Implementação de um Modelo para Previsão de Evasão Escolar no IFSULDEMINAS

Gabriel Soares de LIMA[1]; Paulo M. de ÁVILA[2]; Raquel GILAVERTE[3]

RESUMO

Esse trabalho apresenta um modelo para previsão de evasão escolar no IFSULDEMINAS. Foi implementada uma aplicação Java que através de técnicas de mineração de dados disponibilizadas pelas *APIs* da ferramenta *Weka* permitem detectar padrões e perfis considerando os registros acadêmicos dos discentes. Os resultados obtidos alcançaram um nível de acerto relevante na construção do perfil de alunos com tendência a evasão escolar.

INTRODUÇÃO

A mineração dos dados é uma etapa do processo de Descoberta de Conhecimento em Base de Dados denominado como (KDD). Devido à grande dificuldade de detectar padrões existentes por métodos tradicionais de análise, que em sua maioria são lentos e de alto custo o KDD propõe que a descoberta de conhecimento seja obtida por meio do uso de ferramentas especificas que utilizam diferentes técnicas. Estas ferramentas têm o papel de realizar buscas por padrões válidos e ao mesmo tempo úteis e compreensíveis, de forma compacta dentro de uma base de dados já organizada e preparada para a etapa de mineração que auxilia profissionais de diferentes áreas na realização de análises e na tomada de decisões. Neste trabalho foram utilizados classificadores e regras de associação como técnica de descoberta de conhecimento, juntamente com a ferramenta de mineração WEKA para realizar buscas por padrões existentes na base de dados e que aparentemente são irrelevantes.

MATERIAL E MÉTODOS

[1] Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Câmpus Poços de Caldas. Poços de Caldas/MG, email: gabrielsoaresdelima@hotmail.com.
[2] Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas – Câmpus Poços de Caldas. Poços de Caldas/MG, email: paulo.avila@ifsuldeminas.edu.br.
[3] Universidade de Ribeirão Preto – UNAERP, email: raquelgilaverte@gmail.com.

O trabalho teve o objetivo principal de identificar padrões potencialmente úteis que permitam a extração de indicadores do perfil de aluno com tendência a evasão escolar através da aplicação de técnicas de mineração de dados como: classificação e regras de associação.

Através da análise de histórico de evasão foi possível elaborar um perfil dos alunos com tendência a evasão escolar.

Para a realização deste trabalho foi utilizado a ferramenta a WEKA (*Waikaito Environment for Konowledge Analysis*) (Mark Hall et. al., 2009), um pacote de algoritmos implementado para realização de tarefas que envolvem Mineração de Dados, desenvolvida pela Universidade de *Waikato* - Nova Zelândia e é disponibilizado gratuitamente pelo site http://www.cs.waikato.ac.nz/ml/weka/.

O processo de descoberta de conhecimento ou KDD (*Knowledge Discovery in Database*) é composto de várias etapas, sendo a mineração de dados uma importante fase desse processamento (FAYYAD, 1996). A seguir são apresentadas as etapas envolvidas no processo.

O processo tem início com a coleta das informações, ou seja, os dados brutos que deseja-se extrair algum conhecimento. Em seguida, na segunda etapa, os dados coletados devem ser preparados para um formato adequado e também se deve eliminar eventuais inconsistências. A terceira etapa consiste na fase de transformação dos dados, onde com a utilização dos dados pré-processados e livres de inconsistências ocorre a adequação aos algoritmos de mineração, ou seja, os dados são formatados para um padrão de entrada para ser utilizado por algoritmos de mineração. A quarta etapa é a mineração de dados, onde ocorre a extração dos padrões propriamente ditos. Uma vez descoberto o padrão na fase de mineração tem-se início a quinta etapa, que é a interpretação dos padrões descobertos, a eliminação de padrões com pouca relevância e, se necessário, o retorno a algum passo anterior (ÁVILA, P. M.; ZORZO, S. D, 2009).

Esse projeto explorou diversas fases do processo de KDD, como por exemplo, a coleta de informações na secretaria do câmpus, através dos formulários de evasão, preparação e transformação dos dados eliminando informações nulas ou inconsistentes e pôr fim a fase de mineração e interpretação dos padrões obtidos, etapa essa que foi implementada uma aplicação em Java que faz uso das APIs (*Application Program Interface*) do *software Weka* para geração dos modelos.

A fase de mineração de dados é com certeza a mais complexa, sendo considerada até sinônimo de KDD (HAN; KAMBER, 2006).

O processo de mineração consiste de três fases: escolha da tarefa de mineração, a escolha do algoritmo e a extração de padrões.

A tarefa de mineração é classificada pelo tipo de padrão que se deseja obter e as principais tarefas são: classificação, associação e agrupamento (SCHAFER, 2001). Para o contexto desse trabalho foram utilizados os algoritmos de classificação e associação.

A classificação permite definir novos grupos a partir de dados existentes mediante um modelo ou classificador. Os classificadores podem ser implementados pelo uso de diferentes estratégias de aprendizado de máquina como redes bayesiana, redes neurais, árvores de decisão e regras de classificação. Esse trabalho utilizou dois algoritmos conhecidos para classificação, o algoritmo J48 e o algoritmo *Naive Bayse*.

A associação busca encontrar relacionamento ou padrões frequentes entre os dados. Por exemplo, as regras de associação podem identificar itens de perfil de um usuário baseado em um histórico prévio.

Para que seja possível implementar um modelo é necessário definir variáveis que serão analisadas pelos algoritmos. Nesse projeto foram utilizadas as seguintes variáveis: renda per capita da família, sexo, idade, atividades desenvolvidas além da escolar, distância física do câmpus, formação anterior (escola pública ou privada), tempo em anos que aluno ficou sem frequentar instituição de ensino.

RESULTADOS E DISCUSSÃO

Para realizar esse trabalho foi implementado em Java uma aplicação que faz usos das APIs disponibilizadas na ferramenta *Weka*. Os resultados foram obtidos através da submissão dos dados de histórico de três anos considerando os cursos técnicos subsequentes, cursos integrados e cursos de tecnologia ofertados nesse período. No total foram utilizados 1.487 registros de alunos matriculados com um percentual de 20% de evasão.

A seguir apresentamos algumas análises realizadas através do software implementado nesse trabalho.

Na Figura 1 apresentamos a interface gráfica da aplicação Java. A Aba analise apresentada na figura 1 permite uma análise de qual a probabilidade de um

aluno ou grupo de alunos evadir. Essa análise é feita baseada em um histórico que foi fornecido pela secretaria do Campus.

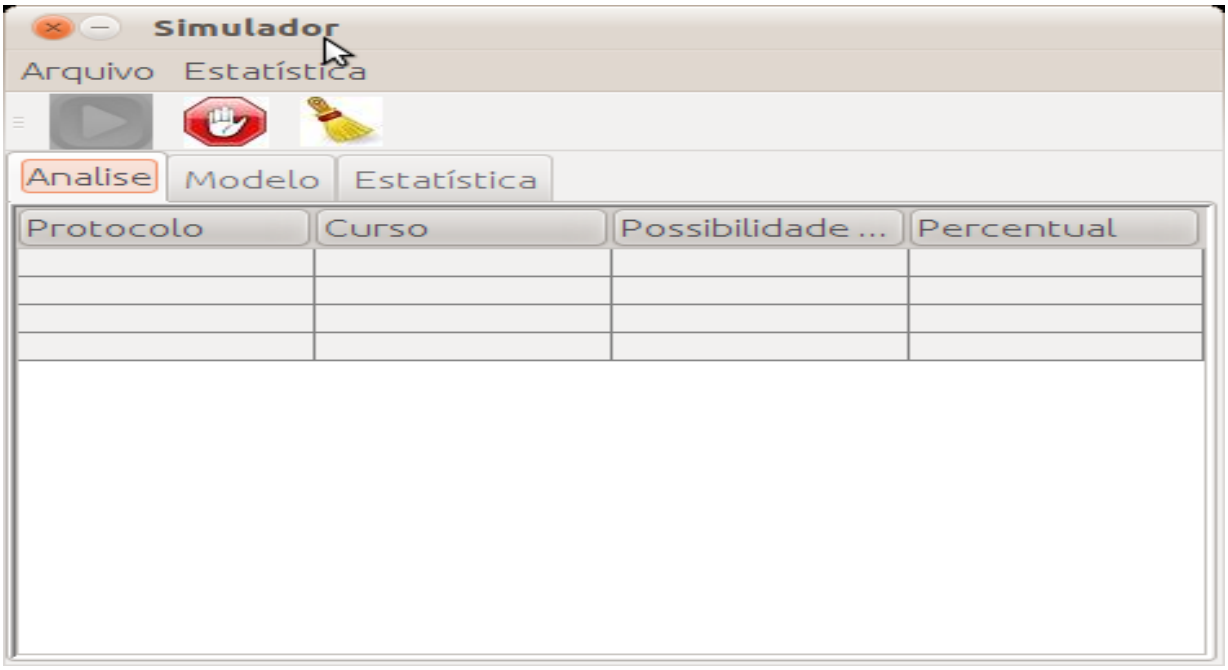

Figura 1: Interface de carga de dados do sistema.

Esse histórico foi processado e gerado um arquivo no padrão ARFF. ARFF é o padrão de arquivo utilizado pela ferramenta *Weka* para realizar a entrada dos dados nos algoritmos. Uma vez que os dados estejam formatados e carregados, a segunda etapa é definir o tipo de algoritmo de mineração será utilizado para criação do modelo. A aplicação implementada permite a escolha de dois algoritmos (J48 ou Naive Bayses).

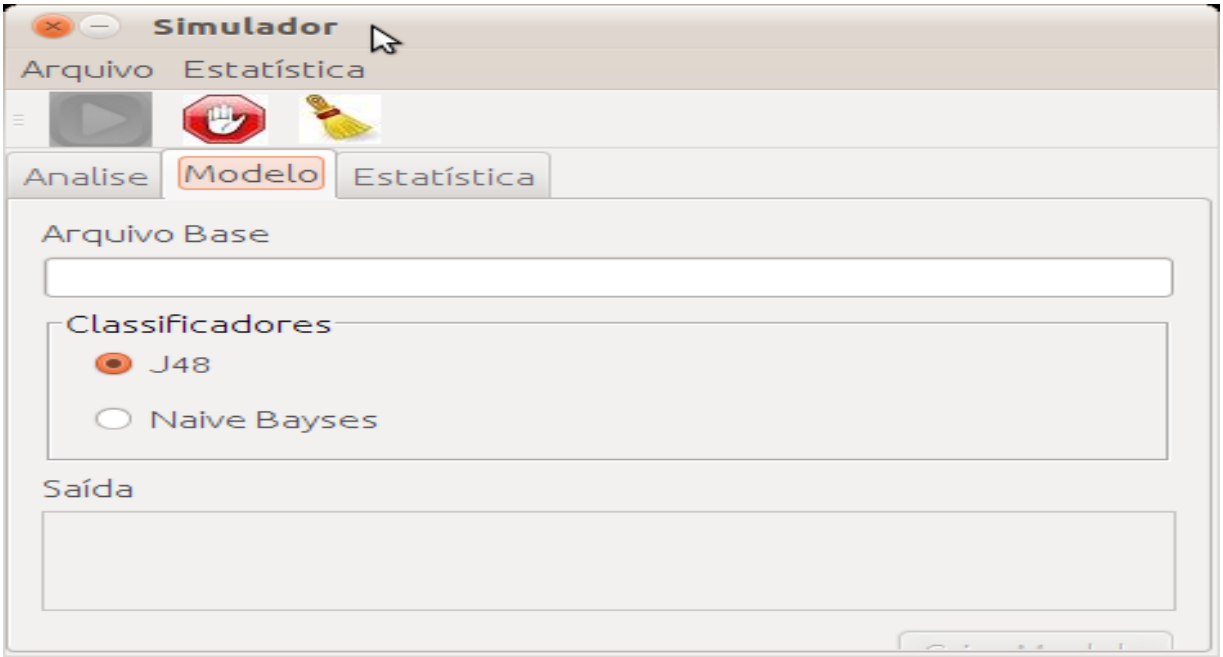

Figura 2: Interface para seleção do Modelo (J48 ou Naive Bayses)

A figura 2 ilustra a interface que permite ao usuário selecionar o algoritmo de mineração que será utilizado para criação do modelo.

Os classificadores possíveis nos softwares são o J48 responsável por criar uma árvore de decisão elencando os motivos reais que levam os alunos a evadirem e o algoritmo *Naive Bayse*s que permite realizar um estudo probabilístico dos alunos produzindo como resposta o grau de probabilidade que cada aluno analisado tem de evadir do Campus.

Na Figura 3 é apresentado o total de alunos analisados, quantos porcentos tem alta probabilidade de evasão e quantos porcentos tem baixa probabilidade de evasão.

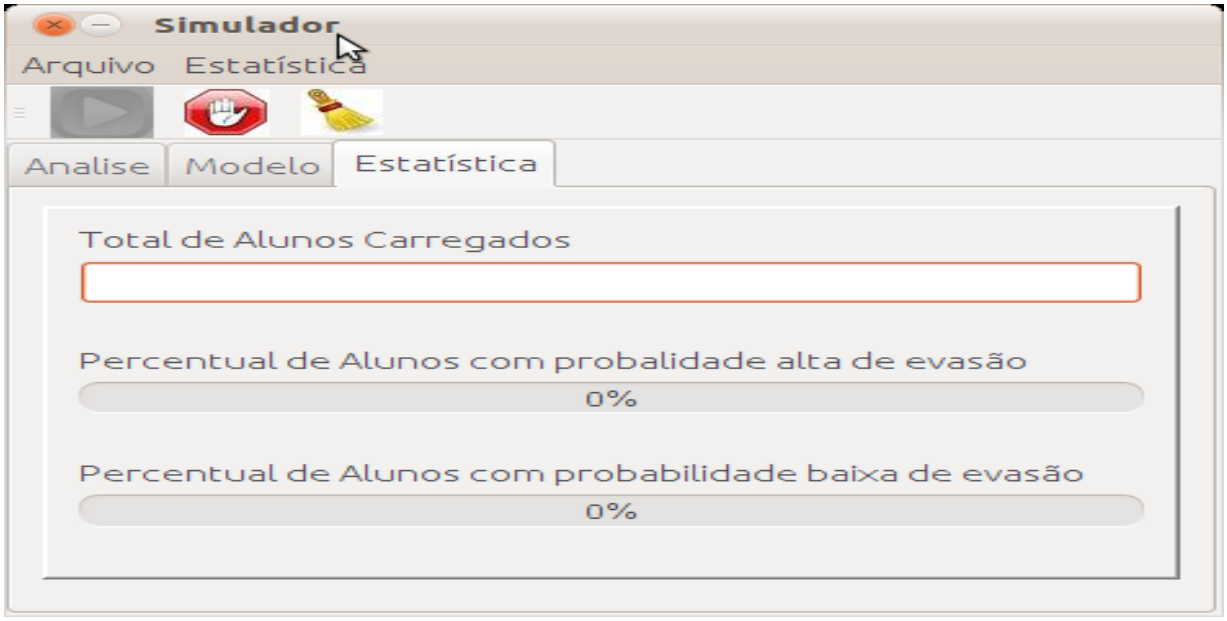

Figura 3: Interface de análise probabilística de alta e baixa evasão.

Foram realizados teste com uma base histórica de três anos no Câmpus. Os resultados finais demonstram um acerto da ferramenta em torno de 85% considerando os dados de teste e o conjunto total dos dados.

CONCLUSÕES

Os testes considerando a base completa de informação apresentaram resultados de acerto em torno de 85%. Os testes foram conduzidos da seguinte forma: primeiro todos os alunos ativos e evadidos foram processados pelo modelo. Era de conhecimento dos pesquisadores quais eram os alunos já evadidos. O objetivo era verificar o índice de acerto do algoritmo após o processamento. O

algoritmo obteve um índice de acerto de 85%, ou seja, dos alunos classificados com alta probabilidade de evasão, 85% efetivamente evadiram. Em linhas gerais o perfil do aluno com forte tendência a evasão é definido como: o aluno não estuda a mais de 5 anos, é maior de 27 anos e trabalha de segunda a sábado em um total de 8 horas diárias. Os alunos que apresentam esse perfil têm uma probabilidade de 88,56% de evadir, considerando os dados históricos do câmpus e o modelo implementado. Vale ressaltar que conforme novos dados são inseridos o modelo deve ser reprocessado, fato que pode ou não alterar o perfil do aluno com forte tendência a evasão.

AGRADECIMENTOS

Os pesquisadores agradecem ao IFSULDEMINAS pela disponibilização dos recursos financeiros através do edital 014/2013.

REFERÊNCIAS BIBLIOGRÁFICAS

ÁVILA, P. M. ; ZORZO, S. D. . **A personalized TV Guide System: An Approach to Interactive Digital Television**. In: IEEE International Conference on Systems, Man, and Cybernetics, 2009, San Antonio. Proceedings of the SMC, 2009.

FAYYAD, U. M.; PIATETSKY-SHAPIRO G.; SMITH, P.: **Knowledge Discovery In Databases: An Overview.** In: KNOWLEDGE DISCOVERY IN DATABASES, eds. G. Piatetsky- Shapiro, and W. J. Frawley, 1996, Cambridge, MA. **Proceedings…**pp 1-36, 1996.

HAN, J.; KAMBER, M.: **Data Mining: Concepts and Techniques**. Morgan Kaufmann, 2001.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. **The WEKA Data Mining Software: An Update**; SIGKDD Explorations, Volume 11, Issue 1, 2009.

SCHAFER, J. B.: MetaLens: **A Framework for Multi-source Recommendations**. 2001.Tese em Ciências da Computação (Doutorado em Ciência da Computação), University of Minnesota, 2001.