MINERAÇÃO DE DADOS EDUCACIONAIS: UM ESTUDO DE CASO APLICADO AO PROCESSO SELETIVO DO IFSULDEMINAS – CÂMPUS MUZAMBINHO

Fernanda Delizete Madeira[1]; Aracele Garcia de Oliveira Fassbinder[2]

INTRODUÇÃO

Data Mining, também chamada de Mineração de Dados, surgiu na década de 80, tornando-se uma das etapas principais dentro do processo para a descoberta de conhecimento em base de dados – KDD (Knowledge Discovery in Database), originada de áreas como Estatística, Inteligência Artificial e Banco de Dados. A técnica é usada para transformar grandes volumes de dados em informações significativas para o planejamento, a gestão e a tomada de decisão.

Embora as aplicações mais comuns de mineração de dados se refiram a clientes, compras e vendas, esta área é ampla e tem sido aplicada no setor educacional. As instituições de ensino tiveram, nos últimos tempos, uma ampliação, tanto de cursos, quanto de vagas. Desta forma, os responsáveis precisam ter a preocupação de acompanhar a permanência desses alunos nos cursos ofertados. Para que o índice de conclusão de cursos aumente é necessário identificar os fatores que levaram ao insucesso dos estudantes.

A maioria das instituições de ensino solicita o preenchimento de uma ficha chamada “Questionário sócio econômico e cultural”, a fim de se obter informações que possam levar ao conhecimento sobre os candidatos ao processo seletivo. O IFSULDEMINAS – Câmpus Muzambinho faz uso dessa prática; entretanto, não são realizadas ações que conseguem, efetivamente, aproveitar destes dados para a tomada de decisões internas.

Sendo assim, este trabalho propõe delinear o perfil do candidato ao processo seletivo de admissão para os cursos de ensino técnico e superior, através da aplicação do Processo de Descoberta de Conhecimento em Bases de Dados, a fim de levantar informações relevantes que tragam subsídios para as tomadas de decisões dentro da instituição de ensino

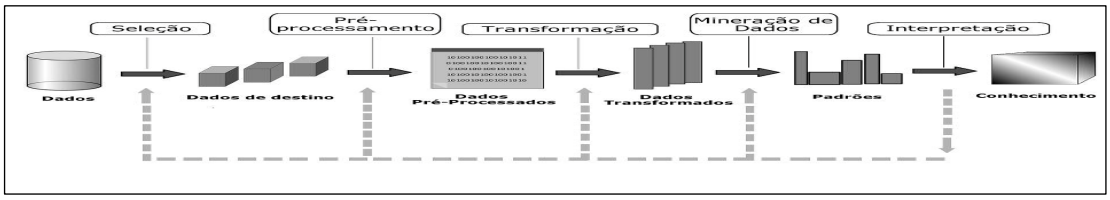
[1]Aluna do Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais - Câmpus Muzambinho. Ciência da Computação. E-mail: fernandanovaresende@hotmail.com
[2]Orientadora do Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais - Câmpus Muzambinho. Ciência da Computação. E-mail: aracele.garcia@gmail.com

PROCEDIMENTOS METODOLÓGICOS

Os dados utilizados no presente trabalho, foram disponibilizados em duas planilhas com o formato .xls, padrão do Microsoft Excel. Tal formato possui cabeçalhos na estrutura compatível para a importação das mesmas para o banco SqlServer 2008. Estas planilhas contêm as respostas dos itens propostos no questionário sócio econômico e cultural aplicado aos candidatos no ato da inscrição, dos anos de 2012 e 2013. Além destes dados, foram disponibilizados, também, alguns dados do cadastro geral dos candidatos, tais como: o sexo, a idade, a cidade e o total de pontos obtidos pelo candidato no vestibular.

A Figura 1 mostra as etapas de KDD da fonte de dados inicial até o processo de extração de conhecimento.

Figura 1 - Etapas do processo de extração do conhecimento.



Fonte - (adaptado de Fayyad et al.,1996).

O processo de KDD pode ser dividido em três etapas operacionais: Pré-Processamento, Mineração de Dados e Pós-Processamento.

Uma breve descrição destas etapas, segundo Boente, Goldsmidt e Estrela (2006) é apresentada a seguir:

a) Pré-Processamento: Compreende na obtenção, limpeza, padronização e transformação e adequação dos dados.

A limpeza inclui a remoção de ruídos, a adequação de valores que estejam fora do contexto. Na base de dados utilizada, foram excluídas da tabela as colunas inscrição e data_nascimento. A coluna inscrição foi excluída por possuir dados irrelevantes e a coluna data_nascimento por já existir uma coluna semelhante que é a coluna idade. Também foram retirados alguns registros que possuíam idades erradas, tais como, (-52, 1, 2, 3, 1015, 1016, 1017, 1018, 1032, 2007, 2013).

A padronização dos dados é deixar os dados que possuem os mesmos valores, ou significados, representados de forma única. Neste contexto, algumas cidades possuíam mais de uma forma de cadastro, como por exemplo: ((Alpinopolis, Alpinópolis), (Conceicao da Aparecida, Conceiçao da Aparecida, Conceição da

Aparecida), (Divinolandia, Divinolândia), entre outras). Estes dados então foram alterados.

A transformação é agregar mais informações aos registros existentes, enriquecendo os dados, para que estes forneçam mais informações para o processo de KDD. Na base de dados utilizada, os atributos das colunas idade, cidade, total_pontos e curso foram transformados em conjuntos maiores de dados. No quais ficaram representados da seguinte forma: Idade (Menos de 17 anos, 17 anos, Entre 18 e 22 anos, Entre 23 e 30 anos, Entre 31 e 40 anos, Acima de 40 anos). Cidade (Em Muzambinho, Até 50km de Muzambinho, Entre 50 e 100km de Muzambinho, Entre 100 e 150km de Muzambinho, Entre 150 e 200km de Muzambinho, Acima de 200km de Muzambinho). Total_Pontos (Menos de 20 pontos no vestibular, Entre 20 e 40 pontos no vestibular, Entre 40 e 60 pontos no vestibular, Entre 60 e 80 pontos no vestibular, Acima de 80 pontos no vestibular).

Com as tabelas criadas e devidamente modificadas, foi finalizado o processo de pré-processamento.

b) Mineração de Dados: Data Mining é a etapa do KDD que consiste na aplicação de algoritmos específicos, que extraem padrões a partir dos dados (DANTAS, E.R. et al., 2012 apud FAYYAD et al., 1996). Originalmente, essa técnica deriva das áreas de Estatística, Inteligência Artificial e Banco de Dados, e tem como objetivo explorar grande quantidade de dados na busca de padrões consistentes.

Várias ferramentas apoiam a etapa de Mineração de dados, tais como, Clementine, Enterprise Miner, Intelligent Miner, Weka.

A ferramenta Weka, escolhida para dar suporte a esta etapa, possui interface gráfica amigável, possibilita a utilização de recursos via API's, é distribuída gratuitamente (característica que as outras ferramentas citadas não possuem), e é muito citada por diversos autores como uma excelente ferramenta de suporte à mineração de dados.

As regras são geradas no seguinte formato:

1. internet=Diariamente classe_pontos=ENTRE_60-80_PONTOS 145 ==> micro=Tem em casa e usa regularmente 135 conf:(0.93)

Este símbolo ==> faz a divisão entre o antecedente e o consequente. O número que aparece antes do símbolo ==> indica o suporte da regra, neste caso é o 145. O número que aparece no final da regra indica quantas vezes o consequente aparece para cada ocorrência do antecedente, ou seja, 135. E o número final, entre

parênteses, é o valor da confiança, que é calculado a partir das transações em comum, ou seja, 135/145= 0,93, representando 93%.

c) Pós-Processamento: A etapa de pós-processamento compreende a visualização, análise e interpretação da etapa de mineração. Nessa etapa, o analista/especialista em KDD verifica os resultados obtidos na etapa anterior e faz uma análise para a transformação do conhecimento em novas alternativas de uso de informações. Os padrões extraídos podem ser simplificados, avaliados, visualizados ou simplesmente documentados para o usuário final.

RESULTADOS E DISCUSSÕES

A principal regra gerada foi através da seleção dos atributos referentes às seguintes questões do Questionário Sócio Econômico aplicado em 2012:

- Qual o motivo principal da escolha do curso para o qual você se inscreveu?
- Por que escolheu o IFSULDEMINAS?

Juntamente com essas duas perguntas foram selecionadas o atributo classe_cidade, no qual foi transformado a partir da relação de cidade.

Após realizar a etapa de Mineração de Dados, as seguintes relações foram geradas:

1.ifsuldeminas=Realização pessoal classe_cidade=ENTRE_100KM-150KM_MUZAMBINHO 41 ==> motivo=Realização pessoal 36 conf:(0.88)

2.ifsuldeminas=Realização pessoal classe_cidade=ENTRE_50KM-100KM_MUZAMBINHO 115 ==> motivo=Realização pessoal 96 conf:(0.83)

3.ifsuldeminas=Possibilidades no mercado de trabalho classe_cidade=EM_MUZAMBINHO 73 ==> motivo=Possibilidades no mercado de trabalho 50 conf:(0.68)

4.ifsuldeminas=Possibilidades no mercado de trabalho classe_cidade=ATE_50KM_MUZAMBINHO 120 ==> motivo=Possibilidades no mercado de trabalho 78 conf:(0.65)

Com essas regras pode-se observar que: a regra 1 e 2 indica que de 83 a 88% dos candidatos que escolheu o IFSULDEMINAS por realização pessoal e que moram entre 50 e 150 km de Muzambinho também escolheu o curso por realização pessoal. A regra 3 indica que, 68% dos candidatos que escolheu o IFSULDEMINAS

por possibilidade no mercado de trabalho, moram em Muzambinho e escolheram o curso também por possibilidade no mercado de trabalho. A regra 4 indica que, 65% dos candidatos que escolheu o IFSULDEMINAS, por possibilidade no mercado de trabalho, moram até 50km de Muzambinho e escolheram o curso também por possibilidade no mercado de trabalho.

Com base nessas regras acima, conclui-se que os candidatos que vêm de longe, numa distância de 50 a 150 km, vêm pela qualidade do IFSULDEMINAS; já os que moram em Muzambinho ou em até 50 km de distância, procuram o IFSULDEMINAS por possibilidade no mercado de trabalho.

CONSIDERAÇÕES FINAIS

Com as experiências adquiridas na realização deste trabalho, foi possível perceber que a primeira etapa no processo de descoberta de conhecimento em banco de dados, que é o pré-processamento, é de extrema importância e exige uma atenção especial, pois a forma que se conclui esta etapa irá influenciar diretamente no sucesso das etapas posteriores. Quanto mais os dados forem transformados e enriquecidos, mais padrões de associações são encontrados.

O objetivo principal deste trabalho foi extrair conhecimentos interessantes com a finalidade de delinear o perfil dos candidatos ao processo seletivo do IFSULDEMINAS, através da aplicação da técnica de mineração de dados. Com os resultados obtidos, percebemos que faltam informações para que seja possível traçar com mais precisão o perfil dos candidatos e a partir destas informações, implementar ações visando melhorar a qualidade de ensino, diminuindo evasões. Sugere-se uma adaptação no atual questionário sócio econômico e cultural.

REFERÊNCIAS

BOENT, A.N.P; GOLDSCHIMIDT, R.R; ESTRELA, V.V. **Uma metodologia de suporte ao processo de conhecimento em bases de dados**. In V SIMPÓSIO DE EXCELÊNCIA EM GESTÃO E TECNOLOGIA, 2008. Resende (RJ). Disponível em: <http://www.boente.eti.br/publica/seget2008kdd.pdf> Acesso em: 12 dez. 2012.

DANTAS, E.R. et al. **O Uso da Descoberta de Conhecimento em Base de Dados para Apoiar a Tomada de Decisões**. UNIPÊ Centro Universitário de João Pessoa, João Pessoa. Disponível em <http://www.aedb.br/seget/artigos08/331_331_Artigo_SEGET_EJDR_Versao_Final_010808.pdf> Acesso em: 13 dez. 2012.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery: An Overview.** Knowledge Discovery and Data Mining, Menlo Park : AAAI Press, 1996.