Trabalho apresentado no XLI CNMAC, Unicamp - Campinas - SP, 2022.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics
Preprint

Árvore de decisão aplicada ao estudo da evasão de alunos em um curso de graduação

Dirceu Scaldelai[1], Felipe de Oliveira Teixeira[2], Gabriel F. Pereira[3], Gislaine A. Periçaro[4], Solange R. dos Santos[5]
UNESPAR/ Campus de Campo Mourão, PR

Com o crescente volume de dados o entendimento de métodos que envolvam matemática e estatística tornam essenciais para tomar decisões, tirar conclusões ou, ainda, fazer previsões. Nesse cenário as ferramentas da *Statistical Learning* podem ser empregadas nas tomadas de decisões e previsões em diversas situações do mundo moderno. Tais ferramentas são divididas em duas classes: aprendizado supervisionado e não supervisionado [1, 4].

Esse trabalho tem por objetivo apresentar os resultados da aplicação de uma das técnicas de aprendizado supervisionado denominada de árvore de decisão, na identificação de fatores que levaram acadêmicos a evadirem do curso de Matemática da UNESPAR - Campus Campo Mourão. Para tanto, consideramos os dados sociais, econômicos e educacionais extraídos do formulário de matrícula online, respondido pelos acadêmicos ao ingressarem no curso de graduação.

Conforme apresentado em [1, 3] uma árvore de decisão consiste de uma hierarquia de nós conectados por ramos, em que o primeiro nó é denominado nó raiz, os nós finais são as folhas e os nós intermediários são chamados de nós internos ou sub-nós de decisão. Cada nó interno se divide em ramificações, as quais são determinadas de acordo com uma medida apropriada, por exemplo, a entropia e o ganho de informação. As árvores de decisão possuem algumas características interessantes: (i) são bastante fáceis de serem explicadas; (ii) possibilitam a ilustração gráfica e são facilmente interpretadas; (iii) lidam facilmente com preditores qualitativos sem precisar criar variáveis artificiais [2].

Os dados utilizados, disponibilizados pela Pró-Reitoria de Ensino de Graduação – PROGRAD da UNESPAR, consistem em duas bases, a primeira, denominada "perfil do ingressante", referente aos formulários preenchidos por todos os acadêmicos ingressantes no curso de Matemática, enquanto a segunda é composta pela relação de "acadêmicos desistentes" no período de 2018 a 2020. Para construção da árvore de decisão foram consideradas 28 informações do questionário sócio, econômico e educacional (variáveis) de 111 acadêmicos ingressantes.

De acordo com a PROGRAD, um aluno é considerado evadido quando estava matriculado no ano anterior e não efetuou a renovação da matrícula para nenhuma disciplina para o ano seguinte. A partir dessa informação, realizamos um cruzamento das duas bases de dados, identificando 49 acadêmicos evadidos e 62 não evadidos, representados por "S" e "N", respectivamente. Por fim, construímos a árvore de decisão para o problema de classificação (Figura 1) utilizando o algoritmo CART [1] disponível no pacote *rpart* do *software* R .

[1]dirceu.scaldelai@ies.unespar.edu.br
[2]feholi99@gmail.com
[3]gabriel-favaro_pereira@hotmail.com
[4]gpericaro@gmail.com
[5]solange.regina@ies.unespar.edu.br

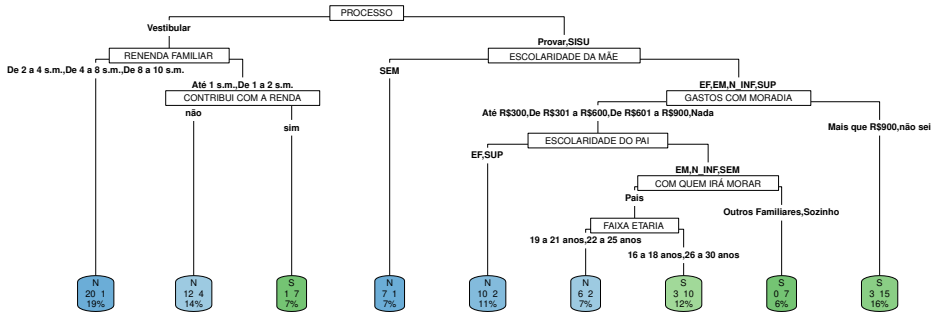

Figura 1: Árvore de decisão para o problema de evasão gerada pelo *software* R.

Com base na Figura 1, notamos que a árvore é composta por um nó raiz, sete nós intermediários e nove folhas. A variável que repartiu o nó raiz foi o processo de entrada, de maneira que a classe "Vestibular" ficou com a ramificação à esquerda enquanto a classe "SISU ou Provar", à direita. Na sequência, a variável que repartiu o nó "Vestibular" foi a renda familiar em salários mínimos (s.m.), de maneira que as classes "De 2 a 4 s.m.", "De 4 a 8 s.m." e "De 8 a 10 s.m." ficassem à esquerda e "Até 1 s.m." e "De 1 a 2 s.m." à direita. Isso significa que a árvore classificou os acadêmicos de acordo com a quantidade de salários mínimos que sua respectiva família recebe. Notamos também, que não existe mais nenhuma ramificação depois do nó à esquerda, ou seja, um acadêmico que entrou no curso pelo "Vestibular" e tem renda familiar acima de dois salários mínimos é muito mais propenso a não evadir do curso. Por outro lado, quando analisamos a ramificação à direita, ainda existe mais uma variável de repartição, a contribuição com a renda, que se ramifica em duas folhas, classificando os acadêmicos que contribuem com a renda como "evadidos" e os que não contribuem como "não evadidos".

Analogamente, também percebemos que as cinco variáveis de repartição à direita da classe "SISU ou Provar", são todas distintas e são distribuídas entre escolaridade da mãe, gasto com moradia, escolaridade do pai, tipo de moradia e faixa etária. No final, a árvore produziu nove folhas, das quais cinco classificaram os acadêmicos como "não evadidos" (N) e as demais como "evadidos" (S). Em cada folha, podemos verificar o número de acadêmicos atribuídos a cada classe, N à esquerda e S à direita, e a porcentagem de acadêmicos alocados na referida folha.

A partir da árvore de decisão obtida, identificamos fatores de relevância na evasão dos acadêmicos, os quais estão associados diretamente ao processo de ingresso na Universidade (Vestibular, SISU, PROVAR), renda familiar, despesas com moradia, com quem irá morar e escolaridade dos pais. Uma vez descoberto os principais fatores, a pesquisa pode ser dirigida para futuros estudos, no intuito de proporcionar maior aprofundamento sobre o tema, fornecendo subsídios para gestores no aprimoramento de políticas educacionais voltadas à permanência estudantil.

Referências

[1] Trevor Hastie et al. **The elements of statistical learning: data mining, inference, and prediction**. Vol. 2. Springer, 2009.

[2] Gareth James et al. **An introduction to statistical learning**. Vol. 112. Springer, 2013.

[3] Oded Z Maimon e Lior Rokach. **Data mining with decision trees: theory and applications**. Vol. 81. World scientific, 2014.

[4] Kevin P Murphy. **Machine learning: a probabilistic perspective**. MIT press, 2012.