Aprendizado de Máquina Aplicado à Análise de Evasão em Cursos de Sistemas de Informação

Patricia D. Santos[1], **Denise H. Goya**[1]

[1]Universidade Federal do ABC (UFABC)
Av. dos Estados, 5001 – Bairro Santa Terezinha – Santo André – SP – Brazil

{patricia.santos,denise.goya}@ufabc.com.br

Abstract. *This study aims to identify relevant attributes for predicting dropout of students in Information Systems courses at Brazilian institutions, by comparing several machine learning approaches, including Random Forest, Adaptive Boost (AdaBoost), K-Nearest Neighbor (KNN), Logistic Regression and Voting Classifier. The results show promising recall (85,6%), accuracy (87,1%) and precision (92,4%) using as little as 14 attributes.*

Resumo. *Este estudo tem por objetivo identificar atributos relevantes para a previsão de evasão de alunos em cursos de Sistemas de Informação de instituições brasileiras, comparando várias abordagens de aprendizado de máquina, incluindo Florestas Aleatórias, AdaBoost, Vizinho mais próximo de K, Regressão Logística e Classificador de Votação. Os resultados mostram cobertura (85,6%), acurácia (87,1%) e precisão (92,4%) promissoras usando apenas 14 atributos.*

1. Introdução

No Brasil, a evasão atinge diferentes cursos e instituições e se caracteriza não só pelo desligamento da universidade, mas também pelo abandono ou trancamento de matrícula do estudante em qualquer etapa do curso ou transferência para outro curso [Solis et al. 2018]. Já a persistência seria o ato de o aluno se manter no curso até a sua diplomação [Tinto and Cullen 1973].

A integração das características dos estudantes com as da universidade são a chave para entender a evasão. Vários tipos de atributos têm sido usados nesse tipo de pesquisa como preditores de abandono, tais como os relacionados às características acadêmicas, financeiras e sociodemográficas [Delen 2010] e escola secundária de origem estudantil. A combinação dessas variáveis possibilita a melhoria da sensibilidade na análise de modelos preditivos e fornece importantes informações sobre fatores individuais e institucionais que podem aumentar a probabilidade de desistência [Solis et al. 2018].

Este trabalho usa vários algoritmos de aprendizado de máquina para prever a evasão de estudantes de cursos de Sistemas de Informação no Brasil. Para isso, foram incluídos atributos relacionados às características sociodemográficas dos alunos, dos programas e das instituições de ensino.

2. Metodologia

Foram utilizados dados referentes aos cursos de Sistemas de Informação coletados pelo Censo da Educação Superior entre os anos de 2014 e 2018 e disponibilizados pelo Inep. Duas perspectivas foram adotadas para prever a evasão:

- Perspectiva 1: considera evasão um aluno que tenha solicitado trancamento de curso, tenha sido transferido para outro curso ou tenha sido desligado da instituição, enquanto persistência é um aluno com matrícula ativa ou que concluiu seus estudos. Um total de 223.687 registros atendeu a esses critérios após o pré-processamanto dos dados.
- Perspectiva 2: a evasão é definida da mesma maneira que na perspectiva anterior; no entanto, a persistência é definida como um aluno que terminou seus estudos. O objetivo aqui é eliminar o ruído dos alunos ativos ao treinar o algoritmo, uma vez que não se sabe de antemão se eles se formarão ou abandonarão. Um total de 94.692 registros atendeu a esses critérios após o pré-processamanto dos dados.

Os atributos utilizados neste estudo foram classificadas em três grupos: dados do aluno, dados do curso e dados da instituição. O grupo de dados dos alunos é composto pelos seguintes atributos: cor/raça, sexo, idade, situação do vínculo, financiamento estudantil, apoio social, atividade extracurricular, tipo de escola onde o aluno concluiu o ensino médio e tempo de permanência no curso. O grupo de dados do curso é composto por turno, nível e grau acadêmico e modalidade de ensino. Por fim, o grupo de dados da IES é composto por categoria administrativa e organização acadêmica.

2.1. Estratégia Empírica

Os experimentos conduzidos neste trabalho consistem da aplicação de algoritmos de Aprendizagem de Máquina do paradigma supervisionado: Florestas Aleatórias (FA), *AdaBoost* (Ada), Vizinho mais próximo de K (KNN), Regressão Logística (RL) e Classificador de Votação (CV) [Mduma et al. 2019]. O objetivo é verificar se classificadores gerados por esses algoritmos são capazes de distinguir entre alunos que concluem seus cursos e alunos propensos a evadir, considerando os atributos descritos no parágrafo anterior.

Para determinar a capacidade preditiva dos atributos e escolher o melhor método de previsão, foram realizadas as seguintes etapas:

1. Divisão: para fins preditivos, os dados foram divididos em duas partes: uma para treinar o modelo (75% dos dados) e a outra para testar sua capacidade preditiva (25% dos dados).
2. Balanceamento: dada a grande diferença de proporção entre as classes avaliadas, após o recorte das partições foi realizado o balanceamento do conjunto de treino empregando um método de *Oversampling*. O método escolhido realiza a geração sintética de instâncias da classe minoritária, *Synthetic Minority Oversampling (SMOTE)*.
3. Treinamento: os parâmetros dos modelos foram estimados nesta etapa. Para determinar quais parâmetros contribuem para uma melhor predição, foi utilizada uma abordagem de validação de folha cruzada com 5 folhas (*5-fold cross validation*). Para estabelecer um mecanismo de comparação entre os modelos, foi utilizado um classificador Base Aleatória *Dummy* (*Baseline*) que decide aleatoriamente o rótulo de cada exemplo, de acordo com as probabilidades de classe (verdadeiro/falso) observadas no conjunto de treinamento. Na Floresta Aleatória foram produzidas cem estimativas em cada folha da validação cruzada, observado o critério de impureza de *Gini*.
4. Validação: para a validação dos classificadores, foram calculadas a acurária, a precisão e a cobertura (*recall*) para cada um dos modelos.

3. Resultados e Discussão

Observa-se que o fenômeno de evasão nos dois recortes experimentais é significativo. No primeiro conjunto, considerando alunos formados/cursando e evadidos, observou-se um total de 34,77% alunos que poderiam ser classificados como evadidos enquanto que no segundo conjunto, considerando apenas alunos formados e evadidos, foi observado um aumento para 81,90% de casos que poderiam ser classificados como evasão.

As Tabelas 1 e 2 apresentam as métricas de desempenho para a validação cruzada do *Baseline*, dos quatro algoritmos estudados RL, Ada, KNN, FA, e do CV. É importante notar que todos os classificadores estudados obtiveram desempenho superior ao *Baseline*.

A acurácia mediu o quanto um determinado classificador acertou, sejam alunos evadidos ou persistentes. Em termos de acurácia, o modelo FA obteve a maior acurácia (70,37%) na Perspectiva 1 e (87,13%) na Perspectiva 2.

A precisão mediu a capacidade de um dos algoritmos de classificar corretamente os alunos que evadiram. Considerando a precisão, o modelo CV obteve melhor desempenho com 70,97% na Perspectiva 1 e foi superior na Perspectiva 2 com 92,45%.

A cobertura mede a capacidade de um modelo de classificar corretamente os alunos evadidos e foi a métrica escolhida para avaliar os modelos neste trabalho, pois entende-se que classificar um determinado aluno erroneamente como retenção é mais prejudicial que classificá-lo erroneamente como evasão. A Tabela 3 apresenta a seleção dos atributos mais importantes utilizando o modelo Floresta Aleatória que obteve o maior grau de cobertura nas duas perspectivas estudadas: 70,14% na Perspectiva 1 e 85,61% na Perspectiva 2.

Tabela 1. Performance das métricas de validação para a Perspectiva 1

Métrica	*Baseline*	RL	Ada	KNN	FA	CV
Acurácia	50,52%	66,19%	68,80%	66,01%	70,37%	69,36%
Cobertura	50,86%	66,34%	70,04%	64,68%	70,14%	65,32%
Precisão	50,51%	64,41%	68,04%	66,41%	70,16%	70,97%

Tabela 2. Performance das métricas de validação para a Perspectiva 2

Métrica	*Baseline*	RL	Ada	KNN	FA	CV
Acurácia	50,44%	72,20%	81,06%	81,99%	87,13%	84,36%
Cobertura	50,25%	72,61%	75,31%	79,39%	85,61%	75,34%
Precisão	50,44%	72,02%	85,67%	83,82%	89,98%	92,45%

Observando-se a Tabela 3, é possível notar que ambos os modelos indicam que idade, tempo de permanência no curso, atividade extracurricular e financiamento estudantil são os atributos mais importantes para determinar a possibilidade de evasão enquanto que turno e apoio social são os menos importantes. Analisando-se os dados dos alunos evadidos, verificou-se que 60,68% dos alunos evadidos possuíam entre 18 e 25 anos, 48,26% evadiram o curso entre o primeiro e o segundo ano, 77,71% não possuíam financiamento estudantil e 89,91% não realizaram nenhum tipo de atividade extracurricular. Realizando a mesma análise entre os alunos formados, essas percentagens mudaram para:

55,93% possuíam entre 18 e 25 anos, 72,87% se diplomaram entre o terceiro e quinto ano do curso, 73,12% não possíam financiamento estudantil e 75,37% não realizaram nenhum tipo de atividade extracurricular.

Tabela 3. Atributos de validação - por ordem de importância

	Perspectiva 1	Importância	Perspectiva 2	Importância
1	idade	27,27%	tempo de permanência	39,38%
2	financimamento estudantil	17,53%	idade	20,73%
3	tempo de permanência	16,72%	financimamento estudantil	6,23%
4	atividade extracurricular	8,24%	atividade extracurricular	5,34%
5	cor/raça	7,83%	organização acadêmica	5,07%
6	organização acadêmica	5,65%	cor/raça	4,77%
7	categoria administrativa	4,11%	categoria administrativa	4,36%
8	sexo	3,48%	grau acadêmico	3,96%
9	escola ensino médio	3,40%	sexo	3,60%
10	grau acadêmico	2,05%	escola ensino médio	3,51%
11	turno	2,03%	turno	1,72%
12	apoio social	1,77%	apoio social	1,09%

4. Considerações Finais e Trabalhos Futuros

Este trabalho buscou identificar características comuns a alunos que evadem dos cursos de Sistemas de Informação do ensino superior brasileiro e verificar a eficácia de diferentes algoritmos de classificação para a previsão da evasão. Em relação às características identificadas a partir dos modelos treinados, após breve análise dos dados da amostra, verificou-se que os alunos que possuem maior propensão a evadir são os que: estão no início do curso e não exercem qualquer atividade extracurricular, tal como estágio não obrigatório, iniciação científica, atividade de extensão ou monitoria. Trabalhos futuros incluem a investigação de métodos para estimar o risco de evasão ao longo do tempo, a utilização de dados demográficos dos estudantes, analisar uma quantidade maior de dados e como a distinção entre alunos de IES públicas e particulares pode impactar os atributos de validação.

Referências

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4):498–506.

Mduma, N., Kalegele, K., and Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal*, 8(14):1–10.

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., and Hernandez, M. (2018). Perspectives to predict dropout in university students with machine learning. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–6. IEEE.

Tinto, V. and Cullen, J. (1973). Dropout in higher education: A review and theoretical synthesis of recent research.