Análise e predição de evasão dos alunos do ensino superior da Universidade Federal de Santa Maria *Campus* Frederico Westphalen por meio da mineração de dados educacionais

Edson Noetzold[1], Solange Pertile[2]

[1]Curso de Bacharelado em Sistemas de Informação
[2]Departamento de Tecnologia da Informação (DTecInf)

Universidade Federal de Santa Maria (UFSM) - Campus Frederico Westphalen - Linha 7 de Setembro, s/n, CEP: 98400-000, BR 386 Km 40- Frederico Westphalen - RS

edsonversusnoetzold@gmail.com, solangepertile@gmail.com

Resumo. *Os elevados índices de evasão escolar constituem uma realidade presente em diversos cursos superiores ofertados no Brasil, o que evidencia a necessidade de investigação dessa problemática responsável por perdas econômicas nas instituições e impactos no cenário global da educação. Este artigo tem como proposta desenvolver um estudo sobre os padrões da evasão escolar no Ensino Superior, com base na análise de dados fornecidos pelo Curso de Sistemas de Informação da Universidade Federal de Santa Maria (UFSM). Esses dados passarão por uma sistemática de tratamento de dados, a fim de apontar indicadores relacionados a fatores que classifiquem possíveis evasões. Serão gerados resultados por meio de gráficos que possam responder a problemática em questão.*

1. Introdução

De acordo com Favero (2006) denomina-se a evasão escolar como o processo de desistência do ensino pelo discente de determinado curso, indiferentemente à porcentagem de participação do aluno nas aulas. Tal problemática encontra-se emergente no cenário atual, como observado nas preocupantes taxas de evasão atuais (ALMEIDA & KAPPEL, 2020). Segundo Silva Filho et al. (2007), esse impasse vem impactando até mesmo o cenário internacional, afetando expressivamente os resultados dos sistemas educacionais e causando perdas nas instituições públicas e privadas. Assim, a motivação para a delimitação da temática do presente artigo reside em analisar o problema apresentado, bem como identificar padrões de alunos propensos a evadir em cursos de graduação, utilizando algoritmos de aprendizagem de máquina na mineração dos dados.

2. Materiais e métodos

A primeira etapa de desenvolvimento do projeto constituiu-se de uma revisão bibliográfica, em que foram investigadas diferentes definições de evasão escolar, dados e causas relacionadas a essa temática, bem como a expansão do acesso ao ensino superior nas últimas décadas. Nesse contexto, foi possível obter uma visão mais ampla acerca do ensino superior no Brasil. Dando continuidade, a revisão abordou a definição e aspectos importantes da mineração da dados, ferramenta principal a ser utilizada na etapa seguinte do projeto. Além disso, pesquisou-se a respeito do *Software Waikato Environment for Knowledge Analysis* (WEKA), software escolhido para a mineração de dados.

Outro ponto importante do processo de revisão foi a procura por trabalhos com propostas relacionadas, o qual abrangeu três trabalhos centrados na mineração de dados com o objetivo de entender as causas da evasão escolar no ensino superior de diferentes instituições. Nesse contexto, o primeiro trabalho considerado foi o de Paz e Cazella (2017), que desenvolveram um estudo de caso com enfoque na identificação do perfil de alunos propensos à evasão escolar, utilizando, para isso, a ferramenta de mineração WEKA, especificamente o algoritmo J48. Indo adiante, o segundo trabalho foi elaborado por Manhães et. al (2012), que utilizou diversas técnicas de mineração de dados com vista em investigar as relações que direcionavam o desempenho acadêmico de alunos evadidos da Universidade Federal do Rio de Janeiro. Por fim, o terceiro trabalho foi desenvolvido por Gonçalves, Da Silva e Cortes (2018), que dispôs da mineração como ferramenta para identificar a tendência à evasão escolar de alunos do Instituto Federal do Maranhão.

Terminada a etapa de obtenção de conhecimento teórico sobre a problemática da evasão no ensino superior, foi iniciada a fase de desenvolvimento da mineração de dados, com o objetivo de descobrir as causas da evasão escolar no curso de Sistemas de Informação da Universidade Federal de Santa Maria. Para tanto, escolheu-se o pacote de Software WEKA, e a tarefa de mineração foi dividida em cinco etapas principais: seleção, pré-processamento e limpeza, transformação, mineração e interpretação dos resultados.

3. Desenvolvimento

Até o presente momento, o projeto se encontra em fase de desenvolvimento, com as fases de seleção, pré-processamento e limpeza concluídas. Nesse contexto, o progresso das etapas são descritas nas próximas seções.

3.1. Seleção dos Dados

Primeiramente, o conjunto de dados foi obtido a partir das informações dos alunos do curso de Sistemas de Informação da Universidade Federal de Santa Maria, *campus* Frederico Westphalen, disponíveis no sistema de controle acadêmico. Não foi utilizado qualquer dado pessoal que pudesse identificar o aluno. A seleção de dados abrangeu cinco planilhas contendo dados referentes a todos os alunos por semestre que mantinham vínculo com a instituição entre 2015 e 2019. Visando garantir a privacidade desses estudantes, seus códigos de identificação (ID) foram substituídos por um código aleatório.

A primeira planilha possuía dados referentes ao ano, semestre e forma de evasão, contidos nos atributos, além do código aleatório de identificação, que estava presente em todas as planilhas. Já a segunda planilha contabilizou as aprovações e reprovações em cada disciplina por aluno, informando a disciplina cursada, seu código, ano e semestre em que foi cursada e situação final do aluno na referida disciplina. A terceira planilha apresentava as médias semestrais dos alunos. Assim, indicou os atributos referentes à média semestral, o número da matrícula, bem como o semestre e ano em que a média foi obtida. A quarta planilha indicava o número de trancamentos no respectivo semestre de cada ano. Por fim, a última planilha continha informações referentes ao ingresso na instituição, como seu ano, semestre e forma de ingresso. Além disso, continha dados sobre a evasão do estudante, como o ano, semestre e a forma que a evasão ocorreu. Por fim, a planilha apresentou outros dados dos alunos pertencentes ao estudo, como seu sexo, data de nascimento, naturalidade, estado de naturalidade e cidade e estado de moradia.

3.2. Pré-processamento e Limpeza dos Dados

Esta etapa teve por objetivo eliminar os dados considerados incompletos ou irrelevantes para o estudo, além de formar novos dados de importância com base na relação entre os já presentes. Foram considerados como evadidos os alunos que evadiram do curso por meio do abandono, transferência interna ou externa e cancelamento. Esse procedimento foi efetuado em uma planilha eletrônica e foi dividido em três etapas menores: limpeza dos dados, integração dos dados e redução dos dados (Han et al. 2011).

Com vista na redução de dados, definiram-se requisitos de pertinência para a utilização das informações obtidas. Nesse contexto, descartaram-se os dados referentes a alunos que não tivessem ao menos uma média semestral. No contexto de integração, os dados de interesse foram integrados em uma nova planilha. Nela, implementou-se o desempenho geral do aluno em cada ano e semestre cursado, indicando o número de reprovações por frequência, o número de disciplinas aprovadas no semestre, a média de notas semestral e a situação do aluno em cada semestre.

A distância entre a instituição de ensino e a moradia do aluno também foi considerada para este estudo, visando identificar se alunos que residem mais distante podem estar mais propensos a evadir. Para isso, foi necessário calcular a distância por uma interface de programação de aplicações do *Google Maps*. Este cálculo foi realizado pela seguinte fórmula:

```
=SUBSTITUIR(FILTROXML(SERVIÇOWEB("https://
maps.googleapis.com/maps/api/distancematrix/xml?
origins=cidade+campus&destinations=cidade+aluno&k
ey=YOUR_API_KEY");"//distance/text");" km";"")
```

Com o objetivo de padronizar algumas informações, as médias semestrais foram convertidas em conceitos representados pelas letras do alfabeto latino “A”, “B”, “C” e “D”, utilizando como base os critérios de conversão definidos pela Instituição de Ensino. Outro conjunto de dados padronizado foi o referente ao número de trancamentos do estudante durante todo o seu período de formação. Assim, definiram-se critérios de conversão para os valores contabilizados por cada estudante, sendo considerado “muito baixo” para os alunos que não trancaram nenhuma matéria, “baixo” para um ou dois trancamentos, “médio” para três trancamentos, “alto” para quatro e cinco trancamentos e “muito alto” para seis e sete trancamentos. Além disso, as informações referentes ao status foram padronizadas de modo que houvesse apenas três possibilidades para o aluno: regular, formado e evadido. Nesse contexto, os status de cancelamento de matrícula, transferência interna ou externa e abandono foram todos classificados como evadido.

Na fase de limpeza dos dados, objetivou-se resolver as possíveis inconsistências encontradas no banco de dados obtido, como foi o caso da presença de informações de moradia desatualizadas, como cidades localizadas demasiadamente distantes para o deslocamento diário até a instituição. Portanto, os resultados obtidos acima de quinhentos quilômetros foram convertidos a zero. A distância referente a moradores da cidade onde a instituição está localizada também recebeu zero. Outra ação relacionada à limpeza de dados foi a exclusão das informações de alunos que obtiveram classificação entre as vagas ofertadas pelo curso, mas não efetuaram sua matrícula.

3.3. Transformação dos Dados

O terceiro passo consistiu na transformação dos dados pré-processados. Em virtude de a segunda etapa ter sido desenvolvida no LibreOffice Calc, o volume de dados foi convertido inicialmente no formato "ODS" (*Open Document Spreadsheet*), que é nativo do software em questão. Todavia, o prosseguimento do processo de mineração de dados tornou necessária uma nova conversão dos dados, visando adequá-los ao software WEKA. Nessa circunstância, o conjunto de dados foi exportado no formato CSV, que possui compatibilidade com o WEKA. O conjunto de dados final está composto por 971 instâncias (alunos por semestre), e 7 atributos, sendo eles: distância entre cidade de moradia e instituição, percentual de reprovação por frequência, número de reprovação por nota, número de disciplinas cursadas, média de trancamentos realizados, média de notas e o status atual que será utilizado para predição.

4. Considerações Finais

Destarte, conclui-se que o embasamento teórico se encontra desenvolvido para a aplicação desta proposta, fato que foi possibilitado pelo método da pesquisa e estudo de projetos já desenvolvidos para a área objetivada. Para a conclusão do presente projeto, ainda se faz necessária a aplicação de algoritmos de mineração de dados sobre o conjunto selecionado e pré-processado, além da interpretação dos resultados devolvidos pelo software. Espera-se que, com a conclusão desta proposta, os resultados obtidos possam auxiliar gestores da educação superior na tomada de decisão ao predizer casos propensos de evasão.

Referências Bibliográficas

FAVERO, Rute Vera Maria. Dialogar ou evadir: Eis a questão!: um estudo sobre a permanência e a evasão na educação a distância. 2006.

DE ALMEIDA TEODORO, Leonardo; KAPPEL, Marco André Abud. Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil. **Revista Brasileira de Informática na Educação**, v. 28, p. 838-863, 2020.

SILVA FILHO, Roberto Leal Lobo et al. A evasão no ensino superior brasileiro. Cadernos de pesquisa, v. 37, n. 132, p. 641-659, 2007.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. Data mining: concepts and techniques. Elsevier, 2011.

PAZ, Fábio; CAZELLA, Sílvio. Identificando o perfil de evasão de alunos de graduação através da Mineração de dados Educacionais: um estudo de caso de uma Universidade Comunitária. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2017. p. 624.

MANHÃES, Laci Mary Barbosa et al. Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. **Anais do VIII Simpósio Brasileiro de Sistemas de Informação**, São Paulo, 2012.

GONÇALVES, Tayná Costa; DA SILVA, Josenildo Costa; CORTES, Omar Andres Carmona. Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do Instituto Federal do Maranhão. **Revista Brasileira de Computação Aplicada**, v. 10, n. 3, p. 11-20, 2018.