PREVISÃO DE ESTUDANTES COM RISCO DE EVASÃO UTILIZANDO TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Laís Matie Hara[1], Márcia Ferreira Cristaldo[1], Leandro de Jesus[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso do Sul – Aquidauana, MS

lmatiehara@gmail.com, [marcia.cristaldo, leandro.jesus]@ifms.edu.br

Resumo

O uso da mineração de dados para área de educação está sendo uma ferramenta crucial para previsão de evasão. A evasão do curso técnico é um fenômeno em crescimento e tornou-se foco de preocupação para gestores educacionais. Entretanto, as características da mesma têm carência de pesquisa e modelos de identificação de seus motivos. Esta pesquisa tem como abordagem aplicar técnicas computacionais para identificar padrões a serem utilizados na análise da evasão de estudantes no curso técnico de informática do Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso do Sul, campus Aquidauana, a fim de auxiliar os gestores educacionais em suas tomadas de decisões. Pretende-se utilizar um método para seleção dos melhores atributos para tarefa de classificação, que considera as classes "haverá evasão" e "não haverá evasão", baseado na seleção e criação de atributos.

Palavras-chave: Evasão, Inteligência Artificial, Previsão.

Metodologia e Desenvolvimento

Os estudos ligados à evasão iniciaram com teorias que buscam explicar a evasão e retenção. Apesar de não existir um conceito coeso, autores como Tinto (1975) abordam o modelo de integração do estudante, frisando que a decisão de evadir surge em função da falta de adaptação com o meio acadêmico/social da instituição, influenciada pelas características individuais e expectativas para a carreira ou curso e intenções e assumidas antes do início do curso. Já segundo Barroso e Falcão (2004) as condições que motivam a evasão escolar são classificadas sob três agrupamentos: 1) econômica; 2) vocacional; e 3) institucional.

O Instituto Federal de Educação de Mato Grosso do Sul (IFMS) oferece cursos técnicos integrados em informática desde 2011, mesmo com a concorrência por vaga, o problema de evasão preocupa a direção do IFMS. A Base de Dados (BD) utilizada foi coletada por questionário composto por 20 questões, no qual os estudantes tiveram a sua identificação preservada, sendo composto por 19 registros de estudantes do semestre 2-2018. Para esta investigação, as características selecionadas como preditoras para análise e identificação da evasão foram perguntas vocacionais.

Atualmente a aplicação das técnicas de mineração de dados é facilitada devido a existência de ferramentas que dispõem de recursos de análise de dados e implementam algoritmos específicos. A ferramenta de mineração de dados Waikato Environment for Knowledge Analysis (Weka) (HALL et al., 2011) foi a escolhida para este trabalho devido a fatores: facilidade de aquisição e disponibilidade.

Foi adotado uma pesquisa quantitativa, pois a abordagem adotada para análise do método proposto de seleção dos melhores atributos para classificação.

O método utilizado para seleção dos melhores atributos:

Figura 1. Diagrama do desenvolvimento do projeto.

Os experimentos foram realizados subsequentemente neste método para a prova de conceito e análise do método propriamente dito. O método utilizado foi composto por 10 etapas:

Pré-processamento. Nesta etapa são realizadas as atividades de extração, limpeza, transformação, carga e atualização dos dados, conforme os procedimentos tradicionais empregados em mineração de dados.

Criação de Atributos. A criação de novos atributos pode capturar informações importantes em um conjunto de forma mais eficiente do que os atributos originais.

Transformação dos dados. Nesta etapa são realizadas as tarefas de normalização, discretização e amostragem dos dados, também seguindo os procedimentos tradicionais empregados na mineração de dados.

Remoção de Valores Discrepantes. Nesta etapa é verificada a necessidade de remoção de valores discrepante (outliers).

Balanceamento de classes. Apesar de a evasão ser um problema nas instituições de ensino, o número de casos de evasão ainda é, em geral, menor em relação ao número de alunos não evadidos. Sendo assim, o problema se caracteriza por desbalanceamento de classes. Este problema faz com que os algoritmos de aprendizagem tendem a ignorar as classes menos frequentes (classes minoritárias) e só considerar nas mais frequentes (classes majoritárias).

Seleção do Subconjunto de Atributos. A seleção do subconjunto de atributos é um método de redução da

dimensionalidade quando são detectados e removidos atributos irrelevantes, fracamente relevantes ou redundantes.

Execução do Classificador em Cada Subconjunto de Atributos. Depois de selecionados os subconjuntos de atributos, os classificadores devem ser avaliados quanto ao desempenho, utilizando-se como medida a acurácia.

Exclusão dos Subconjuntos sem Significância Estatística. O objetivo dessa etapa é descartar o subconjunto de atributos cuja acurácia seja muito inferior à melhor acurácia obtida com o classificador no experimento.

Ordenação e Escolha do Melhores Atributos. Para se obter os melhores atributos, após o descarte dos subconjuntos sem significância estatística.

Aplicação do Algoritmo de Classificação. Nesta etapa foi aplicado o algoritmo de classificação no dataset com os melhores atributos selecionadas.

Resultados e Considerações Finais

Este experimento foi realizado no ambiente *WEKA*, utilizou-se os algoritmos *BestFIrst procurando os melhores atributos. Após encontrado a melhor busca, foi aplicado o algoritmo* **K-Nearest Neighbor** (**KNN**). O classificador K-nn, ou K vizinhos mais próximos, é uma técnica baseada na aprendizagem por analogia, ou seja, comparando uma determinada tupla teste com tuplas de treinamento que são semelhantes. As tuplas de treinamento são descritas por n atributos. Cada tupla representa um ponto em um espaço n-dimensional. Desta forma, todas as tuplas de formação são armazenadas num espaço padrão de n dimensões. Quando uma dada tupla é desconhecida, um classificador k-vizinho mais próximo procura o espaço para as tuplas de treinamento k que estão mais próximas da tupla desconhecida. Estas tuplas de treinamento k são os k "vizinhos mais próximos" da tupla desconhecida (HAN et al., 2011). Para este projeto foi utilizado como saída para o modelo a notas da disciplina de algoritmos, sendo a disciplina com maior índice de reprovação, levando assim a desmotivação do estudante em relação ao curso. O r^2 (Coeficiente de Correlação) obteve 73% de correlação entre as variáveis de entrada do modelo, ou seja, as perguntas do questionário tiveram um alto índice de correlação entre as perguntas e a resposta (nota da disciplina de algoritmo).

Tabela 1. Desempenho da previsão utilizando o algoritmo KNN.

Parâmetros	KNN
6 atributos foram selecionados pelo best-first	
R^2	**69%**
r^2	**73%**
EMA (Erro médio absoluto)	**1,64**
RQEM (Raiz quadrada do erro médio)	**2,05**
RQER (Raiz quadrada do erro relativo)	**68%**
RAR (Erro absoluto relativo)	**63%**

Os resultados das amostras mais representativas, ou seja, 85% da base, serviu como fonte para que fosse gerado um gráfico de dispersão com o respectivo R^2 para o algoritmo KNN. O coeficiente de determinação, também chamado de R^2, é uma medida de ajustamento de um modelo estatístico linear generalizado, como a regressão linear, em relação aos valores observados. O R^2 varia entre 0 e 1, indicando, em percentagem, o quanto o modelo consegue explicar os valores observados. Para este projeto obteve-se 69% de acerto entre valor observado e previsto (Tabela 1).

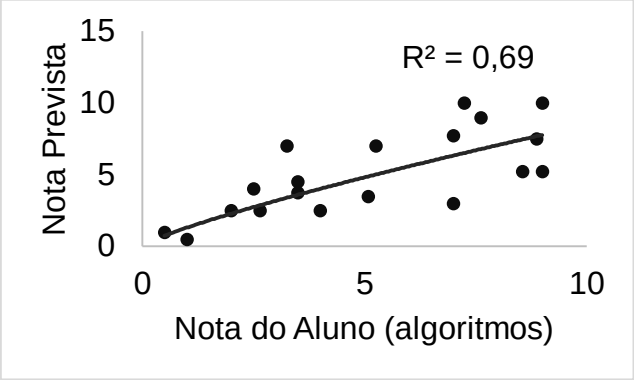

Figura 2. Gráfico de dispersão com as notas estimadas versus nota do aluno (Algoritmo KNN).

Conforme os resultados apresentados, pode-se verificar que as notas dos alunos da disciplina de algoritmos e as notas previstas tem interferência direta nas decisões do estudante "evasão" ou "não evasão", pode-se visualizar que o R^2 (Coeficiente de determinação) 69% de acerto entre a nota prevista dos estudantes baseada no questionário aplicado em relação a nota real. O trabalho está em andamento, mas verificou-se que 1 estudante evadiu após a aplicação deste questionário, no qual o algoritmo acertou na sua nota prevista, baseado nas respostas do questionário.

Conforme revisão da bibliografia, a evasão ainda é um fenômeno em crescimento no IFMS/Aquidauana, justificando a necessidade de se gerar interferências para identificação de padrões para a sua análise. Esse método vai contribuir para o processo de identificação de padrões a serem utilizados na previsão da evasão para apoiar a tomada de decisão, com a finalidade de reduzi-la no ensino técnico em informática do IFMS, campus Aquidauana.

Referências

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: Oportunidades para o Brasil. Revista brasileira da informática na educação, v. 19, n. 2, 2011.

Borges, V. A., Nogueira, B. M., and Barbosa, E. F. (2015). Uma análise exploratória de tópicos de pesquisa emergentes em informática na educação. Revista Brasileira de Informática na Educação, 23(01):85.

BOUCKAERT R., EIBE F. MARK HALL, KIRKBY, R., REUTEMANN, P., Seewald, A., and Scuse, D. (2010) "WEKA Manual for Version 3-6-4". December

WITTEN, I. H.; FRANK, E.; HALL, M. A. Data Mining: Practical Machine Learn Tools and techniques. Morgan Kaufimann Publishers Inc. San Francisco, CA, USA, 3° ed., 570p, 2011.

OPCIONAL: 2ª ´Página

Esta parte não é obrigatória e pode ser excluída, caso os autores assim desejarem. Entretanto, é recomendável que se faça a versão em Inglês desses elementos, até para fins de divulgação mais ampla)

TITLE IN ENGLISH

Abstract: (*Write the English version with the same structure using italic characters*)

Keywords: (*Write the same words in English using italic characters)*