Predição de Alunos com Risco de Evasão: estudo de caso usando mineração de dados

Mariele de Almeida Lanes[1], **Cleber de Souza Alcântara**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

Abstract. *One of the challenges for educational institutions is to reduce high dropout rates in their undergraduate degrees. A widely used solution to achieve this goal is the use of educational data mining in order to identify patterns that assist managers in decision making. In this work, we present a study that aims to identify students who present risk of evasion, starting from their first year in the undergraduate course. The experiments were carried out with information extracted from the FURG academic system. The obtained results show that the potential evader students can be identified with accuracy of 90.7% using the J48 algorithm.*

Resumo. *Um dos desafios das instituições de ensino é reduzir os altos índices de evasão em seus cursos superiores. Uma solução bastante utilizada para atingir esse objetivo é o uso de mineração de dados educacionais, a fim de identificar padrões que auxiliem os gestores na tomada de decisão. Neste trabalho, apresentamos um estudo que visa identificar estudantes que apresentam risco de evasão a partir do seu primeiro ano no curso de graduação. Os experimentos foram realizados com informações extraídas do sistema acadêmico da FURG. Os resultados obtidos mostram que os potenciais alunos evasores podem ser identificados com acurácia de 90,7% usando o algoritmo J48.*

1. Introdução

As instituições brasileiras têm apresentado altos índices de evasão em seus cursos superiores. Esse fato tem causado preocupação aos gestores acadêmicos, que buscam melhorar os indicadores de retenção e conclusão de curso. Nesse contexto, surge a importância de identificar antecipadamente quais estudantes não terão êxito na conclusão do curso, pois essa informação pode auxiliar na tomada de decisões para que essa previsão possa ser modificada [Oliveira et al. 2017].

Para isso, podem ser utilizadas técnicas de Mineração de Dados Educacionais (*Educational Data Mining* - EDM). Essas técnicas são capazes de identificar o perfil desses estudantes a partir de dados de outros estudantes que não foram bem sucedidos na universidade [Baker et al. 2011].

A Universidade Federal do Rio Grande (FURG) é uma das instituições brasileiras que tem enfrentado o desafio de reduzir o elevado número de evasão, pois muitos alunos abandonam seus cursos durante a trajetória acadêmica. Por isso, existe a necessidade do uso de técnicas de EDM para auxiliar na detecção antecipada da evasão escolar, permitindo que suas possíveis causas possam ser analisadas e tratadas prematuramente.

DOI: 10.5753/cbie.sbie.2018.1921

Nesse sentido, este trabalho tem como objetivo identificar o subconjunto dos alunos de graduação da FURG que apresentam risco de evasão. Para isso, foram analisadas diferentes informações dos alunos que evadiram ou concluíram seus cursos de graduação no período de 2012 a 2017, visando gerar um modelo capaz de classificar corretamente novas instâncias de alunos como sendo ou não um possível evasor. Espera-se que o referido modelo auxilie os gestores da FURG no desenvolvimento de ações para reduzir o abandono de curso.

A previsão da evasão escolar é um tema já estudado por diversos pesquisadores [Manhães et al. 2011, Brito et al. 2015, Digiampietri et al. 2016]. No entanto, este trabalho apresenta uma abordagem diferente destes, pois busca compreender a evasão analisando dados de alunos que já cursaram no mínimo um ano do curso de graduação, a partir de dados demográficos e de desempenho acadêmico.

O restante do trabalho está organizado da seguinte forma: na Seção 2 é apresentada a metodologia empregada no estudo; na Seção 3 estão descritos os resultados obtidos; e a Seção 4 apresenta as conclusões e os trabalhos futuros.

2. Metodologia

A base de dados utilizada neste trabalho foi coletada do sistema acadêmico da FURG, contendo informações sobre alunos de 12 cursos de graduação de diferentes áreas do conhecimento. É importante ressaltar que os alunos tiveram a sua identificação preservada, sendo a matrícula o atributo usado para relacionar os dados.

Os dados foram extraídos em diferentes arquivos no formato XLS, e importados para um banco de dados *MySQL*. Para compor o *dataset* final somente foram considerados os alunos que ingressaram pelo ENEM/SISU, que evadiram ou concluíram o curso, e aqueles que já tinham cursado no mínimo um ano do curso. Esse último critério foi utilizado devido ao atributo índice de rendimento acadêmico, que na maioria dos casos se obtém um valor significativo após o primeiro ano no curso de graduação. Como resultado dessa filtragem, obteve-se 916 registros, onde 720 pertencem a classe evadido e 196 a classe concluinte.

Além disso, os atributos numéricos foram discretizados e as faixas de valores foram usadas para criar as seguintes categorias:

- Faixa de idade do aluno no momento da evasão ou conclusão do curso: ID1 (16-20); ID2 (21-25); ID3 (26-30); ID4 (mais de 30 anos).
- Média geral do ENEM: ME1 (400-499); ME2 (500-599); ME3 (600-699); ME4 (maior ou igual a 700).
- Intervalo de tempo entre a conclusão do ensino médio e o ingresso no curso de graduação: II1 (menor que 2 anos); II2 (2-4 anos); II3 (5-9 anos); II4 (maior ou igual a 10 anos).
- Último coeficiente de rendimento: CO1 (menor que 5,0); CO2 (5,0-6,9); CO3 (7,0-8,9); CO4 (maior ou igual a 9,0).

Já para o atributo representante do curso foram criadas 4 categorias de acordo com a área de conhecimento do mesmo, sendo elas: ciências da saúde - AC1 (Educação Física, Enfermagem), ciências exatas - AC2 (Engenharia Civil, Engenharia de Computação, Engenharia Química), ciências humanas - AC3 (História Licenciatura, Pedagogia, Psicologia, Geografia) e ciências sociais aplicadas - AC4 (Administração, Ciências Contábeis,

Direito). Os demais atributos referem-se ao gênero (M ou F), tipo da escola de origem (Pública ou Privada), se o aluno foi ou não bolsista em algum momento da graduação, estado de origem (Rio Grande do Sul ou outro estado). Por fim, a situação final do aluno no curso (evadido ou concluinte) foi usada como classe alvo.

Para o processo de mineração de dados utilizou-se a ferramenta de código aberto *Weka* [Hall et al. 2009]. Após a etapa de pré-processamento, a tarefa de classificação foi realizada utilizando o algoritmo J48 [Quinlan 1993] para processar o *dataset* e gerar uma árvore de decisão. A tarefa de classificação utiliza algoritmos de aprendizagem de máquina para rotular, automaticamente, novas instâncias de uma base de dados com uma determinada classe, aplicando um modelo previamente aprendido baseado no valor dos atributos das instâncias de treinamento [Tan et al. 2005]. Neste trabalho, foi utilizado o algoritmo J48 por apresentar resultados intuitivos e, consequentemente, permitir melhor entendimento por parte dos gestores em relação aos valores dos atributos e a situação final de cada aluno no curso (concluinte ou evadido).

O referido algoritmo foi configurado para usar como critério de poda somente os nós que possuíam a partir de 10 instâncias (parâmetro *minNumObj*), visando evitar que a árvore ficasse muito grande. Além disso, esse parâmetro proporcionou uma melhor acurácia na classificação. Para os demais parâmetros, foi mantida a configuração padrão do *Weka*. O modelo de classificação gerado foi testado utilizando 10 partições na validação cruzada.

3. Resultados Obtidos

A Figura 1 mostra as regras geradas pelo algoritmo J48. Nela, é possível observar que o valor do atributo coeficiente de rendimento é significativo para gerar o modelo de classificação, pois os alunos com coeficiente menor que 7,0 (categorias CO1 e CO2) são classificados na situação "Evadido", representando mais de 60% da amostra, tendo taxa de acerto de 100%. Ou seja, alunos com coeficiente de rendimento baixo tendem a abandonar o curso, independente de outras características.

Já os alunos com coeficiente maior ou igual a 9,0 e que foram bolsistas durante a graduação, independente do período de duração da bolsa, são classificados como "Concluinte", e os não bolsistas como "Evadido". Isso demonstra que, mesmo o aluno possuindo um bom coeficiente de rendimento, o envolvimento em atividades extraclasse é muito importante para a permanência do aluno na universidade.

De acordo com o modelo gerado, para os alunos com coeficiente de rendimento maior ou igual a 7,0 e menor que 9,0 (categoria CO3), independente de ter sido bolsista ou não, a sua faixa de idade influência diretamente na situação final no curso, pois os alunos bolsistas com faixa de idade acima de 20 anos (categorias ID2, ID3 e ID4) são classificados como "Concluinte", mas com idade entre 15 e 20 anos (categoria ID1) tendem a evadir do curso. Em corroboração, os alunos que se enquadram na categoria CO3, não bolsistas, e com faixa de idade acima de 25 anos (categorias ID3 e ID4), são classificados como "Concluinte". Já os alunos da categoria ID1 são classificados como "Evadido" com uma acurácia de 100%. Com isso, percebe-se que os alunos mais jovens, com coeficiente intermediário, estão mais propensos a evasão.

O atributo área do curso é analisado para aqueles alunos com coeficiente de rendimento da categoria CO3, não bolsistas e com idade maior que 20 e menor ou igual a

25 anos (categoria ID2). Nesse caso, são classificados como “Evadido” os alunos da área das ciências da saúde (AC1) ou das ciências humanas (AC3), e como “Concluinte” sendo da área das ciências exatas (AC2) ou ciências sociais aplicadas (AC4). Essa informação demonstra que a área de conhecimento do curso não está diretamente ligada a evasão, pois primeiramente são analisadas outras características dos alunos.

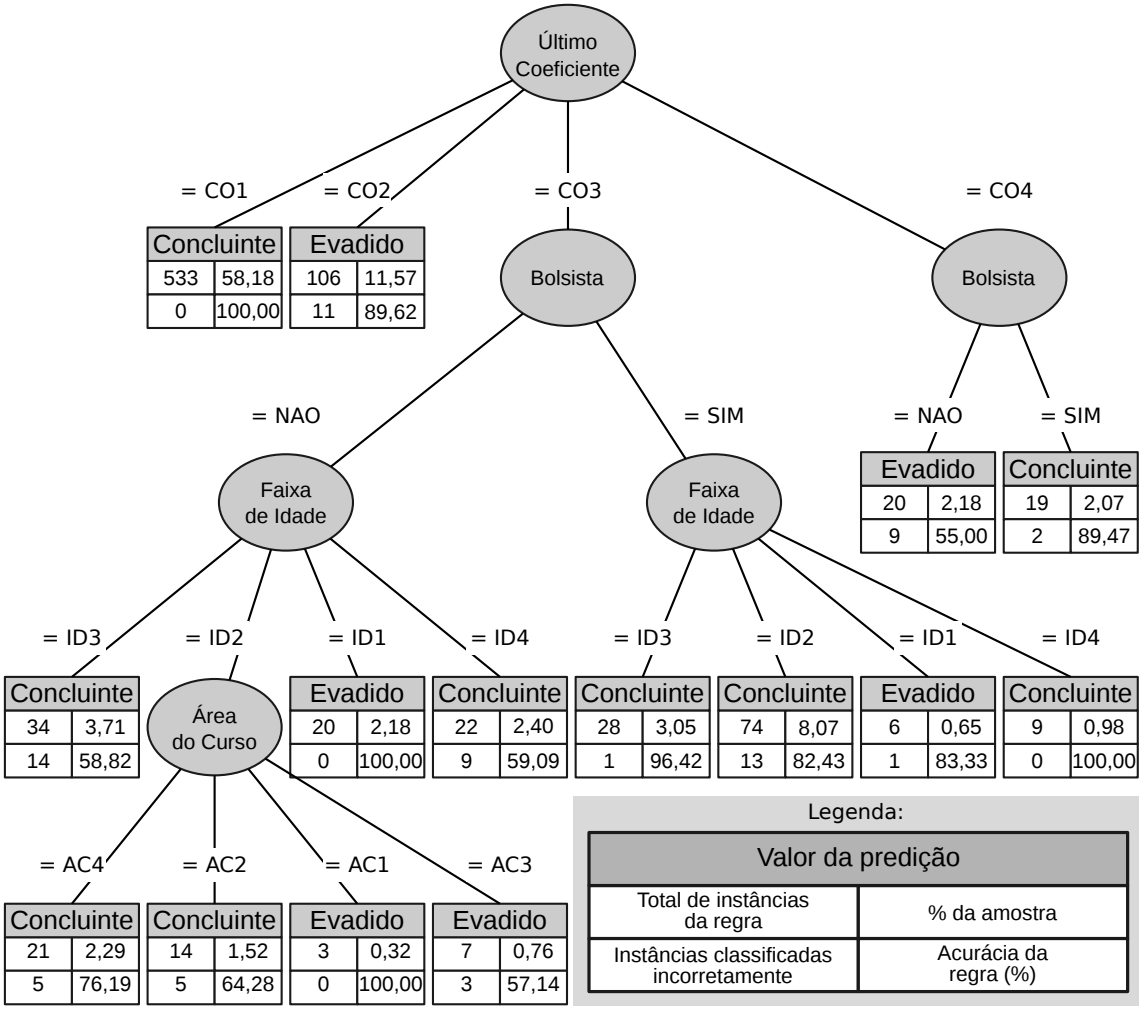

Figura 1. Árvore de decisão gerada pelo algoritmo J48.

Para a amostra de dados e configuração usada nesse experimento, percebe-se que os valores da média do ENEM, o estado de origem do aluno, o tipo de escola de origem, o gênero, e o intervalo para ingresso na graduação não foram considerados relevantes para gerar o modelo de classificação. Ou seja, não são usados diretamente como indicadores para detectar um aluno como sendo ou não um possível evasor. No entanto, observa-se que a faixa de idade apresentou grande influência para essa predição, o que demonstra que a utilização de dados demográficos pode contribuir para a obtenção de modelos cada vez mais precisos.

De acordo com os resultados experimentais, o algoritmo J48 apresentou acurácia de 90,7%, mostrando ser viável a identificação dos alunos que apresentam risco de evasão através do uso de técnicas de mineração de dados.

4. Conclusão e Trabalhos Futuros

O elevado índice de evasão escolar é um sério desafio enfrentado pelas instituições de ensino superior. A identificação precoce de alunos com este perfil permite o planejamento de ações com propósito de evitar o abandono do curso. Neste trabalho, foi utilizado o algoritmo J48 na identificação de alunos com risco de evasão com taxas de acerto de 90,7%. Desse modo, espera-se que esse modelo de classificação proporcione aos gestores educacionais da FURG indicadores que possam ser efetivamente utilizados para reduzir os índices de evasão.

Como trabalhos futuros pretende-se tratar a distribuição desbalanceada entre as classes, e realizar experimentos usando os dados dos demais cursos de graduação da FURG. Dessa forma, poderão ser realizados estudos mais detalhados por curso, bem como estudos comparativos. Além disso, poderá ser verificado a eficiência na predição da evasão com base em outras informações acadêmicas dos alunos, como a frequência, o registro de empréstimos na biblioteca, e o desempenho nas disciplinas.

Referências

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o brasil. *Brazilian Journal of Computers in Education*, 19(02):11.

Brito, D. M., Lemos, M. O., Pascoal, T. A., do Rêgo, T. G., and Araújo, J. G. G. d. O. (2015). Identificação de estudantes do primeiro semestre com risco de evasão através de técnicas de data mining. *Nuevas Ideas en Informática Educativa TISE*, pages 459–463.

Digiampietri, L. A., Nakano, F., and de Souza Lauretto, M. (2016). Mineraçao de dados para identificaçao de alunos com alto risco de evasao: Um estudo de caso. *Revista de Graduação USP*, 1(1):17–23.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Manhães, L. M. B., Da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., and Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 1.

Oliveira, J. J. G., Noronha, R. V., and Kaestner, C. A. A. (2017). Método de seleção de atributos aplicados na previsão da evasão de cursos de graduação. *Revista de Informática Aplicada*, 13(2).

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, USA.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley.