ANÁLISE PREDITIVA DA EVASÃO NO ENSINO SUPERIOR: UM ESTUDO COMPARATIVO ENTRE INSTITUIÇÕES E MODALIDADES

Carlos Eduardo da Paixão Bravin (PIC/UEM), Eduarda Barragan da Silva (PIC/UEM), Valéria Delisandra Feltrim (Orientadora). E-mail: vdfeltrim@uem.br.

Universidade Estadual de Maringá, Centro de Tecnologia, Maringá, PR.

Ciência da Computação, Metodologia e Técnicas da Computação

Palavras-chave: Aprendizado de Máquina; Evasão Acadêmica; Desbalanceamento das classes.

RESUMO

Este trabalho teve como objetivo desenvolver e avaliar modelos de Aprendizado de Máquina para a predição da evasão no ensino superior comparando uma instituição pública e uma privada nas modalidades presencial e a distância. Utilizaram-se bases institucionais, reunindo informações estáticas de ingresso e dinâmicas da trajetória acadêmica, como desempenho e frequência. Essas informações foram modeladas temporalmente por meio de *snapshots*. Aplicaram-se os algoritmos Random Forest e XGBoost, avaliando-se também o impacto do desbalanceamento dos dados. Os resultados indicaram alta capacidade preditiva na instituição pública, com medida-F entre 0,877 e 0,925, mas desempenho inferior a 0,20 na privada, caracterizado por sensibilidade elevada e baixa precisão. Essa diferença pode estar associada à maior proporção de evadidos na instituição pública e à acentuada divergência entre as distribuições dos conjuntos de treinamento e teste na instituição privada, o que também limitou a eficácia do balanceamento de dados. Conclui-se que o desempenho depende fortemente das características de distribuição de cada base, exigindo cautela na comparação direta entre contextos educacionais distintos.

INTRODUÇÃO

A evasão acadêmica pode ser compreendida como o abandono do curso pelo estudante antes de sua conclusão, sem a intenção de retomar os estudos (ARAÚJO, 2012). Além de afetar a trajetória acadêmica e profissional do estudante, esse rompimento representa perdas de recursos e investimentos para as instituições de ensino e para a sociedade.

Nesse contexto, técnicas de Aprendizado de Máquina (AM) apresentam-se como alternativa para identificar, a partir de dados institucionais, padrões associados à evasão, o que permite às coordenações e aos setores de apoio pedagógico atuar de forma preventiva e direcionar melhor os recursos para retenção.

Diante disso, o presente estudo tem como objetivo desenvolver e avaliar modelos de AM para a predição da evasão no ensino superior, a partir de dados de estudantes da Universidade Estadual de Maringá (UEM) e do Centro Universitário

Cidade Verde (Unicive). O estudo dá continuidade à pesquisa desenvolvida por Minache (2024), que investigou a predição de evasão acadêmica na UEM. Como avanço em relação ao estudo anterior, foram incorporadas informações dinâmicas referentes ao percurso acadêmico dos estudantes e a comparação entre diferentes contextos institucionais, contemplando uma instituição pública e uma privada, bem como cursos presenciais e de Educação a Distância (EAD). Busca-se, assim, avaliar a capacidade preditiva dos modelos e identificar quais características são relevantes para a predição da evasão em diferentes contextos educacionais.

MATERIAIS E MÉTODOS

Foram utilizados dados anonimizados da UEM, universidade pública, e do Unicive, instituição privada. Foram recuperados 85.275 registros referentes a 23 anos na UEM e 102.930 registros referentes a 17 anos no Unicive, contemplando estudantes das modalidades presencial e EAD.

Os dados reuniram informações estáticas, relacionadas ao ingresso e às características acadêmicas dos estudantes, e informações dinâmicas, referentes à trajetória acadêmica, como desempenho em disciplinas, frequência e reprovações.

Para a realização dos experimentos, foram gerados diferentes *snapshots*, que representam recortes da trajetória acadêmica dos estudantes em diferentes momentos, conforme a abordagem de Barros (2022).

O estudo foi aprovado pelo Comitê Permanente de Ética em Pesquisa com Seres Humanos da UEM (COPEP; CAAE 89887025.3.0000.0104).

O problema foi modelado como uma classificação binária, na qual a variável-alvo indica a ocorrência ou não de evasão. Na etapa de pré-processamento, os atributos numéricos foram padronizados por meio de *StandardScaler* e os categóricos foram codificados por *One-Hot Encoding*.

Em ambas as instituições, há desbalanceamento de classes, sendo os estudantes evadidos a classe minoritária. Para mitigar esse efeito, avaliou-se a aplicação do SMOTE (*Synthetic Minority Over-sampling Technique*). Utilizaram-se dois algoritmos de AM: Random Forest (RF) *e* XGBoost (XGB). A escolha baseou-se nos resultados apresentados por Minache (2024), mantendo a continuidade com a abordagem adotada no estudo anterior.

Por fim, a avaliação foi realizada sobre os conjuntos de teste, mantendo sua distribuição original de classes. Como métricas, foram utilizadas precisão, sensibilidade e medida-F.

RESULTADOS E DISCUSSÃO

Esta seção apresenta o desempenho dos modelos nas duas instituições (Tabela 1). As métricas utilizadas referem-se à classe dos estudantes evadidos e são apresentadas como média e desvio padrão dos *snapshots*. Os resultados foram agregados por modalidade.

Tabela 1 – Desempenho preditivo por modalidade e modelo na UEM e Unicive

Inst.	Moda.	Mod.	SMOTE	Precisão	Sensi.	Medida-F
Unicive	Presencial	RF	Não	0,095±0,159	0,671±0,325	0,139±0,177
			Sim	0,084±0,155	0,743±0,313	0,126±0,173
		XGB	Não	0,088±0,154	0,722±0,325	0,132±0,175
			Sim	0,086±0,156	0,755±0,310	0,129±0,175
	EAD	RF	Não	0,135±0,212	0,946±0,053	0,191±0,226
			Sim	0,133±0,209	0,950±0,042	0,192±0,224
		XGB	Não	0,127±0,205	0,959±0,031	0,185±0,219
			Sim	0,126±0,203	0,958±0,029	0,185±0,217
UEM	Presencial	RF	Não	0,939±0,083	0,917±0,049	0,924±0,047
		XGB	Não	0,924±0,088	0,931±0,042	0,924±0,049
	EAD	RF	Não	0,876±0,220	0,931±0,102	0,876±0,199
		XGB	Não	0,895±0,218	0,948±0,056	0,896±0,193

Inst. = Instituição; Moda. = Modalidade; Mod. = Modelo; Sensi. = Sensibilidade

Na UEM, os resultados apresentaram desempenho semelhante nas duas modalidades. Apesar dos valores elevados em ambas, observou-se redução das métricas no EAD, especialmente no modelo RF, com redução de 0,048 na medida-F. Além disso, os resultados do EAD apresentaram maior desvio padrão, indicando maior variação do desempenho entre os períodos, enquanto a modalidade presencial se mostrou mais estável.

Diante desses valores, analisou-se o comportamento das características acadêmicas ao longo da base. Nos anos de 2008, 2012, 2016, 2020 e 2024, os estudantes evadidos apresentaram médias de notas inferiores às dos não evadidos, sugerindo associação entre desempenho acadêmico e evasão.

No Unicive, a medida-F foi inferior a 0,20 em todos os cenários, com desempenho superior no EAD em relação ao presencial. A sensibilidade elevada associada à baixa precisão indica que os modelos identificam grande parte dos estudantes que evadem, mas também classificam como evadidos muitos estudantes que permanecem. A dispersão da medida-F foi elevada nas duas modalidades, com desvios padrão superiores às respectivas médias, enquanto a sensibilidade apresentou maior estabilidade no EAD do que no presencial.

Diante do baixo desempenho dos modelos, aplicou-se o SMOTE que aumentou a sensibilidade no presencial, mas reduziu a precisão e a medida-F, enquanto, no EAD, as diferenças entre as condições foram inferiores a 0,002. Esses resultados indicam que o balanceamento do conjunto de treinamento não foi suficiente para melhorar o desempenho dos modelos. Uma possível causa está na divergência entre as distribuições dos conjuntos de treinamento e teste. Em média, a proporção de evadidos no treinamento superou a do teste em 26,9 pontos percentuais no presencial e 56,9 pontos percentuais no EAD, chegando a 31,4% contra 3,1% no segundo semestre de 2020. Esse deslocamento é compatível com a elevada sensibilidade e a baixa precisão, mas não explica isoladamente os

resultados. Embora o EAD tenha apresentado maior divergência entre treinamento e teste, sua medida-F foi levemente superior à observada no presencial.

CONCLUSÕES

O presente estudo reforça a relevância da predição da evasão no ensino superior e do uso de técnicas de AM para sua identificação, avançando em relação a Minache (2024) ao incorporar informações dinâmicas da trajetória acadêmica e comparar contextos institucionais distintos. Na UEM, os modelos XGB e RF apresentaram resultados elevados, observando-se associação entre o desempenho acadêmico e a evasão. No Unicive, por outro lado, o desempenho manteve-se baixo mesmo após a aplicação do SMOTE, possivelmente em razão das divergências entre as distribuições dos conjuntos de treinamento e teste. Essa limitação, decorrente dos dados disponíveis, reforça que a capacidade preditiva depende das características específicas de cada base.

Como trabalhos futuros, pretende-se investigar as características de maior relevância para os modelos e avaliar estratégias para melhorar as métricas das predições na instituição privada, contribuindo para o desenvolvimento de ferramentas de apoio à prevenção da evasão.

AGRADECIMENTOS

Agradecemos ao Programa de Educação Tutorial (PET), financiado pelo FNDE/MEC, pela bolsa concedida ao autor Carlos Eduardo da Paixão Bravin e ao Programa de Apoio à Educação Tutorial (PRÓ-PET), financiado pela Fundação Araucária, pela bolsa concedida à autora Eduarda Barragan da Silva.

REFERÊNCIAS

ARAÚJO, E. de J. M. de. Evasão no PROEJA: Estudo das causas no Instituto Federal de Educação, Ciência e Tecnologia do Maranhão/IFMA. 2012. 116 f. Dissertação (**Mestrado em Educação**) – Universidade Católica de Brasília, Brasília, 2012.

BARROS, B. D. M. Aprendizado de máquina automático aplicado à predição da Evasão no Ensino Superior. 2022. **Dissertação** (Mestrado em Ciência da Computação) – Universidade Federal de Goiás (UFG), 2022.

MINACHE, R. B. Previsão de evasão acadêmica por meio de aprendizagem de máquina. 2024. **Trabalho de Conclusão de Curso** (Bacharelado em Ciência da Computação) - Universidade Estadual de Maringá, Maringá, 2024.