INSTITUTO FEDERAL DE MINAS GERAIS
CAMPUS SÃO JOÃO EVANGELISTA
FERNANDO ELIAS DE OLIVEIRA

BUSCA DE CONHECIMENTO SOBRE O PROCESSO SELETIVO DO INSTITUTO FEDERAL DE MINAS GERAIS – CAMPUS SÃO JOÃO EVANGELISTA: a mineração de dados como tecnologia para desvelar tendências e padrões

São João Evangelista
2013

FERNANDO ELIAS DE OLIVEIRA

BUSCA DE CONHECIMENTO SOBRE O PROCESSO SELETIVO DO INSTITUTO FEDERAL DE MINAS GERAIS – CAMPUS SÃO JOÃO EVANGELISTA: a mineração de dados como tecnologia para desvelar tendências e padrões

Monografia apresentada ao Curso de Sistemas de Informação do Instituto Federal de Minas Gerais, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Ma. Karina Dutra de Carvalho Lemos
Coorientador: Me. José Fernandes da Silva

São João Evangelista

2013

FICHA CATALOGRÁFICA
Elaborada pelo Serviço Técnico da Biblioteca do
Instituto Federal Minas Gerais – Campus São João Evangelista

O48b OLIVEIRA, Fernando Elias, 1989-

Busca de conhecimento sobre o processo seletivo do Instituto Federal de Minas Gerais – Campus São João Evangelista: a mineração de dados como tecnologia para desvelar tendências e Padrões./ Fernando Elias de Oliveira. São João Evangelista, MG: IFMG - Campus São João Evangelista, 2013.
90 p.: il.

Trabalho de Conclusão de Curso - TCC (graduação) apresentado ao Instituto Federal Minas Gerais – Campus São João Evangelista – IFMG, Curso de Bacharelado em Sistemas de Informação, 2013.
Orientadora: Profa. Ma. Karina Dutra de Carvalho Lemos
Coorientador: Prof. Me. José Fernandes da Silva

1. Educação. 2. Informática. 3. Conhecimento. I. Instituto Federal Minas Gerais – Campus São João Evangelista. Curso de Bacharelado em Sistemas de Informação. II. Título.

CDD 371.3078

Fernando Elias de Oliveira

BUSCA DE CONHECIMENTO SOBRE O PROCESSO SELETIVO DO INSTITUTO FEDERAL DE MINAS GERAIS – CAMPUS SÃO JOÃO EVANGELISTA:
a mineração de dados como tecnologia para desvelar tendências e padrões

Monografia apresentada ao Curso de Sistemas de Informação do Instituto Federal de Minas Gerais – Campus São João Evangelista, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Ma. Karina Dutra de Carvalho Lemos (Orientadora) - IFMG

Me. José Fernandes da Silva (Coorientador) - IFMG

Esp. Bruno de Souza Toledo - IFMG

São João Evangelista, 05 de novembro de 2013.

Dedico este trabalho à minha família, pela dedicação, doação, carinho, segurança e amor que sempre destinaram a mim.

AGRADECIMENTOS

A Deus, pela vida.

A todos que contribuíram para realização deste trabalho e que auxiliaram durante o curso, especialmente:

À minha família, pela força e companhia.

Ao professor José Fernandes da Silva, pela auxilio nas enfermidades, aprendizado e orientações nos momentos necessários.

À professora Karina Dutra de Carvalho Lemos, pelo aprendizado, orientação e apoio neste trabalho.

Aos colegas, pelo companheirismo.

RESUMO

Diante da globalização e do avanço tecnológico, a informação tem desempenhando um papel primordial no desenvolvimento e sucesso das organizações. Nesse sentido, a constante produção de dados nestas organizações faz com que aumentem consideravelmente seu volume de dados, dos quais é possível retirar informações que agreguem valor e vantagens competitivas. Assim sendo, o processo de Descoberta de Conhecimento em Banco de Dados (DCBD) visa aplicar a Mineração de Dados, sua principal etapa, para analisar e extrair conhecimento através de padrões consistentes e relacionamentos sistemáticos entre as instâncias de dados. O objetivo deste estudo é buscar informações implícitas na base de dados dos candidatos inscritos nos processos seletivos para os anos letivos de 2011, 2012 e 2013 do Instituto Federal de Minas Gerais – Campus São João Evangelista, através da aplicação do processo de DCBD. Neste trabalho, utilizando-se de uma ferramenta chamada *Waikato Enviroment for Knowledge Analysis* (WEKA), aplicaram-se as técnicas de Regra de Associação, Árvore de Decisão e Agrupamento. Os quesitos que nortearam o desenvolvimento deste trabalho foram a verificação da existência de relação entre o desempenho e a escolha do candidato no processo seletivo com a renda familiar, acesso à informação e localização do candidato. A metodologia para a realização deste estudo descritivo de caráter qualitativo e de natureza aplicada, no qual se utilizou do método de pesquisa hipotético-dedutivo, objetivou-se a aplicação da técnica de mineração de dados através do processo de DCBD. Para a execução da pesquisa utilizou-se os procedimentos bibliográfico, documental, experimental e estudo de caso. A amostra estudada consistiu nos 1985 candidatos inscritos nos três processos seletivos para os cursos técnicos. Os resultados obtidos confirmaram que a escolha por um curso e o desempenho de um candidato está intimamente ligada à renda familiar e à sua localização. O processo de DCBD mostrou-se eficaz para a transformação de dados em conhecimento, disponibilizando informações relevantes, motivando a continuidade da pesquisa e outros trabalhos futuros.

Palavras-chave: Descoberta de Conhecimento. Mineração de Dados. Processo Seletivo.

ABSTRACT

In front of globalization and technological advance, the information has played a primordial role in the development and success of organizations. In this sense, the constant production of data in these organizations makes considerably increase their volume of data, which is possible to obtain information that add value and competitive advantages. Thus, the process of Knowledge Discovery in Database (KDD) aims to apply Data Mining, its main step, to analyze and extract knowledge through consistent patterns and systematic relationships between the data instances. The aim of this study is to find implicit informations in the database of registered candidates in selection processes for the academic years 2011, 2012 and 2013 of the Institute Federal of Minas Gerais - Campus St. John the Evangelist, by applying the KDD process. In this work, using a tool called Waikato Enviroment for Knowledge Analysis (WEKA), we applied the techniques of Rule Association, Decision Tree and Clustering. The questions that guided the development of this work were to verify the existence of relationship between performance and the choice of the candidate selection process by family income, access to information and location of the candidate. The methodology used for this descriptive study of character qualitative and applied nature, in which we used the method of hypothetical-deductive research, aimed to apply the technique of data mining through the process of KDD. For the execution the research used literature procedures, documentary, experimental and case study. The sample consisted of 1985 registered candidates in the three selective processes for technical courses. The results obtained confirmed that the choice of a course and the performance of a candidate is closely linked to family income and its location. The KDD process was effective for the transformation of data into knowledge, providing relevant information, motivating the continuity of research and other future work.

Keywords: Knowledge Discovery. Data Mining. Selection Process.

SUMÁRIO

1. INTRODUÇÃO

Este trabalho tem como tema a busca de conhecimento sobre o processo seletivo do Instituto Federal de Minas Gerais – Campus São João Evangelista, onde se aplica a mineração de dados como tecnologia para desvelar tendências e padrões. Esta pesquisa é uma extensão do projeto "Ilação sobre os perfis de candidatos no processo seletivo do Instituto Federal de Minas Gerais – Campus São João Evangelista" do Programa Institucional de Bolsa de Iniciação Científica (PIBIC), financiado pelo Instituto Federal de Minas Gerais (IFMG), coordenado pelo mestre José Fernandes da Silva e com a participação do bolsista Fernando Elias de Oliveira, estudante do curso de Bacharel em Sistemas de Informação.

As organizações trabalham constantemente com um grande volume de dados provenientes de seus processos. A maioria delas utilizam sistemas informatizados para automatizar as atividades e registrar informações de suas operações. Percebe-se que a velocidade de coleta de informações é maior do que a velocidade de processamento ou análise dessas, o que gera um problema e, simultaneamente, uma contradição. Por possuírem uma grande quantidade de dados, as organizações possuem uma falsa sensação de que estão bem informadas, porém, essas informações de nada servem se não forem analisadas de forma correta e em tempo hábil. De acordo com Carvalho (2005), são poucas as organizações que utilizam essas bases de dados para transformá-las em conhecimento útil para a sua gestão.

Diante desta situação, baseado em Tang, Steinbach, e Kumar (2009), surgiu a área de descoberta de conhecimento, que está sendo amplamente difundida e aperfeiçoada, pelo fato de fornecer resultados promissores a todas as áreas onde pode ser aplicada, despertando interesse, principalmente no meio comercial, científico e acadêmico. As ferramentas de recuperação de informações tradicionais como o SQL não são viáveis para análises de dados ou busca de informações implícitas, desta forma, surgiu, então, a mineração de dados. A mineração de dados é constituída por diversas técnicas primordiais para a aquisição de conhecimento e descoberta de tendências, baseadas em análise de busca de padrões, com o objetivo de descoberta de conhecimento em banco de dados. A ferramenta gratuita de maior destaque para este fim no meio acadêmico e científico é o *software Waikato Environment for Knowledge Analysis* (WEKA). Para maior eficiência na aplicação da mineração de dados, é necessário seguir algum processo de Descoberta de Conhecimento em Banco de Dados (DCBD). Alguns autores abordam a mineração de dados como a principal etapa do processo de DCBD, pois algumas delas devem ser executadas antes da mineração de dados. Outros entendem a mineração de dados como o próprio processo, pois desconsideram as etapas a

serem seguidas anteriormente. No caso dessa pesquisa, é proposta a mineração de dados como uma etapa do processo como alguns autores sugerem.

Frente a essa realidade, as instituições de ensino trabalham com um grande volume de dados, sejam esses correspondentes aos funcionários, aos estudantes ou à própria instituição. Estes dados podem conter algum conhecimento que a instituição não saiba e que seja essencial para auxiliá-la em tomadas de decisões. Nesse sentido, o IFMG possui um amplo conjunto de dados que pode ser utilizado para adquirir conhecimentos que possam auxiliar na sua gestão. Uma forma de obter esse conhecimento é a partir de informações relacionadas aos processos seletivos, até porque os estudantes constituem o seu principal público alvo. Assim, acredita-se que as bases de dados dos processos seletivos podem inferir conhecimentos que propiciem o aperfeiçoamento ou a criação de políticas eficientes, que atenderão estudantes ou candidatos nos processos seletivos da instituição. O processo seletivo realizado pela Comissão Permanente de Vestibular e Exame de Seleção (COPEVES) do IFMG é o processo de seleção ao qual candidatos aos cursos técnicos devem submeter-se para ingressar na instituição.

Diante disso, pergunta-se: com a aplicação da mineração de dados é possível extrair algum padrão na base de dados do processo seletivo para os cursos técnicos do IFMG – Campus São João Evangelista? A partir disto, definiram-se hipóteses que norteariam a aplicação da mineração de dados, pois se acreditava que era possível encontrar padrões que representariam algum comportamento ou fenômeno no processo seletivo do IFMG – Campus São João Evangelista. Inicialmente, quanto à situação financeira, pensou-se que ela era um fator determinante na escolha dos cursos e aprovação dos candidatos. Também, confiava-se que o meio de comunicação que mais tinha relevância com os candidatos aprovados eram os panfletos. E, por fim, esperava-se confirmar que os candidatos que leem mais livros seriam aprovados e que os motivos de aprovações para cada curso poderia conter padrões diferentes.

Portanto, o objetivo deste estudo é buscar informações implícitas na base de dados dos candidatos inscritos nos processos seletivos para os anos letivos de 2011, 2012 e 2013 do IFMG – Campus São João Evangelista, através da aplicação do processo Descoberta de Conhecimento em Banco de Dados, utilizando a ferramenta de mineração de dados WEKA. Para cumprir este objetivo, torna-se necessário estudar técnicas de mineração de dados para descobrir padrões de comportamento dos estudantes que se inscrevem no processo seletivo, identificar características dos candidatos decorrentes da aplicação das técnicas de mineração de dados e perceber a viabilidade da aplicação da mineração de dados nos registros dos processos seletivos.

Assim, justifica-se a aplicação do processo de DCBD no IFMG – Campus São João Evangelista pelo fato de ele possuir uma grande base de dados correspondentes à vida social, cultural e financeira de todos os candidatos que se inscrevem no processo seletivo. A base de dados, resultantes das respostas aos questionários socioeconômicos do processo seletivo, pode, dessa forma, constituir uma fonte rica de informações sobre os estudantes da instituição, e auxiliar nas decisões para a criação de programas para assistência estudantil e *marketing*. O IFMG possui diversos programas de auxílio para estudantes de baixa renda e a COPEVES (Comissão Permanente de Vestibular e Exame de Seleção) necessita de informações dos processos seletivos anteriores para aperfeiçoar os métodos de comunicação para as próximas seleções, possibilitando, inclusive, a COPEVES utilizar algum padrão como um suporte para a estratégia de comunicação nos processos seletivos posteriores.

Assim sendo, de forma generalizada, a caracterização metodológica para a realização deste estudo resume-se em um estudo descritivo de caráter qualitativo e de natureza aplicada, no qual se utilizou do método de pesquisa hipotético-dedutivo e para a execução da pesquisa utilizou-se os procedimentos bibliográfico, documental, experimental e estudo de caso. Para aplicação da técnica de mineração de dados, seguiu-se o processo de DCBD, no qual a população estudada seria os candidatos inscritos nos processos seletivos para os cursos técnicos do IFMG – Campus São João Evangelista. Dessa maneira, quanto aos autores, a organização metodológica deste estudo fundamentou-se, principalmente, em Prodanov e Freitas (2013); já na ancoragem teórica e resultados, baseou-se especialmente em Carvalho (2005), Coelho (2007), Tang, Steinbach, e Kumar (2009) e Russel (2011), já que esses são pesquisadores com experiência na área de mineração de dados.

Com os resultados obtidos, conclui-se que é possível encontrar padrões nos processos seletivos do IFMG, que revelam que, cada curso possui um perfil de candidatos, assim como um perfil para aprovação em cada curso. Além disso, entende-se que a intenção com que o candidato procura um curso vai depender da sua localização e situação financeira dos seus pais, podendo existir diferenças entre os candidatos que moram na zona rural ou urbana de São João Evangelista, o que afetam na escolha e na intenção com que procuram os cursos.

Para a apresentação da pesquisa realizada, estruturou-se este trabalho em cinco sessões, além dessa, introdutória, ficando assim divididas:

Na sessão 2 está a fundamentação teórica sobre a busca de conhecimento em bancos de dados, a descrição do processo de DCBD, a definição de mineração de dados e suas técnicas, apresentação do *software* WEKA e aplicações da mineração de dados.

Na sessão 3, será apresentada a revisão bibliográfica da aplicação da mineração de dados em diversas organizações e instituições.

A sessão 4 apresenta a metodologia adotada, bem como os materiais e métodos utilizados. Dessa forma, foi exposta a caracterização da pesquisa, bem como a unidade de análise e observação.

Na sessão 5 são apresentados os resultados da aplicação das técnicas de mineração de dados e na sexta e última sessão são apresentadas as conclusões, a relevância do estudo e possíveis trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Segundo Lakatos e Marconi (2010), a fundamentação teórica é o levantamento das fontes teóricas, com o objetivo de elaborar a contextualização do estudo e seu embasamento teórico, o qual fará parte do referencial da pesquisa na forma de uma revisão da literatura, diante de análises e reflexões, sobre os dados e informações coletadas. Essas providências indicam até que ponto o tema pesquisado já foi estudado e discutido na literatura pertinente. Perante isto, nesta seção apresenta-se a importância da informação para pessoas e organizações, o embasamento teórico sobre os enfoques distintos aos métodos de busca de informações em bancos de dados, as diferentes abordagens para o Processo de Descoberta de Conhecimento em Banco de Dados (DCBD), ou *Knowledge Discoveryin Databases* (KDD), assim como as etapas deste processo, o conceito de mineração de dados, as principais tarefas e técnicas de mineração de dados, a apresentação de *softwares* para mineração de dados e algumas aplicações de mineração de dados.

Com o surgimento da rede mundial de computadores e com a intervenção da globalização na vida social, pessoas e organizações que antes eram separadas pelas distâncias geográficas ou influências étnicas estão cada vez mais próximas virtualmente. Desta forma, um grupo de pessoas ou um mercado específico não está totalmente isolado das influências, tendências e aspectos de outras culturas ou de organizações. Os recursos da tecnologia da informação auxiliam indivíduos para que se comuniquem de forma instantânea e tornam simples a aquisição de diversos tipos de informações. Com esta facilidade de interação entre diferentes tipos de pessoas, diversos tipos de organizações e destas com pessoas de culturas distintas, a informação se tornou essencial para a interligação e compreensão da comunicação e dos fenômenos da globalização. Diante deste contexto, a informação se tornou a matéria prima de extrema importância para todas as organizações, sejam elas uma empresa, instituição, órgão governamental ou um grupo social. Sob o ponto de vista de Russel (2011), a valorização do sucesso e do prestígio pode ser quantificada a partir da detenção do conhecimento, porque quem possui ciência e habilidade para manipulá-lo serão os detentores da condição essencial para o desenvolvimento e para a competitividade. Toda organização competitiva utiliza-se dos sistemas de informação para automatizar seus processos operacionais e estratégicos, principalmente devido ao advento de tecnologias robustas e acessíveis, que aumentam a facilidade de armazenamento de dados e que permitem que as informações sejam mais facilmente localizadas, compreendidas e melhor utilizadas. Os sistemas de informação, portanto, favorecem que estas organizações possuam registros

completos de suas interações; além disso, estes sistemas tendem a utilizar pelo menos um Sistema Gerenciador de Banco de Dados (SGBD), que proporciona à organização controle na entrada, saída e permanência dos dados, gerenciando o armazenamento de informações operacionais.

Segundo Morais (2010), a maioria das organizações possui grande quantidade de dados, mas não faz processamentos e análises desses dados de forma que sejam produzidas informações que possam auxiliar nas suas tomadas de decisões. Para ele, as pessoas possuem comportamentos que podem influenciar na criação de relações ou produtos e no estabelecimento de serviços. Além disso, na interação envolvendo pessoas podem existir informações para melhor entender, compreender e atendê-las. Portanto, as bases de dados dos diversos sistemas que armazenam dados referentes a pessoas, podem contribuir para o sucesso social e financeiro das organizações e da sociedade.

Deste modo, conforme Thomé (2012), nos diferentes segmentos da sociedade cresce a busca por tecnologias que agreguem valor aos negócios, seja para agilizar as operações ou viabilizar inovações. Assim, com o mercado cada vez mais competitivo, já não basta organizar a produção, reduzir os custos e atender bem. É preciso adquirir conhecimento sobre pessoas, acontecimentos, processos e interpretar as suas expectativas e comportamentos através de suas interações e iterações registradas nos sistemas de informação, já que a informação é a principal essência das organizações, por isso faz-se necessária a aplicação de processos que acelerem a extração de informações de grandes bases de dados.

2.1. Busca de informações em bancos de dados

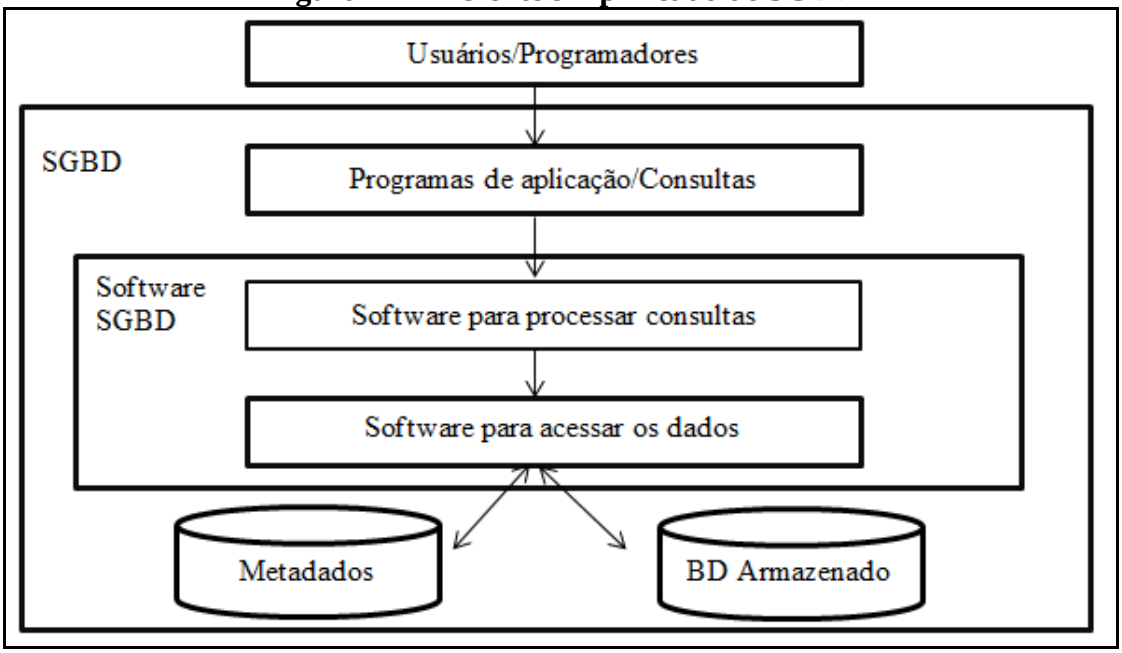
Para Elmasri e Navathe (2011), um SGBD (Sistema Gerenciador de Banco de Dados) disponibiliza uma série de funcionalidades que permite controlar e acompanhar melhor os dados armazenados e ele consiste em uma coleção de dados inter-relacionados e um conjunto de programas para acessá-los. O seu objetivo principal é prover aos usuários uma visão abstrata dos dados, omitindo detalhes de armazenamento físico destes dados, proporcionando um ambiente conveniente e eficiente para definição, armazenamento, recuperação e alteração de dados. Elmasri e Navathe (2011) expõem também que as principais características de um SGBD são:

a) permitir o acesso concorrente às bases de dados;
b) realizar o gerenciamento de transações;

c) permitir criar e aplicar regras de segurança às bases de dados;
d) permitir criar regras que garantam a integridade da base de dados.

Brauner (2003), diz que a facilidade e rapidez de manipulação e gerenciamento dos relacionamentos dos dados em um banco de dados são os principais motivos da difusão dos sistemas gerenciadores de banco de dados, pois, além do próprio banco de dados, um ambiente de um SGBD é composto do *software* para processar e acessar estes dados, juntamente com as consultas e manipulações dos dados realizadas pelos programadores. A Figura 1, abaixo, ilustra um ambiente simplificado de banco de dados.

Figura 1 - Ambiente simplificado de SGBD



Fonte: BRAUNER, 2003, p.6.

Nesse sentido, Elmasri e Navathe (2011) destacam que as tecnologias de armazenamento utilizadas pelos Sistemas Gerenciadores de Bancos de Dados estão presentes em quase todos os lugares onde existem sistemas de informação. Para eles, a maioria das organizações, por menor que sejam, preferem esta tecnologia como uma alternativa mais ágil e segura de armazenarem informações, permitindo ter quantas mais lhes for possíveis. Porém, de acordo com Tang, Steinbach e Kumar (2009), não adianta ter o melhor SGBD do mercado ou os dados simplesmente armazenados, mais que isso, é imperativo ter uma adequada manipulação dessas informações, de acordo com as necessidades da organização.

Frente à necessidade da recuperação de informação em tempo hábil e da busca de conhecimento que auxiliem em tomadas de decisões, criaram-se várias técnicas para análise e

processamento de dados. Dessa maneira, diversos autores como Brauner (2003), Carvalho (2005) e Elmasri e Navathe (2011) apontam que as principais técnicas para a recuperação de informação são as ferramentas de consulta, os métodos de mineração de dados e o processo analítico *On-line* (*Online Analytical Processing* – OLAP). De acordo com os resultados esperados e a situação onde estas três técnicas serão empregadas, elas podem trabalhar de forma integrada, concomitante ou independente.

Ainda conforme Elmasri e Navathe (2011), as ferramentas de consultas, também chamadas de linguagem estruturada de consulta, ou *Structured Query Language* (SQL) oferecem várias funcionalidades para manipulação, definição e controle de dados. Assim, para consultas, inserções, atualizações e remoções de linhas em tabelas do banco de dados, são utilizados os comandos da Linguagem de Manipulação de Dados ou *Data Manipulation Language* (DML); já para criar, alterar ou remover tabelas no banco de dados são utilizados os comandos da Linguagem de Definição de Dados ou *Data Definition Language* (DDL).

Brauner (2003) diz que a maioria dos sistemas gerenciadores de banco de dados incorpora a linguagem SQL para consulta, acesso e manipulação dos dados, levando o usuário direto às informações as quais precisa. Ele descreve que o SQL também permite encontrar informações nos dados de acordo com os padrões pré-definidos pelos criadores do banco de dados na aplicação. Dessa maneira, vários programas são desenvolvidos utilizando ferramentas de consultas SQL para fornecer *feedback* ou recuperar informações aos usuários.

Portanto, baseado em Morais (2010), com as ferramentas de consultas a bancos de dados, é possível responder somente a perguntas com conjecturas antecipadamente definidas. Segundo Brauner (2003), a formulação de hipóteses para a coleta de informação utilizando SQL cria um problema que, em muitas situações, os usuários não as conseguem formular, ou, conseguem formular apenas meias hipóteses. Nesse sentido, os criadores do banco de dados precisam descobrir padrões nos dados para, então, gerar hipóteses, possivelmente corretas, e conseguir, após isso, comprovar o comportamento do restante dos dados em relação ao padrão descoberto.

Nesse sentido, Tang, Steinbach e Kumar (2009) explicam que nas organizações onde existem grades fluxos de informações, não é viável a utilização de apenas o SQL para a recuperação de informação. Portanto, para que uma determinada organização faça da sua base de dados um manancial de informações úteis, ela necessita fazer análises que busquem padrões em diferentes níveis de abstração, criando, assim, uma visão lógica dos dados. Carvalho (2005) expõe que o SQL é inviável para atender a estas especificações de análises de dados e afirma que o OLAP é a ferramenta adequada para a busca de informações em

grandes variedades de dados. Dessa forma, a linguagem de consulta pode atuar como processo auxiliar.

Sob o ponto de vista de Brauner (2003), os sistemas OLAP utilizam-se da extensa indexação dos armazéns de dados, ou *data warehouse*, para possibilitar o acesso e a apresentação gráfica de pedaços dos dados, combinados praticamente de qualquer modo que desejado pelo usuário. O OLAP se caracteriza por fazer análises multidimensionais dos dados armazenados de forma *online*, ou seja, em tempo real. Ele funciona como uma análise interativa de dados que permite visões por meio de associações em diversas dimensões, e possibilita, também, a exibição das informações em mapas e gráficos. Além disso, do OLAP derivam-se análises estatísticas envolvendo medidas ou dados numéricos. Tang, Steinbach e Kumar (2009) dizem que o *data warehouse* e OLAP são tecnologias que se complementam. O *data warehouse* abrange o processo de reunir de forma organizada e eficiente os dados de diferentes fontes, enquanto o OLAP transforma os dados de um *data warehouse* em informações úteis.

Porém, baseado em Carvalho (2005), trabalhar com *data warehouse* e OLAP é um processo caro e necessita de um rigoroso planejamento. Além disto, os sistemas OLAP emitem respostas para algumas perguntas que necessitam saber que dados se encaixam em determinado padrão. Portanto, quando necessita saber quais são os padrões que existem nos dados, o OLAP não apresenta eficiência. De acordo com Penedo e Capra (2012), os sistemas OLAP vêm incorporando técnica de mineração de dados, para que esta ferramenta consiga, também, descobrir padrões em bases de dados, porém, Tang, Steinbach e Kumar (2009) afirmam que as incorporações de ferramentas de mineração de dados em sistemas OLAP ainda não são eficazes, além disso, a mineração de dados não precisa e quase sempre não pode ser feita *online*, por sua complexidade e demora.

As técnicas de mineração de dados permitem ao usuário encontrar padrões sobre os dados armazenados no banco de dados. Estes padrões podem ser comprovados através de consultas SQL e sistemas OLAP. Por isso, é crescente a criação e implementação de técnicas que serão aplicadas para automatizarem o processo de descoberta de padrões e tendências sobre base de dados. Diante deste contexto, Carvalho (2005) afirma que a mineração de dados pode complementar os sistemas OLAP, porém, ela é independente da existência de um *data warehouse* ou um sistema OLAP. Ele Enfatiza que a mineração de dados é um processo barato, diferentemente do que acontece com o OLAP e data *warehouse* e que não exige grande recurso tecnológico das organizações. Zambon e Meirelles (2001) definem o *data*

warehouse como a memória de uma organização e a mineração de dados como a inteligência da mesma.

Russel (2011) escreve que as ferramentas de consulta, OLAP, *data warehouse* e as técnicas de mineração de dados são complementares, e a mineração de dados não substitui nenhuma destas ferramentas, mas oferece condições para que as informações sejam descobertas em grandes volumes de dados. Os padrões encontrados servem para prever futuras tendências e comportamentos, permitindo novos processos de tomada de decisão, baseado, principalmente, no conhecimento implícito, frequentemente desprezado, contido nos bancos de dados.

Conforme afirma Brauner (2003), a mineração de dados de dados pode atuar como sinônimo do processo de DCBD, porém, Murasse e Tsunoda (2010) dizem que a mineração de dados é uma fase do processo de DCBD. Nesse sentido, Russel (2011) descreve que a forma na qual se procura padrões nas bases de dados é deliberada de formas distintas de acordo com cada autor. De acordo com Brauner (2003), quando o autor possui uma abordagem voltada para a área de tecnologia da informação, análise de dados e de negócios, eles tendem a chamar o DCBD de mineração de dados.

Murasse e Tsunoda (2010) falam que, na visão dos profissionais da inteligência artificial, a mineração de dados é uma etapa do processo de DCBD onde se aplicam os algoritmos para inferência dos padrões; já Carvalho (2005) indica que, para melhor apresentação de um estudo, é importante seguir o processo de DCBD e ter a mineração de dados como uma das suas etapas.

2.2. Descoberta de conhecimento em banco de dados

Baseado em Oliveira e Garcia (2004), o processo de descoberta de conhecimento em banco de dados é uma área da inteligência artificial voltada para a inteligência de negócios que pode ser utilizado para auxiliar na descoberta de conhecimento útil nas grandes bases de dados, de forma rápida e confiável. O DCBD aplica métodos interdisciplinares, especialmente métodos estatísticos e de aprendizado de máquina, para extrair conhecimento de alto nível a partir de bases de dados reais. A mineração de dados ou *data mining* é uma das principais tarefas do processo DCBD, que consiste na aplicação de algoritmos com a finalidade de extrair padrões de comportamento em uma base de dados.

Existem diversas abordagens para o processo de busca de conhecimento em banco de dados, mas todas elas possuem um conjunto de etapas que inclui a identificação e

entendimento do problema, descoberta de conhecimento, análises das relações descobertas e avaliação dos resultados. Nesse sentido, de acordo com Russel (2011), independente da abordagem, em todos os processos a principal etapa é aquela na qual se aplicam as técnicas que utilizam algoritmos para extração de conhecimento de uma base de dados estudada, ou seja, a mineração de dados.

Coelho (2007) cita Fayyad e outros, que criaram uma abordagem que determina a iteratividade de etapas de um processo de DCBD e a interatividade do usuário ao processo de descoberta de conhecimento em banco de dados, conforme é demonstrado na Figura 2. O modelo de processo proposto por essa abordagem visa que, a cada etapa, o usuário analise as informações geradas, procure incorporar sua experiência e tome decisões para obter resultados melhores que as etapas anteriores. O processo, então, é composto de cinco etapas: seleção, processamento, transformação, mineração de dados e interpretação. O modelo de processo de DCBD exposto por Fayyad e outros é a base para a elaboração da maioria dos processos utilizados. Como pode ser verificado abaixo, ele considera que é necessário fazer uma análise do problema a ser resolvido pelo processo de descoberta de conhecimento, pois, o perfeito entendimento do problema é importante para definir corretamente os objetivos do processo de DCBD.

Figura 2 - Processo de DCBD de Fayyad e outros

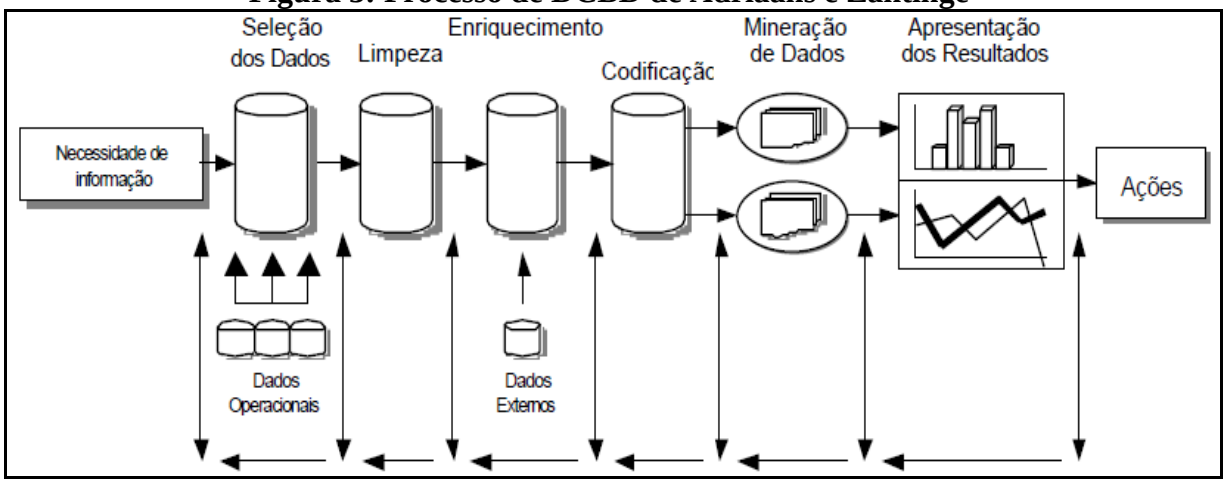
Fonte: FAYYAD e outros *apud* MARTINHAGO, 2005, p. 25.

Tang, Steinbach e Kumar (2009) afirmam, porém, que este modelo de processo criado por Fayyad e outros pode ser modificado de acordo com a necessidade da análise de dados. Já segundo Carvalho (2005), a mineração de dados deve ser repetida diversas vezes em uma organização, porque com um grande fluxo de dados é possível surgir um novo padrão a qualquer instante. Mesmo que uma etapa tenha sido bem planejada e executada, os resultados

podem não ser claros em algumas situações, dessa forma, é necessário retornar a alguma etapa anterior para aplicar uma nova atividade. (RUSSEL, 2011). Diante deste contexto, Russel (2011) sugere que o modelo a ser aplicado em uma instituição que queira continuar com a mineração de dados seja o modelo de Adriaans e Zantinge, que complementa aquele modelo criado por Fayyad e outros.

Segundo Brauner (2003), que também cita Adriaans e Zantinge, esta abordagem do processo de DCBD baseia-se na necessidade de as organizações obterem continuamente novas informações sobre seus dados, sendo reproduzido quando novas necessidades de informações aparecerem. Com isso, nesta abordagem não existe uma etapa específica para entendimento dos dados, pressupondo-se que já exista um conhecimento prévio do domínio do banco de dados e dos objetivos do negócio. O processo de descoberta de conhecimento em banco de dados de Adriaans e Zantinge compreende seis fases: seleção de dados, limpeza de dados, enriquecimento, transformação ou codificação de dados, mineração de dados e o relatório e exibição da informação descoberta, conforme é demonstrado na Figura 3, a seguir:

Figura 3: Processo de DCBD de Adriaans e Zantinge



Fonte: ADRIAANS e ZANTINGE *apud* BRAUNER, 2003, p. 33.

De acordo com Carvalho (2005), quando o processo é repetido para analisar novos dados que complementam os dados já existentes, esse modelo de processo de Adriaans e Zantinge não exige conhecimento do domínio da aplicação do processo de DCBD, pois, parte-se do pressuposto que a estrutura da base de dado já é conhecida e o que pode existir em incomum são as instâncias. A repetição do processo busca considerar os novos dados que ainda não foram analisados e que podem conter algum padrão. Russel (2011) também diz que a nomenclatura das etapas dos dois modelos é diferente, mas os objetivos de ambas são semelhantes. Na abordagem de Fayyad e outros existe a etapa de processamento devido à

importância da preparação dos registros para a mineração de dados, que foi desmembrada em duas etapas (limpeza e enriquecimento) no modelo de Adriaans e Zantinge.

O processo de DCBD de Adriaans e Zantinge é um modelo de processo completo para a maioria das organizações, porém, o responsável pela aplicação da descoberta de conhecimento de banco de dados não precisa executar todas as etapas ou fluxos desta abordagem. (TANG; STEINBACH; KUMAR, 2009).

Assim, diante do exposto, é importante frisar que o modelo a ser utilizado neste estudo é a abordagem de Adriaans e Zantinge, por ela complementar as características da abordagem de Fayyad e outros e por ela ser mais detalhada. Este modelo também atende, de forma mais clara, a proposta e relevância desta pesquisa, na qual se trabalharam as seis etapas do processo: seleção de dados, limpeza, enriquecimento, codificação, mineração de dados e apresentação dos resultados, as quais serão explicadas a partir de agora.

2.2.1. Seleção de dados

De acordo com Coelho (2007), após a definição do domínio, deve-se localizar e escolher quais as fontes de dados estão relacionadas a este domínio para que o conjunto de dados apropriado possa ser selecionado. A seleção de dados também consiste em criar uma nova base de informações utilizando conjuntos de registros de várias bases de dados, onde o novo banco de dados deve estar coerente com o original. Segundo Elmasri e Navathe (2011), as fontes internas, normalmente, são fontes de dados que já estão incorporadas ao sistema de aplicação do domínio em questão e as fontes externas são compostas por outros tipos de fontes que, habitualmente, não são incorporados ao sistema de aplicação, como: documentos, livros, *internet* e informações do especialista do domínio.

Nesta etapa ocorre uma análise de todos os dados operacionais do banco de dados e são selecionados apenas aqueles que são necessários para alcançar os objetivos do processo. Podem ser feitas novas seleções quando houver outra iteração, ou seja, podem-se incluir dados anteriormente descartados, pois o processo é iterativo, permitindo a retomada de qualquer etapa, independente de em qual esteja. Ainda nesta etapa poderá haver a escolha do tipo de técnica de mineração a ser adotada. A questão a ser minerada e a própria técnica a ser trabalhada ajuda a definir qual parte da massa de dados inicial vai ser utilizada e, portanto, selecionada.

2.2.2. Limpeza

Após reunir dados de determinadas fontes, é muito provável que o conjunto de dados venha a conter registros duplicados, erros e dados ausentes. A geração desses ruídos pode ocorrer devido a problemas em migrações de um sistema para outro, quedas de tensão na hora do processamento, desligamento do computador tendo arquivos abertos, falta de tratamento adequado no armazenamento das entradas de dados vindas de sistemas operados por usuários comuns, dentre outros. Em um conjunto de dados constituídos por diversas fontes pode acontecer, por exemplo, que um atributo sexo possua diferentes valores e tipos com o mesmo significado como: “masculino”, “mas”, “m”, “M” ou “1”. Então, é preciso transformar estes valores em um tipo comum para todo o conjunto de dados.

A etapa de limpeza pode ser executada inúmeras vezes, já que é impossível prever, com antecedência, todos os problemas de qualidade existentes na base. Conforme Coelho (2007), alguns estudos mostram que a etapa de limpeza dos dados pode tomar até 80% do tempo necessário para todo o processo de descoberta de conhecimento. Por isso, esta etapa é considerada uma das mais importantes para o sucesso do processo como um todo.

2.2.3. Enriquecimento

Algumas informações podem ser incluídas ao banco de dados para que seja possível atingir os objetivos do processo. Estes dados podem estar disponíveis em outros locais, ou, até mesmo, podem ser gerados a partir de dados existentes no banco de dados e transformados para obtermos a informação.

Desta forma, o enriquecimento normalmente melhora os dados com fontes de informações adicionais. Por exemplo, dados os nomes de cliente e números de telefone, um estabelecimento comercial pode adquirir outros dados sobre idade, renda e avaliação de crédito e anexá-las a cada registro.

2.2.4. Codificação

Sob o ponto de vista de Coelho (2007), antes de aplicar a técnica de mineração de dados é preciso ser realizada a codificação dos registros, com o objetivo de facilitar seu uso pelas técnicas de mineração. É necessário fazer certas adequações no conjunto de dados de

acordo com a técnica de mineração de dados a ser utilizada, pois existem diversos tipos de algoritmos, e cada um necessita de uma entrada específica, além das conversões de dados, criação de novas variáveis e categorização de variáveis contínuas.

A forma que os dados estão armazenados nos bancos de dados pode não ser a representação mais apropriada para a utilização no processo de DCBD. Geralmente, os dados têm sua representação apropriada ao contexto da aplicação. Por exemplo, um atributo com valores literais pode não ser adequado a determinados algoritmos mineradores utilizados na etapa de mineração de dados. Para adequá-lo, pode ser necessário normalizar estes valores dentro de um determinado intervalo. A codificação é um procedimento criativo, já que existem diversas maneiras de codificação, sendo, portanto, difícil descrevê-las, pois cada caso deve ser analisado individualmente e sua codificação pode variar de acordo com a escolha do algoritmo minerador da próxima etapa.

2.2.5. Mineração de dados

Esta etapa consiste na efetiva aplicação das tarefas e das técnicas escolhidas sobre os dados a serem analisados para encontrar os padrões implícitos. Portanto, o sucesso desta etapa depende diretamente da correta realização das etapas anteriores. Para Carvalho (2005), tarefas são classes de problemas que foram definidas através de estudos na área de mineração de dados e técnicas são grupos de soluções (algoritmos) para os problemas propostos nas tarefas, pois, cada uma apresenta várias técnicas e algumas delas podem ser utilizadas para solucionar tarefas diferentes.

Esta é a etapa onde os dados são manipulados para que seja extraído o conhecimento. Desta maneira, ela é a etapa que mais exige dos recursos computacionais. Utilizando inicialmente uma ferramenta de consulta SQL, é possível ter uma visão geral dos dados para, então, partir para uma análise menos trivial. Conforme Coelho (2007), nesta primeira tarefa dentro do processo de DCBD, 80% do conhecimento é extraído e já pode revelar alguma informação interessante. Entretanto, as informações extraídas por estas consultas podem não ser suficientes, surgindo a necessidade de utilizar técnicas avançadas. Os algoritmos utilizados para se criar modelos a partir de dados, normalmente, provêm de áreas como Aprendizado de Máquina, Reconhecimento de Padrões e Estatística. Estas técnicas, muitas vezes, podem ser combinadas para se obter resultados melhores.

2.2.6. Apresentação dos resultados

Finalizada a etapa de mineração de dados, resultam informações num formato específico de acordo com a técnica utilizada. Deve-se levar em conta, porém, que os dados podem estar codificados ou mesmo que o método utilizado na etapa de mineração gere, como saída, informações em algum formalismo ou representação muito específicas. Estes resultados devem ser exibidos de forma clara para que sejam de fácil entendimento para quem irá utilizá-los, pois essas são, geralmente, pessoas que não interpretarão os resultados tão facilmente quanto aquela que conduziu o processo de DCBD.

Como exposto por Coelho (2007), caso o conhecimento adquirido com a aplicação da mineração de dados não seja útil, deve-se, então, retornar às etapas anteriores e tentar refazê-las ou melhorá-las. Esta iteração pode ocorrer até que se seja obtido resultados aceitáveis ou concluir-se que não seja possível extrair conhecimento relevante dos dados. A falta de padrões serve como resultado de que os métodos utilizados não inferem algum conhecimento, mas, com o aumento da base dados é possível adquirir algum resultado viável.

2.3. Mineração de Dados

A mineração de dados, ou *data mining*, consiste na aplicação de algoritmos com a finalidade de extrair informações em uma base de dados. O processo de busca do conhecimento na mineração é realizado através da busca de padrões de comportamento ou relacionamentos íntegros entre as diversas instâncias de dados.

Diversos autores adotam a definição de mineração de dados como sinônimo do DCBD, porém, neste trabalho, como já dito anteriormente, a mineração de dados será abordada como uma tarefa do processo de descoberta de conhecimento em banco de dados.

A mineração de dados pode ser empregada para resolver diversos problemas na medicina, em *marketing*, em bancos, na astronomia, na previsão do volume de vendas, na previsão de mercados financeiros, planejamento de produção industrial, melhoramento de serviços etc. De acordo com Zambon e Meirelles (2001), a mineração de dados é um subprocesso da busca de conhecimento em banco de dados e, antes de ser executada, devem-se ter definidos os objetivos finais, que podem constar dentro das seguintes classes:

a) **previsão ou predição** - Baseados em comportamentos anteriores dos atributos de dados, a mineração de dados pode mostrar como determinadas ações vão se comportar no futuro;
b) **identificação** - Utiliza padrões de dados para identificar a existência de um evento ou uma atividade. Os padrões são determinados a partir de relações específicas entre os diversos atributos de dados;
c) **classificação** - Este objetivo visa que a mineração de dados particione os dados em diferentes classes ou categorias, de acordo com a combinação dos parâmetros, sendo possível a classificação dos dados em diferentes grupos;
d) **otimização** - Constitui um objetivo relevante da mineração de dados, baseado na execução otimizada de atividades que utilizam recursos limitados, como tempo, espaço, dinheiro ou materiais.

A Mineração de dados dispõe de tarefas básicas classificadas nas categorias descritivas, que envolvem a descoberta de padrões interpretáveis por humanos que descrevam os fatos cadastrados na base de dados, e preditivos, que utilizam determinadas variáveis para prever valores desconhecidos de outras variáveis de interesse.

Enquanto uma tarefa está relacionada ao que se pretende buscar nos dados, ou seja, que tipo de padrões deseja-se encontrar, uma técnica, por sua vez, está relacionada a como encontrar os padrões de interesse. De acordo com Carvalho (2005), as principais tarefas da mineração de dados são: associação, classificação e agrupamento.

2.3.1. Classificação

A classificação consiste em examinar as características de um dado e atribuir a ele uma classe pré-definida. Ou seja, esta tarefa objetiva a construção de modelos que permitam o agrupamento de dados em classes. Esta tarefa é considerada preditiva, pois uma vez que as classes são definidas, ela pode prever automaticamente a classe de um novo dado. Por exemplo, uma população pode ser dividida em categorias para avaliação de concessão de crédito com base em um histórico de transações de créditos anteriores. Em seguida, uma nova pessoa pode ser enquadrada, automaticamente, em uma categoria de crédito específica, de acordo com suas características.

Para Martinhago (2005), a classificação é uma das tarefas mais estudadas pela comunidade científica, e a árvore de decisão constitui a principal característica desta tarefa.

As árvores de decisão possuem este nome devido à sua estrutura assemelhar-se a uma árvore, constituindo uma estrutura fácil de entender, já que elas dividem os dados em subgrupos, com base nos valores das variáveis. Conforme Coelho (2007), a construção de uma árvore de decisão é o resultado de uma hierarquia de declarações do tipo "Se existe isto, então existe aquilo", que são utilizadas quando o objetivo da mineração de dados é a classificação dos dados.

Baseado em Carvalho (2005), na árvore, cada nó especifica o teste de um atributo da instância, e cada ramificação corresponde a um dos possíveis valores do atributo. Uma instância é classificada, começando pela raiz da árvore, testando o atributo especificado, movendo-se para um nível abaixo. Este processo é repetido para o nó mais baixo, enraizada pelo novo nó. Uma árvore de decisão utiliza a estratégia chamada "dividir para conquistar", pois divide um problema maior em outros menores.

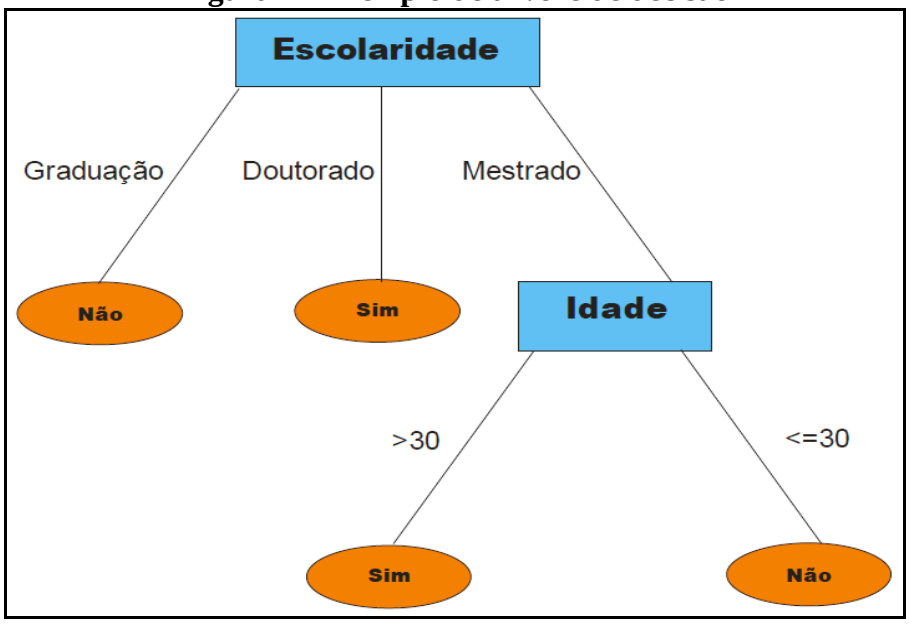
Conforme Russel (2007), as árvores de decisão podem ser aplicadas a um grande conjunto de dados, permitindo bom entendimento e o resultado do algoritmo é de fácil interpretação. Uma árvore de decisão é formada por um conjunto de regras de classificação, onde cada caminho da raiz até uma folha representa uma destas regras. A árvore de decisão deve ser definida de forma que, para cada observação da base de dados, exista um (e apenas um) caminho da raiz até a folha. Para Coelho (2007), um dos principais algoritmos de árvores de decisão é o *J48*. As quatro regras de classificação do Quadro 1, abaixo, compõem a árvore de decisão da Figura 4, logo a seguir.

Quadro 1- Exemplo regra de classificação

Se (Escolaridade = "Graduação") então (Rico = "Não")
Se (Escolaridade = "Doutorado") então (Rico = "Sim")
Se (Escolaridade = "Mestrado") e (Idade = ">30") então (Rico = "Sim")
Se (Escolaridade = "Mestrado") e (Idade = "<=30") então (Rico = "Não")

Fonte: COELHO, 2007, p.14.

Figura 4 - Exemplo de árvore de decisão



Fonte: COELHO, 2007, p.14.

2.3.2. Associação

Estuda um padrão de relacionamento entre itens de dados. Por exemplo, uma análise das transações de compra em um supermercado pode encontrar itens que tendem a ocorrerem juntos em uma mesma compra, como café e leite. Os resultados desta análise podem ser úteis na elaboração de catálogos e *layouts* de prateleiras, de modo que produtos a serem adquiridos na mesma compra fiquem próximos um do outro. Essa tarefa é considerada descritiva, ou seja, ela é usada para identificar padrões em dados históricos. A tarefa de associação visa combinar dados importantes, de forma que, ao descobrir a presença de um dado em uma determinada transação, pressupõe a de outro na mesma transação.

Tang, Steinbach e Kumar (2009) expõem que a regra de associação é uma expressão representada na forma X => Y (X implica em Y), em que X e Y são conjunto de itens na base de dados e $X \cap Y = \emptyset$. X é o antecedente da regra (lado esquerdo) e Y é o consequente da regra (lado direito) e pode envolver qualquer número de itens em cada lado da regra. O significado desta regra é que as transações da base que contém X tendem a conter Y. Um exemplo prático da utilização desta regra é afirmar que 80% dos registros que contêm X também contêm Y.

Para interpretar os resultados da regra de associação, é necessário compreender dois parâmetros: o suporte e a confiança. Eles limitam a quantidade de regras que serão extraídas e descrevem a qualidade delas. Para Tang, Steinbach e Kumar (2009), o suporte determina a

frequência na qual a regra é aplicável a um determinado conjunto de dados, enquanto que a confiança determina a frequência na qual os itens em Y aparecem em transações que contenham X. Nesse sentido, Carvalho (2005) define a confiança como a frequência com que o relacionamento mantém-se verdadeiro na amostra de treinamento e o suporte como a frequência com que a combinação acontece. Assim, uma associação pode se manter 100% do tempo e ter a mais alta confiança, porém, pode ser de pouca utilidade se o suporte ocorrer raramente.

Para ilustrar como se aplicam o suporte e a confiança nas regras de associação, utiliza-se um exemplo de uma cesta de compras. Para analisar os itens adquiridos em cinco compras, conforme o Quadro 2, onde é analisada a regra {Leite, Fraldas} => {Cerveja}, quem comprou leite e fraldas comprou cerveja. Já que o contador de suporte para {Leite, Fraldas, Cerveja} é 2 e o número total de transações é 5, o suporte da regra é 2/5=0,4. A confiança da regra é obtida dividindo-se o contador de suporte para {Leite, Fraldas, Cerveja} pelo contador de suporte de {Leite, Fraldas}. Como há 3 transações que contêm leite e fraldas, a confiança para esta regra é 2/3=0,67. Assim, segundo Carvalho (2005), o principal algoritmo da regra de associação é o *Apriori*.

Quadro 2 - Exemplo da regra de associação com cesta de compras

Itens comprados
{Pão, Leite}
{Pão, Fraldas, Cerveja, Ovos}
{Leite, Fraldas, Cerveja, Cola}
{Pão, Leite, Fraldas, Cerveja}
{Pão, Leite, Fraldas, Cola}

Fonte: RUSSEL, 2011

2.3.3. Agrupamento

No agrupamento, as informações podem ser particionadas em classes de elementos similares. Neste caso, nada é informado ao sistema a respeito das classes existentes. O próprio algoritmo descobre as classes a partir das alternativas encontradas na base de dados, agrupando, assim, um conjunto de objetos em classes com características semelhantes. Por exemplo, uma população inteira de dados sobre tratamento de certa doença pode ser dividida em grupos baseados na semelhança de efeitos colaterais produzidos; ou acessos à web

realizados, por um conjunto de usuários em relação a um conjunto de documentos, podem ser analisados para revelar grupos ou categorias de usuários. Esta tarefa é, então, considerada descritiva.

De acordo com Russel (2011), a análise de agrupamento fornece uma abstração de objetos individuais de dados para os grupos nos quais esses objetos de dados residem. Além disso, algumas técnicas de agrupamento caracterizam cada grupo em termos de um protótipo de grupo, onde um objeto de dados seja representativo dos outros objetos do grupo. Estes protótipos de grupos podem ser usados como a base para uma quantidade de técnicas de análise de dados ou de processamento de dados. Portanto, no contexto da utilidade, a análise de grupos é o estudo de técnicas para encontrar os protótipos de grupos mais representativos. Baseado em Tang, Steinbach e Kumar (2009), o algoritmo *Simple KMeans* é um dos algoritmos de agrupamento de maior eficácia na ferramenta WEKA (*Waikato Environment for Knowledge Analysis)*. Os algoritmos da tarefa de agrupamento, assim como os algoritmos das tarefas de associação e classificação, são técnicas que buscam o resultado ótimo, ou seja, o melhor resultado possível para a análise de uma base de dados.

2.4. Ferramenta de mineração de dados WEKA

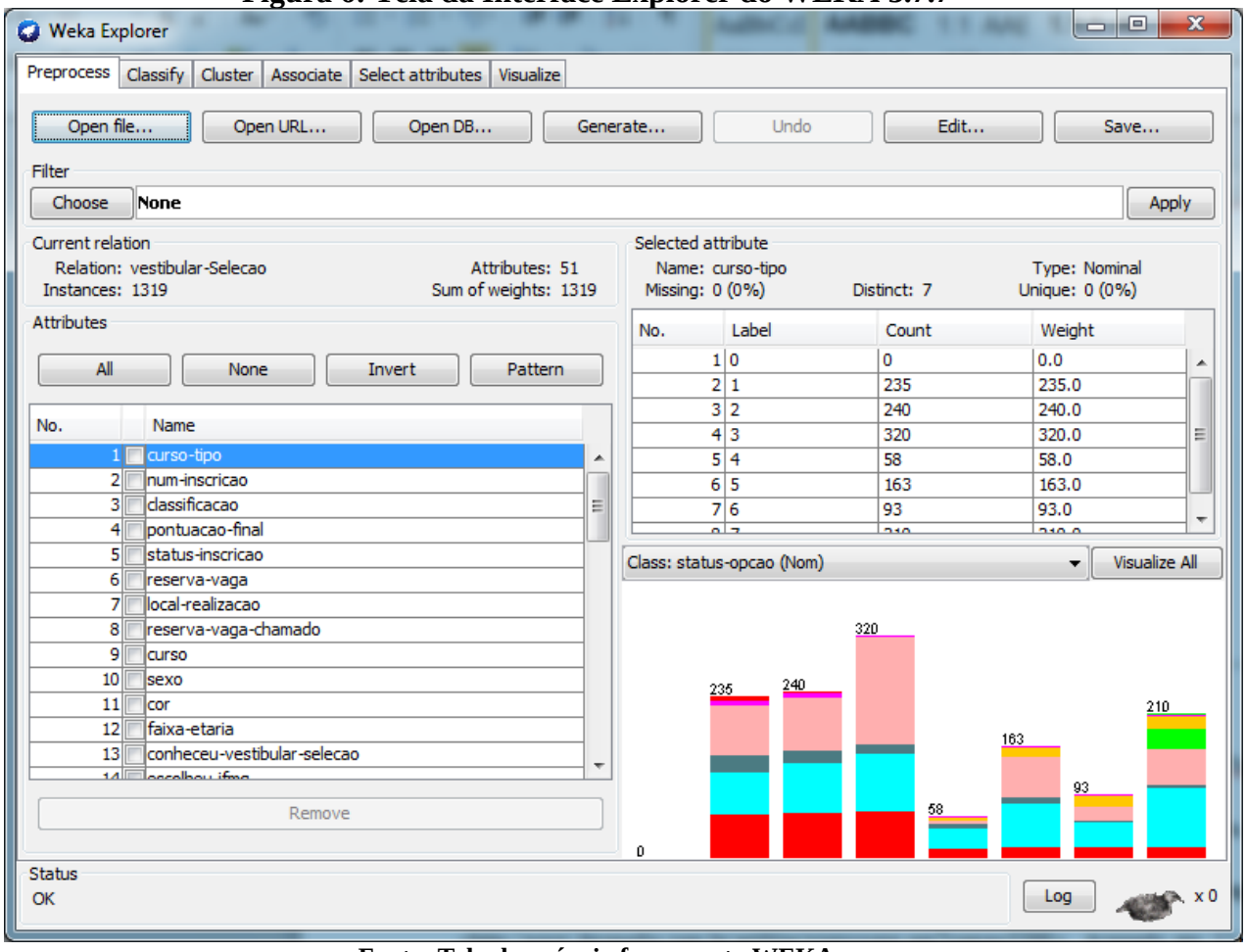
De acordo com Clésio (2013), o WEKA é uma ferramenta gratuita e de código fonte aberto que implementa diversos tipos de algoritmos de aprendizagem de máquina para várias técnicas da mineração de dados, sendo que alguns destes algoritmos utilizam o conceito de inteligência artificial. Esta ferramenta foi desenvolvida por um grupo de pesquisadores da Universidade de *Waikato*, na Nova Zelândia, no ano de 1993. Desde então, ela é atualizada sempre que é criada uma nova versão. O WEKA é implementado na linguagem de programação *Java*, permitindo com que ele seja portável e funcione em diversas plataformas, como: *Windows*, *Linux* e *MacOS.* (MACHINE LEARNING GROUP AT THE UNIVERSITY OF WAIKATO, 2013). Devido às diversas características positivas que detém, o WEKA consolidou-se como a ferramenta gratuita mais popular para mineração de dados, principalmente no meio acadêmico.

Conforme as Figuras 5 e 6, a versão 3.7.7 do WEKA utilizada neste estudo é simples e amigável, podendo ser utilizada até por pessoas que não sejam especialistas na ferramenta.

Figura 5: Tela inicial do WEKA 3.7.7

Fonte: Tela da própria ferramenta WEKA

Figura 6: Tela da Interface Explorer do WEKA 3.7.7



Fonte: Tela da própria ferramenta WEKA

Como pode ser notado nas figuras acima, as principais interfaces de interação do *software* são:

a) *explorer* - É a interface gráfica mais utilizada para executar o processo de mineração de dados. Nela encontram-se as opções de pré-processamento de dados, as técnicas de mineração de dados (classificação, agrupamento e associação), estimação de atributo e visualização de dados processados;
b) *experimenter* - Determinar se um esquema é estatisticamente melhor do que o outro esquema;
c) *knowledgeFlow* - Forma um fluxo de conhecimento para o processamento e análise de dados;
d) *simple CLI* - É a interface utilizada por usuários especialistas em mineração de dados e, na ferramenta, ela é operada através de linhas de comandos.

Para que um arquivo contendo uma base de dados seja lido pelo WEKA, é necessário que ele esteja no formato ARFF. O arquivo ARFF é composto de três partes, conforme pode ser notado na Figura 7, abaixo: a primeira parte (*relation*) contém o nome da relação; a segunda parte (*attribute*) é onde se determina o tipo de atributo e os respectivos valores que irão representar, e a última parte (*data*) consiste nas instâncias que serão mineradas. O WEKA também oferece a opção de importar uma base de dados do MySQL.

Figura 7: Estrutura de um arquivo ARFF

```
@relation Mercado

@attribute leite {y,n}
@attribute pao {y,n}
@attribute bolacha {y,n}
@attribute ovo {y,n}
@attribute manteiga {y,n}
@attribute cafe {y,n}
@attribute suco {y,n}

@data
y,y,?,y,?,?,?
y,?,?,?,?,?,y
?,?,?,?,y,?,y
y,y,?,y,?,?,?
?,?,?,y,?,y,?
?,?,?,?,?,y,?
?,?,?,?,?,y,y
y,y,y,y,?,?,?
```

Fonte: Dados da pesquisa

2.5. Aplicação da mineração de dados

Essa típica tarefa de mineração de dados é usada por grandes lojas de departamentos e administradoras de cartões de crédito, que utilizam os dados das compras dos clientes no passado recente para traçar seus perfis de consumo. Informações como idade, sexo, estado civil, salário, moradia (própria ou alugada), bairro e cidade também são importantes, pois permitem a setorização ainda mais fina dos clientes. Conhecer o perfil de seus clientes é fundamental para que uma empresa possa se manter no mercado. Nesse sentido, conforme Vianna (2007), muito investimento deve ser feito para que o cliente continue fiel à empresa e para que outros sejam conquistados. Para tanto, as empresas precisam realizar os desejos e necessidades do cliente, cuidando do estoque, da distribuição dos produtos nas prateleiras e das promoções criativas, a fim de propiciar compras casadas de produtos.

Apesar de a mineração de dados no Brasil ser mais difundida no meio acadêmico do que no setor comercial ou privado, os resultados obtidos com esta tecnologia têm aumentado sua popularidade nas empresas brasileiras. Devido à sua influência nas tomadas de decisões, a mineração se tornou umas das principais armas das empresas competitivas, permitindo-as identificar fenômenos que antes não conheciam ou que eram definidos como triviais (por falta de conhecimento).

Vale ressaltar, ainda, que a mineração de dados não é empregada apenas para identificar certos comportamentos de pessoas em determinadas ocasiões, pois, inicialmente, ela era aplicada nas empresas principalmente para conhecer o perfil de sua clientela, no intuito de fidelizar clientes como consumidores e fazer com que consumissem mais produtos ou serviços. Com isto, a utilização da mineração de dados difundiu-se para diversas áreas, podendo ser aplicada a inúmeros contextos organizacionais. Desta forma, a mineração de dados pode encontrar, por exemplo, uma doença no DNA de um ser, a partir de manifestações em seu DNA em relação a outros seres vivos analisados. Conforme Carvalho (2005), atualmente as áreas que mais utilizam a mineração de dados e que possuem os resultados mais expressivos, são: *marketing*, finanças, manufatura, saúde, varejo e educação.

2.5.1. Marketing

As aplicações de mineração de dados no marketing envolvem a análise de comportamento do consumidor, baseado nos padrões de compra. Desta forma, é possível descobrir quais clientes possuem maior probabilidade de comprar um produto específico, a

partir do histórico de vendas do produto e dos dados de clientes compradores. Com a identificação dos padrões de comportamento dos consumidores, permite-se a determinação das estratégias de *marketing* que incluem propaganda, local da loja, endereço do cliente, segmentação de clientes, produtos comercializados, catálogos de divulgação, *layouts* da loja e campanhas publicitárias.

Segundo Martinhago (2005), para exemplificar o exposto, a multinacional americana de lojas de departamento Walmart é uma das mais avançadas empresas em mineração de dados e na aplicação de seus resultados ao negócio, principalmente na área de *marketing*. Na intenção de buscar novas informações para aperfeiçoar suas estratégias de mercado, a Walmart investiu na análise das relações entre o volume de vendas e os dias da semana. Ao aplicar as técnicas de mineração de dados nos dados referentes às vendas para consumidores dos Estados Unidos da América, os executivos da Walmart identificaram um hábito curioso e que passava despercebido. O *software* de mineração de dados utilizado pela empresa identificou que nas sextas-feiras as vendas de cervejas cresciam na mesma proporção que as de fraldas. Dentro de várias investigações minuciosas, foi, então, revelado que os pais, ao comprar fraldas para seus filhos, aproveitavam para abastecer o estoque de cerveja para o final de semana. Diante disso, as lojas de departamento da Walmart passaram a exibir a cerveja perto da categoria de fraldas.

2.5.2. Finanças

Em finanças, a aplicação de mineração de dados destina-se a diversas subáreas, incluindo a análise de crédito de clientes, a segmentação de contas a receber e a análise de desempenho de investimentos financeiros. No entanto, ela pode ser utilizada para trabalhar investimentos em ações, títulos e fundos de investimentos, avaliação de opções de financiamento e detecção de fraudes.

Segundo Monteiro e Rocha (2005), o banco Itaú é uma empresa pioneira no Brasil no uso de mineração de dados. Era comum o envio de mais de um milhão de malas diretas com ofertas para os correntistas. A correspondência era composta de diversos serviços e direcionava-se a todos os correntistas, mas no máximo 2% deles respondiam às promoções. Hoje, o banco mantém informações sobre toda a movimentação financeira de mais de cinco milhões de clientes. Assim, através da mineração dessa base de dados, é possível que as cartas sejam direcionadas apenas àqueles clientes que demonstram maior chance de responder à

oferta. A taxa de retorno, então, aumentou de 2% para 30% e houve uma economia de aproximadamente 80% nas despesas com serviços de correio.

2.5.3. Manufatura

Na construção de produtos, a mineração de dados é útil pelo fato de envolver a otimização de recursos, como máquinas, mão de obra e materiais.

De acordo com Silva (2001), por exemplo, a Usiminas, empresa do setor siderúrgico com sede em Belo Horizonte – MG, implantou um sistema que monitora a produtividade de seus funcionários e permite analisar a relação entre dias da semana, horário e produtividade. Ao aplicar o processo de mineração de dados, os gestores da empresa identificaram que certos perfis tendiam a ter maior produtividade em turnos e dias específicos. Diante disso, a Usiminas fez uma reorganização de pessoal, adequando os funcionários, de acordo com o perfil, ao novo quadro de horário de trabalho, aumentando substancialmente a produção da empresa.

2.5.4. Saúde

Em saúde, algumas aplicações envolvem a descoberta de padrões em imagem radiológica, análise de dados experimentais de *chip* de gene para agrupar genes e relacionar sintomas ou doenças, análise de efeitos colaterais de drogas e eficácia de certos tratamentos, otimização de processos em um hospital e o relacionamento de dados de bem estar do paciente com qualificações do médico. Em sua obra, Carvalho (2005) utiliza as técnicas de mineração de dados baseadas em inteligência artificial aplicando-a à área da neurociência, trabalhando com padrões de comportamentos de pessoas com distúrbios mentais para diagnosticar possíveis tratamentos.

2.5.5. Varejo

Vários fatores podem contribuir para a necessidade de previsão de vendas, tais como a manutenção do cliente que não se frustra ao encontrar na loja o que deseja, o menor custo com estoques graças à manutenção de estoques mais ajustados às vendas futuras, a melhor alocação de vendedores em função da previsão das vendas para o futuro período, entre outros.

Os parâmetros importantes a serem considerados quando se analisa a disponibilidade de produtos em uma loja é a capacidade de produção e distribuição da indústria produtora do item, a existência ou não de propaganda realizada pelo produtor do item e o período do ano ou mês, dependendo do produto tratado.

Em seu livro, também a título de exemplificação, Carvalho (2005) descreve o emprego da mineração de dados em uma grande revendedora de automóveis que trabalhava com vários fabricantes nos seus diversos modelos. Observando sua perda de venda e de clientes a cada vez que não possuía o desejado carro em seus estoques, e contrapondo este fato com o alto custo de manutenção de grandes estoques deste produto durável e caro, a empresa resolveu, então, desenvolver um sistema de previsão de vendas. A empresa possuía um banco de dados de vendas de carros nos últimos 5 (cinco) anos e desejava um sistema de previsão capaz de avaliar as vendas 15 dias a frente, pois este era o tempo necessário para encomenda e transporte de novos itens.

Além da informação contida no banco de dados, foi necessário contextualizar cada dado de venda com outras informações, como a existência de propaganda realizada pelo fabricante, se a venda foi realizada em certos períodos do ano mais propícios à compra de automóveis e, também, se ela acontecia ao fim de cada mês, quando há um natural aquecimento das vendas. Como prever o futuro não é nada fácil, a maior quantidade de informação pertinente possível deve ser considerada em qualquer metodologia.

O procedimento acima descrito foi escolhido pelo uso de uma rede neural. O treinamento da rede neural foi feito com dados de quatro anos e meio, deixando os últimos seis meses do período de 5 (cinco) anos de vendas para a testagem da capacidade de previsão do sistema. O aprendizado mostrou-se eficiente, tendo um erro máximo de previsão em algumas semanas de 20%, porém, o erro médio se manteve dentro dos desejados 10%.

Desta forma, o sistema passou a prever as vendas dos próximos 15 dias, fornecendo mais tempo para a encomenda e o transporte do produto. A cada quatro semanas, a rede neural era ensinada de novo, incluindo-se os dados de mais 4 (quatro) semanas ocorridas seis meses antes e testando-se o erro de previsão utilizando-se sempre os últimos 6 (seis) meses de vendas, agora incluindo as últimas quatro semanas recentemente terminadas.

2.5.6. Educação

A mineração de dados educacionais é uma área recente de pesquisa que tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados

em ambientes educacionais. Atualmente, ela vem se estabelecendo como uma forte e consolidada linha de pesquisa, que possui grande potencial para melhorar a qualidade do ensino. Apesar dos esforços de pesquisadores brasileiros, essa área ainda é pouco explorada no país. No terceiro capítulo deste trabalho serão abordadas diversas aplicações do DCBD à educação.

No sentido deste contexto, considera-se que os sistemas de descoberta de conhecimento em banco de dados (DCBD) são utilizados nas organizações por ter uma grande influência nas atividades relacionadas à inteligência de negócio e à tomada de decisão. Com a globalização e a facilidade da troca de informações, as empresas necessitam se tornar mais competitivas para permanecerem no mercado. As informações existentes nas bases de dados das empresas podem representar um valor inestimável, quando utilizadas corretamente em tomadas de decisões. A utilização do processo de descoberta de conhecimento pode ser implícita para os usuários. Desta forma, os utilizadores do processo devem ter um entendimento sólido do negócio a que pretende empregar a mineração de dados e as técnicas que poderão apresentar melhores resultados. Dependendo da literatura, o processo de mineração de dados poderá ser uma fase do processo de DCBD ou ser considerado um sinônimo deste processo.

3. REVISÃO BIBLIOGRÁFICA

A mineração de dados deixou de ser apenas um processo de pesquisa científica no meio acadêmico. Atualmente, ela é aplicada em diversos ambientes em que existam dados, buscando conhecimentos implícitos para otimizar processos ou auxiliar na tomada de decisão. Desta forma, encontram-se diversos autores que relatam o emprego de tal técnica para diversos fins no Brasil.

Em sua dissertação de mestrado, Martinhago (2005) aplicou a mineração de dados para a descoberta de conhecimento sobre o processo seletivo da Universidade Federal do Paraná, onde utilizou a ferramenta WEKA. Nesse trabalho, foi aplicada a técnica de mineração de dados, árvore de decisão, através dos algoritmos de classificação J48 e J48.PART. Para isso, foi utilizada a base de dados do vestibular realizado no ano de 2004, que continha os dados coletados no questionário sócio-educacional preenchido pelos candidatos com o registro das notas obtidas pelo candidato nas provas e na redação, a opção pelo ENEM, a nota do ENEM, a média das notas dos candidatos e o resultado do vestibular. Os resultados da pesquisa constataram que nos cursos mais concorridos, os dados socioeconômicos e culturais do candidato são relevantes para o seu bom desempenho, o que não acontece com os cursos menos concorridos. A respeito do desempenho nas notas, interpretou-se que os candidatos que prestaram vestibular para cursos da área de exatas e foram aprovados obtiveram as melhores pontuações nas disciplinas das outras áreas, enquanto que na área de humanas ocorreu o inverso: as disciplinas da área de exatas contribuíram para a aprovação dos candidatos.

Penedo e Capra (2012) publicaram um trabalho no qual descrevem um processo de descoberta de conhecimento para buscar informações úteis para o ambiente educacional em sistemas utilizados na educação a distância (EAD), com intuito de investigar aquelas que auxiliem na identificação do padrão dos usuários que utilizam o sistema. O processo de DCBD foi aplicado em uma amostra de dados reais referentes ao *log* de acessos do sistema de EAD utilizado na Fundação Centro de Ciências e Educação Superior a distância, do Estado do Rio de Janeiro - Fundação CECIERJ/Consórcio CEDERJ. O banco de dados reunia informações referentes aos acessos efetuados ao sistema no ano de 2010. O padrão apontado pelo estudo demonstrou que a maioria dos usuários que utilizavam o sistema era oriundo de escolas estaduais, possuíam coeficiente de rendimento ruim (entre 0 a 4,9), eram do sexo feminino, e tinham idade acima de 28 anos. Identificou-se, também, a tendência de utilização

das ferramentas do sistema que dizem respeito às disciplinas, sendo essas relacionadas a aplicativos pouquíssimos utilizados.

No estudo de Coelho (2007), é exposto que, na Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), um programa de mineração de dados, depois de examinar milhares de alunos, forneceu a seguinte regra: se o candidato é do sexo feminino, trabalha e teve aprovações com boas notas, então, não faz matrícula nesta instituição. A regra encontrada confirma uma situação que parece estranha, mas, que ninguém havia pensado. Com uma reflexão justifica a regra fornecida pelo programa: de acordo com os costumes do Rio de Janeiro, uma mulher em idade de vestibular, se trabalha, é porque precisa, e, neste caso, deve ter feito inscrição para ingressar na universidade pública gratuita. Se obtiver boas notas, provavelmente foi aprovada na universidade pública, onde efetivará a matrícula.

A partir de um banco de dados com indicadores de desenvolvimento social, Murasse e Tsunoda (2010) aplicaram o processo de descoberta de conhecimento com uso de técnicas de mineração de dados e a ferramenta WEKA. Três hipóteses de correlação entre economia, demografia e saúde com a taxa de mortalidade foram testadas neste trabalho. O banco de dados selecionado suporta um programa com o nome de ”Objetivos de Desenvolvimento do Milênio” (ODM), formado por objetivos, metas e indicadores condescendidos pelo Brasil e por outros 190 países membros das Nações Unidas para melhorar indicadores sociais, ambientais e econômicos. Assim, com o processo de mineração de dados foi identificada uma correlação confiável que representa uma contribuição concreta para direcionamento de ações futuras do programa de desenvolvimento social ODM. O estudo revelou a proporção de empregados no mercado formal, o número de gestantes quanto ao número de consultas pré-natal e de natalidade e se existe correlação com a taxa de mortalidade infantil.

Braz e outros (2009) fizeram um estudo em que utilizaram mineração de dados para a área de segurança pública, no intuito de determinar os locais com maior criminalidade, definir perfis de vítimas e criminosos, identificar a existência de quadrilhas e *serial killers*, detectar quais dias da semana ocorrem mais delitos e, até mesmo, suas causas. Foi utilizado, para tanto, o banco de dados pertencente à Polícia Militar do Estado de Alagoas, cujo sistema é nomeado de “Sistema de Gestão de Ocorrências Policiais” (SISGOP), no qual ficam registrados os boletins das ocorrências policiais em algumas cidades de Alagoas.

Em sua monografia, Coelho (2007) aplicou o processo de DCBD na base de dados sobre os candidatos ao processo seletivo dos vestibulares ocorridos no ano de 2006 da Universidade Federal de Lavras (UFLA). Ele utilizou a ferramenta WEKA, sendo aplicadas as técnicas de mineração visual de dados, árvore de decisão, regras de associação e redes

neurais. Entre os diversos resultados obtidos, descobriu-se que a maioria dos candidatos que concluiu o Ensino Médio em até três anos antes do processo seletivo, não trabalhava, estava tentando ingressar na UFLA há mais de um ano, e já havia sido classificada em, pelo menos, um vestibular. Esta regra foi confirmada utilizando a técnica de redes neurais, pois a atividade de trabalho remunerado apresentou bastante influência na aprovação do candidato.

Em sua obra, Oliveira e Garcia (2004) descrevem a aplicação do processo de DCBD utilizando a técnica de regra de associação na mineração de dados. Foram utilizados os dados relacionados ao questionário socioeconômico e cultural aplicado durante o processo seletivo do Centro Universitário de Formiga, no ano de 2004. Foi deduzido que 43% dos candidatos que moram em uma distância máxima de 100 km de Formiga ficaram sabendo do processo seletivo por meio de panfletos. Outra relação, indicava que 41% dos candidatos que moram em uma distância máxima de 100 km de Formiga escolheram a UNIFOR devido ao conceito de que desfruta a Instituição, enquanto 51% das pessoas que residem na cidade de Formiga a escolheram por estarem mais perto de casa. Acredita-se, assim, que através destes resultados, pode-se tentar melhorar a qualidade das informações divulgadas no processo seletivo.

Carreira e outros (2012) utilizaram o processo de DCBD, aplicando a mineração de dados para obter informações úteis ao processo decisório de políticas públicas para área de ensino no Brasil. Para este trabalho, foi utilizada a base de dados referente ao Exame Nacional do Ensino Médio (ENEM) do ano de 2010 e a ferramenta WEKA. O banco de dados continha, inicialmente, 4.200.000 alunos cadastrados, perfazendo cerca de 1,5 bilhão de dados. O foco da pesquisa buscou associar o desempenho na prova objetiva com situações sócio-econômicas, grau de escolaridade dos pais, acesso à *internet* e o tipo de escola em que o estudante cursou o Ensino Médio (pública ou privada). Os dados analisados foram da Região Sul do País (estados do Rio Grande do Sul, Santa Catarina e Paraná). O trabalho demonstrou que a maioria dos estudantes que tem acesso a vários meios de comunicação não possui um rendimento satisfatório no ENEM (são regulares). Desta forma, pôde-se supor que o estudante ter acesso a diversos meios de informação não implica em melhor desempenho no ENEM; a partir disso, pode-se inferir que o modo de como as informações são utilizadas devem ser estudadas nas escolas.

Diante do exposto, considera-se que a mineração de dados é uma tarefa multidisciplinar que pode atender diversas áreas do conhecimento.

Neste capítulo, portanto, buscou-se expor alguns exemplos de aplicações do processo de busca de conhecimento na área educacional, revelando a variedade de informações que podem ser delineadas com o processo de DCBD.

4. METODOLOGIA

Para Prodanov e Freitas (2013), a metodologia abrange um processo científico de como estudar, compreender e avaliar os vários métodos disponíveis para a realização de uma pesquisa. Em um nível aplicado, com a metodologia é possível examinar, descrever e avaliar métodos e técnicas de pesquisa que possibilitam a coleta e o processamento de informações, tendo em vista o encaminhamento, a resolução de problemas e as questões de investigação. Este processo metodológico é a aplicação de procedimentos e técnicas que devem ser observados para construção do conhecimento, com a intenção de comprovar sua validade e serventia nos diversos domínios da sociedade.

Nesta seção, portanto, são apresentados o percurso metodológico deste estudo, bem como as atividades referentes aos processos utilizados para obter as informações resultantes desta pesquisa, as caracterizações deste estudo, a unidade de análise e observação na qual se fez este estudo, os instrumentos e o processo de coletas de dados e a estratégia de análises dos dados adotada.

4.1. Caracterização

De acordo com Prodanov e Freitas (2013), o método científico é um procedimento para conseguir um fim específico. Como a finalidade da ciência é a busca do conhecimento, ele é um conjunto de procedimentos que deve ser adotado com o objetivo de chegar a algum conhecimento. Também Lakatos e Marconi (2010) afirmam que a utilização de métodos científicos não é exclusiva da ciência, sendo possível usá-los para a resolução de problemas do cotidiano. Nesse sentido, este estudo utiliza o método de pesquisa hipotético-dedutivo, pois, conforme Prodanov e Freitas (2013), esse método trata-se de uma abordagem que se inicia com um problema ou uma lacuna no conhecimento científico, passando pela formulação de hipóteses e por um processo de inferência dedutiva, o qual testa a predição da ocorrência de fenômenos abrangidos pela referida hipótese. Neste método, as hipóteses podem ser modificadas até que não haja oposições entre a teoria e os experimentos.

Assim, fundamentada por Prodanov e Freitas (2013), uma pesquisa deve ter em vista o conhecimento de um ou mais aspectos de determinado assunto. Para tanto, ela deve ser sistemática, metódica e crítica.

O planejamento de uma pesquisa depende do problema que se deseja estudar, sua natureza, a situação temporal e do ambiente em que se encontra e o nível de conhecimento do

pesquisador. Dito isso, segundo Lakatos e Marconi (2010), nenhum tipo de pesquisa é autossuficiente, sendo que, na prática, é necessário mesclar os tipos de pesquisa, acentuando um ou outro tipo.

Baseado em Prodanov e Freitas (2013) e sob o ponto de vista de sua natureza, este estudo classifica-se como uma pesquisa aplicada. Os autores a definem como aquela que objetiva gerar conhecimentos para a prática dirigida à solução de problemas específicos, envolvendo verdades e interesses locais, a partir da aplicação de um conhecimento para se chegar a um novo. Porém, este estudo necessita da pesquisa básica para fundamentar conhecimentos novos e úteis para o avanço da aplicação prática prevista.

Quanto aos objetivos deste estudo, pode-se classificá-lo como uma pesquisa explicativa, já que, de acordo com o exposto por Prodanov e Freitas (2013), o que define a pesquisa explicativa é a maneira de buscar os motivos das ocorrências e suas causas, por meio do registro, da análise, da classificação e da interpretação dos fenômenos analisados. Por visar à identificação dos fatores que definem a ocorrência dos fenômenos, então se crê que esse tipo de pesquisa explica a razão e as circunstâncias dos fenômenos. Porém, conforme Prodanov e Freitas (2013), este estudo também apresenta aspectos da pesquisa descritiva, pois, esta pesquisa também descreve as características de determinada população e procura descobrir a frequência com que um perfil ocorre, suas características, causas e relações com outras ocorrências.

Sob o ponto de vista da forma de abordagem do problema, esta pesquisa pode ser classificada como qualitativa, pois Prodanov e Freitas (2013) também expõem que este tipo de pesquisa busca informação nas palavras dos participantes do estudo, e a base deste tipo de pesquisa pode ser a interpretação dos fenômenos, atribuição de significados a estes fenômenos e análises de documentos. Além disto, ela preocupa-se mais com o processo do que com o produto. Desta forma, este estudo necessita, também, da pesquisa quantitativa para exprimir em números a importância de informações classificadas e analisadas. Ainda de acordo com os autores, a pesquisa quantitativa classifica a relação entre variáveis para garantir a precisão dos resultados, evitando contradições no processo de análise e interpretação. Diante deste contexto, serão utilizados dados secundários, levantados pela Comissão Permanente de Vestibular e Processo Seletivo do IFMG.

Quanto aos procedimentos técnicos de uma pesquisa, acredita-se que esta seja uma forma de adquirir dados necessários para a execução de um estudo, que pode ser traduzido como delineamento. O delineamento é o planejamento da pesquisa, envolvendo diagramação, previsão de análise e interpretação de coleta de dados. Um elemento importante para a

identificação de um delineamento é o que define a forma como será realizada a coleta de dados. Segundo Prodanov e Freitas (2013), podem ser definidos dois grandes grupos de delineamentos: aqueles que se valem das chamadas fontes de papel (pesquisa bibliográfica e pesquisa documental) e aqueles cujos dados são fornecidos por pessoas (pesquisa experimental, pesquisa *ex-postfacto,* levantamento, estudo de caso, pesquisa-ação e pesquisa participante).

O estudo realizado, entre outras, é também uma pesquisa bibliográfica, pois, Severino (2007) diz que a pesquisa bibliográfica é um estudo extremamente importante, decorrente de pesquisas anteriores, e que utiliza dados teóricos já trabalhados por outros pesquisadores. Ela é elaborada a partir de material já publicado, composto principalmente de: livros, revistas, publicações em periódicos, artigos científicos, jornais, boletins, monografias, dissertações e teses, com o objetivo de colocar o pesquisador em contato direto com materiais já escritos sobre o assunto da pesquisa. Desta forma, utilizou-se também dessa pesquisa para adquirir conhecimento acerca do tema estudado.

Severino (2007) diz, ainda, que a pesquisa documental aparentemente pode assemelhar-se à pesquisa documental, mas, a natureza de origem das fontes de dados de ambas são diferentes. Prodanov e Freitas (2013) diz que, nessa tipologia de pesquisa, os documentos são classificados em dois tipos principais: fontes de primeira mão e fontes de segunda mão. Os documentos de primeira mão ainda não sofreram tratamento analítico, como: cartas, diários, base de dados, fotografias e gravações. Os documentos de segunda mão são os que, de alguma forma, já foram analisados, tais como: relatórios de pesquisa, relatórios de empresas e tabelas estatísticas. Portanto, neste estudo, empregou também a pesquisa documental de primeira mão, para realizar a busca de conhecimento implícito na base de dados dos processos seletivos (2011-2013) do IFMG – Campus São João Evangelista. Considerou-se, aqui, portanto, que as informações contidas na base de dados são primordialmente documentos institucionais.

Devido à necessidade de manipular diretamente as variáveis relacionadas com o objetivo do estudo, possibilitando o estudo da relação entre as causas e os efeitos de determinados fenômenos ou comportamento de candidatos nos processos seletivos, o procedimento no qual se balizou este trabalho foi o experimental. (MARCONI; LAKATOS, 2010).

Valeu-se, para tanto, do procedimento de estudo de caso, que, conforme Severino (2007), possibilita angariar dados relevantes na pesquisa, buscando-se a aplicação prática de conhecimentos para a solução do problema de descoberta de conhecimentos implícitos na

base de dados dos processos seletivos e que permite a descoberta de novos aspectos que não foram previstos inicialmente. Abalizado com as ideias de Prodanov e Freitas (2013), algumas características de estudo de caso aplicam-se ao grupo de candidatos inscritos nos processos seletivos dos anos de 2011, 2012 e 2013, porque este trabalho envolveu a análise de informações destes indivíduos de acordo com o assunto da pesquisa, a fim de traçar perfis destes candidatos.

4.2. Unidade de análise e observação

O Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) é composto por doze campi: Bambuí, Betim, Congonhas, Formiga, Governador Valadares, Ibirité (em implantação), Ouro Branco, Ouro Preto, Ribeirão das Neves, Sabará, Santa Luzia (em implantação) e São João Evangelista, além das unidades conveniadas de Pompéu, Piumhi, Oliveira, Bom Despacho, João Monlevade. A instituição também mantém polos de ensino a distância nos municípios de Alfenas, Betim, Cachoeira do Campo e Cataguases, bem como tem parceria para oferta do projeto especial do Proeja FIC nos municípios de Carandaí, Congonhas, Sabará, Iguatama, Perdões, Pompéu e Santa Bárbara. Atualmente, são disponibilizados mais de 60 cursos, divididos entre as modalidades de formação inicial e continuada, ensino técnico, ensino superior e pós-graduação *lato sensu*. (INSTITUTO FEDERAL DE EDUCAÇÃO CIÊNCIA E TECNOLOGIA DE MINAS GERAIS, 2013).

Antes de se tornar um campus do IFMG, no ano de 2008, o IFMG – Campus São João Evangelista era conhecido como Escola Agrotécnica Federal de São João Evangelista (EAFSJE). Nessa época, existiam estudantes matriculados na EAFSJE oriundos de diversas regiões do estado de Minas Gerais e do sul da Bahia, chegando a possuir, também, estudantes do estado do Espírito Santo, São Paulo e até de Angola, país africano. Com a criação de diversos Institutos Federais no Brasil, principalmente nas regiões do Sul e Norte de Minas Gerais, foram ampliadas, para as pessoas dessas regiões, as opções por instituições de ensino profissional. Possivelmente, a proximidade de novas instituições induziu as pessoas que residem mais distante da cidade de São João Evangelista perdessem o interesse pela instituição que se encontra mais distante de suas residências de origem, no centro-nordeste de Minas Gerais.

Nessa perspectiva, o IFMG – Campus São João Evangelista, possui três cursos técnicos: Técnico em Agropecuária, Técnico em Nutrição e Dietética e Técnico em Manutenção e Suporte em Informática. Para selecionar estudantes para estes cursos, a

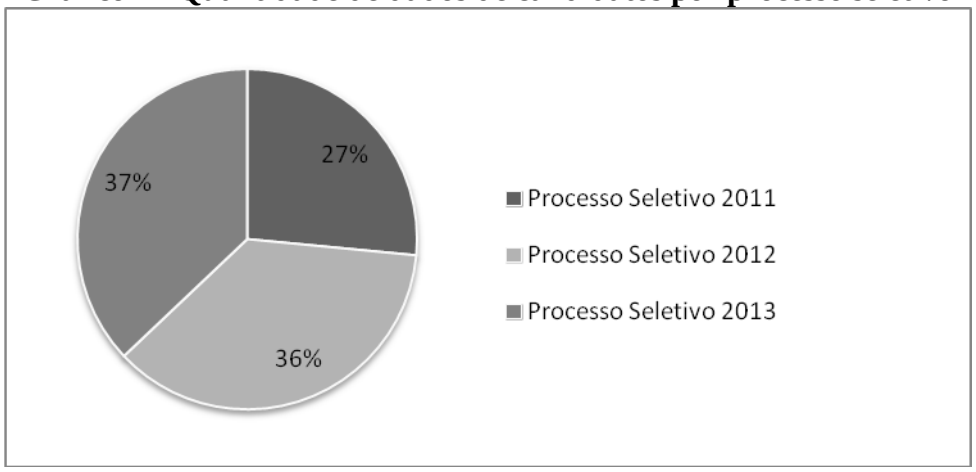
instituição realiza, no final de todo segundo semestre de cada ano, um processo seletivo. E para se inscrever nesse processo seletivo, os candidatos respondem a um questionário socioeconômico e cultural. Portanto, este estudo visou buscar padrões de comportamentos dos candidatos que se inscreveram nos processos seletivos para os anos letivos de 2011, 2012 e 2013. O objeto de estudo não abrangeu os questionários respondidos pelos candidatos aos cursos superiores, porque o interesse é descobrir padrões dos candidatos aos cursos técnicos integrados ou concomitantes ao Ensino Médio. Optou-se, assim, pelos três últimos processos seletivos, pelo motivo que somente um processo seletivo não possuiria uma grande quantidade de dados para ser analisada, conforme indica Carvalho (2005).

Atualmente, o IFMG possui dez campi em atividade e todos eles estão situados em regiões distintas, sendo que o perfil dos candidatos pode ser diferente em todos os campi. Conforme Carvalho (2005), o ideal é que se faça a aplicação da mineração de dados em uma unidade e, posteriormente, se aplique às demais, até obter o resultado íntegro de toda a instituição. Desta forma, pretende-se aplicar o processo de descoberta de conhecimento em banco de dados apenas no IFMG – Campus São João Evangelista.

No intuito de acompanhar as características referentes ao processo seletivo da instituição, a COPEVES (Comissão Permanente de Vestibular e Exame de Seleção) utiliza planilhas eletrônicas para analisar os dados dos candidatos inscritos. As ferramentas utilizadas para fazer o levantamento estatístico dos dados do processo seletivo geram informações que podem ser estudadas e medidas por meio de histogramas, gráficos de linhas, gráficos de setores, média aritmética e média ponderada. Porém, o levantamento estatístico aplicado pelo IFMG não é o processo ideal para se obter todas as informações úteis que se encontram na base de dados dos processos seletivos da instituição. Baseado em Carvalho (2005), estas ferramentas estatísticas fornecem informações quantitativas e não permitem predizer informações ou encontrar comportamentos correspondentes ao perfil dos candidatos, sendo possível representar apenas a quantidade de uma ocorrência, mas não as relações entre elas. Assim, uma maneira de enriquecer as informações é incluir alguma técnica que encontre padrões de comportamento.

O banco de dados abrangia todas as questões do questionário socioeconômico respondidas por todos os candidatos que se inscreveram nos três últimos exames de seleção. Inicialmente, A base de dados dos processos seletivos adquirida com a COPEVES continha dados de 3560 candidatos, distribuídos conforme o Gráfico 1, a seguir.

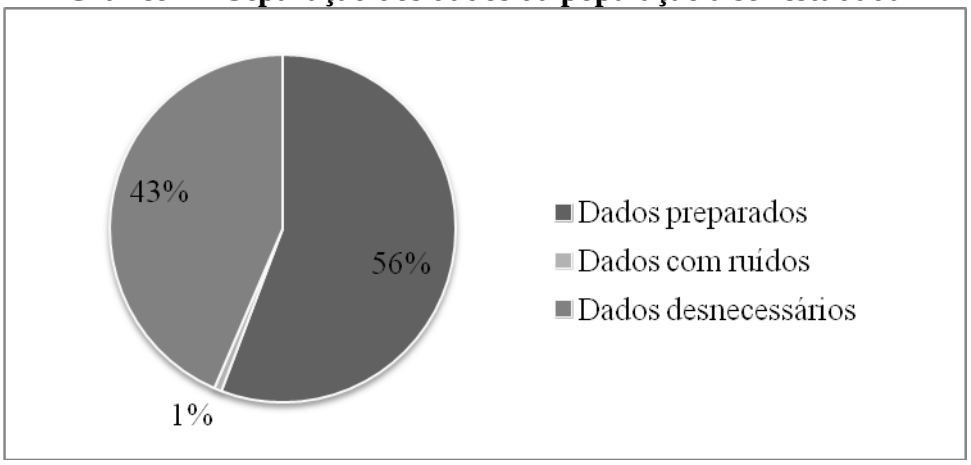
Gráfico 1 - Quantidade de dados de candidatos por processo seletivo



Fonte: Dados da pesquisa

Conforme o Gráfico 2, abaixo, com a necessidade de retirar os dados com ruídos e que pertenciam aos candidatos aos cursos superiores, eliminou-se 44% dos dados inúteis ao estudo.

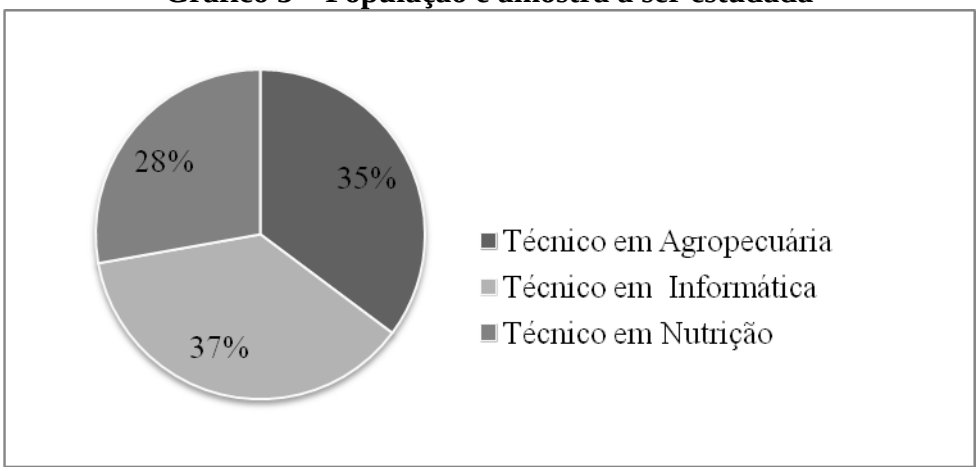
Gráfico 2 – Separação dos dados da população a ser estudada



Fonte: Dados da pesquisa

Conforme o Gráfico 3, restaram, portanto, dados de 1985 candidatos aos três cursos técnicos. Como se trabalhou com os dados de todos os candidatos inscritos no processo seletivo para os cursos técnicos, a amostra também corresponderá ao valor da população. Foram analisados, então, 699 candidatos do curso Técnico em Agropecuária, 734 do curso Técnico em Manutenção e Suporte em Informática e 552 do curso Técnico em Nutrição e Dietética.

Gráfico 3 – População e amostra a ser estudada



Fonte: Dados da pesquisa

4.3. Instrumento de coleta de dados

Como o interesse deste estudo foi a busca de conhecimento implícito na base de dados dos processos seletivos do IFMG – Campus São João Evangelista, definiu-se que a técnica adequada seria a mineração de dados. Para este estudo, então, foi adotada a abordagem do processo de descoberta de conhecimento em banco de dados definida por Adriaans e Zantinge. A abordagem desses autores, porém, não exige que se tenha uma ampla experiência na aplicação do processo de DCBD e permite retornar a etapas anteriores para modificar as atividades de uma etapa. Assim, por trabalhar com três tarefas de mineração de dados, foi necessário retornar o processo para adaptar a base de dados para cada tarefa. (BRAUNER, 2003).

Como esta pesquisa visou à busca de conhecimento implícito sobre a base de dados do processo seletivo do IFMG – Campus São João Evangelista, foram utilizadas as respostas dos candidatos aos questionários socioeconômicos e cultural criado pela COPEVES, conforme Anexos A, B e C deste trabalho, com isso, utilizou-se dados secundários. Os questionários utilizados para os três processos seletivos apresentam disparidades quanto a algumas questões. Desta forma, na intenção de homogeneizar os questionários, buscou-se criar um novo, que pode ser visto no Anexo D desta pesquisa. Ele possui indagações comuns aos questionários dos três processos seletivos e o resultado da situação do candidato quanto à aprovação no curso inscrito. A partir desse novo questionário, recuperaram-se todas as informações pertinentes ao mesmo.

Vale ressaltar que o acesso a estes dados foi liberado mediante assinatura de um termo de compromisso ao IFMG, que ressalvava a utilização das informações apenas para fins

acadêmicos, conforme é demonstrado no Anexo E. Os dados para aplicar o processo de DCBD foram disponibilizados pela COPEVES do IFMG em uma planilha eletrônica. No entanto, para buscar as informações nessa base de dados foi selecionada a ferramenta de mineração de dados WEKA (*Waikato Enviroment for Knowledge Analysis*), devido a esta ferramenta ser um *software* livre e gratuito, além de sua qualidade e precisão ser referenciada por diversos autores e por existir um vasto acervo bibliográfico que ensina como utilizá-la. Será, portanto, utilizada a interface *Explorer* do WEKA, devido à sua facilidade de uso.

As técnicas de mineração de dados utilizadas para a busca de conhecimento sobre o processo seletivo foi a técnica de regra de Associação, Agrupamento e Classificação. Estas técnicas conseguiriam encontrar padrões satisfatórios para a descoberta de conhecimento na base de dados estudada, pois são as principais técnicas de mineração de dados, segundo os autores citados na fundamentação teórica deste trabalho. Além disso, estas três regras são implementadas na ferramenta WEKA com os principais algoritmos para busca de padrões na mineração de dados, sendo essas regras constantemente atualizadas, de acordo com a melhora ou criação de novos algoritmos para a própria ferramenta. (MACHINE LEARNING GROUP AT THE UNIVERSITY OF WAIKATO, 2013).

Para cada técnica foi utilizado um algoritmo. Assim, para a regra de associação optou-se pelo algoritmo *Apriori*, por ele ser, segundo o referencial teórico, o principal algoritmo de associação existente na ferramenta WEKA e por ser o algoritmo com melhores resultados nessa técnica. Já para a técnica de classificação, optou-se pelo algoritmo J48, que, segundo o referencial teórico, este algoritmo possui alto grau de otimização e é bem definido para a criação de árvores na técnica de classificação. Por sua vez, o algoritmo *SimpleKMeans* foi utilizado na técnica de agrupamento, porque ele permite definir a quantidade de agrupamentos e segundo o referencial teórico, ele possui alto grau de confiabilidade para a criação de classes que formaram os agrupamentos.

Os recursos materiais e imateriais que foram utilizados para a execução deste estudo foram providenciados pelo autor, como: computador, *internet*, os *softwares* de planilha eletrônica, editor de texto e de mineração de dados. O computador utilizado possuía, como características mínimas, um sistema operacional *Windows Seven*, processador *core* 2 *duo* e quatro gigabytes de memória RAM. Devido à existência de uma grande quantidade de conteúdo relacionado ao processo de DCBD na *internet*, esta foi utilizada para adquirir e complementar o acervo bibliográfico necessário para o estudo e para encontrar os *softwares* empregados neste trabalho.

4.4. Coleta de dados

Para aumentar o conhecimento teórico e prático em relação ao tema trabalhado, realizou-se a revisão bibliográfica na primeira fase deste estudo. Inicialmente, foram estudados os conceitos do processo de DCBD e as funcionalidades da ferramenta WEKA, dando ênfase à etapa de mineração de dados. Para entender como o processo de descoberta de conhecimento em banco de dados vem sendo empregado em algumas organizações, estudou-se relatos de instituições públicas e privadas que empregaram este processo para auxiliar na tomada de decisão. Para realizar a pesquisa bibliográfica e documental, foram utilizados diversos materiais referentes ao processo de DCBD, como artigos científicos, livros, tutoriais, sítios, monografias e teses de pós-graduação.

A segunda fase do projeto constituiu-se na aquisição da base de dados referente aos processos de seleção de estudantes para o primeiro semestre dos anos de 2011, 2012 e 2013. A base de dados consistia em todas as questões dos questionários socioeconômicos respondidas pelos candidatos no exame de seleção. Os dados para aplicar o processo de DCBD foram disponibilizados pela COPEVES do IFMG em uma planilha eletrônica, no formato *xlsx*, sendo que, para trabalhar com esses dados, foi utilizado, inicialmente, o *software Microsoft Excel* 2010, planilha eletrônica, que permite visualizar e manipular todos os dados exportados para o formato *xlsx*.

No banco de dados dos três últimos processos seletivos do IFMG – Campus São João Evangelista existiam 3560 (três mil e quinhentos e sessenta) registros de estudantes que se inscreveram para ingressar na instituição. Esta base de dados apresentava alguns campos inconsistentes ou em branco, além de alguns dados referentes aos candidatos aos cursos superiores, que são irrelevantes para a pesquisa, como é mostrado na Figura 8, a seguir. Diante disso, esta fase compreendeu a limpeza desses registros desnecessários para o processo de DCBD no processo seletivo, restando, então, 1985 registros de candidatos aos cursos técnicos.

Figura 8 – Dados dos processos seletivos antes da tarefa de limpeza

139	Tecnologia em Silvicultura	23	64.5	inscrição confirmada
140	Técnico em Agropecuária - Integrado	148	10	inscrição confirmada
141				
142	Bacharelado em Agronomia			inscrição confirmada
143	Bacharelado em Agronomia	21	95.5	inscrição confirmada
144	Licenciatura em Matemática	16	85.5	inscrição confirmada

Fonte: Dados da pesquisa

Para fazer estas manipulações e a homogeneização, utilizou-se o *software* de planilha eletrônica *Excel*. Através deste *software*, então, retiraram-se as *tuplas* que não eram comuns nos três questionários e executou-se a tarefa de limpeza. Para que os dados ficassem fáceis de visualizar, reduziram-se respostas (dados) que apresentavam grande quantidade de caracteres, para tanto, respostas para onde o candidato mora, como "Moro em cidade próxima, até 50 km" foram reduzidas para "Até 50 km", por exemplo, e respostas de onde os pais trabalham, como "Na agricultura, no campo, em fazenda ou na pesca", reduziu-se a "Na agricultura. No campo". Os dados do tipo número foram convertidos para dados nominais. Neste caso, onde o conteúdo de um campo era o número "2" converteu-se para "dois", pois, o algoritmo *Apriori* trabalha com tipos nominais. Para que algumas árvores fossem melhor visualizadas, devido ao tamanho, campos como "Trabalhador do setor informal" foi modificado para "Tr. informal".

No formulário do processo seletivo de 2011 no qual constavam 24 (vinte e quatro) questões, existiam indagações não presentes nos outros dois questionários utilizados, do tipo: com que frequência os candidatos acessavam músicas e danças, e se ele participava de grupos políticos. Já no questionário do processo seletivo de 2012, também constavam 24 (vinte e quatro) questões e tinham como incomum a interrogação de qual a frequência com que o candidato acessa informações no rádio. O questionário do processo seletivo de 2013, por sua vez, apresentava 35 questões, as quais indagavam informações que não estavam presentes nos outros questionários anteriores, como, por exemplo, a quantidade de livros que o candidato lia por ano, a quantidade de livros que o candidato possuía em casa, frequência com que acessa revistas, presença da mãe e do pai, tipo de escola onde o candidato estudou os ensinos primário, fundamental e médio, a sua participação na renda familiar e faixa etária do candidato.

Quanto ao enriquecimento dos dados, juntaram-se as respostas dos questionários socioeconômicos aos resultados obtidos pelos candidatos na prova de seleção. Com isso, constituiu-se uma relação na qual dispôs as principais informações dos candidatos comuns

nos três formulários. Com a utilização de dados de apenas um formulário, de acordo com Carvalho (2005), iriam existir poucos dados para definir algum padrão, por este motivo, optou-se por utilizar as respostas dos três questionários de forma homogênea, de acordo com o Anexo D deste trabalho.

As etapas de limpeza, enriquecimento e separação dos dados para o estudo, demandaram desprenderam muito tempo para serem cumpridas. Após ter definido quais os dados dos processos seletivos a serem trabalhados, definiu-se como trabalhar com eles. Para isto, codificou-se os dados e os adaptou com informações adicionais para os formatos eletrônicos exigidos para utilização no *software* de mineração de dados, conforme é demonstrado na Figura 9, a seguir. Diante deste contexto, foram utilizadas vinte questões, em comum nos três questionários, para constituírem o novo banco de dados que foi adequado aos requisitos de funcionamento da ferramenta WEKA, conforme é demonstrado abaixo:

Figura 9 - Transformação dos dados para o formato, ARFF, aceito pelo WEKA

```
@relation exame-Selecao-completo

@attribute Curso {'Agropecuária','Informática','Nutrição'}
@attribute Status_Opção {'Aprovado','Excedente'}
@attribute Conheceu_Vest {'Banner/folhetos','Outdoor','Outros meios','Rádio','Televisão','Website (inter
@attribute Escolheu_IFMG {'É próximo de minha residência','Pela qualidade de ensino prestada','Por falta
@attribute Escolheu_Curso {'Melhor possibilidade no mercado','Por falta de opção','Por influência de ter
@attribute Distancia_IFMG {'Distrito município','Até 50 KM','Acima de 50 KM', 'Mesma cidade'}
@attribute Renda_Familiar {'acima de 10 salários','Até um salário mínimo','De 1 a 3 salários','de 3 a 5
@attribute Estudou_Escola_Publica {'Nunca','Parcialmente','Sempre'}
@attribute Instrucao_Pai {'Analfabeto','fundamental (1 a 4)','fundamental (5 a 8)','médio','pós-graduaçã
@attribute Instrucao_Mae {'Analfabeta','fundamental (1 a 4)','fundamental (5 a 8)','médio','pós-graduaçã
@attribute Trabalho_Pai {'Aposentado','Ausente','Desempregado','Funcionário público ou militar','Na agri
@attribute Trabalho_Mae {'Aposentada','Ausente','Desempregado','Funcionário público ou militar','Na agri
@attribute Local_Acesso_Internet {'Em casa', 'Escola', 'Lan house','Na casa de parentes', 'Na casa de vi
@attribute Acessa_Cinema_Teatro {'Muito', 'Pouca', 'Nunca'}
@attribute Acessa_Internet {'Muito', 'Pouca', 'Nunca'}
@attribute Acessa_Livros {'Muito', 'Pouca', 'Nunca'}
@attribute Acessa_Televisão {'Muito', 'Pouca', 'Nunca'}
@attribute Pessoas_Em_Casa {'Um','Dois','Três','Quatro','Cinco', 'Mais que cinco', 'Moro sozinho'}
@attribute Tem_Filhos {'Sim', 'Não'}
@attribute Recebe_Beneficio {'Sim', 'Não'}

@data

'Informática','Aprovado','Outros meios','Por indicação de terceiros','Por influência de terceiros','Até
'Agropecuária','Excedente','Outros meios','Por indicação de terceiros','Sempre quis este curso','Acima d
'Nutrição','Aprovado','Website (internet)','Pela qualidade de ensino prestada','Sempre quis este curso',
'Agropecuária','Excedente','Banner/folhetos','Porque é gratuito','Melhor possibilidade no mercado','Acim
'Nutrição','Excedente','Website (internet)','Pela qualidade de ensino prestada','Sempre quis este curso'
'Agropecuária','Excedente','Banner/folhetos','Porque é gratuito','Sempre quis este curso','Até 50 KM','A
'Agropecuária','Aprovado','Outros meios','Pela qualidade de ensino prestada','Sempre quis este curso','M
```

Fonte: Dados da pesquisa

No processo de codificação definiu-se que a relação (base de dados) teria o nome de *exame-selecao-completo* e que todos os dados seriam do tipo nominal (*string*), para que fosse utilizado pelas três técnicas, principalmente pela associação, que trabalha apenas com dados nominais. Para realizar a limpeza e o enriquecimento da base de dados, utilizou-se o resultado dos candidatos nas provas do exame de seleção, que foram transformados em um arquivo de

texto com a extensão *arff*, extensão exigida pelo *software* WEKA. A ferramenta utilizada para fazer o pré-processamento dos dados foram os *softwares* editores de textos Notepad ++ e o bloco de notas, ambos instalados no ambiente *Windows Seven.*

Para a realização da etapa de mineração de dados, embasou-se na fundamentação teórica, buscando responder à pergunta norteadora dessa pesquisa e analisando as informações de acordo com as hipóteses criadas. Por meio das respostas à questão-problema, foi, então, possível implementar as estratégias de aplicação das técnicas de mineração de dados. Os padrões buscados com estas técnicas objetivam identificar a probabilidade de um item, correlacionado à presença de outro item ocorrer, e de encontrar grupos comportamentais com classes semelhantes, dentro da base de dados.

Na etapa de execução do processo de mineração, como já dito anteriormente, utilizou-se o *software* WEKA (Waikato Enviroment for Knowledge Analysis) para minerar a base de dados dos processos seletivos. Utilizando as técnicas de mineração de dados foi, então, realizada uma análise de toda a base de dados, sem desconsiderar nenhum campo, com intuito de identificar alguma informação implícita que não estivesse prevista nas hipóteses levantadas. Em seguida, o processo de mineração de dados deste projeto trabalhou com cada hipótese em específico. Portanto, nestas análises foram desconsideradas as instâncias que não poderiam influenciar nas informações buscadas para responder a uma hipótese.

Com a regra de agrupamento, procurou-se trabalhar com todas as *tuplas*, pois, nesta técnica, a grande quantidade de atributos não atrapalha na visualização das informações, permitindo, assim, que sejam criados grupos bem definidos, que foram gerados pelos grupos de resultado do exame, grupos dos cursos e grupos da distância de que moram do IFMG – Campus São João Evangelista. Procurou-se, assim, trabalhar com dois ou três agrupamentos, porque, conforme Carvalho (2005), com poucos agrupamentos obtêm-se grupos bem definidos e confiáveis. Procurou-se, ainda, utilizar as informações obtidas com a técnica de agrupamento para subsidiar, juntamente com as hipóteses, a aplicação das técnicas de classificação e associação.

Com a técnica de classificação, buscou-se utilizar poucas *tuplas*, conforme Tang, Teinbach e Kumar (2009) indicam: que para análises de árvores de decisão que não contenham uma grande (milhões) quantidade de dados. As árvores geradas com a aplicação da regra de classificação baseada nas hipóteses, também não havia a necessidade de trabalhar com os dados dos 20 (vinte) atributos do questionário socioeconômico e cultural ao mesmo tempo. Vale ressaltar que a ferramenta WEKA oferece a função de excluir os atributos que não se deseja utilizar para a aplicação de uma determinada técnica de mineração de dados.

Já com a técnica de associação, testou-se a sua aplicação utilizando todos os atributos, pois, Carvalho (2005) diz que a regra de associação é a técnica mais susceptível a encontrar conhecimentos implícitos, derivados de associações desprezadas no cotidiano. Posteriormente, eliminou-se da base de dados para a análise os dados que possuíam maior ocorrência entre as respostas dos candidatos, pois estes dados forçavam a ocorrência de associações de confiança muito altas e sem informações úteis. Neste caso, é provável que quase todos os candidatos não possuam filhos e assistam muito televisão. Desta forma, essa relação apareceu constantemente, que se o candidato assiste muito televisão, ele não tem filhos e se ele não tem filhos, ele assiste muito televisão, ou, que se o candidato é aprovado, ele assiste muito televisão e se excedente, ele assiste muito televisão também. Por isso, desconsideraram-se os dados muitos repetitivos, como frequência que assiste televisão e se tem filhos.

4.5. Estratégia de análises dos dados

Na etapa de execução do processo de mineração, portanto, utilizou-se o *software* WEKA para minerar a base de dados dos processos seletivos. Utilizando-se as técnicas de mineração de dados, realizou-se análise completa da base de dados sem desconsiderar nenhum campo, com intuito de identificar alguma informação implícita que não esteja nas hipóteses levantadas. Em seguida, o processo de mineração de dados deste projeto trabalhou com cada hipótese em específico. Portanto, em algumas análises foram desconsideradas algumas instâncias que não poderiam influenciar nas informações buscadas para responder a uma hipótese.

Como se seguiu o processo de DCBD da abordagem de Adriaans e Zantinge, a estratégia de análise de dados são as táticas especificadas por este processo. A ferramenta WEKA expõe os resultados da análise de cada técnica de forma diferente, nas regras de associação utilizadas, sendo criadas árvores de decisão; já nas técnicas de associação e agrupamento empregadas, os resultados são apenas em forma de textos. Os resultados da mineração de dados são, assim, interpretados e descritos, de acordo com as informações obtidas com as hipóteses e com as informações não planejadas, que podem surgir com a mineração de dados. Dessa maneira, foi criada uma análise crítica dos correlacionamentos entre as informações obtidas, sendo que alguns padrões encontrados foram confirmados com a aplicação de outra técnica.

Como a visualizações de resultados do WEKA para as técnicas de agrupamento e associação são na forma de texto e estes textos possuem uma limitação gráfica de fonte, eles serão inseridos como texto e não como uma Figura. Também algumas visualizações das árvores de decisões, criadas pela técnica de classificação, costumam ser de difícil visualização devido ao seu tamanho, por isso elas foram adaptadas para ficarem visíveis nos resultados deste estudo.

Na apresentação dos resultados, quando se fala de pais dos candidatos, refere-se às pessoas que o candidato respondeu como seu pai e sua mãe, de acordo com os campos dos formulários criados pela COPEVES (Comissão Permanente de Vestibular e Exame de Seleção). Desta maneira, de acordo com quem o candidato indicou ser o seu pai, na tarefa de classificação, procurou-se referir ao substantivo pai no singular.

No caso das regras de associação que se encontra em formatos textuais, essas são explicadas na subseção das regras de associação do capítulo de análise dos resultados em formato de alíneas numeradas, onde são descritas as confianças das regras em relação ao suporte dos dados.

Conforme Tang, Steinbach e Kumar (2009), é inviável a utilização das árvores de decisão com muitos nodos na tarefa de classificação em uma base de dados que não seja muito grande (dezenas de milhares de registros), pois isso pode criar diversos problemas, entre eles, os mais comuns são a replicação de sub-árvores, o *overfitting* e o *underfitting*. O *overfitting* é quando a árvore fica muito grande, fazendo com que a taxa de erro de teste aumente, e o *underfitting* é o aumento dos erros de teste e de aprendizagem do algoritmo, quando se aplica uma pequena árvore em uma base de dados grandiosa (dezenas de milhares de dados) ou uma árvore como muitos nós em uma base de dados grande (milhares de dados). Como os atributos correspondem aos nós da árvore, utilizou-se poucos atributos para cada árvore, buscando evitar, principalmente, o *overfitting*.

5. RESULTADOS

Nesta sessão são apresentados os testes e os resultados obtidos com a aplicação de técnicas de mineração de dados, por meio da utilização da ferramenta WEKA, no intuito de obter perfis dos candidatos que participaram dos processos seletivos do IFMG-SJE (Instituto Federal de Minas Gerais – Campus São João Evangelista), realizados no mês de dezembro dos anos de 2011, 2012 e 2013, para a seleção de estudantes dos respectivos anos. Os resultados apresentados foram obtidos a partir da aplicação das técnicas de mineração de dados: Classificação (Árvore de Decisão), Regras de Associação e Agrupamento.

Diante deste contexto, com a realização deste trabalho buscou-se encontrar informações que possibilitem analisar o perfil dos candidatos aos processos seletivos do IFMG – Campus São João Evangelista, bem como encontrar regras capazes de induzir algum conhecimento útil para a instituição.

5.1. Regra de Associação

Conforme Carvalho (2005) propõe, procurou-se trabalhar com os padrões que possuíam confiança acima de 70 %. Empregando a técnica de associação, na qual foi utilizado o algoritmo *Apriori*, as seguintes regras foram geradas:

a) Desc_Curso1=Técnico em Agropecuária Desc_Sit_Resultado_Candidato_Concurso=Aprovado 209 ==> distancia-cidade-IFMG=moro em outra cidade acima de 50 KM 175 <conf:(0.83)>:
   - Esta regra indica que 83% das pessoas que se candidataram ao curso Técnico em Agropecuária e foram aprovados no processo seletivo, são candidatos que moram em cidades com distância acima de 50 km de São João Evangelista. Vale ressaltar que os valores 209 e 175 correspondem ao suporte para o cálculo da confiança, que neste caso é 83%.

b) Curso=Técnico em Manutenção e Suporte em Informática Distancia_IFMG=Moro em outra cidade acima de 50 KM 397 ==> Status_Opcao=Excedente 317 <conf:(0.80)>:
   - Esta regra indica que 80% das pessoas que se candidataram ao curso Técnico em Manutenção e Suporte em Informática e que moram em alguma cidade

acima de 50 Km de distância de São João Evangelista, são excedentes, ou seja, não foram aprovados no processo seletivo.

c) Curso=Técnico em Manutenção e Suporte em Informática Distancia_IFMG=Moro em cidade próxima. até 50 KM Acessa_Televisão=Muito 97 ==> Status_Opção=Aprovado 89 <conf:(0.92)>:
   - Esta regra indica que 92% das pessoas que se candidataram ao curso técnico em Manutenção e Suporte em Informática moram em cidades com distância de até 50 km de São João Evangelista e assistem muito televisão, são aprovados no processo seletivo.

d) Status_Opção=Aprovado 728 ==> Escolheu_IFMG=Pela qualidade de ensino prestada 657 <conf:(0.9)>:
   - Esta regra indica que 90% dos candidatos que foram aprovados escolheram o IFMG pela qualidade de ensino prestada.

e) Acessa_Internet=Muito Local_Acesso_Internet=Em casa Acessa_Televisão=Muito Acessa_Livros=Pouca Acessa_Cinema_Teatro=Pouca 304 ==> Status_Opção=Excedente 273 <conf:(0.9)>:
   - Esta regra indica que 90% dos candidatos que acessam frequentemente a internet em casa, assistem muito televisão, acessam poucos livros e que frequentam o cinema poucas vezes, não são aprovados no processo seletivo.

f) Acessa_Internet=Muito Local_Acesso_Internet=Em casa Acessa_Livros=Muito 216 ==> Status_Opção=Aprovado 199 <conf:(0.92)>:
   - Esta regra indica que 92% dos candidatos que acessam muito a internet em casa e que acessam muitos livros são aprovados no processo seletivo.

g) Instrucao_Mae=Superior 388 ==> Trabalho_Mae=Funcionário público ou militar 307 <conf: (0.79)>:
   - Esta regra indica que 79% das mães com curso superior são funcionárias públicas ou militares.

h) Conheceu_Vest=Banner/folhetos Escolheu_Curso=Sempre quis este curso Acessa_Televisão=Muito 103 ==> Moro na mesma cidade onde está o campus 87 <conf: (0.85)>:
   - Esta regra indica que 85% dos candidatos que conheceram o processo seletivo através de *banners* e folhetos, escolheram um curso porque sempre o quiseram e assistem muito televisão, são candidatos da cidade de São João Evangelista.

i) Conheceu_Vest=Banner/folhetos Escolheu_Curso=Sempre quis este curso Acessa_Internet=Muito Acessa_Televisão=Muito 117 ==> Moro em outra cidade acima de 50 KM 98 <conf: (0.84)>:
   - Esta regra indica que 84% dos candidatos que conheceram o processo seletivo através de *banners* e folhetos, escolheram um curso porque sempre o quiseram, acessam muito *internet* e assistem muito televisão, são candidatos de cidades próximas à cidade de São João Evangelista.

j) Curso=Técnico em Agropecuária – Integrado Acessa_Internet=Pouca Renda_Familiar=Até um salário mínimo 127 ==> Status_Opção=Aprovado 102 <conf: (0.80)>:
   - Esta regra indica que 80% dos candidatos que se inscreveram para o curso Técnico em Agropecuária, acessam pouca internet e possuem renda familiar de até um salário mínimo, são aprovados no processo seletivo.

k) Renda_Familiar= de 3 a 5 salários 80 ==> Curso=Técnico em Manutenção e Suporte em Informática 80 <conf: (1)>:
   - Esta regra indica que 100% dos candidatos em que a renda familiar é de 3 a 5 salários mínimos e que acessam muito a *internet*, se inscrevem para o curso Técnico Manutenção e Suporte em Informática.

l) Distancia_IFMG=Moro na mesma cidade onde está o campus Escolheu_IFMG=Pela qualidade de ensino prestada 85 ==> Curso=Técnico em Nutrição e Dietética 71 <conf:(0.84)>:
   - Esta regra indica que 84% dos candidatos que moram em São João Evangelista e que optaram pelo IFMG pela qualidade de ensino prestada, se inscrevem para o curso Técnico em Nutrição em Dietética.

m) Distancia_IFMG=Moro em outra cidade acima de 50 KM Curso=Técnico em Agropecuária – Integrado Escolheu_IFMG=Pela qualidade de ensino prestada 327 ==> Escolheu_Curso=Melhor possibilidade no mercado 272 <conf: (0.83)>:
   - Esta regra indica que 83% dos candidatos que moram em cidades distantes de São João Evangelista – acima de 50 km, que optam pelo curso Técnico em Agropecuária e escolheram o IFMG pela qualidade de ensino prestada, são os candidatos que escolhem este curso por melhor possibilidades no mercado.

Com a tarefa de associação, percebeu-se, então, que a regra mais interessante é a que diz que alunos que optam pelo curso de manutenção e suporte são provenientes de famílias

com maior renda no processo seletivo, moram em cidades vizinhas de São João Evangelista e possuem maior acesso à informação, pois, contraria a ocorrência no curso técnico em agropecuária, cujos candidatos aprovados possuem perfil de baixa renda familiar.

5.2. Agrupamento

A aplicação desta técnica teve como objetivo criar agrupamentos que permitem identificar o perfil dos candidatos. Para a técnica de agrupamento utilizado, empregou-se o algoritmo de agrupamento *SimpleKMeans*.

Conforme se pode verificar na Figura 10, abaixo, criou-se três grupos (ou *cluster*): o *cluster* 0 (zero) representa o curso Técnico em Agropecuária - com 584 pessoas, o *cluster* 1 (um) representa o curso Técnico em Nutrição e Dietética – com 697 pessoas, e o *cluster* 2 (dois) representa o curso Técnico em Manutenção e Suporte em Informática – com 704 pessoas. Os valores abaixo do número que identifica o agrupamento representam a quantidade de pessoas que estão neste grupo.

Figura 10 - Agrupamento por curso

Attribute	0 (584)	1 (697)	2 (704)
Status_Opção	Aprovado	Excedente	Excedente
Conheceu_Vest	Outros meios	Outros meios	Outros meios
Escolheu_IFMG	Pela qualidade de ensino prestada	Pela qualidade de ensino prestada	Pela qualidade de ensino prestada
Escolheu_Curso	Sempre quis este curso	Sempre quis este curso	Sempre quis este curso
Distancia_IFMG	Moro em outra cidade acima de 50 KM	Moro em outra cidade acima de 50 KM	Moro em outra cidade acima de 50 KM
Renda_Familiar	Até um salário mínimo	Até um salário mínimo	De 1 a 3 salários
Estudou_Escola_Publica	Sempre	Sempre	Sempre
Instrucao_Pai	fundamental (1 a 4)	fundamental (1 a 4)	médio
Instrucao_Mae	fundamental (1 a 4)	fundamental (1 a 4)	médio
Trabalho_Pai	Na agricultura. no campo	Na agricultura. no campo	Na indústria. no comércio
Trabalho_Mae	Funcionário público ou militar	Trabalhador do setor informal	Funcionário público ou militar
Local_Acesso_Internet	Em casa	Lan house	Em casa
Acessa_Cinema_Teatro	Nunca	Nunca	Pouca
Acessa_Internet	Muito	Pouca	Muito
Acessa_Livros	Muito	Muito	Muito
Acessa_Televisão	Muito	Muito	Muito
Pessoas_Em_Casa	Quatro	Cinco	Quatro
Tem_Filhos	Não	Não	Não
Recebe_Beneficio	Não	Não	Não

Fonte: Dados da Pesquisa

Portanto, baseado nas informações geradas pelo agrupamento da Figura 10, nota-se que a forma de conhecimento do processo seletivo (*Conheceu_Vest*) é "outros meios", então, conclui-se que os meios de comunicação como folhetos, *banners*, *internet*, televisão, rádio e *outdoor* possuem pouca influência para o conhecimento do processo seletivo do IFMG. Outro comportamento constante e generalizado é que os candidatos buscam ingressar no IFMG pela qualidade de ensino prestada, sempre quiseram o curso em que se inscreveram, moram em outra cidade acima de 50 km (cento e cinquenta quilômetros) de distância da cidade de São João Evangelista, sempre estudaram em escola pública e assistem muito televisão, além de não possuírem filhos e acessar muitos livros.

Neste agrupamento pelos três cursos, ainda é possível encontrar diferenças significativas, fazendo um paralelo. Os candidatos aos cursos Técnico em Agropecuária e Técnico em Nutrição e Dietética possuem renda familiar de até um salário mínimo, nunca foram em cinemas e seus pais possuem Ensino Fundamental incompleto, enquanto os candidatos ao curso Técnico em Manutenção e Suporte em Informática possuem renda familiar de 1 a 3 salários mínimos, vão ao cinema esporadicamente e seus pais possuem Ensino Médio.

O pai de candidatos ao curso Técnico em Agropecuária e o curso Técnico em Nutrição e Dietética tendem a trabalhar na agricultura ou no campo (*Trabalho_Pai*). Já os pais dos candidatos ao curso Técnico em Manutenção e Suporte em informática trabalham na indústria ou no comércio. Por sua vez, as mães dos candidatos ao curso Técnico em Agropecuária e o curso Técnico em Manutenção e Suporte em Informática são funcionárias públicas ou militares, enquanto as mães dos candidatos do curso Técnico em Nutrição e Dietética trabalham no setor informal.

Neste contexto, verifica-se que os candidatos aos cursos Técnico em Agropecuária e ao curso Técnico em Manutenção e Suporte em informática acessam frequentemente a *internet* em casa e possuem uma família composta por quatro pessoas. Já os candidatos ao curso Técnico em Nutrição e Dietética acessam *internet* em *lan houses* e o fazem esporadicamente, sendo que suas famílias são compostas por cinco pessoas.

Conforme Figura 11, a seguir, criou-se dois grupos (ou *cluster*), sendo que o *cluster* 0 (zero) representa os candidatos aprovados - com 1144 pessoas, e o *cluster* 1 (um) representa os candidatos excedentes – com 841 pessoas. Nestes agrupamentos percebe-se a divergência entre candidatos aprovados e excedentes. Os candidatos aprovados possuem renda familiar de um a três salários mínimos e os excedentes possuem renda familiar menor que um salário mínimo. O pai de candidatos aprovados possui Ensino Fundamental e trabalha na indústria ou

no comércio, e as mães possuem Ensino Médio e trabalham como funcionárias públicas ou militares, porém, o pai e mãe de candidatos excedentes possuem Ensino Fundamental incompleto e trabalham, respectivamente, na agricultura ou no campo e no setor informal.

Figura 11 – Agrupamento de candidatos aprovados e reprovados

Cluster centroids:

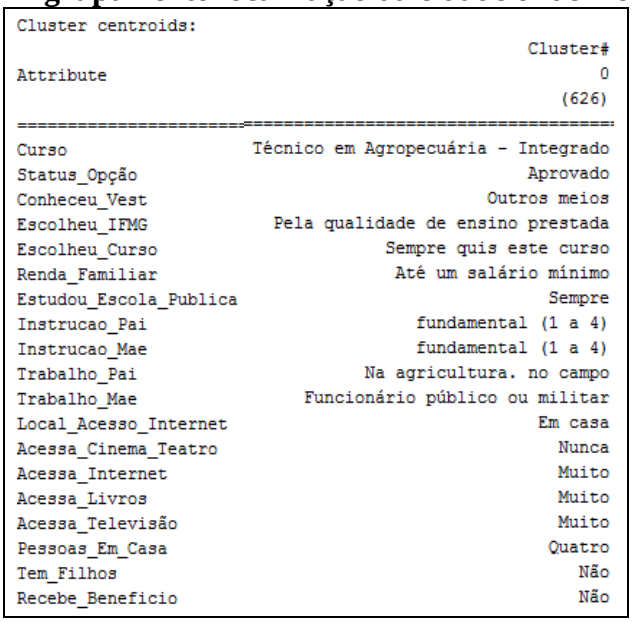
Attribute	Cluster# 0 (1144)	1 (841)
Curso	Técnico em Agropecuária - Integrado	Técnico em Manutenção e Suporte em Informática
Conheceu_Vest	Outros meios	Outros meios
Escolheu_IFMG	Pela qualidade de ensino prestada	Pela qualidade de ensino prestada
Escolheu_Curso	Sempre quis este curso	Sempre quis este curso
Distancia_IFMG	Moro em outra cidade acima de 50 KM	Moro em outra cidade acima de 50 KM
Renda_Familiar	De 1 a 3 salários	Até um salário mínimo
Estudou_Escola_Publica	Sempre	Sempre
Instrucao_Pai	fundamental (1 a 4)	fundamental (1 a 4)
Instrucao_Mae	médio	fundamental (1 a 4)
Trabalho_Pai	Na indústria. no comércio	Na agricultura. no campo
Trabalho_Mae	Funcionário público ou militar	Trabalhador do setor informal
Local_Acesso_Internet	Em casa	Lan house
Acessa_Cinema_Teatro	Pouca	Nunca
Acessa_Internet	Muito	Pouca
Acessa_Livros	Muito	Muito
Acessa_Televisão	Muito	Muito
Pessoas_Em_Casa	Quatro	Cinco
Tem_Filhos	Não	Não
Recebe_Beneficio	Não	Não

Fonte: Dados da pesquisa

Vale ressaltar, ainda com relação à Figura 11, que os candidatos aprovados possuem *internet* em casa, vão ao cinema esporadicamente, acessam *internet* frequentemente e possuem grupo familiar composto por quatro pessoas. Já os candidatos excedentes tendem a acessar *internet* em *lan house*, nunca foram ao cinema, acessam *internet* poucas vezes e possuem grupo familiar composto por cinco pessoas.

Já conforme Figuras 12 e 13, a seguir, que apesar de estarem separadas apresentam a mesma aplicação do agrupamento, foram gerados três agrupamentos: o *cluster* 0 (zero), com 626 pessoas, corresponde a quem mora em cidades em um distância acima de 50 km de São João Evangelista, o *cluster* 1(um), com 657 pessoas, corresponde a quem mora em São João Evangelista e o *cluster* 2 (dois), com 702 pessoas, corresponde a quem mora em cidades próximas de São João Evangelista, até 50km.

Figura 12 – Agrupamento localização da cidade onde mora, parte 1

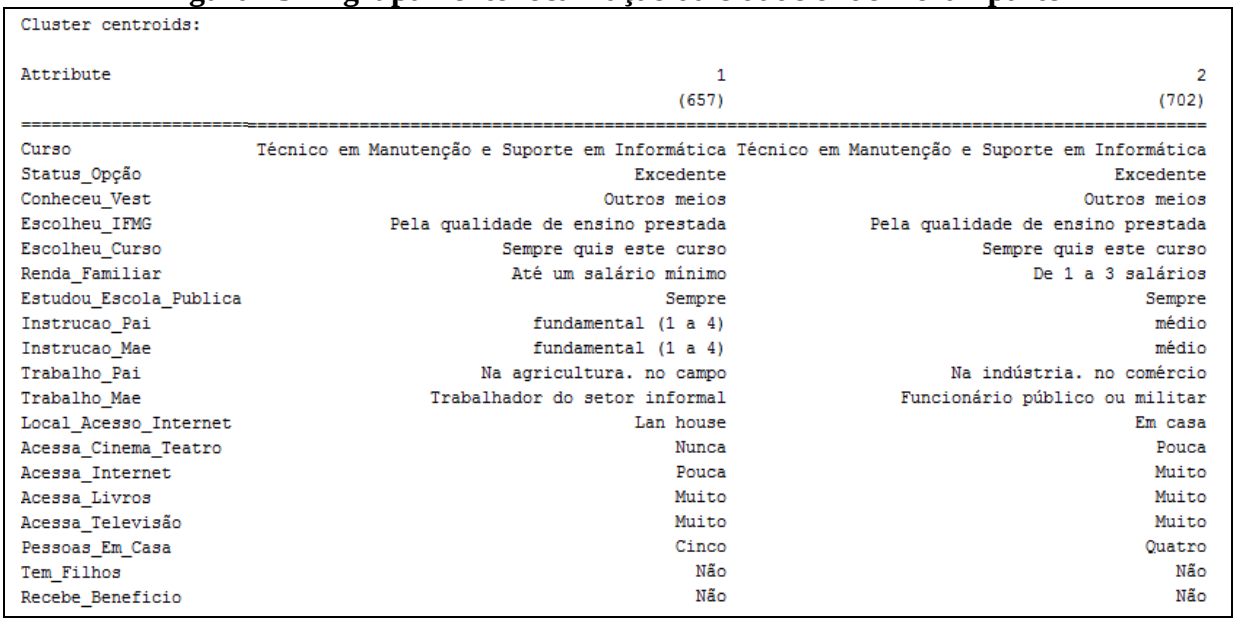


```
Cluster centroids:
                                                Cluster#
Attribute                                              0
                                                   (626)
=======================================================
Curso                    Técnico em Agropecuária - Integrado
Status_Opção                                      Aprovado
Conheceu_Vest                                  Outros meios
Escolheu_IFMG           Pela qualidade de ensino prestada
Escolheu_Curso                    Sempre quis este curso
Renda_Familiar                   Até um salário mínimo
Estudou_Escola_Publica                            Sempre
Instrucao_Pai                      fundamental (1 a 4)
Instrucao_Mae                      fundamental (1 a 4)
Trabalho_Pai                  Na agricultura. no campo
Trabalho_Mae            Funcionário público ou militar
Local_Acesso_Internet                          Em casa
Acessa_Cinema_Teatro                             Nunca
Acessa_Internet                                  Muito
Acessa_Livros                                    Muito
Acessa_Televisão                                 Muito
Pessoas_Em_Casa                                  Quatro
Tem_Filhos                                          Não
Recebe_Beneficio                                    Não
```

Fonte: Dados da pesquisa

Desta forma, nota-se que o *cluster* 0 (Fig. 12) mostra que o perfil dos candidatos que moram distantes da cidade onde situa o campus do IFMG, são aprovados, possuem renda familiar inferior a um salário mínimo, os pais e mães possuem Ensino Fundamental incompleto, possuem *internet* em casa e procuram fazer o curso Técnico em Agropecuária.

Figura 13 - Agrupamento localização da cidade onde mora - parte 2



```
Cluster centroids:

Attribute                                                1                                                2
                                                     (657)                                            (702)
=============================================================================================================
Curso                  Técnico em Manutenção e Suporte em Informática Técnico em Manutenção e Suporte em Informática
Status_Opção                                       Excedente                                        Excedente
Conheceu_Vest                                   Outros meios                                     Outros meios
Escolheu_IFMG            Pela qualidade de ensino prestada                Pela qualidade de ensino prestada
Escolheu_Curso                     Sempre quis este curso                           Sempre quis este curso
Renda_Familiar                    Até um salário mínimo                            De 1 a 3 salários
Estudou_Escola_Publica                             Sempre                                           Sempre
Instrucao_Pai                       fundamental (1 a 4)                                            médio
Instrucao_Mae                       fundamental (1 a 4)                                            médio
Trabalho_Pai                   Na agricultura. no campo                         Na indústria. no comércio
Trabalho_Mae                 Trabalhador do setor informal             Funcionário público ou militar
Local_Acesso_Internet                          Lan house                                        Em casa
Acessa_Cinema_Teatro                              Nunca                                            Pouca
Acessa_Internet                                   Pouca                                            Muito
Acessa_Livros                                     Muito                                            Muito
Acessa_Televisão                                  Muito                                            Muito
Pessoas_Em_Casa                                   Cinco                                            Quatro
Tem_Filhos                                          Não                                              Não
Recebe_Beneficio                                    Não                                              Não
```

Fonte: Dados da pesquisa

Conforme a Figura 13, acima, os candidatos que habitam em São João Evangelista (*cluster* 1) são excedentes, possuem renda familiar menor que um salário mínimo, os pais e

mães possuem Ensino Fundamental incompleto e trabalham, respectivamente, na agricultura e no setor informal. Estes candidatos também não possuem *internet* em casa e possuem grupo familiar de cinco pessoas.

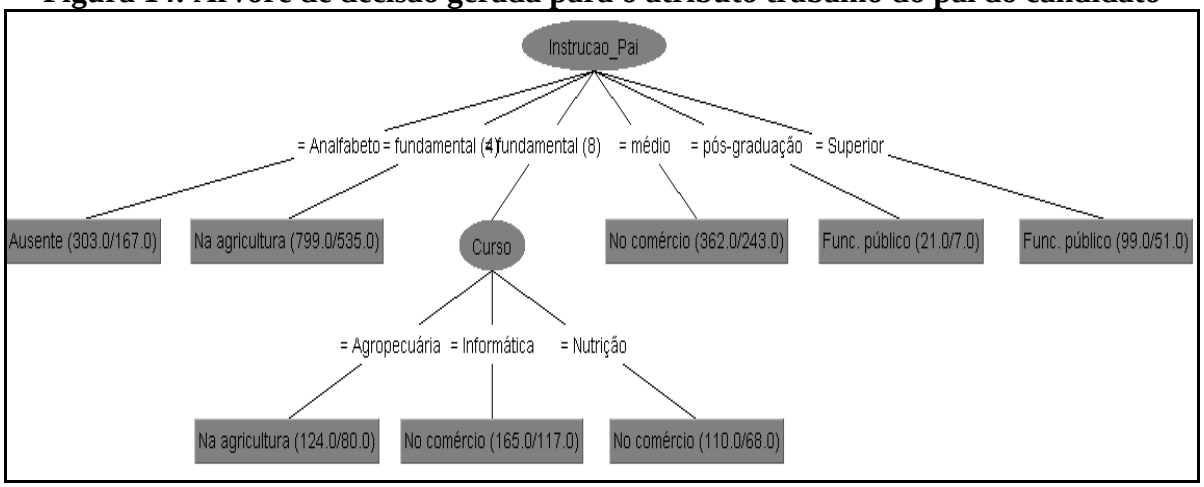
Já os candidatos de cidades vizinhas (*cluster* 2) (Fig.13), procuram o curso Técnico em Manutenção e Suporte em Informática, possuem renda familiar de 1 a 3 salários mínimos, os pais e mães possuem Ensino Médio e trabalham, respectivamente, na indústria ou no setor público. Estes candidatos acessam muito vários meios de comunicação e possuem *internet* em casa.

5.3. Classificação

A aplicação da regra de classificação utilizando o algoritmo J48 na base de dados do processo seletivo do IFMG tem como objetivo gerar árvores de decisão, para categorizar os dados analisados.

A árvore da Figura 14, por exemplo, foi gerada a partir dos dados referentes ao "curso", "trabalho pai", "instrução pai", orientada pelo trabalho do pai do candidato.

Figura 14: Árvore de decisão gerada para o atributo trabalho do pai do candidato



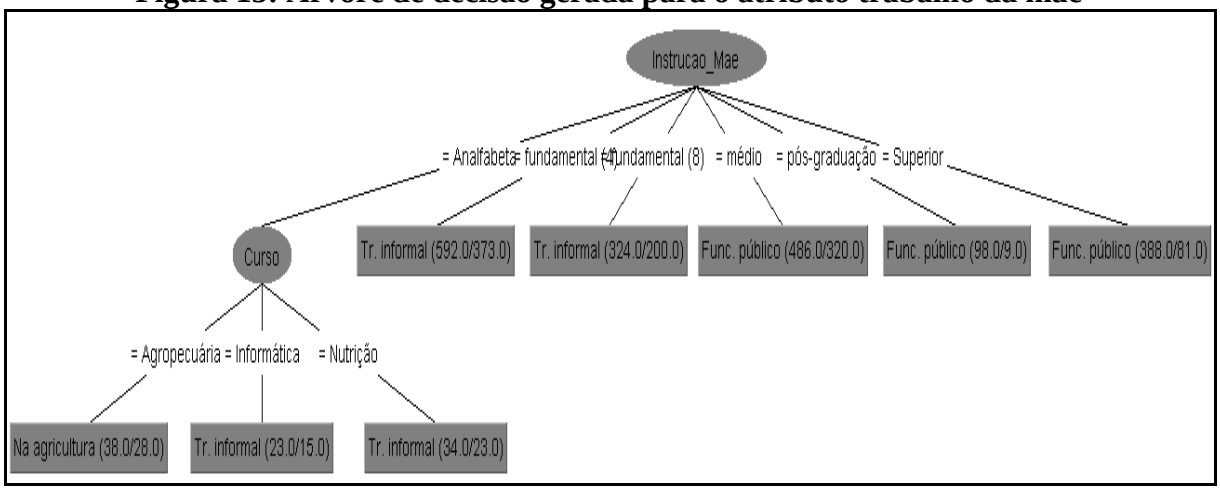
Fonte: Dados da pesquisa

De acordo com a categoria criada pela árvore acima, nota-se que 303 dos pais que são analfabetos 167 são ausentes. Já dos 799 pais que possuem Ensino Fundamental incompleto, 535 trabalham na agricultura. Nota-se, também, que 80 pais de candidatos que possuem o Ensino Fundamental completo e que seus filhos procuram o curso Técnico em Agropecuária, trabalham na agricultura e no campo. Porém, 185 pais que possuem Fundamental completo,

cujos filhos candidatam aos cursos Técnico em Manutenção e Suporte em Informática e Técnico em Nutrição e Dietética, trabalham na indústria ou no comércio. (Fig. 14).

Já a árvore da Figura 15, a seguir, foi gerada a partir dos dados referentes ao "curso", "trabalho mãe" e "instrução da mãe", orientada pelo trabalho da mãe. De acordo com a árvore gerada, 66 mães são analfabetas e trabalham no setor informal e 410 são funcionárias públicas; o restante são mães que possuem Ensino Fundamental completo ou incompleto e trabalham no setor informal.

Figura 15: Árvore de decisão gerada para o atributo trabalho da mãe

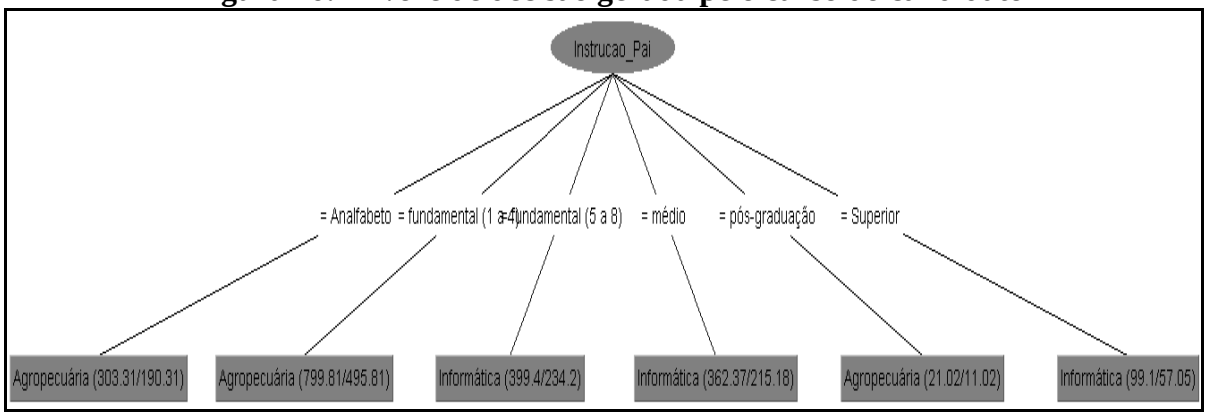


Fonte: Dados da pesquisa

De acordo com esta árvore, portanto, percebe-se que as mães que possuem Pós-graduação, tendem a ser trabalhadoras do setor público. As mães que possuem Ensino Fundamental incompleto ou completo, de acordo com a categorização gerada, tendem a trabalhar no setor informal. Nota-se, ainda, que a quantidade de mães analfabetas de candidatos ao curso Técnico em Agropecuária é superior aos outros cursos.

A Figura 16, abaixo, por sua vez, exibe uma árvore gerada pelos atributos "curso" e "instrução do pai", orientada pelo curso. Com a árvore gerada, percebe-se que os candidatos cujos pais são analfabetos, possuem Ensino Fundamental incompleto ou possuem Pós-graduação, irão inscrever no processo seletivo para o curso Técnico em Agropecuária. Já os candidatos cujos pais possuem Ensino Fundamental completo, Ensino Médio ou curso superior, irão se candidatar ao curso Técnico em Manutenção em Informática. (Fig. 16)

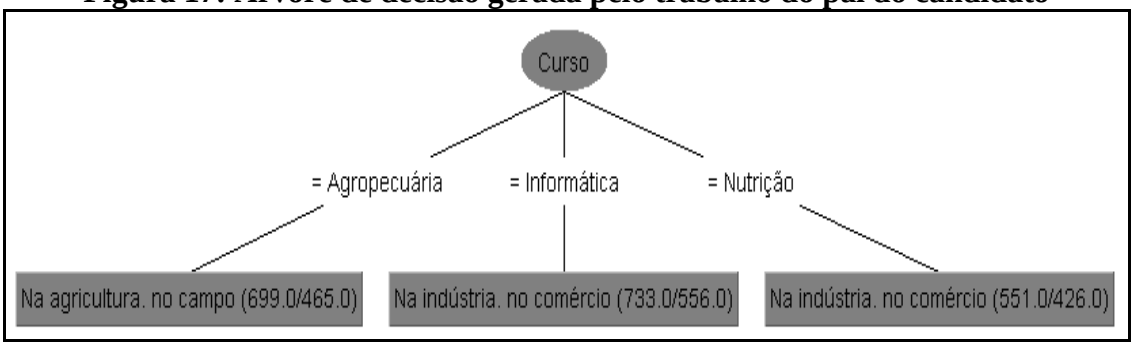
Figura 16: Árvore de decisão gerada pelo curso do candidato



Fonte: Dados da pesquisa

A Figura 17, mais abaixo, demonstra uma árvore gerada pelos atributos "curso" e "trabalho do pai", orientada pelo trabalho do pai. Diante desta árvore, percebe-se que os pais de 465 (67%) candidatos ao curso Técnico em Agropecuária trabalham na agricultura e no campo. Diante disso, 978 ou (80%) dos pais de candidatos ao curso Técnico em Manutenção e Suporte em Informática e ao curso Técnico em Nutrição e Dietética, trabalham na indústria e no comércio.

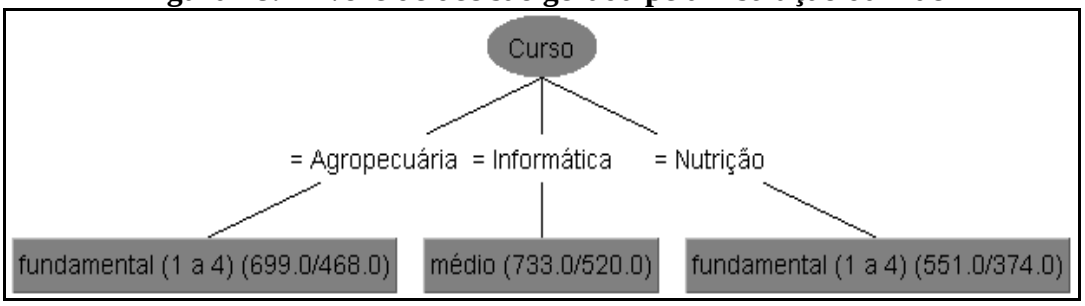
Figura 17: Árvore de decisão gerada pelo trabalho do pai do candidato



Fonte: Dados da pesquisa

Conforme Figura 18, a seguir, percebe-se a influência da escolaridade da mãe do candidato na escolha do curso. A árvore foi gerada pelos atributos "curso" e "Instrução_Mãe", orientada por "Instrução_Mãe". Diante disso, nota-se que 468 das mães dos candidatos ao curso Técnico em Agropecuária e 374 das mães dos candidatos ao curso Técnico em Nutrição e Dietética, possuem Ensino Fundamental incompleto. Já 520 mães dos candidatos ao curso Técnico em Manutenção e Suporte em Informática possuem Ensino Médio, como pode ser verificado abaixo:

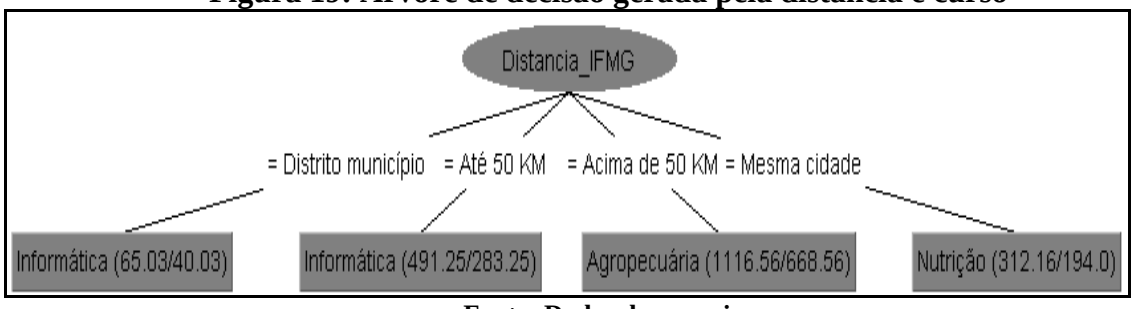
Figura 18: Árvore de decisão gerada pela instrução da mãe



Fonte: Dados da pesquisa

Já a árvore da Figura 19 foi gerada pelos atributos "curso", "Conheceu_Vestibular*"* e *"*Distancia_IFMG*"*, sendo orientada pelo curso. Nota-se que 62% dos candidatos que moram em distritos do município de São João Evangelista optam pelo curso Técnico em Manutenção e Suporte em Informática. Dos candidatos que são da cidade de São João Evangelista, 62% se inscrevem para o curso Técnico em Nutrição e Dietética. Os 58% das pessoas que moram em cidades próximas de até 50 km de São João Evangelista são candidatos ao curso Técnico em Manutenção e Suporte em Informática. Assim sendo, nota-se que 668 candidatos ao curso Técnico em Agropecuária são de cidades com distância acima de 50 km de São João Evangelista. (Fig. 19).

Figura 19: Árvore de decisão gerada pela distancia e curso

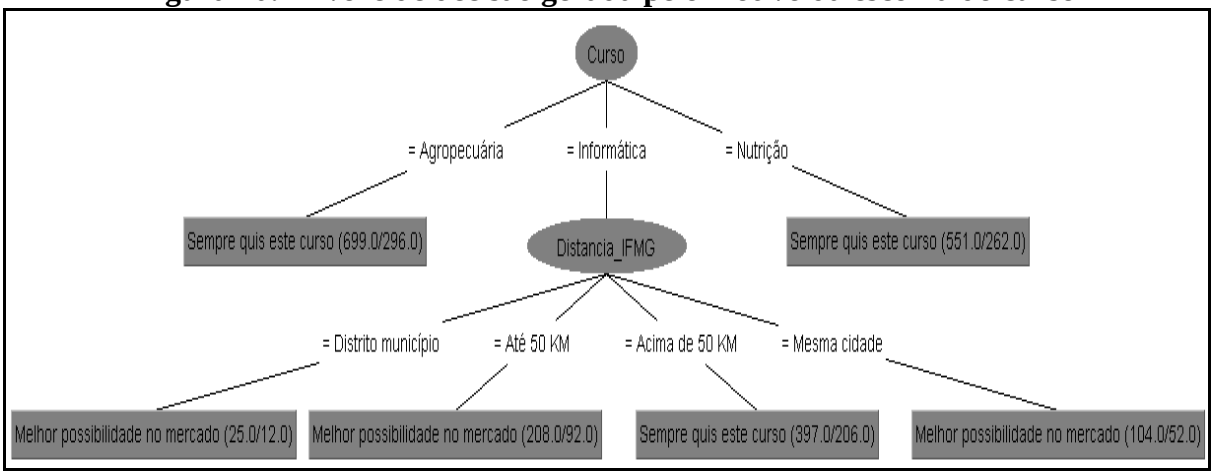


Fonte: Dados da pesquisa

Por sua vez, a árvore da Figura 20 foi gerada pelos atributos "curso", "Escolheu_IFMG", "Escolheu_Curso", "Distancia_IFMG", sendo orientada por "Escolheu_Curso". Com esta classificação, nota-se que 296 candidatos alegam que sempre quiseram o curso Técnico em Agropecuária e 262 alegam que sempre quiseram o curso Técnico em Nutrição e Dietética. Dos estudantes que se inscreveram para o curso Técnico em Manutenção e Suporte em Informática, 48% moram em distritos de São João Evangelista e buscam o curso por melhores possibilidades de mercado, 44% dos estudantes que moram em cidades distantes até 50 km escolheram o curso também por melhores oportunidades de

mercado, assim como 50% dos candidatos que moram em São João Evangelista, que colocaram a mesma opção como resposta a porque escolheram o curso. Já 52% dos candidatos que moram acima de 50 km de São João Evangelista escolheram o curso porque sempre o quiseram. (Fig. 20).

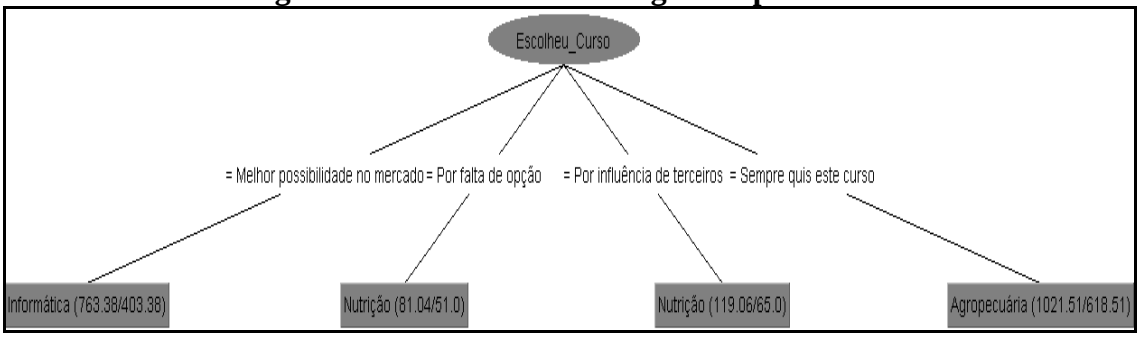
Figura 20: Árvore de decisão gerada pelo motivo da escolha do curso



Fonte: Dados da pesquisa

Conforme Figura 21, a seguir, gerou-se, pelos atributos “curso” e “Escolheu_Curso”, uma árvore, sendo orientada pelo curso. Com a categoria gerada por esta árvore, nota-se que 53% ou 403 dos candidatos que escolheram o curso por melhores possibilidades no mercado estão inscritos para o curso Técnico em Manutenção e Suporte em Informática. Dos candidatos que escolheram o curso por falta de opção, 63% são do curso Técnico em Nutrição e Dietética e dos candidatos que escolheram o curso por influência de terceiros, 55% são também candidatos deste curso. Também se percebe que 61 % ou 618 candidatos que escolheram o curso porque o sempre quisera, optam pelo curso Técnico em Agropecuária, como é demonstrado na Figura 21:

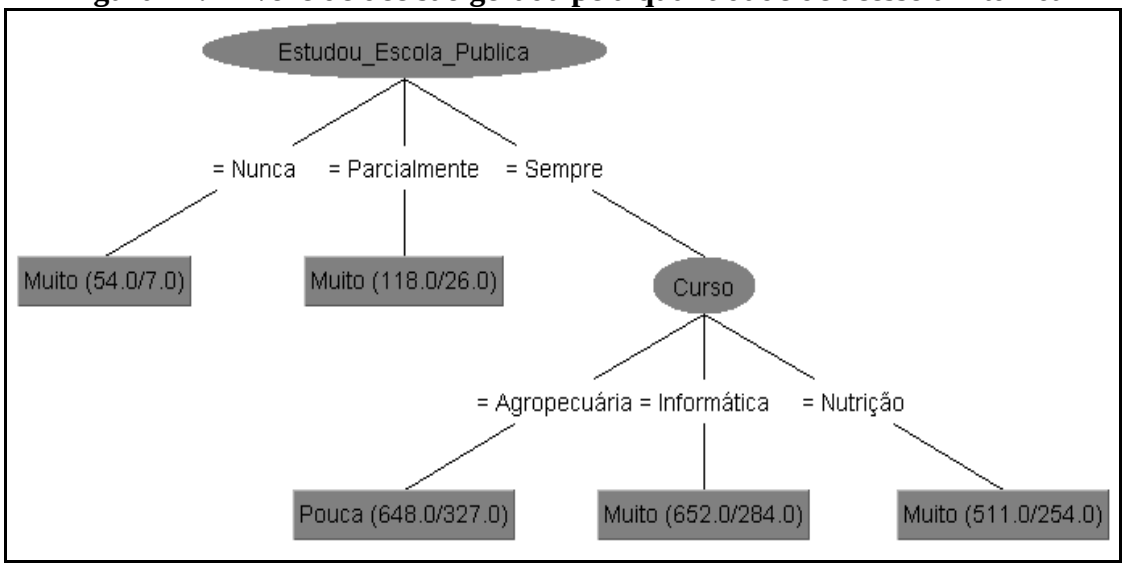
Figura 21: Árvore de decisão gerada pelo curso



Fonte: Dados da pesquisa

Baseado na árvore da Figura 22, abaixo, gerada pelos atributos “curso”, “Estudou_Escola_Pública”, “Acessa_Internet”, foi orientada pela frequência que acessa a *internet*. Dos estudantes que sempre estudaram em escola pública e optaram pelo curso Técnico em Agropecuária, 51% acessam pouco a *internet*. Os candidatos que sempre estudaram em escola pública do curso Técnico em Manutenção e Suporte em Informática 44% acessam muito a *internet*. Já no curso de Técnico em Nutrição e Dietética, 50% dos candidatos que sempre estudaram em escola pública, também acessam muito a *internet*.

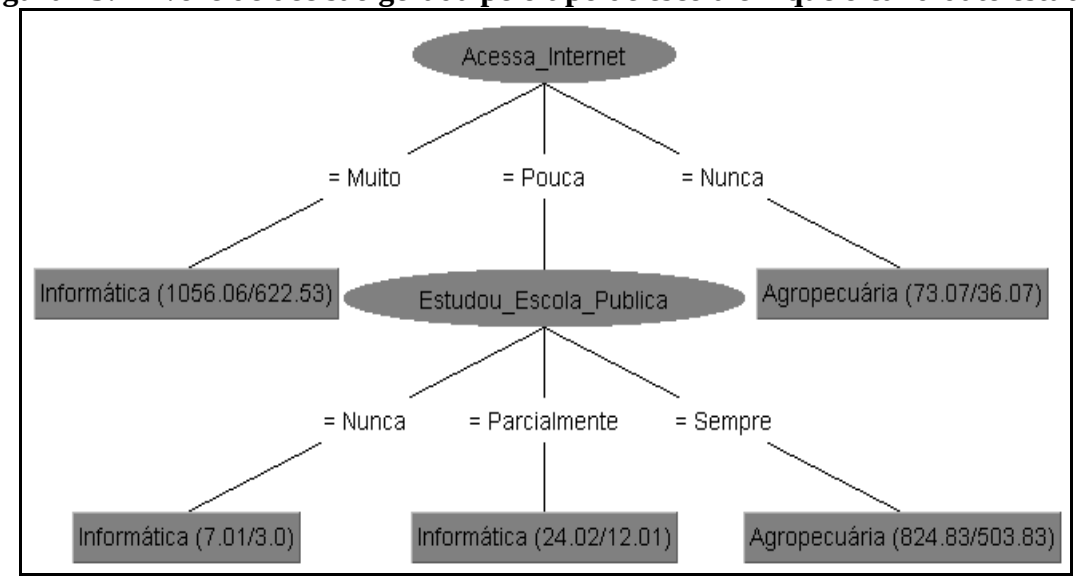
Figura 22: Árvore de decisão gerada pela quantidade de acesso a *internet*



Fonte: Dados da pesquisa

A árvore da Figura 23, por sua vez, foi gerada pelos atributos “curso”, “Estudou_Escola_Pública”, “Acessa_Internet”, sendo orientada pela frequência com que acessa *internet*. Dos candidatos que acessam pouca *internet* e que já estudaram em escola particular, 49% deles optam pelo curso Técnico em Manutenção e Suporte em Informática. Já dos candidatos que acessam pouca *internet* e sempre estudaram em escola pública, 503 deles são do Curso Técnico em Agropecuária. (Fig. 23).

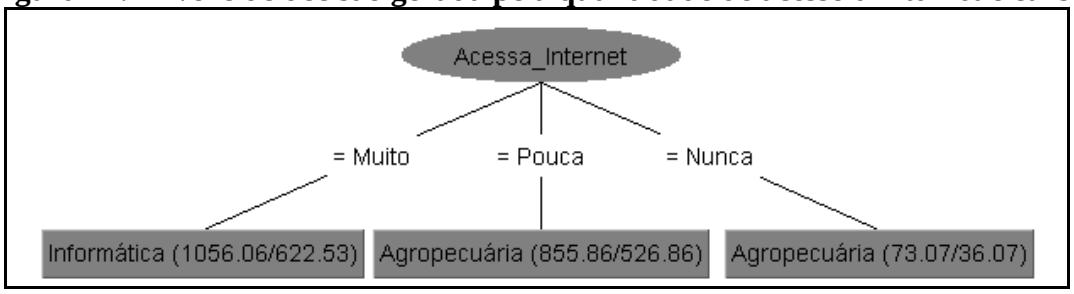
Figura 23: Árvore de decisão gerada pelo tipo de escola em que o candidato estudou



Fonte: Dados da pesquisa

Na Figura 24, abaixo, tem-se a árvore que foi gerada pelos atributos "curso", "Acessa_Internet", sendo orientada pelo curso do candidato. Diante desta árvore, percebe-se que 59% dos candidatos ao curso Técnico em Manutenção e Suporte em Informática acessam muito a *internet*. Já 62% dos candidatos ao curso Técnico em Agropecuária acessam pouca *internet*, enquanto apenas 36 candidatos nunca acessaram *internet*.

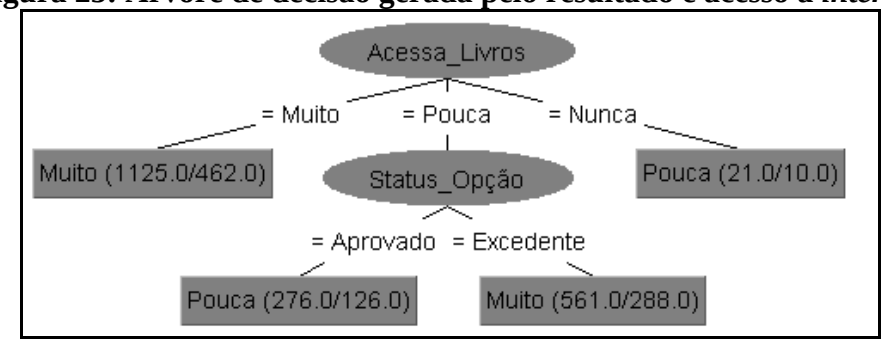
Figura 24: Árvore de decisão gerada pela quantidade de acesso a *internet* e curso



Fonte: Dados da pesquisa

Por fim, a árvore na Figura 25, foi gerada pelos atributos "status_Opção", "Acessa_Internet", "Acessa_Livros", sendo orientada pela frequência com que o candidato acessa a *internet*. Com esta árvore, percebe-se que 47 % dos candidatos que acessam pouco a *internet* são aprovados e acessam poucos livros. Nota-se, também, que 52% dos candidatos que acessam muito a *internet* não são aprovados e acessam poucos livros. Ainda é possível perceber que 462 dos 1125 candidatos que acessam muito a *internet*, acessam muitos livros também. (Fig. 25).

Figura 25: Árvore de decisão gerada pelo resultado e acesso a *internet*



Fonte: Dados da pesquisa

5.4. Discussões sobre resultados

Na aplicação do algoritmo J48 da ferramenta WEKA (*Waikato Enviroment for Knowledge Analysis)*, pôde-se constatar regras para futuras ações do IFMG, referentes ao perfil dos candidatos que prestaram o processo seletivo nos anos de 2011, 2012 e 2013. Observou-se que nos cursos Técnico em Agropecuária e Técnico em Manutenção e Suporte em Informática, os dados socioeconômicos e culturais do candidato são relevantes para o seu bom desempenho e escolha do curso. A mesma característica já não aparece como determinante no curso Técnico em Nutrição e Dietética. Confirmou-se, ainda, que a análise dos resultados obtidos por meio das regras geradas pela ferramenta em questão, na base de dados do processo seletivo, poderá facilitar os gestores do IFMG na implementação de ações pedagógicas e administrativas para melhorar assistência e marketing do processo seletivo da instituição, após detectado o perfil dos candidatos que buscam cada curso e suas situações socioeconômicas.

A utilização da ferramenta WEKA para a mineração dos dados foi útil, demonstrando informações ocultas em base de dados, reafirmando a necessidade do conhecimento de tais informações pela instituição. A técnica de agrupamento foi, nesse sentido, fundamental para compreender e confirmar as informações geradas pelas técnicas de associação e classificação, pois a maioria dos padrões foi identificada por mais de uma tarefa de mineração de dados.

O algoritmo *Apriori* desta ferramenta transformou em regras os dados preparados para este trabalho, com informações claras, criando afinidades e dependências entre os dados. Assim, entende-se que descobrir e utilizar uma ferramenta que possa apontar soluções de forma clara e simples aos usuários na descoberta do conhecimento é um bom começo para se compreender a importância do estudo da mineração de dados.

Conforme Coelho (2007), dentre as varias técnicas existentes para a análise de dados, as técnicas estatísticas são as mais próximas às técnicas de mineração de dados. Nesse sentido, Carvalho (2005) cita que grande parte das análises feita pelas técnicas de mineração de dados era realizada pelas técnicas estatísticas. Entretanto, estes autores contemplavam que quase tudo do que é feito com a mineração de dados poderia ser executado com análises estatísticas. Porém, Russel (2011) enfatiza que o que está atraindo vários analistas para a mineração de dados é a facilidade relativa com que podem ser obtidos conhecimentos mais elaborados em relação às aproximações estatísticas tradicionais.

De acordo com Tang, Steinbach, e Kumar (2009), em relação às regras obtidas com a aplicação das técnicas de mineração de dados, é interessante destacar os seguintes resultados:

a) algumas regras descobertas em uma técnica foram confirmadas ou complementadas com a utilização de outra técnica. Por exemplo, quando se utilizou a técnica de associação, descobriu-se que a maioria dos candidatos que são de cidades vizinhas a São João Evangelista vai escolher o curso Técnico em Manutenção e Suporte em Informática. Esta regra foi confirmada utilizando as técnicas de classificação e agrupamento;
b) verificou-se que o ato de o estudante ter muito acesso apenas à *internet* ou apenas muito acesso a livros não influencia em sua aprovação, pois cerca de 44% dos candidatos aprovados acessam poucos livros ou pouca *internet*, e a principal fonte de informação destes ainda é a televisão. Os estudantes que acessam muito a *internet* e acessam muitos livros, possuem grande probabilidade de serem aprovados;
c) a escolaridade dos pais influencia na escolha do curso, sendo que nos cursos de Técnico em Nutrição e Dietética e Técnico em Manutenção e Suporte em Informática os pais possuem maior escolaridade. A escolaridade de grande parte deles é Ensino Fundamental e Médio completos;
d) os meios de comunicação que mais influenciam os estudantes no conhecimento do vestibular não é televisão, *internet*, *outdoor*, *banner* ou folhetos. A opção marcada pelos candidatos é “outros”, o que leva a levantar uma hipótese que seja a influência de parentes e amigos que já estudaram na instituição, uma vez que a maioria destes candidatos é de cidades distantes (mais que 50 km) de São João Evangelista, de onde provinham a maior parte dos alunos que já estudaram na instituição. Portanto, os *banner* e folhetos possuem grande influência nos

candidatos da cidade de São João Evangelista, mas não nos de cidades mais distantes;

e) em relação ao agrupamento da Figura 10, percebe-se que o resultado final dos candidatos ao curso Técnico em Agropecuária é aprovado (*Status_Opção*), porém, esta informação é obtida devido ao grande número de vagas (140 vagas), que eram destinadas ao curso nos processos seletivos de 2011 e 2012 e à pouca concorrência;

f) no curso Técnico em Agropecuária encontrou-se uma regra: ela mostra que os estudantes aprovados no curso Técnico em Agropecuária possuem renda familiar abaixo de um salário mínimo, acessam muito *internet* em casa, acessam poucos livros, moram em cidades distantes acima de 50 km de São João Evangelista, procuram o curso porque sempre quiseram este curso e os pais trabalham no campo. Esta regra pode ser confirmada com as três técnicas empregadas.

Pode-se inferir, portanto, diante do exposto, que as hipóteses foram primordiais para nortearem a aplicação da técnica de mineração de dados, já que a maioria das hipóteses foi confirmada. Acreditava-se, inicialmente, que os perfis dos candidatos poderiam estar ligados à distância em que moram do IFMG – Campus São João Evangelista, pois grande parte dos aprovados é de candidatos que possuem maiores rendas e moram até 50 km (cento e cinquenta quilômetros) de distância de São João Evangelista. Porém, descobriu-se que o perfil de candidatos aprovados nos curso Técnico em Agropecuária são os candidatos de menores rendas, moram acima de 50 km de São João Evangelista e acessam muito a *internet*. Descobriu-se, também, que a origem dos candidatos ao curso Técnico em Nutrição e Dietética é de São João Evangelista. Outra informação é a de que o meio de comunicação de maior abrangência para conhecimento do processo seletivo é definido como outros meios pelos candidatos. Por sua vez, os candidatos que leem muitos livros e não acessam com frequência outro meio de comunicação possuem pouca chance de serem aprovados, enquanto os candidatos que leem muitos livros, acessam muito *internet* e assistem muito televisão possuem grande chance de serem aprovados.

6. CONSIDERAÇÕES FINAIS

Com este trabalho, procurou-se entender algumas das principais ferramentas de busca de conhecimento em banco de dados e a relação de integração que pudesse existir entre elas, percebendo-se que a escolha por uma determinada ferramenta ou técnica vai depender do problema que se busca resolver. A variedade de ferramentas de análise de informação deve-se à grande necessidade que as organizações ou instituições possuem em adquirir informações de seus processos operacionais e clientes. Instituições e organizações estão cada vez mais cientes que, por si só, o armazenamento de dados não lhes trazem segurança ou informação útil. Nesse sentido, nota-se que a técnica de mineração de dados se difundiu entre as organizações comerciais, acadêmicas e sociais, devido à facilidade de recuperação de informação implícita que esta técnica oferece.

Porém, nem sempre a aplicação de uma técnica de mineração de dados poderá resolver os problemas de uma empresa ou uma proposta de pesquisa. A forma de como os dados se encontram e a tarefa específica que será executada são fatores importantes que influenciam na escolha das técnicas de mineração de dados. Assim, no processo de descoberta de conhecimento em banco de dados, todas as etapas, desde a preparação dos dados até a extração de conhecimento, são de extrema importância e exigem que a mesma atenção seja atribuída para cada uma delas. O sucesso de uma etapa depende, portanto, exclusivamente, do bom desenvolvimento das etapas anteriores.

O sucesso da mineração de dados deve-se, também, ao baixo custo de investimento que ela exige, em comparação com outras técnicas como a ferramenta OLAP. Além disso, ela é muito flexível e pode atender áreas da biologia, química, administração, educacional e estatística. Normalmente as instituições de ensino trabalham com diversos tipos de pessoas, podendo ser de etnias e situações econômicas diferentes, todavia, estas pessoas podem ter comportamentos semelhantes que podem ser representados por padrões que permitam a quebra de preconceitos ou tomadas de decisões que auxiliem em um ambiente de ensino. Acredita-se, portanto, que uma boa instituição de ensino deve conhecer seus estudantes, procurando, de forma contínua, obter informações necessárias para fornecer-lhes as melhores condições de ensino possíveis.

Nesse sentido, buscou-se utilizar a técnica de mineração de dados no processo seletivo do IFMG – Campus São João Evangelista, no intuito de encontrar informações implícitas e que poderiam ser úteis à instituição. Concluiu-se que é possível a extração de conhecimento implícito no processo seletivo do IFMG – Campus São João Evangelista. Diante deste

contexto, compreendeu-se que através dos questionários respondidos por candidatos é possível encontrar padrões que identifiquem os estudantes que procuram um curso específico e os perfis de candidatos aprovados ou reprovados. Estes questionários são de grande valia para a instituição, principalmente para o *marketing*, pois permitem à instituição mapear corretamente os perfis dos candidatos para cada curso, região, renda familiar e meio de comunicação de maior eficiência para atraí-los.

Com relação especificamente à pesquisa, pode se afirmar que todos os objetivos deste estudo foram alcançados e a maioria das hipóteses, confirmada. Deste modo, conseguiu-se caracterizar a influência das características socioeconômicas e culturais na escolha e aprovação dos candidatos aos cursos oferecidos pelo IFMG – Campus São João Evangelista, já que se observou que existem diversos padrões nos dados do processo seletivo. Além disso, foi possível aplicar todas as técnicas de mineração de dados definidas, com resultados satisfatórios, percebendo-se a viabilidade do uso da mineração de dados nos processos seletivos da instituição. A descoberta de todas estas informações deve-se ao estudo do processo de DCBD e das técnicas de mineração de dados para descobrir padrões de comportamento no processo seletivo do IFMG.

Este estudo possui contribuições para o meio acadêmico, social e pessoal, sendo importante salientar que este trabalho poderá ser utilizado como um referencial teórico para outros trabalhos relacionados à busca de conhecimento em banco de dados, além de viabilizar a aplicação das técnicas de mineração de dados nos processos seletivos dos demais campi do IFMG ou de processos seletivos posteriores ao ano de 2013 na instituição. Com este estudo, também é possível que o IFMG tenha um *marketing* direcionado a cada nicho de candidatos e possa complementar sua comunicação no repertório das teorias utilizadas pelos profissionais responsáveis pelos processos seletivos e de atendimentos estudantis. Acredita-se, portanto, que este trabalho foi de grande valia para que se aprofundasse no tema de busca de conhecimento sobre banco de dados, abalizando a criação de diversas ideias que visam a aplicação destes métodos em outras bases de dados, como a mineração de dados em redes sociais, bioinformática e análise de rendimentos estudantis.

Diante do estudo realizado, sugere-se, para trabalhos futuros, a criação de uma estrutura de armazenamento de dados para instituição, como um *data warehouse*, para que seja fornecidos os dados pré-formatados, agilizando o processo de descoberta de conhecimento em banco de dados, evitando, assim, o moroso processo de limpeza e codificação dos dados. Dando continuidade ao que já foi feito, indica-se também a criação de um modelo de visualização de dados dos estudantes enquanto matriculados na instituição,

sendo esta busca de informações realizada através das técnicas de mineração de dados. Com a criação do armazém de dados torna-se possível auxiliar o processo de mineração de dados com um sistema OLAP. Estes modelos iriam buscar conhecimento útil nos históricos escolares dos estudantes nos diversos períodos de estudo, podendo descobrir prováveis padrões que influenciem sua vida acadêmica na instituição.

REFERÊNCIAS

BRAUNER, Daniela Francisco. **O processo de descoberta de conhecimento em banco de dados**: um estudo de caso sobre os dados da UFPEL. 2003. 75f. Monografia (Conclusão do curso) – Universidade Federal de Pelotas, Instituto de Física, Pelotas.

BRAZ, Lucas *et al.* **Aplicando mineração de dados para apoiar a tomada de decisão na segurança pública do estado de Alagoas**, Maceió, v. 13, n.2, maio/ago. 2009. Disponível em: <http://www.grow.ic.ufal.br/article/pid=0003232&nm=pt>. Acesso: em 03 fev. 2013.

CARREIRA, Suely da Silva *et al*. Aplicação de data mining na base de dados do processo seletivo do exame nacional do ensino médio - ENEM 2010. In: CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO, 2, 2012, Curitiba. **Aplicação da data mining em sistemas de informação gerenciais**. Ponta Grossa: Universidade Estadual de Maringá, 2012. p.79–91.

CARVALHO, Luís Alfredo Vidal de. **Data mining**: a mineração de dados no marketing, medicina, economia, engenharia e administração. São Paulo: Erica, 2005. 225 p.

CLÉSIO, Flávio. **Mineração de Dados**, 2013. Disponível em: <http://mineracaodedados.wordpress.com/tag/weka/>. Acesso em: 12 set. 2013.

COELHO, Éden de Oliveira Pinto. **Descoberta de conhecimento sobre o Processo seletivo da UFLA**. 2007. 56f. Monografia (Conclusão do curso) – Universidade Federal de Lavras, Departamento de Ciência da Computação, Lavras.

ELMASRI, Ramez; NAVATHE, Shamkant. **Sistemas de banco de dados**. Trad. Daniel Vieira. 6. ed. São Paulo: Pearson Addison Wesley, 2011. 788 p.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS. **Histórico.** 2013. Disponível em: <http://www.ifmg.edu.br/index.php/institucional/historico>. Acesso em: 06 out. 2013.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica**. 7 ed. São Paulo: Atlas, 2010. 297 p.

MACHINE LEARNING GROUP AT THE UNIVERSITY OF WAIKATO. **Data Mining software in Java.** 2013. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/index.html>. Acesso em: 12 set. 2013.

MARTINHAGO, Sergio. **Descoberta de conhecimento sobre o processo seletivo da UFPR**. 2005. 125f. Dissertação (Mestrado em Ciências) – Universidade Federal do Paraná, Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Curitiba.

MONTEIRO, Marcos de Souza; ROCHA, Vinícius Carvalho. **Descoberta de conhecimento na base de dados do processo seletivo seriado da UFPA – 2004, usando regras de associação**. 2005. 75p. Monografia (Conclusão do curso) – Universidade Federal do Pará, Departamento de Ciência da Computação, Belém.

MORAIS, Bruno Carlos Sales de. **Extração de conhecimento da plataforma lattes utilizando técnicas de mineração de dados**: estudo de caso POLI/UPE. 2010. 54f. Monografia (Conclusão do curso) – Escola Politécnica de Pernambuco/ Universidade de Pernambuco, Departamento de Ciência da Computação, Recife.

MURASSE, Carlos; TSUNODA, Denise. **Descoberta de conhecimento a partir de uma base de indicadores de desenvolvimento social utilizando WEKA**. Trabalho apresentado no Seminário de Tecnologias Inteligentes, promovido pelo Serviço Federal de Processamento de Dados, Curitiba, 2010.

OLIVEIRA, Aracele Garcia de; GARCIA, Denise Ferreira. **Mineração da base de dados de um processo seletivo universitário**. 2004. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos/v3.2/vol3.2.htm>. Acesso em: 02 fev. 2013.

PENEDO, Janaína; CAPRA, Eliane. Mineração de dados na descoberta do padrão de usuários de um sistema de educação a distância. In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO, 8, 2012, Rio de Janeiro. **Anais...** Rio de Janeiro: UNIRIO, 2012. p. 396-407.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. 2 ed. Novo Hamburgo: Feevale, 2013. 277 p.

RUSSEL, Mattebew. **Mineração de dados da Web Social**. Trad. Rafael Zanolli. São Paulo: Novatec Editora; Sebastopol, CA: O'Really, 2011. 358 p.

SEVERINO, Antônio Joaquim. **Metodologia do trabalho científico**. 23 ed. rev. São Paulo: Cortez, 2007. 257 p.

SILVA, Marcos Pereira da. **Mineração de dados - conceitos, aplicações e experimentos com WEKA**. 2001. 63f. Monografia (Conclusão do curso) – Escola Regional de Informática, Departamento de Ciência da Computação, Rio de Janeiro.

TANG, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao DATA MINING**: Mineração de Dados. Trad. Acauan P. Fernandes. Rio de Janeiro: Ciência Moderna, 2009. 900 p.

THOMÉ, Ageu Costa. **Redes neurais, uma ferramenta para KDD e data mining**. 23 abril 2012. Disponível em: <http://www.tic.com/ti/inteligencai-artifical/redes-neurais-uma-ferramenta-para-KDD-e-data-mining>. Acesso em: 01 fev. 2013.

VIANNA, Ricardo. Mineração de dados: introdução e aplicações. Curitiba, v. 2, n.10, set. 2007. **Artigo SQL Magazine**, ed 10, 2007. p. 47-58.

ZAMBON, Antônio Carlos; MEIRELLES, J. L. **A evolução do processo decisório e as novas ferramentas de apoio à decisão**: data warehouse, OLAP e data mining. 2001. 111f. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de São Carlos, Programa de Pós-Graduação em Engenharia de Produção, São Carlos.

ANEXO A – QUESTIONÁRIO SOCIOECONÔMICO E CULTURAL DO PROCESSO SELETIVO 2011, DA COMISSÃO PERMANENTE DE VESTIBULAR E EXAME DE SELEÇÃO – IFMG

COMISSÃO PERMANENTE DE VESTIBULAR E EXAME DE SELEÇÃO

PROCESSO SELETIVO 2011

1- Como tomou conhecimento do Vestibular / Exame de Seleção do IFMG?

( ) Televisão
( ) Internet (website)
( ) Rádio
( ) Banner/folhetos
( ) Outdoor
( ) Outros meios.
Especificar:_____

2- Por que você escolheu o IFMG para estudar?

( ) Porque é gratuito.
( ) É próximo a minha residência.
( ) Por falta de opção.
( ) Por indicação de terceiros
( ) Pela qualidade de ensino prestada.

3- Por que você escolheu o curso para o qual está se inscrevendo?

( )Sempre quis este curso.
( )Por influência de terceiros.
( ) Melhor possibilidade no mercado.
( ) Por falta de opção.

4- Qual a distância entre a sua cidade e o campus do IFMG para qual está se inscrevendo?

( ) Moro na mesma cidade onde está o campus.
( ) Moro em algum distrito ou zona rural do município onde está o campus o qual está se inscrevendo.
( ) Moro em cidade próxima, até 50 km.
( ) Moro em outra cidade acima de 50 km.

5- Quantos membros de sua família moram em sua casa (incluindo você)?

( ) Moro sozinho
( ) 2
( ) 3
( ) 4
( ) 5
( ) Mais que 5. Quantos? _____

6- Você tem filhos?

( ) Não
( ) Sim. Quantos?_____

7- Quantos filhos menores de 6 anos você tem?

( ) Nenhum
( ) 1
( ) 2
( ) 3
( ) 4
( ) 5 ou mais.

8- Qual a renda total de sua família? (Em salário mínimo)

( ) 1
( ) 2
( ) 3
( ) 4
( ) 5
( ) 6
( ) 7
( ) 8
( ) mais que 8.

9- Você ou algum membro de sua família recebe algum tipo de benefício social do governo (ex: bolsa família, auxílio escola, Benefício de Prestação Continuada - BPC)?
( ) Sim ( ) Não

10- Você estudou em escola pública?
( ) Sempre ( ) Parcialmente ( ) Nunca

11- Escolaridade de seu pai:
( ) Analfabeto ( ) 2º grau completo ( ) Outro
( ) 1º grau incompleto ( ) Superior incompleto especificar:_____
( ) 1º grau completo ( ) Superior completo
( ) 2º grau incompleto ( ) Não sei

12- Escolaridade de sua mãe:
( ) Analfabeto ( ) 2º grau completo ( ) Outro
( ) 1º grau incompleto ( ) Superior incompleto especificar:_____
( ) 1º grau completo ( ) Superior completo
( ) 2º grau incompleto ( ) Não sei

13- Em que seu pai trabalha atualmente?
( ) Na agricultura, no campo, em fazenda ou na pesca.
( ) Na indústria, no comércio, transporte ou outros serviços ou profissional liberal.
( ) Funcionário público ou militar.
( ) Trabalhador do setor informal (sem carteira assinada ou trabalha em casa)
( ) Aposentado
( ) Ausente
( ) Outro. Qual? _____

14- Em que sua mãe trabalha atualmente?
( ) Na agricultura, no campo, em fazenda ou na pesca.
( ) Na indústria, no comércio, transporte ou outros serviços ou profissional liberal.
( ) Funcionário público ou militar.
( ) Trabalhador do setor informal (sem carteira assinada ou trabalha em casa)
( ) Aposentada
( ) Ausente
( ) Outro. Qual? _____

15- Quanto ao seu trabalho:
( ) não trabalho, mas pretendo trabalhar durante o curso
( ) trabalho e pretendo continuar trabalhando durante o curso
( ) deixarei de trabalhar para estudar
( ) não trabalho e não pretendo trabalhar durante o curso

16- Em qual local acessa, com mais frequência, a internet?
( ) Em casa ( ) na casa de vizinhos ( ) Escola
( ) No serviço ou amigos ( ) Não tenho acesso
( ) Na casa de parentes ( ) Lan house

17- Participa ou participou de grupos como grêmio, associações profissionais ou comunitárias, partido político?
( ) Sim ( ) Não

18- Classifique o grau de acesso à seguinte atividade: Cinema e teatro
( ) Muito ( ) Pouca ( ) Nunca
19- Classifique o grau de acesso à seguinte atividade: Música e Dança
( ) Muito ( ) Pouca ( ) Nunca

20- Classifique o grau de acesso à seguinte atividade: Exposições artísticas
( ) Muito ( ) Pouca ( ) Nunca

21- Classifique o grau de acesso às informações no seu dia a dia: Tv e rádio
( ) Muito ( ) Pouca ( ) Nunca

22- Classifique o grau de acesso às informações no seu dia a dia: Internet
( ) Muito ( ) Pouca ( ) Nuca

23- Classifique o grau de acesso às informações no seu dia a dia: Livro, revistas e jornais
( ) Muito ( ) Pouca ( ) Nunca

24- Classifique o grau de acesso às informações no seu dia a dia: escola
( ) Muito ( ) Pouca ( ) Nunca

ANEXO B – QUESTIONÁRIO SOCIOECONÔMICO E CULTURAL DO PROCESSO SELETIVO 2012, DA COMISSÃO PERMANENTE DE VESTIBULAR E EXAME DE SELEÇÃO – IFMG

COMISSÃO PERMANENTE DE VESTIBULAR E EXAME DE SELEÇÃO
PROCESSO SELETIVO 2012

1- Como tomou conhecimento do Vestibular / Exame de Seleção do IFMG?

( ) Televisão
( ) Internet (website)
( ) Rádio
( ) Banner/folhetos
( ) Outdoor
( ) Outros meios.
Especificar:_____

2- Por que você escolheu o IFMG para estudar?

( ) Porque é gratuito.
( ) É próximo a minha residência.
( ) Por falta de opção.
( ) Por indicação de terceiros
( ) Pela qualidade de ensino prestada.

3- Por que você escolheu o curso para o qual está se inscrevendo?

( )Sempre quis este curso.
( )Por influência de terceiros.
( ) Melhor possibilidade no mercado.
( ) Por falta de opção.

4- Qual a distância entre a sua cidade e o campus do IFMG para qual está se inscrevendo?

( ) Moro na mesma cidade onde está o campus.
( ) Moro em algum distrito ou zona rural do município onde está o campus o qual está se inscrevendo.
( ) Moro em cidade próxima, até 50 km.
( ) Moro em outra cidade acima de 50 km.

5- Quantos membros de sua família moram em sua casa (incluindo você)?

( ) Moro sozinho
( ) 2
( ) 3
( ) 4
( ) 5
( ) Mais que 5. Quantos? _____

6- Você tem filhos?

( ) Não
( ) Sim. Quantos?_____

7- Quantos filhos menores de 6 anos você tem?

( ) Nenhum
( ) 1
( ) 2
( ) 3
( ) 4
( ) 5 ou mais.

8- Qual a renda total de sua família? (Em salário mínimo)

( ) 1
( ) 2
( ) 3
( ) 4
( ) 5
( ) 6
( ) 7
( ) 8
( ) mais que 8.

9- Você ou algum membro de sua família recebe algum tipo de benefício social do governo (ex: bolsa família, auxílio escola, Benefício de Prestação Continuada - BPC)?
( ) Sim ( ) Não

10- Você estudou em escola pública?
( ) Sempre ( ) Parcialmente ( ) Nunca

11- Escolaridade de seu pai:
( ) Analfabeto ( ) 2º grau completo ( ) Outro
( ) 1º grau incompleto ( ) Superior incompleto especificar:_____
( ) 1º grau completo ( ) Superior completo
( ) 2º grau incompleto ( ) Não sei

12- Escolaridade de sua mãe:
( ) Analfabeto ( ) 2º grau completo ( ) Outro
( ) 1º grau incompleto ( ) Superior incompleto especificar:_____
( ) 1º grau completo ( ) Superior completo
( ) 2º grau incompleto ( ) Não sei

13- Em que seu pai trabalha atualmente?
( ) Na agricultura, no campo, em fazenda ou na pesca.
( ) Na indústria, no comércio, transporte ou outros serviços ou profissional liberal.
( ) Funcionário público ou militar.
( ) Trabalhador do setor informal (sem carteira assinada ou trabalha em casa)
( ) Aposentado
( ) Ausente
( ) Outro. Qual? _____

14- Em que sua mãe trabalha atualmente?
( ) Na agricultura, no campo, em fazenda ou na pesca.
( ) Na indústria, no comércio, transporte ou outros serviços ou profissional liberal.
( ) Funcionário público ou militar.
( ) Trabalhador do setor informal (sem carteira assinada ou trabalha em casa)
( ) Aposentada
( ) Ausente
( ) Outro. Qual? _____

15- Quanto ao seu trabalho:
( ) não trabalho, mas pretendo trabalhar durante o curso
( ) trabalho e pretendo continuar trabalhando durante o curso
( ) deixarei de trabalhar para estudar
( ) não trabalho e não pretendo trabalhar durante o curso

16- Em qual local acessa, com mais frequência, a internet?
( ) Em casa ( ) na casa de vizinhos ou amigos ( ) Escola
( ) No serviço ( ) Não tenho acesso
( ) Na casa de parentes ( ) Lan house

17- Participa ou participou de grupos como grêmio, associações profissionais ou comunitárias, partido político?
( ) Sim ( ) Não

18- Classifique o grau de acesso à seguinte atividade: Cinema e teatro
( ) Muito ( ) Pouca ( ) Nunca
19- Classifique o grau de acesso à seguinte atividade: Música e Dança
( ) Muito ( ) Pouca ( ) Nunca

20- Classifique o grau de acesso à seguinte atividade: Exposições artísticas
( ) Muito ( ) Pouca ( ) Nunca

21- Classifique o grau de acesso às informações no seu dia a dia: Tv e rádio
( ) Muito ( ) Pouca ( ) Nunca

22- Classifique o grau de acesso às informações no seu dia a dia: Internet
( ) Muito ( ) Pouca ( ) Nunca

23- Classifique o grau de acesso às informações no seu dia a dia: Livro, revistas e jornais
( ) Muito ( ) Pouca ( ) Nunca

24- Classifique o grau de acesso às informações no seu dia a dia: escola
( ) Muito ( ) Pouca ( ) Nunca

ANEXO C – QUESTIONÁRIO SOCIOECONÔMICO E CULTURAL DO PROCESSO SELETIVO 2013, DA COMISSÃO PERMANENTE DE VESTIBULAR E EXAME DE SELEÇÃO – IFMG

COMISSÃO PERMANENTE DE VESTIBULAR E EXAME DE SELEÇÃO
PROCESSO SELETIVO 2013

1- Curso:

( ) Formação Inicial e Continuada
( ) Curso Técnico Integrado
( ) Curso Técnico Subsequente
( ) Curso Técnico Concomitante
( ) Curso Superior – Bacharelado
( ) Curso Superior – Tecnológico
( ) Curso Superior – Licenciatura
( ) Curso de Pós-Graduação

2- Sexo:

( ) M
( ) F

3- Cor/raça:

( ) Branca
( ) Preta
( ) Parda
( ) Amarela
( ) Indígena

4- Faixa etária:

( ) Até 14 anos
( ) de 15 a 17 anos
( ) de 18 a 19 anos
( ) de 20 a 24 anos
( ) de 25 a 29 anos
( ) de 30 a 39 anos
( ) de 40 a 49 anos
( ) acima de 50 anos

5- Como tomou conhecimento do Vestibular / Exame de Seleção do IFMG?

( ) Televisão
( ) Internet (website)
( ) Rádio
( ) Banner/folhetos
( ) Outdoor
( ) Outros meios.
Especificar:_____

6- Por que você escolheu o IFMG para estudar?

( ) Porque é gratuito.
( ) É próximo a minha residência.
( ) Por falta de opção.
( ) Por indicação de terceiros
( ) Pela qualidade de ensino prestada.

7- Por que você escolheu o curso para o qual está se inscrevendo?

( )Sempre quis este curso.
( )Por influência de terceiros.
( ) Melhor possibilidade no mercado.
( ) Por falta de opção.

8- Qual a distância entre a sua cidade e o campus do IFMG para qual está se inscrevendo?

( ) Moro na mesma cidade onde está o campus.
( ) Moro em algum distrito ou zona rural do município onde está o campus o qual está se inscrevendo.

( ) Moro em cidade próxima, até 50 km.
( ) Moro em outra cidade acima de 50 km.

9- Com quem você mora atualmente? (Permitidas mais de uma marcação)

( ) Pais
( ) Cônjuge
( ) Companheiro(a)
( ) Filhos
( ) Sogros
( ) Parentes
( ) Amigos
( ) Empregados domésticos
( ) Outros
( ) (ou) Sozinho(a)

10- Quantos membros de sua família moram em sua casa (incluindo você)?

( ) Moro sozinho
( ) 2
( ) 3
( ) 4
( ) 5
( ) Mais que 5. Quantos? _____

11- Residência:

( ) Própria
( ) Alugada pelos pais
( ) Própria dos pais
( ) Cedida
( ) Alugada por você

12- Área de procedência:

( ) Urbana
( ) Rural

13- Você tem filhos?

( ) Não
( ) Sim. Quantos?_____

14- Quantos filhos menores de 6 anos você tem?

( ) Nenhum
( ) 1
( ) 2
( ) 3
( ) 4
( ) 5 ou mais.

15- Qual a renda total de sua família? (Em salário mínimo)

( ) 1
( ) 2
( ) 3
( ) 4
( ) 5
( ) 6
( ) 7
( ) 8
( ) mais que 8.

16- Você ou algum membro de sua família recebe algum tipo de benefício social do governo (ex: bolsa família, auxílio escola, Benefício de Prestação Continuada - BPC)?

( ) Sim
( ) Não

17- Qual a sua participação na vida econômica do grupo familiar?

( ) Não trabalho e sou sustentado por minha família e/ou outras pessoas
( ) Trabalho e sou sustentado parcialmente por minha família e/ou outras pessoas
( ) Trabalho e sou responsável apenas por meu próprio sustento
( ) Trabalho, sou responsável por meu próprio sustento e ainda contribuo parcialmente para o sustento da família
( ) Trabalho e sou o principal responsável pelo sustento da família
( ) Outra situação

18- Quanto ao seu trabalho:
( ) não trabalho, mas pretendo trabalhar durante o curso
( ) trabalho e pretendo continuar trabalhando durante o curso
( ) deixarei de trabalhar para estudar
( ) não trabalho e não pretendo trabalhar durante o curso

19- Antes de se matricular no IFMG, até o 5º ano (4ª série), você estudou?
( ) sempre em escola pública
( ) parte em escola pública, parte em particular
( ) parte em escola pública, parte em escola particular com bolsa
( ) escola particular com bolsa
( ) sempre em escolar particular
( ) Ainda não conclui.

20- Antes de se matricular no IFMG, até o 9º ano (8ª série), você estudou?
( ) sempre em escola pública
( ) parte em escola pública, parte em particular
( ) parte em escola pública, parte em escola particular com bolsa
( ) escola particular com bolsa
( ) sempre em escolar particular
( ) Ainda não conclui.

21- Antes de se matricular no IFMG, até o 3º ano Ensino Médio, você estudou?
( ) sempre em escola pública
( ) parte em escola pública, parte em particular
( ) parte em escola pública, parte em escola particular com bolsa
( ) escola particular com bolsa
( ) sempre em escolar particular
( ) Ainda não conclui.

22- Situação do pai:
( )Presente ( )Ausente ( ) Falecido

23- Grau de instrução do pai:
( ) Analfabeto
( ) 1º grau incompleto
( ) 1º grau completo
( ) 2º grau incompleto
( ) 2º grau completo
( ) Superior incompleto
( ) Superior completo
( ) Não sei
( ) Outro especificar:_____

24- Em que seu pai trabalha atualmente?
( ) Na agricultura, no campo, em fazenda ou na pesca.
( ) Na indústria, no comércio, transporte ou outros serviços ou profissional liberal.
( ) Funcionário público ou militar.
( ) Trabalhador do setor informal (sem carteira assinada ou trabalha em casa)
( ) Aposentado
( ) Ausente
( ) Outro. Qual? _____

25- Situação da mãe:
( )Presente ( )Ausente ( )Falecida

26- Grau de instrução da mãe:

( ) Analfabeto
( ) 1º grau incompleto
( ) 1º grau completo
( ) 2º grau incompleto
( ) 2º grau completo
( ) Superior incompleto
( ) Superior completo
( ) Não sei
( ) Outro especificar:___

27- Em que sua mãe trabalha atualmente?

( ) Na agricultura, no campo, em fazenda ou na pesca.
( ) Na indústria, no comércio, transporte ou outros serviços ou profissional liberal.
( ) Funcionária público ou militar.
( ) Trabalhadora do setor informal (sem carteira assinada ou trabalha em casa)
( ) Aposentada
( ) Ausente
( ) Outro. Qual? _____

28- Em qual local acessa com mais frequência a internet?

( ) Em casa
( ) No serviço
( ) Na casa de parentes
( ) na casa de vizinhos ou amigos
( ) Lan house
( ) Escola
( ) Não tenho acesso

29- Com que frequência você tem acesso a estes meios de informação?

29.1 Jornais

( ) Diariamente ( ) Quase diariamente ( ) Às vezes ( ) Raramente ( ) Nunca

29.2 Revistas

( ) Diariamente ( ) Quase diariamente ( ) Às vezes ( ) Raramente ( ) Nunca

29.3 Televisão

( ) Diariamente ( ) Quase diariamente ( ) Às vezes ( ) Raramente ( ) Nunca

29.4 Internet

( ) Diariamente ( ) Quase diariamente ( ) Às vezes ( ) Raramente ( ) Nunca

29.5 Livros

( ) Diariamente ( ) Quase diariamente ( ) Às vezes ( ) Raramente ( ) Nunca

29.6 Rádio

( ) Diariamente ( ) Quase diariamente ( ) Às vezes ( ) Raramente ( ) Nunca

30- Quantos livros didáticos você possui em casa?

( ) 10 livros ou menos ( ) De 11 a 30 livros ( ) De 31 a 50 livros ( ) mais de 50 livros

31- Quantos livros em média você costuma ler por ano?

( ) Nenhum
( ) Um livro
( ) De 2 a 5 livros
( ) De 6 a 10 livros
( ) De 11 a 15 livros
( ) De 16 a 20 livros
( ) De 21 a 30 livros
( ) Mais do que 30 livros

32- Com que frequência você frequenta...

32.1 Cinema

( ) Semanalmente
( ) Ao menos 1 vez por mês
( ) Ao menos 1 vez por ano
( ) Menos que uma vez por ano
( ) Nunca

32.2 Teatro

( ) Semanalmente
( ) Ao menos 1 vez por mês
( ) Ao menos 1 vez por ano
( ) Menos que uma vez por ano
( ) Nunca

32.3 Museu

( ) Semanalmente
( ) Ao menos 1 vez por mês
( ) Ao menos 1 vez por ano
( ) Menos que uma vez por ano
( ) Nunca

32.4 Parques

( ) Semanalmente
( ) Ao menos 1 vez por mês
( ) Ao menos 1 vez por ano
( ) Menos que uma vez por ano
( ) Nunca

32.5 Shows/concertos

( ) Semanalmente
( ) Ao menos 1 vez por mês
( ) Ao menos 1 vez por ano
( ) Menos que uma vez por ano
( ) Nunca

32.6 Bares/danceterias

( ) Semanalmente
( ) Ao menos 1 vez por mês
( ) Ao menos 1 vez por ano
( ) Menos que uma vez por ano
( ) Nunca

ANEXO D – QUESTIONÁRIO SOCIOECONÔMICO E CULTURAL, ADEQUADO COM AS QUESTÕES COMUNS NO PROCESSO SELETIVO DE 2011, 2012 E 2013, DA COMISSÃO PERMANENTE DE VESTIBULAR E EXAME DE SELEÇÃO - IFMG

COMISSÃO PERMANENTE DE VESTIBULAR E EXAME DE SELEÇÃO

1- Curso:
( ) Técnico em Agropecuária
( ) Técnico em Manutenção e Suporte em Informática
( ) Técnico em Nutrição e Dietética

2- Status Opção:
( ) Aprovado
( ) Excedente

3- Como conheceu o Processo Seletivo/Vestibular?
( ) Banner/Folhetos
( ) Outdoor
( ) Outros Meios
( ) Rádio
( ) Televisão
( ) Website (internet)

4- Porque você escolheu o IFMG?
( ) É próximo de minha residência
( ) Pela qualidade de ensino prestada
( ) Por falta de opção
( ) Por indicação de terceiros
( ) Porque é gratuito

5- Por que você escolheu o curso para o qual está se inscrevendo?
( ) Sempre quis este curso.
( ) Por influência de terceiros.
( ) Melhor possibilidade no mercado.
( ) Por falta de opção.

6- Qual a distância entre a sua cidade e o campus do IFMG para qual está se inscrevendo?
( ) Moro na mesma cidade onde está o campus.
( ) Moro em algum distrito ou zona rural do município onde está o campus o qual estou me inscrevendo.
( ) Moro em cidade próxima, até 50 km.
( ) Moro em outra cidade acima de 50 km.

7- Qual a renda da sua família?
( ) Acima de 10 salários
( ) Até um salário mínimo
( ) De 1 a 3 salários
( ) De 3 a 5 salários
( ) De 5 a 10 salários

8- Você estudou em Escola Pública?
( ) Nunca
( ) Parcialmente
( ) Sempre

9- Qual a instrução de seu pai?

( ) Analfabeto
( ) Fundamental (1 a 4)
( ) Fundamental (5 a 8)
( ) Ensino Médio
( ) Ensino Superior
( ) Pós-graduação

10- Qual a instrução da sua Mãe?

( ) Analfabeto
( ) Fundamental (1 a 4)
( ) Fundamental (5 a 8)
( ) Ensino Médio
( ) Ensino Superior
( ) Pós-graduação

11- Em que seu pai trabalha atualmente?

( ) Na agricultura, no campo, em fazenda ou na pesca.
( ) Na indústria, no comércio, transporte ou outros serviços ou profissional liberal.
( ) Funcionário público ou militar.
( ) Trabalhador do setor informal (sem carteira assinada ou trabalha em casa)
( ) Aposentado
( ) Ausente
( ) Desempregado
( ) Outro.

12- Em que sua Mãe trabalha atualmente?

( ) Na agricultura, no campo, em fazenda ou na pesca.
( ) Na indústria, no comércio, transporte ou outros serviços ou profissional liberal.
( ) Funcionária público ou militar.
( ) Trabalhadora do setor informal (sem carteira assinada ou trabalha em casa)
( ) Aposentada
( ) Ausente
( ) Desempregada
( ) Outro.

13- Em qual local acessa, com mais frequência, a *internet*?

( ) Em casa
( ) No serviço
( ) Na casa de parentes
( ) na casa de vizinhos ou amigos
( ) Lan house
( ) Escola
( ) Não tenho acesso

14- Com que frequência você acessa Cinema e/ou Teatro

( ) Muito ( ) Pouca ( ) Nunca

15- Com que frequência você acessa *Internet*

( ) Muito ( ) Pouca ( ) Nunca

16- Com que frequência você acessa Livros

( ) Muito ( ) Pouca ( ) Nunca

17- Com que frequência você acessa Televisão

( ) Muito ( ) Pouca ( ) Nunca

18- Quantas pessoas moram em sua casa, incluindo você?

( ) 1 ( ) 4 ( ) Moro sozinho
( ) 2 ( ) 5
( ) 3 ( ) Mais que 5

19- Você possui filhos?

( ) Sim ( ) Não

20- Você ou algum membro de sua família recebe algum tipo de benefício social do governo (ex: bolsa família, auxílio escola, Benefício de Prestação Continuada - BPC)?

( ) Sim ( ) Não

ANEXO E – TERMO DE RESPONSABILIDADE E COMPROMISSO PARA USO, GUARDA E DIVULGAÇÃO DE DADOS E ARQUIVOS DE PESQUISA, FIRMADO ENTRE O PESQUISADOR E A COPEVES DO IFMG

COMISSÃO PERMANENTE DE VESTIBULAR E EXAME DE SELEÇÃO - IFMG

Título do Projeto: Ilação sobre os perfis de candidatos no processo seletivo do Instituto Federal de Minas Gerais – Campus São João Evangelista

Nome completo do solicitante/pesquisador responsável ou participante: Fernando Elias de Oliveira
RG: MG 16.489.384 CPF: 100.751.916-95
Endereço: Rua Cônego Davino nº: 589
Bairro: Centro cidade: São João Evangelista
CEP: 39705-000 Estado de Minas Gerais

O solicitante/pesquisador responsável ou participante, retro qualificado, se declara ciente e de acordo:

a) de todos os termos do presente instrumento, assumindo toda e qualquer responsabilidade por quaisquer condutas, ações ou omissões que importem na inobservação do presente e consequente violação de quaisquer das cláusulas abaixo descritas, bem como por outras normas previstas em lei, aqui não especificadas, respondendo, de forma ilimitada, irretratável, irrevogável e absoluta perante a fornecedora dos dados e arquivos em eventuais ações regressivas, bem como perante terceiros eventualmente prejudicados por sua não observação.

b) de que os dados e arquivos a ele fornecidos deverão ser usados, guardados e preservados em sigilo e que eventual divulgação dos dados deverá ser feita em estrita observação aos princípios éticos de pesquisa, resguardando-se, ainda, aos termos da Constituição Federal de 1988, especialmente no tocante ao direito à intimidade e à privacidade dos consultados, sejam eles pacientes ou não.

c) de que as informações constantes nos dados ou arquivos a ele disponibilizados deverão ser utilizadas apenas e tão somente para a execução e pesquisa do projeto acima descrito, sendo vedado o uso em outro projeto, seja a que título for, salvo expressa autorização em contrário do responsável devidamente habilitado do setor.

d) de que eventuais informações a serem divulgadas, serão única e exclusivamente para fins de pesquisa científica, sendo vedada a caracterização da instituição e o uso das informações para publicação em quaisquer meios de comunicação de massa que não guardem compromisso ou relação científica, tais como televisão, jornais, periódicos e revistas, entre outros aqui não especificados.

e) sem prejuízo dos termos do presente, que deverão ser respeitadas as normas da Resolução 196/96 e suas complementares na execução do projeto em epígrafe.

São João Evangelista, 04 de abril de 2013.

_____

Nome e assinatura do pesquisador responsável ou participante