

PREVISÃO DE EVASÃO DE ALUNOS DA FATEC DE SÃO JOSÉ DOS CAMPOS UTILIZANDO MINERAÇÃO DE DADOS

M.C.M. Cesar^{1,*}; E. M. Carneiro¹; E. Sakaue¹

¹ Faculdade de Tecnologia de São José dos Campos - Professor Jessen Vidal
Av. Cesare Mansueto Giulio Lattes, 1350 - Eugênio de Melo, São José dos Campos/SP,
CEP.: 12247-014, Brasil.
Telefone: (12) 3905-2423

*mari_carvalhomello@hotmail.com

RESUMO: A evasão no ensino superior é um tema de grande importância, sendo um problema que assola as universidades no Brasil, trazendo desperdício de recursos de ordem econômica e social investidos em alunos que evadem precocemente nos primeiros semestres dos cursos nos quais estão matriculados. Este trabalho avaliou se os dados atualmente disponíveis na Fatec Prof. Jessen Vidal são suficientes para predição de evasão de alunos. Para tanto, foram construídos classificadores e realizados testes utilizando dados recolhidos em semestres anteriores. Como sugerem os resultados, os atributos disponíveis não são suficientes para essa previsão, sendo necessária a utilização de informações mais abrangentes.

PALAVRAS-CHAVE: evasão; mineração; artificial.

ABSTRACT: Higher education drop-outs is a subject of major importance and a problem that affects a great number of Brazilian universities, leading to waste of economic and social resources invested in students that drop out prematurely in early semesters of the course they enrolled. This paper evaluated if available data from Fatec Prof. Jessen Vidal is enough for student drop out prediction. To this end, classifiers were built and tests using previous semesters data were performed. As the results revealed, current attributes aren't enough to perform prediction, so a new set of information is needed.

KEYWORDS: drop out; mining; artificial.

1. INTRODUÇÃO

Fenômeno comum em estabelecimentos de ensino, a evasão pode ocorrer como um evento onde um aluno decide desligar-se da universidade ou mesmo um evento de exclusão, por parte da deficiência e carência de orientadores para nortear o estudante em busca de sua profissão [3]. Para que seja possível uma análise profunda do fenômeno, e, por conseguinte, utilizar isso para traçar uma solução, a complexidade do problema necessita de atenção ao ilustrar diversos fatores como a frustração do aluno com o ensino superior, a insatisfação com o conteúdo curricular dos

cursos, desencanto das expectativas do mercado de trabalho e decepção com a escolha realizada para profissão [2]. Além disso, são encontrados outros fatores de grande importância, como o desperdício de dinheiro público investido no aluno evadido, ao se tratar de uma universidade comunitária, e também o baixo nível de formação acadêmica da população brasileira [8].

Muitas variáveis contribuem para que o aluno desligue-se da universidade e frustrar seus sonhos de concluir um curso de graduação que lhe traria um emprego de qualidade e uma fonte de renda importante para transformar sua vida e garantir sua sobrevivência, mesmo em crises ou

problemas econômicos que o país enfrentar [6]. Por isso, faz-se essencial a busca de uma solução para prever e identificar o aluno que pode evadir, com o objetivo de diminuir a ocorrência do problema.

A FATEC de São José dos Campos apresenta um índice preocupante de desistência de alunos, principalmente no primeiro e no segundo semestre. Em algumas ocasiões a porcentagem de desistentes chega a ser maior que 50%, em específico para os eixos de informática e aeronáutica. Existe a necessidade de um meio para apontar os estudantes com risco de evasão, para fornecer informações e parâmetros mais precisos sobre os alunos para os projetos de permanência existentes, visando um aumento de graduandos.

O objetivo deste trabalho consiste em utilizar técnicas de aprendizagem de máquina em uma base de dados da FATEC - SJC para prever se as informações atuais são suficientes para gerar um classificador capaz de prever a evasão de alunos recém matriculados.

Este artigo está organizado da seguinte forma: Na Seção 2 são introduzidos os trabalhos relacionados; na Seção 3 apresenta-se a metodologia utilizada; na Seção 4 se discutem os resultados e a conclusão e os futuros trabalhos estão na Seção 5.

2. MATERIAIS E MÉTODOS

Na gama de estudos existentes sobre o tema, vários autores buscam combater a evasão. Brito et al. [1] propõe prever o desempenho dos estudantes do curso de Ciência da Computação da UFPB e realiza testes de correlação e de predição entre as notas do vestibular e as disciplinas de exatas com índice mais alto de reprovação no primeiro semestre do curso, utilizando a performance de vários algoritmos para encontrar o mais viável e performático entre eles.

Similarmente, Kantorski et al. [4] busca prever a evasão dos alunos com

informações socioeconômicas, pessoais e sociais, criando modelos com aprendizado de máquina e combinando-os para prever a evasão no curso de Administração. O resultado na previsão da situação do curso do aluno (em curso ou evadido) chega a 95% e a previsão de alunos que não solicitaram rematrícula é de 73%.

Manhães et al. [5] obteve uma taxa média de 75% a 80% de sucesso na predição de alunos com risco de evasão utilizando as notas das matérias mais cursadas no início do curso de Engenharia da Escola Politécnica da UFRJ, utilizando mineração de dados. Outras informações relevantes foram utilizadas, como a situação final do aluno nas disciplinas e o valor do coeficiente de rendimento acumulado no período.

Thammasiri et al. [9] prevê estudantes com risco de evasão com alta taxa de precisão para base de dados onde as quantidade de exemplos regulares não são equilibradas de acordo com os irregulares, com a utilização de técnicas de balanceamento de dados em conjunto com algoritmos de classificação.

Os trabalhos apresentados sugerem que a utilização de Aprendizado de Máquina é adequada ao problema de evasão no ensino superior e pode ser aplicada para previsão de evasão de alunos. Diferente da abordagem de outros artigos, o trabalho utiliza informações socioeconômicas e pessoais, não dependendo de atributos como nota das matérias mais cursadas, coeficiente de rendimento, nota das disciplinas com maior índice de reprovação e nota geral do curso. Os alunos da FATEC SJC costumam evadir nos primeiros semestres, muitas vezes antes de realizar as primeiras provas do curso, por isso essas informações acabam não sendo viáveis para o projeto.

Na metodologia da investigação conduzida, foram utilizados os passos do processo de Knowledge Discovery in Databases (KDD).

Como entrada para o processo foram utilizados dados divididos por período de

ingresso e por curso, com o objetivo de comparar a variação de desempenho de algoritmos de inteligência artificial em diferentes semestres dos anos de 2013 (1º e 2º semestre), 2014 (1º e 2º semestre) e 2015 (1º semestre). A criação de um modelo utiliza um determinado período de ingresso de um curso, enquanto o teste é realizado utilizando um período posterior do mesmo curso. Inicialmente, a esses dados foi aplicado o processo de seleção de atributos do KDD, que limitou os atributos utilizados em: idade do aluno no momento da matrícula, cor de pele, situação do curso, CEP, estudante de escola pública, sexo, escolaridade dos pais e renda familiar.

Continuando com o processo de KDD, ocorreu a limpeza de dados que retirou registros duplicados ou com valores de atributos vazios.

Na fase seguinte, de transformação, colunas numéricas foram discretizadas e colunas contendo valores discretos foram agrupadas, com a finalidade de melhorar a performance dos algoritmos de classificação.

Para a fase de mineração de dados utilizou-se a ferramenta WEKA (*Waikato Environment for Knowledge Analysis*) [11], desenvolvida na Universidade de Waikato na Nova Zelândia. Como classificadores, foram escolhidos dois algoritmos: *J48 Pruned* e *NaiveBayes*, devido à facilidade de interpretação dos modelos gerados, requisito de extrema importância para a Fatec, devido à necessidade de entender os fatores específicos que levam à evasão de alunos.

Foram escolhidos, como métricas de qualidade para avaliação dos resultados obtidos, os indicadores de sensibilidade e precisão. Um modelo é considerado adequado quando ambos indicadores possuem valores superiores a 50%.

3. RESULTADOS E DISCUSSÃO

Os testes realizados envolveram diferentes combinações de períodos e cursos.

Somente um curso apresentou um resultado adequado, apenas para uma única combinação de períodos (modelo treinado com o 1º semestre de 2014 e testado com o 1º semestre de 2015), apresentando sensibilidade de 64,70% e precisão de 78,57%, apresentado na Tabela 1. Suas respectivas matrizes de confusão podem ser observadas na Tabela 2 e na Tabela 3. Esse resultado não se reproduz para outras combinações de períodos, indicando que o modelo gerado não é genérico o suficiente para utilização.

Tabela 1. Taxa de precisão e de sensibilidade para modelo treinado com o 1º semestre de 2014 e testado com o 1º semestre de 2015

Algoritmo	Taxa de Precisão	Taxa de Sensibilidade
NaiveBayes	64,70%	78,57%
J48 – Pruned	58,33%	50%

Tabela 2. Matriz de confusão para o algoritmo NaiveBayes no modelo treinado com o 1º semestre de 2014 e testado com o 1º semestre de 2015

Naive Bayes		
a	b	<< Classificado como
11	3	a = cancelado
6	12	b = em curso

Tabela 3. Matriz de confusão para o algoritmo J48 pruned no modelo treinado com o 1º semestre de 2014 e testado com o 1º semestre de 2015

J48 – Pruned		
a	b	<< Classificado como
7	7	a = cancelado
5	13	b = em curso

4. CONCLUSÃO

Este trabalho buscou utilizar a técnica de KDD, que envolve mineração de dados e inteligência artificial, para conseguir prever se a base de dados atual da FATEC – Prof. Jessen Vidal, que continha especificamente informações socioeconômicas e pessoais, era suficiente para prever evasão de alunos.

Na comparação dos valores, se observa um desempenho superior do algoritmo *NaiveBayes* em relação ao J48. De modo geral, o conjunto de teste com maior equilíbrio entre taxas de precisão e de sensibilidade foi o apresentado na Tabela 1.

Diversas dificuldades foram geradas pela baixa qualidade dos dados disponíveis que, além de apresentarem uma quantidade limitada de atributos, continham registros duplicados para um mesmo aluno, informações conflitantes e muitos valores de atributos vazios, inviabilizando a utilização dessas linhas para os testes.

Com base nos experimentos realizados, pode-se concluir que se faz necessária a coleta de informações adicionais durante o período de matrícula. Como resultado deste trabalho, a partir do primeiro semestre letivo de 2017, um conjunto de informações adicionais passou a ser coletado, com o objetivo de melhorar a capacidade de previsão de evasão de alunos. Constam nas novas informações coletadas: empregabilidade dos pais, que possui um peso importante para a classificação das instâncias [7], e dados sobre empregabilidade do estudante, que refletem em seu desempenho e podem contribuir para sua evasão [10].

5. REFERÊNCIAS BIBLIOGRÁFICAS

- [1] BRITO, D. M. et al. *Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina*. Simpósio Brasileiro de Informática na Educação-SBIE. p. 882, 2014.
- [2] BRUNS, M. A. D. T. *Não era bem isto o que eu esperava da Universidade: um estudo*

de escolhas profissionais. Tese de Doutorado, Faculdade de Educação da Unicamp, 1992.

- [3] BUENO, J. L. O. *A evasão de alunos*. Paidéia. Ribeirão Preto, n. 5, p. 9-16, 1993.

[4] KANTORSKI, G. et al. *Predição da Evasão em Cursos de Graduação em Instituições Públicas*. Simpósio Brasileiro de Informática na Educação-SBIE. Vol. 27, p. 906, 2016.]

[5] MANHÃES, L. M. B. et al. *Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados*. Simpósio Brasileiro de Informática na Educação-SBIE. 2011.

[6] MEC, *Comissão Especial de Estudos. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas*. Avaliação, Campinas, v. 1, n. 2, p. 55-65, 1996.

[7] PASCOAL, T. et al. *Evasão de estudantes universitários: diagnóstico a partir de dados acadêmicos e socioeconômicos*. Simpósio Brasileiro de Informática na Educação-SBIE. Vol. 27, No. 1, 2016.

[8] PEREIRA, F. C. B. et al. *Determinantes da evasão de alunos e os custos ocultos para as instituições de ensino superior: uma aplicação na Universidade do Extremo Sul Catarinense*. 2003.

[9] THAMMASIRI, D., et al. *A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition*. Expert Systems with Applications, v. 41, n. 2, p. 321-330, 2014.

[10] VICKERS, M.; LAMB, S.; HINKLEY, J. *Student Workers in High School and Beyond: The Effects of Part-Time Employment on Participation in Education, Training and Work*. ACER Customer Service, Private Bag 55, Camberwell, Victoria 3124 Australia, 2003.

[11] WITTEN, I. H. et al. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, Fourth Edition, 2016.