PREDIÇÃO DA EVASÃO NO ENSINO SUPERIOR BRASILEIRO: UMA ABORDAGEM DE APRENDIZADO DE MÁQUINA COM DADOS DA PNAD

Túlio Oliveira Prates, Analucia Vieira Fantin, Marcos Vinicio Wink Junior

INTRODUÇÃO

A evasão no ensino superior é um problema relevante para a política educacional brasileira por envolver perdas individuais, sociais e econômicas, reduzindo oportunidades no mercado de trabalho e a formação de capital humano, além de representar possível desperdício de recursos públicos. O problema também apresenta forte componente distributivo: nos dados utilizados, a taxa de evasão é de 29,44%, alcançando 49,75% entre indivíduos de famílias com renda domiciliar per capita de até um quarto do salário mínimo e caindo para 15,11% entre aqueles com renda de três a cinco salários mínimos.

A literatura associa a evasão a fatores acadêmicos, socioeconômicos, familiares e institucionais (AINA et al., 2021). Mais recentemente, técnicas de aprendizado de máquina têm ampliado a análise desses fatores (KOCSIS; MOLNÁR, 2024). Entretanto, muitos estudos utilizam bases de universidades específicas, limitando a generalização dos resultados (Vaarma; LI, 2024; EEGDEMAN et al., 2022; BERENS et al., 2018).

Este trabalho utiliza microdados da Pesquisa Nacional por Amostra de Domicílios (PNAD), do IBGE, para predizer a evasão no ensino superior brasileiro por meio de Random Forest e interpretar os fatores associados às previsões utilizando valores SHAP. A contribuição está em combinar dados domiciliares nacionalmente representativos com aprendizado de máquina interpretável, buscando maior potencial de generalização para políticas de permanência. O estudo também avalia a robustez à retirada da idade e discute particularidades do contexto brasileiro, especialmente a relação entre emprego e evasão.

DESENVOLVIMENTO

A pesquisa utiliza microdados da Pesquisa Nacional por Amostra de Domicílios (PNAD), do Instituto Brasileiro de Geografia e Estatística (IBGE), reunindo os suplementos de educação de 2022, 2023 e 2024. O pré-processamento manteve indivíduos que concluíram ou abandonaram um curso superior, definindo a variável dependente como 1 para evasão e 0 para conclusão. A amostra final contém 7.621 observações e variáveis socioeconômicas e demográficas, além de uma proxy do ambiente educacional familiar baseada na maior escolaridade observada no domicílio.

A amostra foi restrita a indivíduos de 18 a 24 anos e com renda domiciliar per capita de até cinco salários mínimos. A restrição de renda elevou o F1-score em aproximadamente quatro pontos percentuais, enquanto o recorte etário busca reduzir problemas de temporalidade, uma vez que a PNAD não informa a data exata de conclusão da graduação. Entre os diplomados da amostra, 89,34% possuem entre 22 e 24 anos.

Para a classificação, foi utilizado o algoritmo Random Forest, com hiperparâmetros ajustados por busca em grade e validação cruzada em cinco partes, priorizando o F1-score ponderado. A amostra foi dividida de forma estratificada entre treinamento, com 6.097 observações (80%), e teste, com 1.524 (20%). O desempenho foi avaliado por precisão, recall, F1-score e AUC-ROC.

A interpretação das previsões foi realizada pelo método SHAP (SHapley Additive exPlanations), que permite identificar a importância e a direção da contribuição das variáveis para a probabilidade **prevista** de evasão. Por fim, foi estimada uma especificação sem a variável idade para avaliar a robustez dos resultados e verificar a dependência do modelo em relação a esse preditor.

RESULTADOS

O Random Forest apresentou capacidade satisfatória de discriminação entre evasores e concluintes. No conjunto de teste, obteve precisão de 0,53, recall de 0,62, F1-score de 0,57 e AUC-ROC de 0,73. O recall superior à precisão indica maior capacidade de identificar indivíduos que efetivamente evadiram, ainda que com maior ocorrência de falsos positivos, característica potencialmente relevante para políticas de permanência.

A análise por SHAP destaca a importância de características socioeconômicas, familiares e demográficas. A baixa renda domiciliar está associada a maior probabilidade prevista de evasão: indivíduos no quintil inferior apresentam contribuição média positiva de 1,44 ponto percentual, enquanto aqueles no quintil superior apresentam contribuição negativa de 1,34 ponto. A escolaridade familiar também se mostra relevante: a presença de outro membro do domicílio com ensino superior completo reduz a previsão de evasão em aproximadamente 1,46 ponto percentual. Homens apresentam maior risco previsto que mulheres.

Um resultado particular refere-se ao emprego. Diferentemente de parte da literatura internacional, estar empregado está associado a menor previsão de evasão (-0,25 ponto percentual), enquanto não estar empregado apresenta contribuição positiva (0,73 ponto). Uma possível interpretação relaciona-se à elevada participação das instituições privadas no ensino superior brasileiro e à necessidade de renda para financiar os custos de permanência. O resultado, contudo, representa uma associação preditiva, e não causal.

A exclusão da idade, potencialmente afetada pela construção da variável de resultado, reduziu moderadamente o desempenho do modelo: no teste, o F1-score passou de 0,57 para 0,52 e a AUC-ROC de 0,73 para 0,68. Entretanto, os principais preditores permaneceram semelhantes, indicando que o poder preditivo não depende exclusivamente dessa variável.

Em conjunto, os resultados mostram que informações socioeconômicas de uma pesquisa domiciliar nacional podem contribuir para identificar perfis associados à evasão. A combinação de Random Forest e SHAP permite conciliar capacidade preditiva e interpretabilidade, embora a ausência de informações acadêmicas, o risco de temporalidade e os recortes de renda e idade recomendem cautela. Os achados devem, portanto, ser interpretados como associações preditivas com potencial para subsidiar políticas de permanência e orientar investigações futuras.

CONSIDERAÇÕES FINAIS

Os resultados indicam que variáveis socioeconômicas, familiares, demográficas e educacionais disponíveis na PNAD possuem capacidade relevante para predizer a evasão no ensino superior brasileiro. A análise por Random Forest e SHAP destaca renda domiciliar, escolaridade familiar, sexo e características relacionadas ao trabalho entre os principais preditores, ressaltando o papel das desigualdades socioeconômicas na permanência estudantil. Destaca-se a associação entre emprego e menor risco previsto de evasão, resultado que pode refletir particularidades do ensino superior brasileiro e a importância da renda do trabalho para a permanência. A análise sem idade preservou o padrão dos principais preditores, embora a ausência de indicadores acadêmicos e o risco de causalidade reversa recomendem cautela. Em conjunto, o estudo evidencia o potencial de pesquisas domiciliares nacionais combinadas a métodos de aprendizado de máquina interpretáveis para subsidiar políticas de permanência, enquanto futuras pesquisas poderão incorporar dados longitudinais e acadêmicos para ampliar a precisão e a compreensão dos fatores associados à evasão.

Palavras-chave: evasão no ensino superior; aprendizado de máquina; Random Forest; SHAP; permanência estudantil.

REFERÊNCIAS BIBLIOGRÁFICAS

AINA, C. et al. The Determinants of University Dropout: a Review of the Socio-economic Literature. Socio-Economic Planning Sciences, v. 79, 2021, 101102. DOI: 10.1016/j.seps.2021.101102.

BREIMAN, L. Random Forests. Machine Learning, v. 45, 2001, p. 5-32.

EEGDEMAN, I. et al. Identifying false positives when targeting students at risk of dropping out. Education Economics, 2022, p. 1-13. DOI: 10.1080/09645292.2022.2067131.

KOCSIS, Á.; MOLNÁR, G. Factors influencing academic performance and dropout rates in higher education. Oxford Review of Education, v. 51, 2024, p. 1-19. DOI: 10.1080/03054985.2024.2316616.

VAARMA, M.; LI, H. Predicting student dropouts with machine learning: An empirical study in Finnish higher education. Technology in Society, v. 76, 2024, 102474. DOI: 10.1016/j.techsoc.2024.102474.