Universidade Federal do Rio Grande - FURG
23ª Mostra da Produção Universitária – MPU
Rio Grande/RS, Brasil, 27 a 29 de novembro de 2024
ISSN: 2317-4420

Preditores socioeconômicos da evasão universitária: uma aplicação de aprendizado de máquina com *Random Forest* e valores SHAP para a FURG

Guimarães, Matheus.

TILLMANN, Eduardo
math.guisimoni@gmail.com
Universidade Federal do Rio Grande

Palavras-chave: Modelagem preditiva; Evasão Escolar; Socioeconômico; Inteligência Artificial; Modelos Interpretáveis.

1 INTRODUÇÃO

A evasão escolar é um problema persistente no Brasil em todos os níveis educacionais. Em 2023, a taxa de evasão no ensino presencial foi de 30,7% (Mapa do Ensino Superior, 2023). De acordo com Tinto (1975) este fenômeno é complexo e pode envolver uma combinação de fatores pessoais, institucionais e sociais. Estudos recentes evidenciam que entre as principais causas estão fatores socioeconômicos, como dificuldades financeiras, falta de apoio educacional adequado, infraestrutura escolar insuficiente e disparidades sociodemográficas (Instituto Ayrton Senna, 2023). Este trabalho propõe a construção de um modelo de Machine Learning para classificar o risco de evasão dos estudantes da Universidade Federal do Rio Grande com base em características socioeconômicas. Utilizando Python, o classificador *Random Forest* (Breiman, 2001) e valores SHAP (Scott et al, 2017) foram empregados para identificar as variáveis mais impactantes. O modelo apresentou precisão de 83% nos dados de ajuste e 71% na validação. As variáveis com maior impacto na evasão foram a cor do indivíduo, idade e o tipo de escola frequentada.

2 METODOLOGIA

Para atingir o objetivo do estudo, foi utilizado o modelo *Random Forest*, um método *ensemble* - onde se constrói múltiplas árvores de decisão classificadoras e, em cada árvore, é emitido um "voto" para uma classe específica, e a classe que recebe o maior número de votos é selecionada como a predição final para a observação ***i***. Conforme Breiman (2001), dado $h(x)$ árvores classificadoras, a função margem do *ensemble* é dado por:

$$mg(X,Y) = \frac{1}{K}\sum_{k=1}^{K} I\left(h_k(X) = Y\right) - max(j \neq Y)\frac{1}{K}\sum_{k=1}^{K} I\left(h_k(X) = j\right) \quad (1)$$

Universidade Federal do Rio Grande - FURG
23ª Mostra da Produção Universitária – MPU
Rio Grande/RS, Brasil, 27 a 29 de novembro de 2024
ISSN: 2317-4420

A Equação 1 representa a diferença entre a proporção média dos votos para a classe correta e a proporção média máxima dos votos para a outra classe incorreta (em um problema binário), funcionando como uma medida de confiança do modelo para a previsão certa.

Além disso, para identificar as variáveis que mais influenciaram o modelo a classificar os indivíduos que evadiram, utilizou-se a metodologia SHAP (Shapley Additive Explanations). Proposto por Lundberg e Lee (2017), o método SHAP é uma aplicação dos valores de Shapley da teoria dos jogos cooperativos. De acordo com Roth (1988), o valor de Shapley é uma solução que atribui a cada jogador (neste caso, uma variável) um valor único, representando sua contribuição ao resultado final do jogo (ou da predição). No que tange os dados, estes foram retirados do sistemas FURG e abrangem o primeiro e o segundo semestre de 2012 e 2013.

3 **RESULTADOS E DISCUSSÃO**

O modelo *Random Forest*, através de um processo iterativo, obteve seu melhor desempenho através de 500 árvores de decisão. Nesse parâmetro obteve-se 83,34% de acurácia no conjunto de ajuste (80% dos dados) e 70,69% no conjunto de validação (20% dos dados). Foi considerada a capacidade de generalização do modelo, isto é, manter uma margem de erro para cobrir o viés de variável relevante omitida. Na tabela a seguir, para os dados de teste, destaca-se a precisão e a revocação - a primeira mede a proporção de previsões corretas sobre o total de previsões de uma determinada classe, e a segunda mede a proporção de previsões exatas sobre o total de observações de uma classe em questão (Martin, 2008).

Tabela 1 – Precisão e revocação

Classe	***Precisão***	***Revocação***
Formado	62%	66%
Evadido	78%	74%

As variáveis que foram identificadas como fatores de risco, que nesse contexto, é definida como variáveis socioeconômicas que aumentam a probabilidade do indivíduo de evadir, foram:

Tabela 2 – Fatores de risco

Variável	***Acréscimo médio na probabilidade de evadir***
Dbranca	-9,05%
Idade>40	+2,71%

ISSN: 2317-4420

Aux Aliment	-1,40%
Escola púb	+1,39%
Mulher	+1,35%
Aux Transp	-1,30%
Casa própria	+0,09%

A variável de maior impacto, uma *dummy* que toma valor 1 se o indivíduo é branco e 0 se for preto, pardo ou indígena, evidencia que se o indivíduo for branco, há uma redução média de 9,05% em evadir. Outras variáveis significativas como ter mais de 40 anos, vir de escola pública, ser mulher e ter casa própria também aumentam as chances. Observa-se também que ser beneficiado(a) com os auxílios de alimentação e transporte disponibilizados pela Pró-reitoria de Assuntos Estudantis traz um impacto positivo quanto às chances de concluir a graduação.

O modelo criado evidencia uma tendência de menor probabilidade de evasão quando o indivíduo está coberto por privilégios sociais - no sentido exposto por Black e Stone (2005), em que esses se dão com base na educação, idade, classe social, gênero, raça, deficiência, entre outros. Paralelamente, observa-se que os benefícios estudantis impactam positivamente na permanência dos indivíduos na graduação. Esse resultado vai de encontro com estudo realizado por Ferreira (2022) que demonstra o impacto positivo dos auxílios estudantis na permanência dos estudantes no ensino superior e em seus rendimentos acadêmicos.

4 CONSIDERAÇÕES FINAIS

Para uma análise mais robusta e maior capacidade de previsão e interpretabilidade do modelo, é necessário obter dados mais recentes acerca da evasão na Universidade Federal do Rio Grande. Porém, somente com os dados disponíveis e os resultados obtidos, evidencia-se a importância da utilização dos algoritmos de aprendizado de máquina para prever os estudantes que estão em zona de risco de evasão para, assim, formular ações para mitigar os efeitos de suas vulnerabilidades socioeconômicas em sua permanência no ensino superior.

5 REFERÊNCIAS

MAPA DO ENSINO SUPERIOR 2023. *SEMEPS*, 2023. Disponível em: https://www.semesp.org.br/mapa/edicao-13/. Acesso em: 25 de agosto de 2024.

INSTITUTO AYRTON SENNA. *Abandono escolar: principais causas e formas de enfrentamento*. Disponível em: https://institutoayrtonsenna.org.br/abandono-escolar/#:~:text=Dificuldades%20financ

ISSN: 2317-4420

eiras%3A%20a%20necessidade%20de,podem%20levar%20ao%20abandono%20escolar. Acesso em: 25 de agosto de 2024.

K. C. CORDOVA, R. E. KINAI, M. J. S. C. FELLEGI, and A. S. G. TESSMANN. *Evaluating Variance in Random Forest Classifier Performance*. arXiv, 2017. Disponível em: https://arxiv.org/abs/1705.07874. Acesso em: 25 de agosto de 2024.

ROTH, Alvin E. (Ed.). The Shapley Value: Essays in Honor of Lloyd S. Shapley. Cambridge: Cambridge University Press, 1988. Disponível em: https://www.cambridge.org/core/books/shapley-value/D3829B63B5C3108EFB62C4009E2B966E. Acesso em: 26 ago. 2024.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874, 2017. Disponível em: https://arxiv.org/abs/1705.07874. Acesso em: 26 ago. 2024.

TINTO, Vincent. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. Review of Educational Research, v.45, n.1, p.89-125, 1975. Disponível em: https://www.jstor.org/stable/1170024. Acesso em: 27 ago. 2024.

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, 2(1), 37-63. Disponível em: https://www.researchgate.net/publication/228529307_Evaluation_From_Precision_Recall_and_F-Factor_to_ROC_Informedness_Markedness_Correlation. Acesso em: 28 ago. 2024

BLACK, Linda L.; STONE, David. Expanding the Definition of Privilege: The Concept of Social Privilege. *Journal of Multicultural Counseling and Development*, v. 33, n. 4, p. 243-255, out. 2005.

Cassiano Roberto Ferreira Julião, Luiz Ismael Pereira, Marco Aurélio Marques Ferreira. O impacto do programa nacional de assistência estudantil no desempenho dos discentes brasileiros de baixa renda. Revista Gestão Universitária na América Latina - GUAL, [1] v. 15, n. 1, janeiro de 2022.