ANÁLISE DA PREDIÇÃO DE EVASÃO DE ALUNOS DE ALUNOS DA UFPEL UTILIZANDO MINERAÇÃO DE DADOS EDUCACIONAIS

ALEXANDRE GOMES DA COSTA[1]; TIAGO THOMPSEN PRIMO[2]; JÚLIO C. B. MATTOS[3]

[1]Universidade Federal de Pelotas – alexandre.costa@inf.ufpel.edu.br
[2]Universidade Federal de Pelotas – tiagoprimo@gmail.com
[3]Universidade Federal de Pelotas – julius@inf.ufpel.edu.br

1. INTRODUÇÃO

O uso constante de Tecnologia da Informação e Comunicação (TICs) em diversas áreas vêm gerando um grande volume de dados. Tecnologias como redes sociais, AVAs, aplicativos embarcados, sensores e sistemas de informação em geral são alguns exemplos de recursos que vem aumentando o número de dados das mais diversas naturezas (Goldschmidt et al. 2015).

A evasão é um problema que atinge não só as Instituições de Ensino Superior (IES) privadas, mas também as públicas. Segundo dados do INEP (2018), em 2017, o índice de matrículas desvinculadas em todo o Brasil foi de 16,41%. Já para as IES públicas esse índice no mesmo período foi de 11,56%.

Ao comparar os índices da Universidade Federal de Pelotas com os do INEP (2018) estes índices não mudam muito. Em 2017 o índice de matrículas desvinculadas na UFPel foi de 11,53%, que está abaixo do índice de 11,56% apresentado pelo Inep (2018). O problema é que se olhar para a taxa de evasão geral os dados não parecem tão alarmantes, mas se analisarmos a taxa de evasão de um único cursos essa taxa se torna preocupante.

Embora com ajuda de ferramentas computacionais, analisar essa crescente quantidade de dados não é um trabalho humanamente viável. Uma técnica que vem sendo largamente utilizada para analisar grandes quantidades de dados é a Mineração de Dados (MD) que é uma das etapas do processo de KDD. Baker et al. (2010) define MDE como a área de investigação científica centrada no desenvolvimento de métodos para fazer descobertas dentro dos tipos de dados que vêm de ambientes educacionais e usando esses métodos para entender melhor as questões relacionadas aos alunos e a aprendizagem deles.

Neste trabalho é explorada a utilização de MDE visando classificar e identificar perfis de alunos com tendência a evadir utilizando apenas dados acadêmicos dos três primeiros semestres de um curso presencial. Tarefas de classificação são largamente utilizadas para fazer a predição de alunos em risco de evasão escolar como apresentado em Manhães et al. 2011, Rigo et al. 2014, Detoni et al. 2015, Queiroga et al. 2015, Hasbun et al. 2016, Kantorski et al. 2016 e Lanes and Alcantara 2018.

A principal diferença deste trabalho é a aplicação de técnicas de visualização de MDE em cima de dados reais de alunos de um curso de graduação presencial. Diferentemente dos outros trabalhos utilizamos dados apenas dos semestres iniciais

Como uma de suas motivações, este trabalho apresenta a modelagem dos dados produzidos pelos sistemas acadêmicos, para que estes se transformem em informações, e consequentemente conhecimento, sobre o perfil dos estudantes da universidade. Assim, este trabalho pretende auxiliar nas políticas de combate aos índices de evasão apresentados na Universidade Federal de Pelotas, principalmente nos cursos de exatas e engenharias. Além disso, isto é

possibilitado pois existe uma grande quantidade de dados históricos de cursos de graduação presenciais da UFPel.

Neste trabalho são apresentados os dados preliminares desta pesquisa através dos dados acadêmicos dos alunos, de um curso específico, do curso de Ciência da Computação que foi escolhido inicialmente por possuir elevados índices de evasão. Para este propósito, foi analisado e coletado dados dos 3 primeiros semestres de alunos do curso de 2000 a 2020, para responder às seguintes questões de pesquisa: (Q1) Quais são os atributos que mais influenciam no processo de evasão dos estudantes em cursos de computação?; (Q2) Quais os classificadores e técnicas podem ser utilizados nessa tarefa?

2. METODOLOGIA

Este trabalho seguiu a metodologia CRISP-DM que tem como principal objetivo fornecer uma direção para conduzir o processo de KDD (Goldschmidt et al. 2015). Uma breve descrição das seis fases da metodologia é apresentada:

- Compreensão do negócio: Esta é a fase de identificação do problema a ser resolvido. Esta fase compreende também uma descrição do *background*, dos objetivos e também dos critérios de sucesso (Goldschmidt et al. 2015).
- Compreensão dos dados: É a fase responsável por fazer a análise exploratória de dados (AED). Esta fase tem que dizer como os dados foram adquiridos, qual o seu formato, qual foi a quantidade de dados, descrever cada atributo selecionado, fazer visualizações dos dados e além disso qualquer informação pertinente aos dados.
- Preparação dos dados: Compreende as atividades de pré-processamento dos dados para a próxima fase. Normalmente se faz a seleção, limpeza, formatação dos dados, e ainda gera-se novos atributos derivados dos atributos existentes.
- Modelagem: Corresponde a fase de aplicação dos algoritmos de mineração de dados selecionados sobre os dados preparados. É a etapa de mineração de dados do processo de KDD (Goldschmidt et al. 2015). Nessa fase é criado um modelo para testar a sua qualidade e validade. É comum usar a taxa de erro como medida de qualidade do modelo em aprendizado supervisionado (Melo and Viglioni 2007).
- Avaliação: Consiste em avaliar o modelo gerado, examinando os passos seguidos e validando se realmente foram alcançados os objetivos elencados na fase de compreensão do negócio (Melo and Viglioni 2007). A partir da avaliação é possível propor revisões das fases anteriores e redefinir os próximos passos (Goldschmidt et al. 2015).
- Desenvolvimento: É a fase onde se faz o planejamento e acompanhamento a serem realizadas com o modelo gerado pelas fases anteriores (Goldschmidt et al. 2015). Esta fase não faz parte deste trabalho.

Neste trabalho foram utilizados os dados de alunos do Cobalto, que é o sistema de gestão da UFPel. O Cobalto é um sistema integrado de gestão que faz a gestão acadêmica da universidade. Este sistema possui dados históricos de alunos, importados do sistema anterior da IES, denominado GoI.

Os dados extraídos do sistema correspondem aos alunos do curso de Ciência da Computação que ingressaram entre os anos 2000 e 2020. O número total de alunos nas situações de Cursando, Evadido, Formado e Retido foi de 1516. Após todo o processo de preparação dos dados, o conjunto de alunos do estudo se limitou a 760 alunos e 18 atributos selecionados. Destes alunos, 67,5%

(513) estavam na situação de evasão do curso e um total de 32,5% (247) estavam na situação de conclusão do curso de 2000 a 2020.

3. RESULTADOS E DISCUSSÕES

Para a implementação das rotinas mencionadas foram utilizadas as ferramentas *Cloud Google Colab*, *Python*, *Scikit-learn* e *pandas*. Para fazer o treinamento e teste do conjunto de dados foi utilizado o método de Validação Cruzada com K Conjuntos Estratificada que se assemelha a Validação Cruzada com K Conjuntos, porém quando gera os subconjuntos mantém a mesma proporção do atributo classe (Goldschmidt et al. 2015).

Foi utilizado o método "*cross_val_predict*" da biblioteca *Scikit-learn* que por padrão faz estratificação dos conjuntos de dados e também o número de conjuntos foi configurado para dez.

Neste trabalho os dados foram executados pelos algoritmos *Logistic Regression*, *Decision Tree* e *RandomForest*. Os algoritmos *Decision Tree* e *RandomForest* foram escolhidos por gerarem modelos de fácil interpretação e *Logistic Regression* foi escolhido pelo seu desempenho e relevância encontrado em outros trabalhos.

O algoritmo que obteve os melhores resultados foi o *Logistic Regression*, mas a diferença dos resultados comparada com o algoritmo de *RandomForest* não é estatisticamente significativa. Nossa principal preocupação foi de melhorar a taxa de *Recall*, pois quanto mais alta essa taxa menor será a taxa de falsos negativos (FN), que representa os alunos que foram classificados como não evadidos e na verdade são alunos que evadiram.

Foi possível observar que o atributo com maior correlação com o atributo classe é a média do terceiro semestre, ela é inversamente proporcional ao atributo classe. Essa mesma ordem de importância dos atributos é percebida nos algoritmos *Random Forest* e *Logistic Regression*, já o algoritmos *Decision Tree* considera a média dos três primeiros semestres como o atributo mais importante para a classificação.

Tabela 1. Resultado da execução dos algoritmos

Algoritmo	Accuracy	Precision	Recall	F1-score	AUC
Decision Tree	82,24%	86,78%	86,97%	86,73%	79,83%
Random Forest	88,03%	91,10%	91,45%	90,82%	95,25%
Logistic Regression	88,29%	91,24%	92,17%	90,43%	95,53%

4. CONCLUSÕES

Este trabalho apresentou os resultados para a predição da evasão de alunos através de dados dos três primeiros semestres do curso de Ciência da Computação da UFPel de 2000 a 2020. Para fazer essa classificação foi utilizado o processo de KDD e algoritmos de aprendizagem de máquina.

Para responder a primeira questão de pesquisa, foi gerada uma matriz de correlação dos atributos e foi possível perceber que dois dos três algoritmos obtiveram a média do terceiro semestre como o atributo mais importante para fazer a predição do conjunto de dados.

Para a segunda questão, foi apresentada uma acurácia de 88,29% para a predição de alunos em risco de evadir utilizando os dados pessoais e acadêmicos, considerando dados de 760 alunos. O melhor resultado foi do algoritmo *Logistic Regression* com *Recall* de 92,17% e *Precision* de 91,24%.

Como trabalhos futuros pretende-se expandir a análise deste modelo de predição para outros cursos, primeiramente da área de exatas e engenharias e após demais cursos. Desta forma, será possível analisar a viabilidade ou não deste modelo em outros cursos e caso necessário adaptá-lo.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Baker, R. et al. (2010). Data mining for education. International Encyclopedia of Education, 7(3):112–118.

Detoni, D., Cechinel, C., and Araújo, R. (2015). Modelagem e predição de reprovação de acadêmicos de cursos de educação a distância a partir da contagem de interações. Revista Brasileira de Informática na Educação, 23(3).

Goldschmidt, R., Bezerra, E., e Passos, E.. Data mining: conceitos, técnicas, algoritmos, orientações e aplicações. Rio de Janeiro-RJ:Elsevier, pages 56–60.

Hasbun, T., Araya, A., and Villalon, J. (2016). Extracurricular activities as dropout prediction factors in higher education using decision trees. In Advanced Learning Technologies, IEEE 16th International Conference on, pages 242–244.

INEP (2018). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Sinopses estatísticas da educação superior - graduação.

Kantorski, G., Flores, E. G., Schmitt, J., Hoffmann, I., and Barbosa, F. (2016). Predição da evasão em cursos de graduação em instituições públicas. In Simpósio Brasileiro de Informática na Educação-SBIE, volume 27, page 906.

Lanes, M. and Alcântara, C. (2018). Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. In Simpósio Brasileiro de Informática na Educação - SBIE, volume 29, page 1921.

Manhães, L. M. B., da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., Zimbrão, G., 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. Anais do XXII SBIE - XVII WIE, pages 150–159.

Melo, G.; Viglioni, C. (2007). Metodologia Para Previsão de Demanda Ferroviária.

Queiroga, E., Cechinel, C., and Araújo, R. (2015). Um estudo do uso de contagem de interações semanais para predição precoce de evasão em educação a distância. In Anais dos Workshops do Congresso Brasileiro de Informática na Educação, volume 4, page 1074.

Rigo, S. J., Cambruzzi, W., Barbosa, J. L., and Cazella, S. C. (2014). Aplicações de mineração de dados educacionais e learning analytics com foco na evasão escolar: oportunidades e desafios.Revista Brasileira de Informática na Educação, 22(1):132–146.