Métricas como Recall indicam a capacidade de capturar corretamente os casos de evasão, enquanto o F1-Score equilibra precisão e sensibilidade — sendo, portanto, a métrica principal adotada para a seleção do melhor modelo em cada experimento. Tal escolha se justifica pela necessidade de manter um bom compromisso entre minimizar falsos negativos (não detectar um estudante que irá evadir) e evitar falsos positivos excessivos.

[2]Código-fonte disponível em: `https://github.com/fcsimao/AutoML`

4.1. Desempenho da Pipeline Tradicional

A abordagem tradicional supervisionada, com XGBoost como melhor modelo, apresentou os resultados consolidados na Tabela 1.

Tabela 1. Desempenho com aprendizado tradicional (100% dos dados rotulados)

Métrica	Valor
Acurácia	87,04%
Precisão	83,96%
Recall	73,70%
F1-Score	78,50%
AUC-ROC	89,00%
Kappa	69,29%

O modelo demonstrou desempenho robusto e consistente, especialmente em Acurácia e Precisão. A acurácia de 87,04% indica alto poder geral de classificação, enquanto a precisão de 83,96% sugere baixa taxa de falsos positivos, evitando intervenções desnecessárias. O recall de 73,70% revela boa capacidade de identificar evasores, embora com espaço para melhoria. O F1-Score (78,50%) confirma o equilíbrio entre os acertos, e o AUC-ROC de 89,00% evidencia excelente separabilidade entre classes. O coeficiente Kappa (69,29%) indica forte concordância ajustada ao acaso, conferindo robustez estatística ao modelo.

4.2. Desempenho da Pipeline de Aprendizado Ativo

A pipeline ativa foi iniciada com 20% dos rótulos disponíveis, utilizando *entropy sampling* para seleção iterativa de instâncias. Com apenas 40% de dados rotulados, obteve os resultados descritos na Tabela 2.

Tabela 2. Desempenho com aprendizado ativo (40% dos dados rotulados)

Métrica	Valor
Acurácia	84,33%
Precisão	74,66%
Recall	77,46%
F1-Score	76,03%
AUC-ROC	90,00%
Kappa	64,40%

Apesar do uso reduzido de dados rotulados, o modelo ativo demonstrou desempenho competitivo. Destaca-se o recall de 77,46%, superior ao modelo tradicional, indicando maior sensibilidade à identificação de evasores — uma característica essencial para aplicações educacionais. O AUC-ROC de 90,00% reforça sua capacidade discriminativa mesmo sob rotulagem limitada.

4.3. Análise Comparativa

A Figura 1 ilustra as métricas comparativas entre ambas as abordagens.

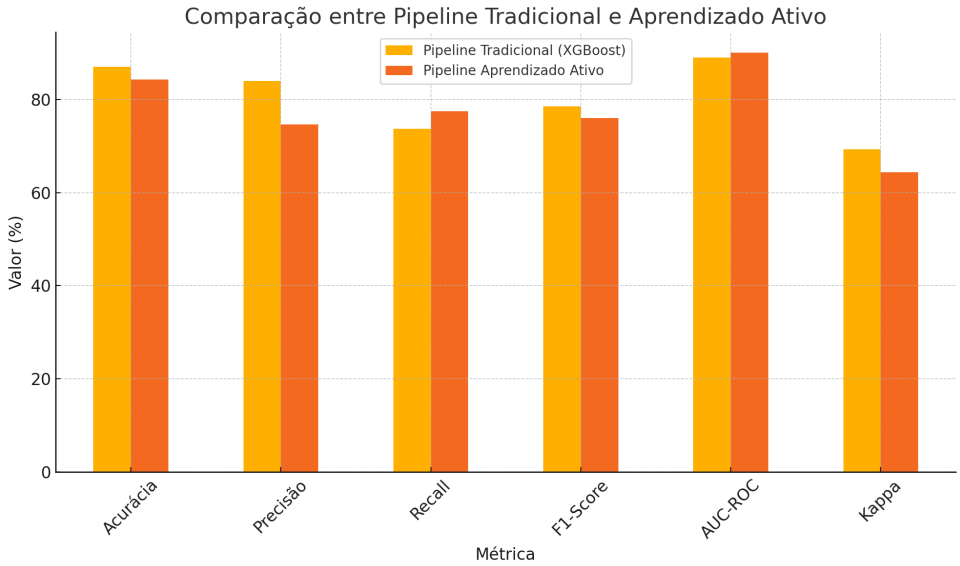

Figura 1. Comparação de desempenho entre aprendizado tradicional e ativo.

A pipeline supervisionada obteve desempenho superior em Acurácia (+2,7pp) e Precisão (+9,3pp), o que a torna adequada para cenários com dados completos e baixo custo de rotulagem. Por outro lado, a pipeline de aprendizado ativo superou o modelo tradicional em Recall (+3,8pp) e AUC-ROC (+1pp), além de manter um F1-Score competitivo (76,03% vs 78,50%), utilizando apenas 30% dos rótulos.

Esses resultados demonstram o potencial do aprendizado ativo em cenários com recursos limitados para anotação. Mesmo com um número reduzido de exemplos rotulados, os modelos mantiveram boa capacidade de detecção de estudantes evadidos e apresentaram desempenho comparável ao obtido pela abordagem supervisionada tradicional.

5. Conclusões, Limitações e Trabalhos Futuros

Os resultados obtidos evidenciam que ambas as abordagens investigadas — tradicional supervisionada e aprendizado ativo — são viáveis para a predição da evasão escolar, apresentando desempenhos quantitativamente semelhantes em métricas como F1-Score e AUC-ROC, com valores próximos a 0,80 na maioria dos experimentos. O principal diferencial da abordagem ativa está na eficiência no uso de dados rotulados: mesmo com apenas 40% dos exemplos anotados, os modelos mantiveram desempenho competitivo, especialmente na detecção da classe minoritária (evasores), com *recall* médio superior a 0,70.

Essa capacidade de identificar estudantes em risco com menos dados rotulados é especialmente relevante em contextos educacionais, onde o custo da anotação é elevado. No domínio da evasão escolar, falsos negativos tendem a ser mais críticos que falsos positivos, por representarem oportunidades perdidas de intervenção. Já o excesso de falsos positivos pode sobrecarregar os serviços de apoio, exigindo um equilíbrio entre sensibilidade e precisão. Nesse contexto, o uso do F1-Score como métrica principal revelou-se adequado por ponderar ambos os aspectos.

Apesar dos avanços, este estudo apresenta algumas limitações. A principal refere-se à generalização dos modelos, uma vez que os dados utilizados provêm de uma única instituição de ensino. Além disso, embora o aprendizado ativo reduza a necessidade de rotulagem, sua eficácia depende da representatividade da amostra inicial e da qualidade do feedback fornecido ao longo do processo de iteração. A ausência de variáveis qualitativas mais profundas — como aspectos motivacionais, emocionais e psicossociais — limita a sensibilidade do modelo frente à complexidade multifatorial do fenômeno da evasão. Ainda, embora a interface facilite a visualização dos resultados, ela requer mediação técnica para correta interpretação, o que pode dificultar sua adoção por gestores com menor familiaridade com ferramentas analíticas.

Como perspectivas futuras, recomenda-se explorar abordagens híbridas que integrem aprendizado ativo a métodos semi-supervisionados ou de *transfer learning*, visando ampliar a aplicabilidade do modelo em contextos com escassez ou heterogeneidade de dados. A validação institucional dessas soluções em ambientes reais, com fluxos contínuos e dados em tempo real, também se mostra fundamental para sua consolidação.

Outra direção promissora envolve a incorporação de variáveis comportamentais, como engajamento acadêmico, participação extracurricular e fatores emocionais, que podem ser capturados por meio de autoavaliações, monitoramento digital ou acompanhamento psicopedagógico. A inclusão desses atributos tende a melhorar a sensibilidade preditiva, permitindo a detecção mais precoce e contextualizada de riscos de evasão.

Por fim, destaca-se a importância do desenvolvimento de interfaces explicativas e acessíveis, capazes de traduzir a lógica dos modelos para os profissionais da educação. Ferramentas como SHAP, LIME e *dashboards* integrados a plataformas de *Business Intelligence* (BI) são fundamentais para promover transparência, fomentar a confiança institucional e subsidiar ações pedagógicas fundamentadas em evidências.

Referências

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, pages 2623–2631. Association for Computing Machinery.

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. (2020). Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Bitencourt, W. A., Silva, D. M., and Xavier, G. C. (2021). Pode a inteligência artificial apoiar ações contra evasão escolar universitária? *Ensaio: Avaliação e Políticas Públicas em Educação*, 29(111).

Cabral, L. (2023). Assessing algorithmic fairness: A comparison of traditional machine learning and active learning methods. Relatório de pesquisa pós-doutoral PD-2025-001, CESAR School. Relatório de pesquisa pós-doutoral.

Costa, M. J. S., Silva, A. A., and Andrade, F. C. S. (2020). Aprendizado ativo para predição da evasão escolar com uso eficiente de dados rotulados. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 343–352. SBC.

de Araujo, C. L., Santos, Q. P., Ribeiro, H. M. L., do Nascimento de Freitas, E. B., and Coutinho, D. J. G. (2025). Evasão escolar: Causas e impactos da evasão escolar no brasil e no mundo. *Revista Ibero-Americana de Humanidades, Ciências e Educação — REASE*, 11(1).

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3:1157–1182.

Luque, A., Carrasco, A., Martín, A., and de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Nguyen, Q.-H., Nguyen, M.-T., Pham, V.-T., Dinh, D.-P., Seo, E. C., and Chung, T.-M. (2022). Deep active learning with semi-supervised training for covid-19 detection from chest ct images. *Electronics*, 11(18):2893.

Oliveira, R. d. S. and Medeiros, F. P. A. d. (2024). Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação. *Revista Brasileira de Informática na Educação (RBIE)*, 32:1–21.

Pimentel, M. S., da Silva, C. B., and Gomes, F. F. B. (2023). Análise de dados com machine learning: Classificação de alunos em risco de evasão escolar utilizando modelos de machine learning. *Apoena - Revista de Educação e Pesquisa*, 9(2):45–63.

Rimal, Y., Sharma, N., and Alsadoon, A. (2024). The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. *Multimedia Tools and Applications*.

Rodrigues, F. D. S., Viana, W. O., Figueiredo, K. L., dos Santos, R. C., da Silva, A. G. P., and Zárate, L. E. (2019). Evaluating machine learning classifiers for predicting student dropout in higher education using imbalanced data. *Education Sciences*, 9(4):275.

Settles, B. (2012). *Active Learning*, volume 6 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers.

Silva, J. J. d. (2022). Uma comparação de técnicas de aprendizado de máquina para predição de evasão de estudantes no ensino público superior. Dissertação de mestrado, Universidade de São Paulo.

Teodoro, L. d. A. and Kappel, M. A. A. (2020). Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no brasil. *Revista Brasileira de Informática na Educação*, 28(0).