Faculdade Presidente Antônio Carlos de Teófilo Otoni - Dezembro de 2018

USO DE MINERAÇÃO DE DADOS PARA DESCOBERTA DE CONHECIMENTO: ESTUDO DE CASO DO VESTIBULAR DA UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI (UFVJM)

Marcus Vinícius Barbosa Costa[1]
Áthila Rocha Trindade[2]

Resumo

A Mineração de Dados (MD) une conhecimentos das áreas de Banco de Dados, Estatística e Inteligência Artificial com o objetivo de extração de conhecimento útil em bases de dados, cada vez maiores. Para isso, ela conta com diferentes técnicas que, quando aplicadas adequadamente, podem revelar conhecimentos de grande valia para a gestão do negócio. Diante disso, neste trabalho foram analisadas as bases de dados referentes aos anos de 2002 a 2009 do processo seletivo unificado da Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), com o objetivo de extração de conhecimento acerca do perfil socioeconômico dos candidatos (potenciais futuros alunos da instituição). Foi utilizado o algoritmo da MD *Apriori*, que implementa a extração de conhecimento através da obtenção de regras de associação. Os resultados obtidos comprovam a eficácia desta técnica para descoberta de conhecimento em grandes bases de dados, o que permitiu à universidade conhecer melhor o perfil socioeconômico dos candidatos no período mencionado.

Palavras - chave: Mineração de Dados - Descoberta de conhecimento - Regras de associação - *Apriori*.

Abstract

Data Mining (DM) is a technique whose aim is to more fully exploit database information by combining knowledge from the fields of database, statistics and artificial intelligence. When properly applied, Data Mining can provide valuable information to business management. The aim of the present study was to apply DM techniques to entry exams at a Brazilian federal university (UFVJM) in order to extract knowledge about the socio-economic profiles of the candidates (who can be students of this University). The study utilized a DM algorithm called Apriori, which works by determining association rules. The results of the study confirmed the effectiveness of this data-extraction method for large databases, thus providing the university with a better understanding of the socio-economic profiles of the candidates.

Keywords: Data Mining - Knowledge Discovery - Association rules - Apriori.

Introdução

Nas últimas décadas, houve um aumento gigantesco no volume de dados produzidos pelas diversas organizações. A evolução das tecnologias de *software* e *hardware* e a redução constante nos custos de armazenamento de dados contribuíram para que, cada vez mais, empresas adquirissem e utilizassem computadores e dispositivos de armazenamento para a execução de suas atividades, culminando num acúmulo de dados cada vez maior.

Diante disso, a capacidade das organizações de coletar e armazenar estes dados tornou-se muito superior à sua capacidade de analisá-los e compreendê-los. Tal cenário gerou uma demanda por técnicas e ferramentas computacionais que permitissem a extração de conhecimento a partir desses volumes crescentes de dados. Os esforços para satisfazer esta necessidade levaram ao surgimento de uma nova área de pesquisa em banco de dados conhecida como Mineração de Dados.

Embora os dados representem uma fonte de conhecimentos valiosos para a gestão de qualquer negócio, ainda não são totalmente explorados por diversas organizações. A Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM) é um exemplo disso. A UFVJM possuía como principal forma de ingresso o processo seletivo unificado (vestibular tradicional). No momento da inscrição cada candidato deveria responder, via internet, um questionário socioeconômico, criando-se, assim, bases de dados sobre esses candidatos. Porém, tais bases nunca foram submetidas a qualquer tipo de análise, o que pode fazer com que os gestores da UFVJM deixem de conhecer aspectos importantes do contexto social, econômico e cultural dos seus futuros alunos.

Este artigo está estruturado em 8 seções. Na segunda e terceira seções serão apresentados, respectivamente, o objetivo e as razões para o desenvolvimento deste trabalho. Na seção 4 serão demonstrados os trabalhos relacionados. Um estudo sobre MD e Regras de associação será feito na Seção 5. A seção 6 apresentará a metodologia adotada. Os resultados serão mostrados na seção 7 e as conclusões serão apresentadas na seção 8.

1. Objetivo

Este trabalho teve como objetivo aplicar uma técnica de Mineração de Dados, conhecida como *Apriori*, nas bases de dados dos vestibulares unificados da UFVJM no período de 2002 a 2009, de forma a obter conhecimentos para subsidiar o processo de tomada de decisão da instituição em nível da gestão acadêmica. Por exemplo, analisando as notas dos vestibulandos podem ser identificadas deficiências dos alunos de um determinado curso ou área; e isto pode ser levado à coordenação dos cursos para que sejam discutidas ações no sentido de minimizar o problema, como adaptação de ementas de disciplinas, programas especiais de monitoria, entre outros. Tal conhecimento foi baseado em informações acerca dos candidatos, a saber: sua origem, tipo da escola em que cursou o ensino médio, curso escolhido e desempenho no vestibular (notas) em função do tipo da escola.

2. Justificativa

A cada dia, mais organizações vêm coletando e armazenando grandes quantidades de dados resultantes de suas operações diárias. Tais dados são de fundamental importância, pois representam a matéria-prima do processo decisório, o que justifica a necessidade de se utilizar técnicas e ferramentas para analisá-los e estudá-los. Apesar disso, a maior parte das organizações ainda não usa esse grande volume de dados para convertê-los em conhecimentos que possam ser utilizados em suas próprias atividades. Nesse contexto se insere a UFVJM que, embora possua bases com quantidades significativas de dados, nunca as submeteram a uma análise através da Mineração de Dados. Tal situação é que motivou o desenvolvimento deste trabalho.

3. Trabalhos relacionados

São inúmeros os casos de aplicação da Mineração de Dados para apoio à gestão de negócios. No trabalho de Oliveira e Garcia (2004) foram utilizados a tarefa de Regras de Associação e o algoritmo *Apriori* para realizar uma análise

sobre os dados relacionados ao questionário sócio-econômico-cultural da Universidade de Formiga (UNIFOR), aplicado durante processo seletivo da referida instituição. Os conhecimentos obtidos permitiram traçar o perfil dos candidatos e melhorar a qualidade das informações divulgadas, atingindo um público ainda maior. Lemos, Steiner e Nievola (2005), realizaram uma avaliação da base de dados de uma agência do Banco do Brasil através das técnicas de redes neurais e árvores de decisão. Os resultados obtidos facilitaram muito o processo de análise de crédito da agência, permitindo uma maior agilidade na classificação de novas empresas como prováveis adimplentes ou inadimplentes. Por fim, munidos dos dados da Secretaria de Saúde do Paraná e utilizando a tarefa de classificação, Vianna *et al* (2010) desenvolveram um projeto que tornou possível traçar o perfil da mortalidade infantil no estado em função de aspectos como sexo, idade da mãe, tipo de parto, número de consultas pré-natais, entre outros, contribuindo para uma redução significativa no número de óbitos.

4. Mineração de dados e regras de associação

Segundo Han e Kamber (2005), o termo Mineração de Dados pode ser definido como sendo o processo de descoberta de conhecimento interessante a partir de grandes quantidades de dados armazenados em bases de dados, *data warehouses* (armazém ou depósito de dados) e outros repositórios. Para revelar tal conhecimento, a MD combina um conjunto de tarefas e técnicas. De acordo com Amo (2004) e Gonçalves (2010), a tarefa especifica o tipo de conhecimento a ser minerado da base de dados, sendo as regras a forma mais comum de representação deste conhecimento. Cada tarefa possui um conjunto de técnicas que correspondem aos métodos (algoritmos) que serão empregados. A tarefa de Regras de Associação e a técnica *Apriori* são o objeto de estudo deste trabalho.

A regra de associação é uma expressão na forma X→Y (lê-se "X implica em Y" ou "se X então Y"), onde X é denominado antecedente e Y consequente da regra e ambos representam conjuntos de itens distintos. O objetivo dessa tarefa é encontrar conjuntos de itens que ocorram simultaneamente em transações de uma base de dados e formar regras a partir desses conjuntos. As regras de associação buscam por relações do tipo: “Um cliente que compra o item A frequentemente compra também o item B” (SCHONHORST, 2010, p.11).

As regras de associação são avaliadas com base em medidas que mensuram o quanto essas regras são relevantes para o usuário. Tais medidas, denominadas medidas de interesse objetivas, incluem o suporte, a confiança e o *lift* (GONÇALVES, 2005, p. 5). Para exemplificar e conceituar estas três medidas considere a Tabela 1:

Tabela 1 – Transações de compra

N° da transação	Produtos comprados
1	biscoito, cerveja, chá, salaminho
2	cerveja, pão, queijo
3	café, pão, queijo
4	chá, café, pão, queijo
5	café, cerveja, pão, salaminho
6	café, pão, refrigerante

Fonte: Extraído de Gonçalves (2010)

Uma possível regra de associação baseada nestas transações seria: "quem compra café, também compra pão" ou {café}→{pão}. Nesse caso, café é o antecedente e pão o consequente da regra. O suporte de uma regra de associação representa a porcentagem de transações da base de dados que contém o antecedente e o consequente simultaneamente em relação ao número total de transações (GONÇALVES, 2010, p. 24). No exemplo em questão, o suporte da regra é de 66% (4/6). A confiança determina em quantas transações das quais o antecedente aparece, o consequente também ocorre. A confiança é calculada como sendo o suporte da regra dividido pelo suporte do antecedente (GONÇALVES, 2010, p. 24). Nesse caso, a confiança é de 100% (4/4), pois para as transações envolvendo compra de café (4), em 100% delas (4) também aparece o produto pão.

O *lift* indica o quanto mais frequente torna-se o consequente quando o antecedente ocorre. Calcula-se o *lift* dividindo-se a confiança da regra pelo suporte do consequente, ou seja, (4/4)/(4/5). O valor de *lift* obtido foi maior que 1 (1,2), o que revela que a compra de café contribuiu positivamente (elevou a chance de compra) para a compra de pão. Caso este valor fosse menor que 1, dir-se-ia que a compra de café contribuiu negativamente (diminuiu a chance de compra) para a compra de pão e, caso viesse a ser igual a 1, a compra de café não teria influência

sobre a compra de pão. Quanto maiores forem os valores do suporte, da confiança e do *lift* mais interessante se torna a regra (GONÇALVES, 2005, p. 3-6).

A intuitividade e a simplicidade de representação são as características marcantes das regras de associação. Dentre os seus diferentes tipos de algoritmos, o *Apriori* tornou-se o mais conhecido e utilizado, pois permitiu reduzir significativamente o número de combinações de itens a serem realizadas, bem como o cálculo de suporte de cada combinação (GONÇALVES, 2010, p.29). Tais fatores justificam a escolha de ambos para uso em processos de MD.

5. Metodologia

As bases de dados utilizadas neste trabalho foram cedidas pelo órgão regulador dos processos seletivos unificados da UFVJM, a COPESE (Comissão Permanente de Processos Seletivos). Para que cada uma delas fosse submetida à aplicação da técnica de Mineração de Dados, foi necessário realizar um processo de preparação das bases de dados que consistiu, basicamente, nos seguintes passos:

1. Primeiramente, foram selecionadas as tabelas cujo conteúdo atenderia o objetivo proposto inicialmente. A primeira tabela selecionada armazenava os códigos referentes às respostas fornecidas no questionário socioeconômico; a segunda registrava os dados pessoais e outros referentes à inscrição do candidato, como curso escolhido, tipo de deficiência, se recebeu isenção, entre outros; a terceira armazenava as notas de todos os candidatos e a quarta somente dos candidatos que foram aprovados no processo seletivo unificado.

2. Como a tabela que armazenava as informações socioeconômicas continha apenas o código referente às respostas, foi criada uma tabela para cada pergunta do questionário, a qual passou a armazenar as respostas para uma determinada pergunta de forma descritiva, facilitando a posterior análise dos dados. Um procedimento similar a este foi realizado para a tabela de notas com a adoção de conceitos descritivos ao invés de valores numéricos de acordo com o aproveitamento obtido em cada disciplina, sendo: péssimo (de 0 a 39% de acertos), regular (de 40 a 59% de acertos), bom (de 60 a 69% de acertos), muito bom (de 70 a 89% de acertos) e ótimo (de 90 a 100% de acertos). A Tabela a seguir exemplifica como foi associada a pergunta "Em que local você reside?" aos

seus respectivos dados:

Tabela 2 – Exemplo de associação das respostas do questionário para a pergunta "Em que local você reside?"

Código	Resposta
1	Belo Horizonte
2	Diamantina
3	Teófilo Otoni
4	Na região do Vale do
5	Na região do Vale do Mucuri
6	Outra região de Minas Gerais
7	Outros estados

3. Feito isso, todas as tabelas foram interligadas conforme mostra o esquema da figura seguinte:

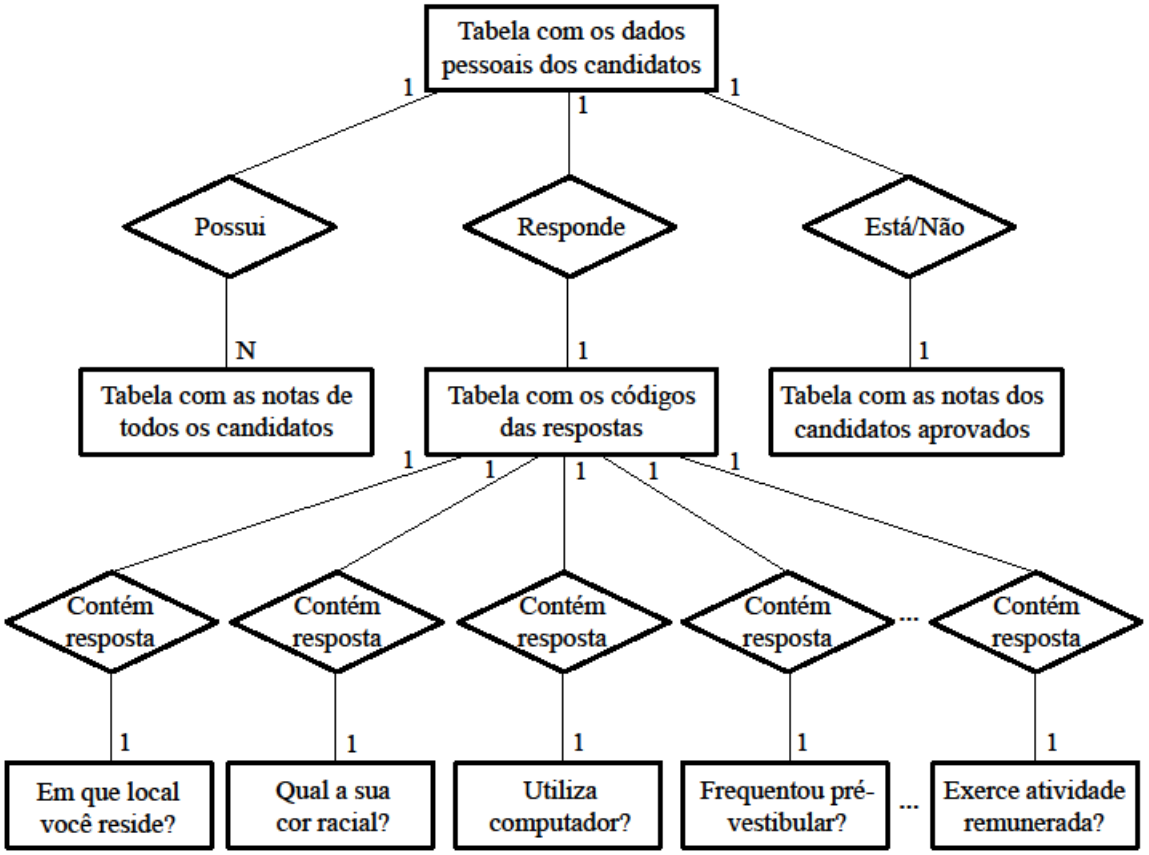


Figura 1 – Esquema da relação entre as tabelas utilizadas para Mineração de Dados

Após a preparação, a técnica de Mineração de Dados foi aplicada sobre cada base de dados. Para a realização dos experimentos computacionais foi utilizada a ferramenta de Mineração de Dados WEKA (http://www.cs.waikato.ac.nz/ml/weka).

O WEKA (*Waikato Environment for Knowledge Analysis*) é um *software* livre e gratuito desenvolvido pela Universidade de Waikato na Nova Zelândia. Inicialmente, o WEKA foi voltado para a investigação de técnicas de aprendizagem de máquina, enquanto sua aplicação inicial foi direcionada para a agricultura, uma área de destaque na economia do país. O WEKA conta com uma série de algoritmos para MD e vem sendo amplamente utilizado em trabalhos de aplicação de técnicas de MD (SILVA, 2004, p.14).

Para iniciar os experimentos, o WEKA foi conectado ao sistema gerenciador de banco de dados MySQL (http://www.mysql.com) e vários testes de geração de regras de associação foram realizados. Para cada teste executado, a ferramenta gerou as regras com a seguinte estrutura: "antecedente → consequente <valor da confiança> <valor do *lift*>".

Após as execuções do algoritmo para cada combinação de atributos, os resultados foram analisados em busca de regras que poderiam proporcionar algum conhecimento útil a respeito do perfil socioeconômico dos candidatos. Tais resultados serão mostrados a seguir.

6. Resultados

Em razão do alto número de regras obtidas, somente aquelas julgadas mais relevantes foram selecionadas. Como critério de seleção, somente foram coletadas as regras que obtiveram, simultaneamente, suporte mínimo de 1%, confiança mínima de 50% e *lift* mínimo de 1.01. Os valores dessas medidas foram ajustados empiricamente. Nas tabelas seguintes serão mostrados os resultados para alguns semestres de diferentes anos de realização do vestibular, o que se justifica em razão do extenso número de regras geradas em cada análise. No geral, a situação se manteve similar nos semestres não apresentados.

A Tabela 3 apresenta os resultados da análise dos atributos local de residência e tipo da escola. Conforme mostra a primeira regra de 2005/1 (Diamantina 392 → Escola Estadual 302), dos 392 candidatos residentes em Diamantina, 302 cursaram o ensino médio em Escola Estadual com uma confiança (conf) de 77%. Conforme o valor do *lift*, a chance de um candidato de Diamantina ter cursado o ensino médio em escola estadual é 1.61 vezes maior. O valor de suporte (sup) indica que esse fato ocorreu em 10% das transações da base de dados. De forma geral, essa análise mostra que a maioria dos candidatos

residentes em Diamantina, Teófilo Otoni e demais cidades do Vale do Jequitinhonha e do Vale do Mucuri cursaram o ensino médio em escola estadual, enquanto os candidatos de Belo Horizonte e outras regiões de Minas Gerais o fizeram, essencialmente, na rede particular de ensino.

Tabela 3 – Regras de associação envolvendo os atributos local de residência e tipo da escola

	Consequente da regra (tipo da escola)							
	Escola Estadual (EE)				**Escola Particular (EP)**			
	Descrição da regra	**Sup**	**Conf**	***Lift***	**Descrição da regra**	**Sup**	**Conf**	***Lift***
2005/1	Diamantina 392 → EE 302	0.1	0.77	1.61	Belo Horizonte 596 → EP 339	0.12	0.57	1.21
	Vale do Jequitinhonha 201→ EE 146	0.05	0.73	1.52	Outra Região de MG 1496 → EP 800	0.27	0.53	1.14
2007/2	Vale do Mucuri 129 → EE 111	0.04	0.86	1.46	Outra Região de MG 1114 → EP 612	0.23	0.55	1.53
	Diamantina 443 → EE 376	0.14	0.85	1.45	Belo Horizonte 374 → EP 200	0.07	0.53	1.49
	Teófilo Otoni 379 → EE 309	0.11	0.82	1.39				
	Vale do Jequitinhonha 279→ EE 210	0.08	0.75	1.28				
2008/1	Diamantina 598 → EE 494	0.14	0.83	1.35	Belo Horizonte 431→ EP 237	0.07	0.55	1.33
	Vale do Jequitinhonha 428→ EE 356	0.1	0.83	1.35	Outra região de MG 1218 → EP 665	0.19	0.55	1.61
	Vale do Mucuri 194→ EE 159	0.05	0.82	1.33				
	Teófilo Otoni 591 → EE 463	0.13	0.78	1.28				
2009/2	Diamantina 469 → EE 404	0.18	0.86	1.37	Belo Horizonte 221 → EP 122	0.05	0.55	1.74
	Vale do Mucuri 144 → EE 116	0.05	0.81	1.28	Outra região de MG 742 → EP 389	0.17	0.52	1.65
	Vale do Jequitinhonha 270 → EE 214	0.09	0.79	1.26				
	Teófilo Otoni 350 → EE 273	0.12	0.78	1.24				

Na Tabela 4 são mostrados os resultados da análise dos atributos local de residência e curso. Conforme pode-se perceber, a maior parte dos candidatos aos cursos de Educação Física e Bacharelado em Humanidades são de Diamantina. Já a maior parte dos candidatos aos cursos de Administração, Ciências Econômicas, Ciências Contábeis e Serviço Social são de Teófilo Otoni. Em contrapartida, a maioria dos candidatos aos cursos de agrárias (Agronomia,

Engenharia Florestal e Zootecnia), bem como Odontologia e Farmácia, ambos da área de saúde, são de outra região do estado de Minas Gerais.

Tabela 4 – Regras de associação envolvendo os atributos local de residência e curso escolhido

	Consequente da regra (local de residência)							
	Diamantina (D) e Teófilo Otoni (TO)				**Outra região de Minas Gerais (OR)**			
	Descrição da regra	**Sup**	**Conf**	***Lift***	**Descrição da regra**	**Sup**	**Conf**	***Lift***
2005/1					Engenharia Florestal 212 → OR 139	0.05	0.66	1.18
					Farmácia/Análise Clínica 254 → OR 159	0.05	0.63	1.13
					Agronomia 199 → OR 121	0.04	0.61	1.1
					Odontologia 418 → OR 241	0.08	0.58	1.04
					Zootecnia 196 → OR 111	0.04	0.57	1.02
					Farmácia/Industrial 192 → OR 107	0.04	0.56	1.01
2007/2	Administração 177 → TO 112	0.04	0.63	4.54	Zootecnia 85 → OR 60	0.02	0.71	1.72
	Ciências Econômicas 72 → TO 41	0.02	0.57	4.09	Agronomia 129 → OR 79	0.03	0.61	1.49
	Ciências Contábeis 104 → TO 58	0.02	0.56	4	Odontologia 389 → OR 232	0.09	0.6	1.46
	Serviço Social 190 → TO 100	0.04	0.53	3.78	Engenharia Florestal 189 → OR 108	0.04	0.57	1.39
					Farmácia 378 → OR 191	0.07	0.51	1.23
2008/1	Administração 281 → TO 176	0.05	0.63	3.67	Zootecnia 103 → OR 66	0.02	0.64	1.82
	Ciências Contábeis 138 → TO 82	0.02	0.59	3.48	Agronomia 122 → OR 72	0.02	0.59	1.68
	Serviço Social 288 → TO 167	0.05	0.58	3.4	Odontologia 474 → OR 275	0.08	0.58	1.65
	Ciências Econômicas 114→ TO 64	0.02	0.56	3.29	Engenharia Florestal 220 → OR 117	0.03	0.53	1.51
	Educação Física 124 → D 67	0.02	0.54	3.13				
2009/2	Administração 140 → TO 87	0.04	0.62	4	Zootecnia 61 → OR 36	0.02	0.59	1.79
	Serviço Social 105 → TO 60	0.03	0.57	3.68	Odontologia 283 → OR 157	0.07	0.55	1.68
	Ciências Contábeis 78 → TO 43	0.02	0.55	3.55	Engenharia Florestal 127 → OR 67	0.03	0.53	1.6
	Ciências Econômicas 56 → TO 29	0.01	0.52	3.33				
	Bacharelado em Humanidades 186 → D 121	0.05	0.65	3.13				
	Educação Física 59 →	0.02	0.63	3.01				

	D 37							

A Tabela 5 exibe os resultados envolvendo os atributos tipo da escola e curso escolhido. De acordo com a Tabela, a maior parte dos candidatos aos cursos de Odontologia e Farmácia cursou o ensino médio em escola particular. Já a maioria dos candidatos de Nutrição, Fisioterapia, Serviço Social, Ciências Econômicas, Ciências Contábeis, Educação Física, Administração, Sistemas de Informação, Ciências Biológicas e Bacharelado em Humanidades cursou o ensino médio em escola estadual. Quanto aos cursos da área de agrárias, Engenharia Florestal e Agronomia tiveram maior ocorrência de alunos oriundos de escola estadual. Já Zootecnia apresentou inicialmente maioria de candidatos oriundos de escola estadual, mas a partir de 2007/1 a maioria passou a ser de escola particular.

Tabela 5 – Regras de associação envolvendo os atributos tipo da escola e curso escolhido

	Consequente da regra (tipo da escola)							
	Escola Estadual (EE)				**Escola Particular (EP)**			
	Descrição da regra	**Sup**	**Conf**	***Lift***	**Descrição da regra**	**Sup**	**Conf**	***Lift***
2003/1	Zootecnia 232 → EE	0.04	0.51	1.11	Farmácia 224 →	0.04	0.54	1.28
	Engenharia Florestal 191 → EE 97	0.03	0.51	1.11	Odontologia 475 → EP251	0.08	0.53	1.24
	Nutrição 364 → EE 184	0.06	0.51	1.1				
2005/2	Nutrição 323 → EE 184	0.08	0.57	1.24	Odontologia 417	0.1	0.59	1.22
	Fisioterapia 348 → EE181	0.08	0.52	1.14	Farmácia 225 → EP 120	0.05	0.53	1.11
2007/1	Serviço Social 163 → EE 135	0.04	0.83	1.58	Odontologia 465 → EP 310	0.09	0.67	1.59
	Ciências Contábeis 82 → EE 62	0.02	0.76	1.44	Zootecnia 118 → EP 67	0.02	0.57	1.36
	Ciências Econômicas 73 → EE 53	0.02	0.73	1.39	Farmácia 349 → EP 181	0.05	0.52	1.24
	Educação Física 145→ EE 100	0.03	0.69	1.32				
	Administração 166 → EE 112	0.03	0.67	1.29				
	Sistemas de Informação 178 → EE 116	0.04	0.65	1.24				
	Ciências Biológicas 199 → EE 108	0.03	0.54	1.04				
	Nutrição 174 → EE 92	0.03	0.53	1.01				
2009/2	Serviço Social 105→ EE	0.04	0.88	1.39	Odontologia 283	0.07	0.59	1.86
	Educação Física 59 →	0.02	0.83	1.32				

	EE 49							
	Ciências Contábeis 78 → EE 64	0.03	0.82	1.3				
	Administração 140 → EE 113	0.05	0.81	1.28				
	Bacharelado em Humanidades 186 → EE 150	0.07	0.81	1.28				
	Ciências Econômicas 56 → EE 45	0.02	0.8	1.28				
	Ciências Biológicas 112 → EE 81	0.04	0.72	1.15				
	Nutrição 83 → EE 60	0.03	0.72	1.15				
	Fisioterapia 87 → EE 61	0.03	0.7	1.11				
	Agronomia 75 → EE 48	0.02	0.64	1.02				

Por fim, analisando a Tabela 6, pode-se verificar que a maior parte dos candidatos oriundos de escola municipal e estadual obteve conceito péssimo na maioria das disciplinas. A ocorrência de conceitos envolvendo escola federal foi pequena, mas com predominância do conceito péssimo. Por representar um número pequeno, as regras correspondentes a esse tipo de escola não foram incluídas na tabela referente a esta análise. O conceito regular, embora menos frequente, está associado tanto às escolas estaduais quanto particulares. Já os conceitos bom, muito bom e ótimo estão sempre associados a alunos de escola particular.

Tabela 6 – Regras de associação envolvendo os atributos tipo da escola e notas

	Regras para Escola Estadual (EE) e				**Regras para Escola Particular (EP)**			
	Descrição da regra	**Sup**	**Conf**	***Lift***	**Descrição da regra**	**Sup**	**Conf**	***Lift***
2003/1	EE 1393 → Biologia=Péssimo 823	0.27	0.59	1.09	Biologia=Muito bom 272 → EP 141	0.05	0.52	1.22
	EE 1393 → Matemática=Péssimo 769	0.25	0.55	1.12	Matemática=Ótimo 148 → EP 78	0.03	0.53	1.24
	EE 1393 → Física=Péssimo 1130	0.37	0.81	1.07	Química=Ótimo 245 → EP 135	0.04	0.55	1.3
	EE 1393 → História=Péssimo 848	0.28	0.61	1.1	Física=Muito bom 85→ EP 52	0.02	0.61	1.44
	EE 1393 → Língua=Péssimo 759	0.25	0.54	1.19	Geografia=Ótimo 153 → EP 82	0.03	0.54	1.26
	Química=Regular 1253 → EE 627	0.21	0.5	1.09	História=Ótimo 156 → EP 82	0.03	0.53	1.24
	Português=Regular 1065 → EE 528	0.17	0.5	1.08	Língua=Muito bom 526 → EP 301	0.1	0.57	1.35

	EM 76 → Matemática=Péssimo 41	0.01	0.54	1.09	Português=Ótimo 239 → EP 125	0.04	0.52	1.23
	EM 76 → Física=Péssimo 59	0.02	0.78	1.02				
2005/2	Biologia=Péssimo 814 → EE 457	0.19	0.56	1.23	Biologia=Ótimo 107 → EP 72	0.03	0.67	1.4
	EE 990 → Matemática=Péssimo 761	0.32	0.77	1.12	Matemática=Muito bom 136 → EP 101	0.04	0.74	1.55
	EE 990 → Química=Péssimo 597	0.25	0.6	1.17	Matemática=Regular 340 → EP 177	0.07	0.52	1.09
	EE 990 → Física=Péssimo 898	0.38	0.91	1.02	Química=Muito bom 213 → EP 140	0.06	0.66	1.37
	EE 990 → Geografia=Péssimo 540	0.23	0.53	1.15	Química=Regular 498 → EP 257	0.11	0.52	1.08
	História=Péssimo 834 → EE 449	0.19	0.54	1.18	Física=Bom 60 → EP 38	0.02	0.63	1.32
	Língua=Péssimo 762 → EE 416	0.17	0.55	1.19	Física=Regular 161 → EP 88	0.04	0.55	1.14
	EM 46 → Matemática=Péssimo 35	0.01	0.76	1.11	Geografia=Muito bom 265 → EP 162	0.07	0.61	1.28
	EM 46 → Química=Péssimo 29	0.01	0.63	1.22	História=Muito bom 373 → EP 225	0.04	0.52	1.21
	EM 46 → Física=Péssimo 43	0.02	0.93	1.05	Língua=Ótimo 86 → EP 68	0.03	0.79	1.65
2007/1	EE 1730 → Biologia=Péssimo 957	0.29	0.55	1.19	Biologia=Ó timo 160 → EP 100	0.03	0.63	1.49
	EE 1730 → Matemática=Péssimo 1120	0.34	0.65	1.1	Matemática=Muito bom 242 → EP 145	0.04	0.6	1.43
	EE 1730 → Química=Péssimo 1246	0.38	0.72	1.11	Química=Muito bom 243 → EP 143	0.04	0.59	1.41
	EE 1730 → Física=Péssimo 1374	0.42	0.79	1.06	Física=Muito bom 130 → EP 82	0.02	0.63	1.51
	Língua=Regular 1258 → EE 684	0.21	0.54	1.04	Geografia=Muito bom 789 → EP 448	1.14	0.57	1.36
	Português=Regular 1732 → EE 919	0.28	0.53	1.01	Língua=Ótimo 98 → EP 69	0.02	0.7	1.68
	EM 74 → Química=Péssimo 50	0.02	0.68	1.04	Português=Muito bom 222 → EP 119	0.04	0.54	1.28
	EM 74 → Física=Péssimo 59	0.02	0.8	1.06				
	EM 74 → História=Péssimo 38	0.01	0.51	1.22				
2009/2	EE 1419 → Biologia=Péssimo 821	0.36	0.58	1.111	Biologia=Muito bom 246 → EP 130	0.06	0.53	1.67
	EE 1419 → Matemática=Péssimo 1016	0.45	0.72	1.07	Matemática=Muito bom 111 → EP 72	0.03	0.65	2.04

	EE 1419 → Física=Péssimo 927	0.41	0.65	1.08	Física=Muito bom 124 → EP 68	0.03	0.55	1.73
	EE 1419 → Geografia=Péssimo 741	0.33	0.52	1.13	Geografia=Muito bom 313 → EP 159	0.07	0.51	1.6
	EE 1419 → História=Péssimo 783	0.35	0.55	1.04	Português=Muito bom 57 → EP 31	0.01	0.54	1.71
	Química=Regular 811 → EE 520	0.23	0.64	1.02				
	Língua=Regular 785 → EE 504	0.22	0.64	1.02				
	Português=Regular 1041 → EE 709	0.31	0.68	1.08				
	EM 53 → Matemática=Péssimo 39	0.02	0.74	1.1				
	EM 53 → Química=Péssimo 30	0.01	0.57	1.21				
	EM 53 → Física=Péssimo 34	0.02	0.64	1.06				
	EM 53 → História=Péssimo 30	0.01	0.57	1.06				

7. Conclusões

Após a análise dos resultados obtidos com a aplicação da Mineração de Dados nas bases de dados do processo seletivo unificado da UFVJM, foi possível concluir que há uma diferença no perfil dos candidatos quanto à localização e o tipo de escola onde estudaram: candidatos de Diamantina, Teófilo Otoni e Vales do Jequitinhonha e Mucuri concluíram o ensino médio em escola estadual. Já os candidatos de Belo Horizonte e outras regiões de Minas Gerais concluíram o ensino médio em escola particular.

Há também uma diferença no perfil dos candidatos quanto à localização e alguns cursos escolhidos: Bacharelado em Humanidades e Educação Física são, em sua essência, formados por candidatos de Diamantina. Os cursos de Administração, Ciências Econômicas, Ciências Contábeis e Serviço Social possuem maioria de candidatos de Teófilo Otoni. Já os cursos de Agronomia, Engenharia Florestal, Zootecnia, Odontologia e Farmácia possuem maioria de candidatos provenientes de outras regiões de Minas Gerais.

Consequentemente existe uma diferença no perfil dos candidatos quanto ao tipo de escola e os cursos escolhidos: Farmácia e Odontologia têm maioria de candidatos de escola particular. Nutrição, Fisioterapia, Serviço Social, Ciências

Econômicas, Ciências Contábeis, Educação Física, Administração, Sistemas de Informação, Ciências Biológicas e Bacharelado em Humanidades possuem maioria de candidatos de escola estadual.

O desempenho em geral é ruim. Os bons desempenhos (bom, muito bom e ótimo) estão sempre associados aos candidatos oriundos de escola particular. Para as demais escolas, têm-se maioria de candidatos com conceito péssimo, salvo os poucos casos de ocorrência de conceito regular em escola estadual.

Os resultados obtidos podem auxiliar setores da universidade. Por exemplo, identificadas as deficiências dos alunos com relação a desempenho na maioria das disciplinas, ações poderiam ser tomadas em conjunto com a comunidade regional para amenizar tal problema, na forma de cursos preparatórios para o ENEM (atualmente única forma de ingresso na universidade) ou mesmo programas de nivelamento para os calouros.

Os resultados envolvendo os cursos da área de saúde e agrárias atestam que é preciso que as autoridades educacionais dos Vales do Jequitinhonha e Mucuri se reúnam e verifiquem a possibilidade de se realizar ações que aumentem as chances de acesso dos estudantes dos Vales aos cursos das áreas de saúde e agrárias, já que estes são compostos, em sua essência, por candidatos de outras regiões de Minas Gerais. Os resultados são úteis ainda para que as escolas procurem identificar quais os fatores que levam os alunos (em sua maioria) a não se mostrarem preparados para realizar o vestibular da UFVJM com sucesso.

Trabalhos futuros interessantes podem surgir em virtude da adoção, por parte do governo, de outras formas de ingresso na universidade, como o sistema de cotas; no sentido de analisar tal sistema de ingresso no que se refere ao seu caráter de instrumento de inclusão social de parcelas menos favorecidas da população no ambiente universitário.

Referências

AMO, S. de. **Técnicas de Mineração de Dados.** Disponível em: <http://www.deamo.prf.ufu.br/arquivos/JAI-cap5.pdf > Acesso em: 26/10/2011.

GONÇALVES, E. C. **Mineração de Dados (Data Mining)**. 2010.

GONÇALVES, E. C. **Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas**. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos/v4.1/art04.pdf> Acesso em: 18/10/2011.

HAN, J.; KAMBER, M. **Data Mining:** Concepts and Techniques. 2. ed. 2005.

LEMOS, E. P.; STEINER, M.T. A.; NIEVOLA, J. C. **Análise de crédito bancário por meio de redes neurais e árvores de decisão**: uma aplicação simples de *data mining*. Disponível em: <http://www.rausp.usp.br/download.asp?file=V4003225.pdf> Acesso em: 11/09/2011.

OLIVEIRA, A. G. de; GARCIA, D. F. **Mineração da Base de Dados de um Processo Seletivo Universitário**. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos/v3.2/art07.pdf> Acesso em: 27/08/2011.

SCHONHORST, G. B. **Mineração de Regras de Associação Aplicada à Modelagem dos Dados Transacionais de um Supermercado**. Dissertação (Pós-Graduação em Engenharia de Produção) - Universidade Federal de Itajubá, Itajubá, 2010.

SILVA, M. P. dos S. **Conceitos, Aplicações e Experimentos com Weka**. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf> Acesso em: 13/10/2011.

VIANNA, R. C. X. F. *et al*. **Mineração de dados e características da mortalidade infantil**. Disponível em: <http://www.scielosp.org/pdf/csp/v26n3/11.pdf> Acesso em: 12/09/2011.