



## Resumo Expandido

<b>Título da Pesquisa:</b> Ilação sobre os perfis de candidatos no processo seletivo do Instituto Federal de Minas Gerais – Campus São João Evangelista		
<b>Palavras-chave:</b> Descoberta de Conhecimento. Mineração de Dados. Processo Seletivo.		
<b>Campus:</b> São João Evangelista	<b>Tipo de Bolsa:</b> PIBIC	<b>Financiador:</b> IFMG
<b>Bolsista (as):</b> Fernando Elias de Oliveira		
<b>Professor Orientador:</b> José Fernandes da Silva		
<b>Área de Conhecimento:</b> Ciência da Computação		

### Resumo:

Diante da globalização e do avanço tecnológico, a informação tem desempenhando um papel primordial no desenvolvimento das organizações e a Mineração de Dados é uma das principais tecnologias para a extração desse tipo de informação. 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 Descoberta de Conhecimento em Banco de Dados (DCBD). Neste trabalho, utilizando-se de uma ferramenta chamada WEKA, aplicaram-se as técnicas de Regra de Associação, Árvore de Decisão e Agrupamento. 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-se a continuidade da pesquisa e outros trabalhos futuros.

### INTRODUÇÃO:

Diante da globalização e do avanço tecnológico, a informação tem desempenhando um papel primordial no desenvolvimento das organizações e a Mineração de Dados é uma das principais tecnologias para a extração desse tipo de informação. 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), para maior eficiência na aplicação da mineração de dados é necessário seguir algum processo de Descoberta de Conhecimento em

Banco de Dados (DCBD). Nesse sentido, o IFMG possui um amplo conjunto de dados que pode ser utilizado para adquirir conhecimentos que auxiliem na sua gestão. Uma forma de obter esse conhecimento é a partir de informações relacionadas aos processos seletivos, porque os estudantes constituem o seu principal público alvo. As bases de dados dos processos seletivos podem inferir conhecimentos que propiciem o aperfeiçoamento de políticas eficientes, que atenderão estudantes nos processos seletivos da instituição. No processo seletivo realizado pela Comissão Permanente de Vestibular e Exame de Seleção (COPEVES) do IFMG, os candidatos 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. Acreditou-se 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 pelos 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 muitos livros seriam aprovados e que os motivos de aprovações para cada curso poderia conter padrões diferentes.

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 foi 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. Justifica-se a aplicação do processo de DCBD no IFMG – Campus São João Evangelista pelo fato dele 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 constituir uma fonte rica de informações sobre os estudantes da instituição, que podem auxiliar nas decisões para a criação de programas para assistência estudantis e marketing.

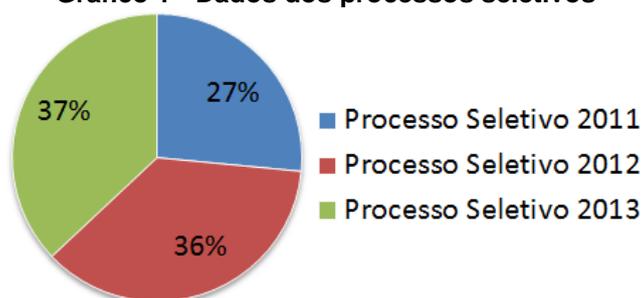
## **METODOLOGIA:**

Assim sendo, 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 estuda foi os candidatos inscritos nos processos seletivos para os cursos técnicos do IFMG – Campus São João Evangelista.

O banco de dados consistia em todas as questões do questionário socioeconômico respondidas por todos os candidatos que se inscreveram nos três exames de seleção. A base de dados dos processos

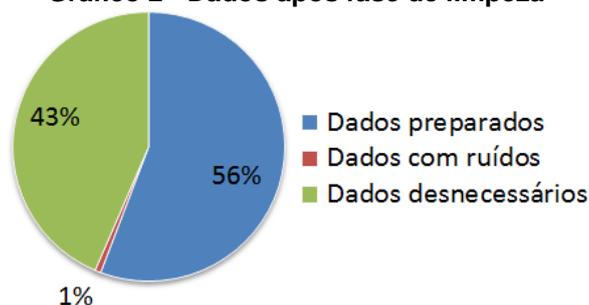
seletivos adquirida com COPEVES abrangia inicialmente dados de 3560 candidatos, distribuídos conforme o Gráfico 1. Conforme o Gráfico 2, com a necessidade de retirar os dados com ruídos e que pertenciam aos candidatos aos cursos superiores, eliminou-se 44% dos dados inúteis ao estudo, restando conforme o Gráfico 3, dados de 1985 candidatos aos três cursos técnicos. Portanto, como se trabalhou com os dados de todos os candidatos inscritos no processo seletivo para os cursos técnicos, a população será igual a 1985 indivíduos e a amostra também corresponderá ao valor da população. Analisaram-se 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 1 - Dados dos processos seletivos**



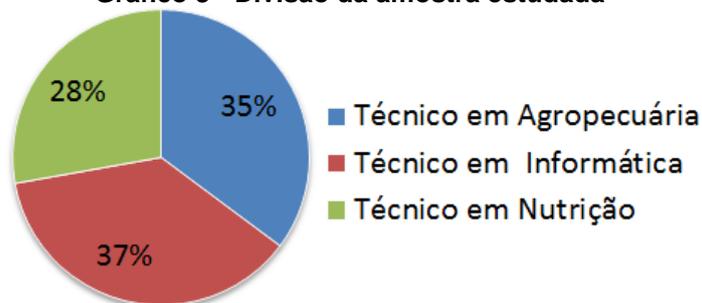
Fonte: Dados da pesquisa

**Gráfico 2 - Dados após fase de limpeza**



Fonte: Dados da pesquisa

**Gráfico 3 - Divisão da amostra estudada**



Fonte: Dados da pesquisa

Como esta pesquisa visou a busca de conhecimento implícito sobre a base de dados do processo seletivo do IFMG – São João Evangelista, foi utilizado as respostas dos candidatos aos questionários socioeconômicos e cultural criado pela COPEVES, com isso, utilizou-se dados secundários. Os questionários utilizados para os três processos seletivos apresentam disparidades quantos a algumas questões, desta forma, na intenção de homogeneizar os questionários, buscou-se criar um novo questionário, ele possui

indagações comuns aos três questionários e o resultado da situação do candidato quanto à aprovação no curso inscrito. A partir do novo questionário, recuperaram-se todas as informações pertinentes ao mesmo. Com a utilização do modelo de processo de DCBD de Adriaans e Zantinge, buscou-se utilizar as tarefas de associação, agrupamento e classificação para aplicar a mineração de dados. Para a tarefa de associação empregou-se o algoritmo *Apriori*, na tarefa de agrupamento utilizou-se o algoritmo *SimpleKmeans* e para a tarefa de classificação utilizou o algoritmo de árvore de decisão *J48*, conforme Russel (2011).

A aquisição da base de dados, referente aos processos de seleção de estudantes para o primeiro semestre dos anos de 2011, 2012 e 2013, foi realizado com a COPEVES do IFMG. A base de dados consistia em todas as questões do questionário socioeconômico respondidas pelos candidatos ao exame de seleção. O acesso a estes dados foi liberado mediante assinatura de um termo de compromisso ao IFMG, que ressaltava a utilização das informações apenas para fins acadêmicos.

## RESULTADOS E DISCUSSÕES:

Na tarefa de associação, conforme Carvalho (2005) procurou-se trabalhar com os padrões que possuíam confiança acima de 70 %. De acordo com o Quadro 1, 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%. Conforme Quadro 2, 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. Com o Quadro 3, nota-se 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. No Quadro 4, percebe-se 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. Através do Quadro 5, inferiu-se que 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.

### Quadro 1 – Regra de associação para candidatos aprovados

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)>
--

Fonte: Dados da pesquisa

### Quadro 2 – Regra de associação para candidatos aprovados

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)>
--

Fonte: Dados da pesquisa

### Quadro 3 – Regra de associação para candidatos excedentes

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)>

Fonte: Dados da pesquisa

### Quadro 4 – Regra de associação para candidatos aprovados no curso, Técnico em Agropecuária

Curso=Técnico em Agropecuária – Integrado Acessa\_Internet=Pouca Renda\_Familiar=Até um  
salário mínimo 127 ==> Status\_Opção=Aprovado 102 <conf: (0.80)>

Fonte: Dados da pesquisa

### Quadro 5 – Regra de associação para candidatos salário acima de três salários

Renda\_Familiar= de 3 a 5 salários 80 ==> Curso=Técnico em Manutenção e Suporte em  
Informática 80 <conf: (1)>

Fonte: Dados da pesquisa

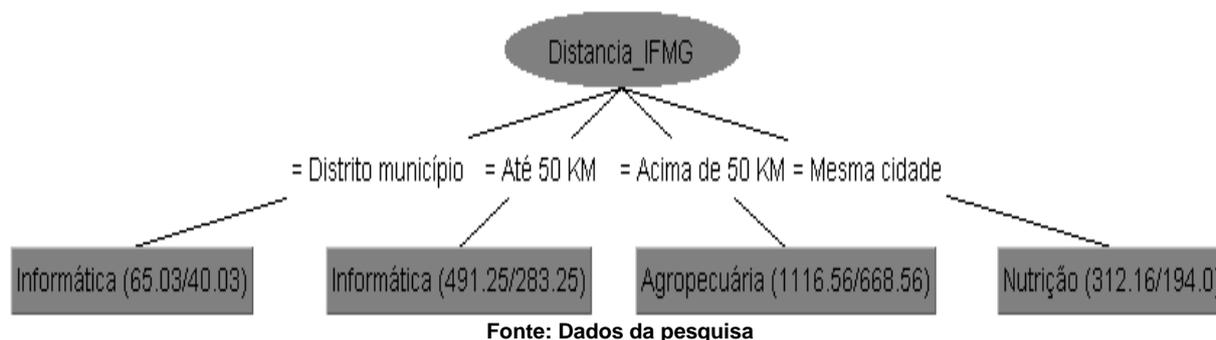
Baseado em Tang, Steinbach, e Kumar (2009), com a tarefa de agrupamento, conforme Figura 1, e a tarefa de classificação, conforme, Figura 2 e Figura 3 notou-se que os candidatos que moram distantes da cidade onde situa o campus do IFMG, são aprovados, possuem renda familiar inferior a um salário mínimo, os pais e mães possuem ensino fundamental incompleto e trabalham no campo ou como funcionários públicos, possuem internet em casa e procuram fazer o curso Técnico em Agropecuária.

Figura 1 – Agrupamento dos candidatos quanto ao curso (Agro, Nutri., Manu.)

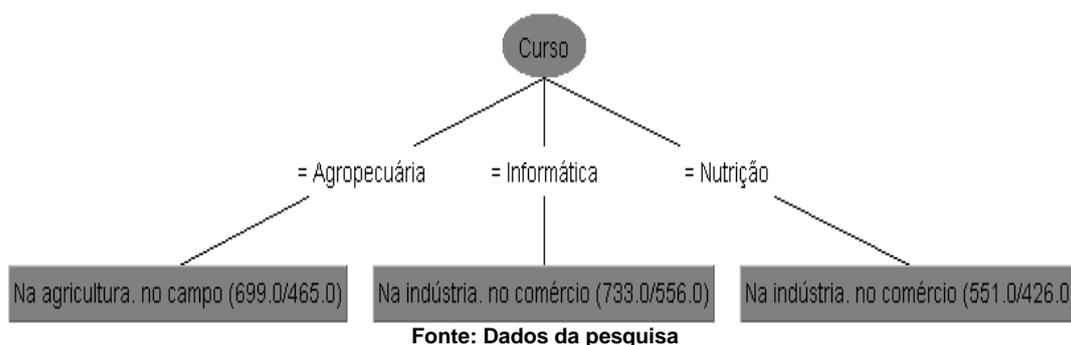
Attribute	0 (584)	1 (697)	2 (704)
Status_Opção	Aprovado	Excedente	Excedente
Conheceu_Vest	Outros meios	Outros meios	Outros meios
Escolheu_IFMG	Pela qualidade de ensino prestada	Pela qualidade de ensino prestada	Pela qualidade de ensino prestada
Escolheu_Curso	Sempre quis este curso	Sempre quis este curso	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

Fonte: Dados da pesquisa

**Figura 2 - Classificação quanto ao curso**



**Figura 3 - Classificação quanto ao trabalho pai**



As hipóteses foram primordiais para nortear a aplicação da técnica de mineração de dados, a maioria das hipóteses foi confirmada. Acreditava-se que os perfis dos candidatos poderiam estar ligados a distância em que moram do IFMG – Campus São João Evangelista, grande parte dos aprovados são 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 (cento e cinquenta quilômetros) 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. O meio de comunicação de maior abrangência para conhecimento do processo seletivo é definido como outros meios pelos candidatos. 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.

## CONCLUSÕES:

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

Todos os objetivos deste estudo foram alcançados e a maioria das hipóteses foi 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. Observou-se 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 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. É 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 no IFMG – Campus São João Evangelista. Com este estudo, também é possível que a instituição tenha um *marketing* direcionado a cada nicho de candidatos e possa complementar no repertório das teorias utilizadas pelos profissionais responsáveis pelos processos seletivos e de atendimentos estudantis. 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 aplicações destes métodos em outras bases de dados, como a mineração de dados em redes sociais, bioinformática e análise de rendimentos estudantis.

#### **REFERÊNCIA BIBLIOGRÁFICA:**

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

RUSSEL, Mattebew A. **Mineração de dados da Web Social**. Tradução: Rafael Zanolli. São Paulo: Novatec Editora; Sebastopol, CA: O'Really, 2011. Mining the social web. 358 p.

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