UNIVERSIDADE FEDERAL DO PAMPA

FRANCISCO CARVALHO PEREIRA

DESCOBERTA DE CONHECIMENTO A PARTIR DE DADOS DAS ELEIÇÕES MUNICIPAIS DAS REGIÕES DO BRASIL

Bagé
2013

FRANCISCO CARVALHO PEREIRA

DESCOBERTA DE CONHECIMENTO A PARTIR DE DADOS DAS ELEIÇÕES MUNICIPAIS DAS REGIÕES DO BRASIL

Trabalho de Conclusão de Curso apresentado ao curso de Especialização em Sistemas Distribuídos com Ênfase em Banco de Dados da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Especialista.

Orientador: Milton Roberto Heinen

Coorientador: Ana Paula Lüdtke Ferreira

Bagé
2013

FRANCISCO CARVALHO PEREIRA

DESCOBERTA DE CONHECIMENTO A PARTIR DE DADOS DAS ELEIÇÕES MUNICIPAIS DAS REGIÕES DO BRASIL

Trabalho de Conclusão de Curso apresentado ao curso de Especialização em Sistemas Distribuídos com Ênfase em Banco de Dados da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Especialista.

Trabalho de Conclusão de Curso defendido e aprovado em: 8 de agosto de 2013.

Banca examinadora:

_____

Prof. Dr. Milton Roberto Heinen
Orientador
UNIPAMPA

_____

Prof.ª Me. Sandra Dutra Piovesan
UNIPAMPA

_____

Prof.ª Me. Carlos Michel Betemps
UNIPAMPA

“Não sou obrigado a vencer, mas tenho o dever de ser verdadeiro. Não sou obrigado a ter sucesso, mas tenho o dever de corresponder à luz que tenho”.

Abraham Lincoln

RESUMO

A atual situação política brasileira fornece ao cidadão o amplo direito de voto, sustentada por um regime democrático caracterizado pela liberdade política. Essa liberdade traz a possibilidade de haver múltiplos partidos e candidatos, fazendo com que o eleitor necessite buscar informações sobre os concorrentes, muitas vezes nos meios de comunicação, ou no próprio ambiente onde vive. Após o término de cada eleição, são gerados muitos dados, que ao serem totalizados, indicam quais candidatos foram eleitos. Mas esses dados não são suficientes para dizer por que um candidato foi eleito, ou seja, qual fator influenciou na escolha do seu nome na urna. Para saber quais são esses fatores é preciso analisar outros tipos de dados, além daqueles gerados no dia da eleição. Esses dados são disponibilizados pelo TSE (Tribunal Superior Eleitoral), onde é possível saber idade, profissão, grau de instrução, estado civil, entre outras informações sobre cada candidato. Tendo em mãos esses dados, é possível relacionar com os dados da eleição e saber o quanto que cada característica influencia no resultado final. Para isso, uma ferramenta útil para extrair esse conhecimento é a mineração de dados, pois lida com grandes bases de dados, buscando padrões não conhecidos anteriormente. A proposta do presente trabalho é utilizar o algoritmo de classificação de dados J48, com auxilio da ferramenta WEKA, para realizar essa tarefa, podendo classificar os dados referentes às eleições para prefeito do ano de 2012, das cinco regiões do Brasil, selecionando um Estado de cada região. Os resultados apontaram que fatores como experiência política do candidato, despesa de campanha e idade influenciaram diretamente no resultado da eleição, além de outros fatores como total de bens e grau de instrução.

Palavras-chave: mineração de dados, classificação de dados, descoberta de padrões.

ABSTRACT

The current Brazilian political situation provides all citizens the right to vote, sustained by a democratic regime characterized by political freedom. This freedom allows multiple parties and candidates, making the elector need to acquire information about competitors, often in the media, or in the environment where they live. After the end of each election are generated a lot of data, which, when summarized, indicate which candidates were elected. But these data are not enough to say why a candidate was elected, i.e., which factors influenced the choice of his name on the ballot. To know which are these factors it is necessary to analyze other types of data beyond those generated on election day. This information is provided by TSE (Supreme Electoral Court), where it is possible to know age, occupation, education level, marital status, and other information about each candidate. Using these data it is possible to find out know how these characteristics influences the final result. For this purpose data mining is a useful tool because it deals with large databases, searching for patterns not previously known. The purpose of this work is to use the J48 data classification algorithm, available at the WEKA software tool to classify data on mayoral elections of the year 2012 in five regions of Brazil by selecting a state of each region. The results show that factors such as political experience of the candidate, campaign spending and age directly influenced the outcome of elections, as well as other factors such as total assets and the level of education of the candidate.

Keywords: *data mining, data classification, pattern discovery.*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ARFF – *Attribute-Relation File Format* (Formato de Arquivo Atributo-Relação)

GPL – *General Public License* (Licença Pública Geral)

IBGE – Instituto Brasileiro de Geografia e Estatística

KDD **–** *Knowledge Discovery Database* (Descoberta de Conhecimento em Banco de Dados)

WEKA – *Waikato Environment for Knowledge Analysis* (Ambiente de Waikato para Análise de Conhecimento)

SUMÁRIO

1 INTRODUÇÃO

A cada eleição que ocorre no Brasil, novos candidatos pleiteiam cargos públicos, buscando conquistar o interesse do eleitorado e alcançar a tão sonhada carreira política. Além disso, os eleitores se tornam cada vez mais críticos, devido ao aumento da disponibilidade de informações em relação aos candidatos, fazendo com que a votação seja uma escolha ainda mais criteriosa.

Além das fontes mais comuns de divulgação eleitoral, a internet é cada vez mais utilizada. De acordo com Figueira (2013), a internet cria espaços conversacionais não hierárquicos, mudando a experiência de campanha, divulgando de certa maneira opiniões políticas através das mídias sociais.

As informações sobre os candidatos, portanto, são buscadas facilmente, o que tornam a vida política ou profissional do candidato uma grande influência na escolha do mesmo. Além deste fator importante, outros vêm a tona, como o total de investimento na campanha política que, muitas vezes, proporciona um maior gasto em propaganda, e as características do candidato, como idade, sexo e profissão, que podem influenciar na escolha do mesmo.

A eleição é descrita por Oliveira (2012) como uma representação do processo eleitoral, ou seja, é o relato do modo como o eleitor votou, e sobre em quais segmentos socioeconômicos, ideológicos ou religiosos um candidato alcançou mais votos. O autor também afirma que decifrar o fenômeno eleitoral vai além deste aspecto, representa identificar as causas que motivam ou levam o indivíduo ou o grupo deles a tomar esta importante decisão. Identificando o comportamento nas eleições, fenômenos sociais, mais precisamente os eleitorais, são elucidados.

Tendo em vista a quantidade de dados disponíveis de cada eleição torna-se, então, interessante avaliar o quanto que cada fator citado acima é determinante para o resultado final. Para auxiliar nessa avaliação, a mineração de dados é fundamental, pois através dela pode-se extrair estas informações.

A tarefa de mineração de dados exige a definição de um escopo para estudo, ou seja, a definição de quais dados serão utilizados, como, por exemplo, elegendo uma categoria de candidato, ou selecionando por unidade federativa, a fim de reduzir a quantidade de dados, facilitando a análise dos mesmos.

A proposta é, portanto, utilizar os dados dos candidatos a prefeito, por se tratar de um cargo que exige que o eleito tenha uma proporção maior de votos, se comparado a outros

cargos, como, por exemplo, vereador. Além disso, o cargo de prefeito possui um menor número de candidatos por cidade, o que reduz a quantidade de dados a serem analisados.

Além de definir qual tipo de candidatura a ser estudada, se fez necessário delimitar ainda mais os dados, pois considerando a quantidade de municípios do país, torna-se inviável a análise de todos os dados disponíveis. Portanto, definiu-se como alvo 5 (cinco) estados brasileiros, Rio Grande do Sul, São Paulo, Bahia, Pará e Goiás, sendo um de cada região do Brasil, incluindo portanto, todos seus municípios. Após esta definição, foram minerados os dados dos estados escolhidos, fazendo uma comparação entre os resultados apontados.

Tendo em mente a complexidade que o processo de mineração de dados envolve, torna-se importante a utilização de uma ferramenta apropriada, que possua algoritmos para extrair o conhecimento pretendido. Para cumprir este objetivo foi escolhida a ferramenta WEKA[1] (do inglês *Waikato Environment for Knowledge Analysis*) desenvolvida pela Universidade de Waikato, da Nova Zelândia.

Conforme cita Witten and Frank (2005), a ferramenta WEKA fornece uma variedade de funções para transformar um conjunto de dados, incluindo algoritmos de vários tipos, permitindo processar um conjunto de dados, baseando-se em um esquema de aprendizagem, e analisar os resultados e desempenho, tudo sem necessitar escrever nenhum código.

Com essa ferramenta pode-se extrair conhecimento da base de dados em estudo, podendo saber o quanto que fatores como idade, sexo, profissão e total de investimento em campanha foram determinantes para a eleição dos candidatos.

[1] Disponível em http://www.cs.waikato.ac.nz/ml/weka/

1.1 Objetivo Geral

Encontrar possíveis padrões em dados eleitorais de candidatos a prefeito das eleições do ano de 2012 no Brasil, indicando possíveis tendências relevantes, utilizando técnicas de mineração de dados disponíveis na ferramenta WEKA.

1.2 Objetivos Específicos

- Filtragem e padronização dos dados disponíveis;
- Processamento e captação de resultados utilizando WEKA;
- Análise dos resultados obtidos.

1.3 Estrutura do Trabalho

O presente trabalho possui mais 5 seções, sendo elas:

- Seção 2 (Situação política do Brasil): exibe uma breve descrição sobre a situação política do Brasil, bem como aspectos importantes referentes aos eleitores e os fatores que influenciam o voto.
- Seção 3 (Mineração de dados): apresenta conceitos de alguns autores sobre mineração de dados, incluindo principais métodos e algoritmos.
- Seção 4 (Desenvolvimento): nesta seção são exibidos os passos do desenvolvimento da proposta do trabalho.
- Seção 5 (Metodologia): neste capítulo são identificados os tópicos do desenvolvimento e como que cada etapa foi realizada.
- Seção 6 (Conclusão): nesta etapa as conclusões finais são apresentadas, baseando-se nos resultados obtidos.

2 SITUAÇÃO POLÍTICA DO BRASIL

Nos dias atuais o Brasil vive um período democrático, marcado principalmente pelo amplo direito de voto por parte do cidadão. Esse cenário passou a ser construído pelo movimento popular conhecido por *Diretas Já*, um dos maiores da história do Brasil, surgido a partir da *Emenda Dante de Oliveira*, que culminou com a constituinte que elaborou, posteriormente, a constituição de 1988, instituindo a democracia Brasileira. Antes, em 15 de janeiro de 1985, foi eleito para presidência da república, Tancredo Neves, marcando o início da chamada *Nova República* (LIMA, 2012).

De acordo com Felisbino et al. (2012), o Brasil completou em 2008, 22 anos de governo civil e 19 anos de experiência democrática, fundamentada pela Constituição Federal de 1988. De acordo com o autor, nos dias de hoje, com a atual maturidade política, não há temeridade em relação ao retorno do regime militar, pois os membros das elites políticas e o povo sabem que a democracia está acima de qualquer outro regime e acreditam ser a melhor opção para o país.

2.1 A liderança política e os cargos públicos

Um dos fatores que mais influenciam na probabilidade de um candidato ocupar um cargo político é a posição hierárquica social ou posição social de origem do candidato, como por exemplo, a última profissão antes da ocupação do cargo. Independente do posto considerado, podendo ser prefeito, gestor, deputado ou vereador, grande parte destes cargos são ocupados por membros das classes superiores da sociedade, como grandes produtores rurais, industriais e grandes empresários, executivos e intelectuais diversos. Desse modo, a representação parlamentar projeta uma imagem invertida da estrutura social, considerando que mais de três quartos dos deputados originam-se da parte mais favorecida, social e culturalmente da população (GAXIE, 2012).

Outro fator relevante a ser observado é o diz respeito à quantidade de candidatos que cada cidade possuiu nas últimas eleições. Rodrigues (2011) cita que em 2004, 14 pessoas apresentaram candidatura para prefeito na eleição de primeiro turno da cidade de São Paulo, e no mesmo ano, apenas um candidato concorreu à eleição da cidade de Bom Jardim da Serra (SC), e obviamente, se tornou vencedor da mesma.

Uma das explicações possíveis para este fato, segundo Rodrigues (2011), é que a estrutura institucional determina o número de candidatos, fazendo com que a presença de

eleições diretas e turno único influenciem no número de concorrentes, justificativa esta frequentemente mencionada para a grande ocorrência de eleições com apenas dois candidatos para a presidência dos Estados Unidos. Por outro lado, conforme cita também o autor, as eleições que tem possibilidade de haver segundo turno, como para governador e presidente do Brasil, atrairiam um número maior de candidatos ao pleito.

2.2 Fatores que influenciam o voto

O comportamento do eleitor nas urnas pode ser explicado, baseando-se na ciência política, seguindo três teorias, a sociológica, a psicológica e a econômica. A teoria econômica ou racial indica que o cidadão age nas urnas de acordo com motivações pessoais, como, por exemplo, buscando bem-estar e melhoria na sua situação econômica. Esse comportamento é fruto do baixo interesse em política, que implica em pouco gasto na obtenção de informação sobre essa temática (FIGUEIRA, 2013).

Figueira (2013) cita que, na abordagem sociológica, diferentemente da anterior, o voto é resultado do contexto social do votante, ligado principalmente às interações com a sociedade onde vive e seus relacionamentos políticos. Pode-se dizer então que a diferença entre as duas teorias encontra-se na decisão do eleitor. Na teoria econômica o eleitor age racionalmente, tendo em vista sua situação individual, já na teoria sociológica, a decisão baseia-se na combinação de seus anseios com os desejos da sociedade onde vive. Figueira (2013) cita também a teoria psicológica, que, segundo o autor, baseia-se na racionalidade de baixa-informação, o que significa que o eleitor acrescenta no seu conhecimento político aquilo que é exibido nas campanhas eleitorais, no cotidiano, nas experiências da vida e nas relações que possui em sociedade.

3 MINERAÇÃO DE DADOS

A tarefa de mineração de dados é definida de diversas formas na literatura atual, entre as definições encontradas, aquelas consideradas mais relevantes foram destacadas.

A mineração de dados oferece a capacidade de visualizar os dados sobre uma nova luz, descobrindo associações e padrões não descobertos anteriormente. Simplificando, a mineração de dados nada mais é que o processo de extração de conhecimento novo ou padrões desconhecidos de um conjunto de dados existente (HAGOOD, 2012).

Para Joseph, Sadath e Rajan (2013) a mineração de dados pode ser definida como um ramo da ciência da computação que lida com o processamento de grandes conjuntos de dados e descobrimento de padrões nos mesmos. O principal objetivo da mineração de dados, de acordo com o autor, é extrair informação para, posteriormente, transformar a mesma em uma estrutura compreensível que possa ser utilizada no futuro.

Um fato relevante, para Bacardit e Llorà (2013), que deve ser considerado é que atualmente vivemos na era *petabyte*[2], caracterizada pela grande geração de informações. De acordo com o autor, necessitamos cada vez mais da análise, processamento e transformação de dados. Segundo o mesmo, a utilização de ferramentas robustas de mineração de dados se faz necessário para alcançar este objetivo.

Atualmente a proliferação da informação também é um fato presente no cotidiano. Luiz *et al.* (2012) explica que essa proliferação ocorre principalmente pelo grande avanço da tecnologia computacional, o que aumenta a capacidade de geração e armazenamento de dados. O autor destaca também que a tarefa de extrair conhecimento das bases de dados se torna algo inviável, sem o auxílio de alguma ferramenta de mineração de dados apropriada.

Segundo Sundar, Latha e Chandra (2012), os primeiros estudos referentes à mineração de dados surgiram a mais de duas décadas, mas o seu potencial só está sendo explorado agora. Para o autor, mineração de dados é a combinação entre análise estatística, aprendizado de máquina e tecnologia de banco de dados, a fim de extrair padrões e relacionamentos que não foram percebidos anteriormente.

[2] Unidade de armazenamento de dados equivalente a 1024 terabytes, sendo 1.152.921.504.606.846.976 bytes (Digerati, 2009).

3.1 Descoberta de conhecimento em banco de dados

A descoberta de conhecimento em banco de dados, ou KDD (*Knowledge Discovery Database*) pode ser definida como um processo não trivial de identificar novos padrões, potencialmente úteis e finalmente compreensíveis nos dados (PADHY, MISHRA E PANIGRAHI, 2012).

Para conseguir realizar a tarefa de extrair conhecimento em banco de dados é preciso seguir alguns passos, entre eles pode-se citar: seleção, pré-processamento, transformação, mineração de dados (MD) e interpretação dos resultados. A etapa de mineração de dados, especificamente, pode ser dividida nas tarefas de classificação, regressão, associação, formação de agrupamentos (*clustering*, em inglês) e detecção de anomalias (*outliers*) (LUIZ *ET AL.*, 2012).

Para Gupta, Kumar e Sharma (2011), o processo de descoberta de conhecimento em banco de dados ocorre de outra forma, seguindo os seguintes passos:

- Limpeza dos dados: é a fase onde os dados irrelevantes são removidos;
- Integração de dados: nesta etapa múltiplas fontes de dados, podendo ser heterogêneas, são combinadas em um único conjunto de dados;
- Seleção de dados: momento onde os dados utilizados para análise são escolhidos da coleção de dados;
- Transformação de dados: fase conhecida também como consolidação de dados, onde os dados selecionados são transformados no formato apropriado para o processo de mineração;
- Mineração de dados: é o passo crucial no processo de KDD, onde técnicas são utilizadas para extrair padrões potencialmente uteis nos dados;
- Avaliação de padrões: neste momento do processo, os padrões estritamente interessantes são identificados com base em medidas previamente indicadas; e
- Representação do conhecimento: no final, o conhecimento descoberto é representado visualmente para o usuário. Nesta fase, técnicas de visualização são usadas para ajudar os usuários a entender e interpretar os resultados da mineração de dados.

3.2 Técnicas de mineração de dados

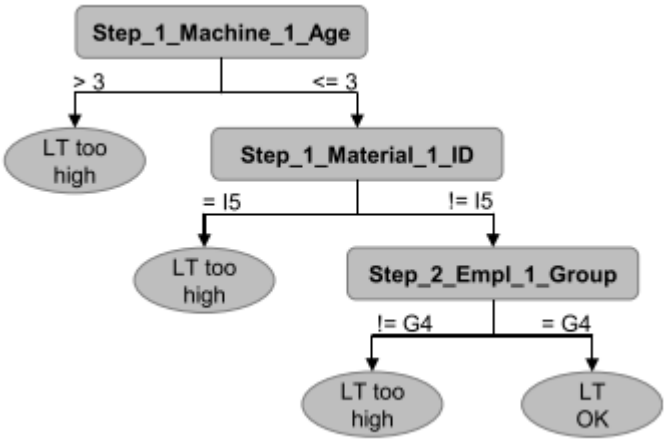
A escolha do método adequado para realizar a mineração dos dados selecionados é um passo importante para alcançar o conhecimento pretendido. Entre as diversas técnicas existentes, podem-se classificar as mesmas em clássicas, como: estatística, *clustering* e vizinhança mais próxima (*Nearest Neighborhood*, em inglês), e aquelas da nova geração, que são as árvores de decisão, redes neurais e regras de associação. (JOSEPH, SADATH E RAJAN, 2013).

3.2.1 Classificação

De acordo com Baradwaj e Pal (2011) a classificação é a técnica de mineração de dados mais aplicada atualmente, que utiliza um conjunto de exemplos pré-classificados, criando um modelo que consegue classificar a população em geral. Esta abordagem, segundo o autor, utiliza-se de árvores de decisão ou algoritmos de classificação baseados em rede neural.

Conforme cita Gröger, Niedermann e Mitschang (2012), as árvores de decisão podem ser consideradas um modelo que possui como principal característica a fácil, compreensível e intuitiva interpretação devido a sua representação em forma de árvore (conforme é exibido na figura 1). Sua estrutura é composta de nós, ou folhas, que representam um teste, onde cada nó é o resultado desse teste, exibindo um rótulo de classe.

Figura 1 – Árvore de decisão



Fonte: Gröger, Niedermann e Mitschang (2012)

3.2.2 Algoritmo J48

O algoritmo J48 foi desenvolvido por *J. Ross Quinlan,* e representa uma versão do algoritmo C4 (Kaur, Mohan e Sandhu, 2012). Ele utiliza dois métodos de poda[3], no primeiro é conhecido como substituição de subárvore, ou seja, os nós de uma árvore de decisão podem ser substituídos por uma folha, basicamente, diminuindo o número de testes ao longo de determinado percurso. O segundo tipo é denominado de captação de sub-árvore. Neste caso, diferente do anterior, um nodo pode ser movido para cima em direção à raiz da árvore, substituindo os outros nós do caminho (KAUR, MOHAN E SANDHU, 2012).

Conforme afirma Patil e Sherekar (2013), o J48 é um algoritmo que cria uma árvore binária de decisão simples. Abaixo encontra-se a estrutura do algoritmo (figura 2), seguido de uma breve explicação do autor.

Figura 2 - Algoritmo J48

```
Algoritmo J48:
ENTRADA:
        D //Dados de treinamento
SAÍDA
        T//Árvore de decisão
DTBUILD (*D)
{
T=ϕ; T= Cria nó raiz e rótulo com atributo de divisão;
T= Adiciona arc para o nó raiz para cada divisão de predicado e rótulo;
Para cada arc faz
        D= Banco de dados criado pela aplicação de divisão de predicado para D;
                Se o ponto de parada atingido por este caminho, então,
                        T'= cria o nó folha e rótulo com classe apropriada
                Se não
                        T'= DTBUILD(D);
                T= adiciona T' para arc;
}
```

Fonte: (Baseado no algoritmo apresentado por Patil e Sherekar, 2013)

[3] O processo de poda remove as ramificações que não tem relevância para o modelo de classificação, selecionando a sub-árvore com menor taxa de erro estimada (Oliveira, 2001).

Durante a construção da árvore são ignorados os valores em falta, ou seja, o valor de um item pode ser previsto de acordo com os valores dos atributos para os outros registros. O objetivo é dividir os dados em serie, com base nos valores de atributos para esse item que se encontram na amostra (PATIL E SHEREKAR, 2013).

No processo de construção da árvore de decisão J48, sempre que um conjunto de itens é encontrado, o algoritmo identifica o atributo que melhor discrimina os diversos exemplos, ou seja, entre os valores possíveis, caso houver qualquer valor que não encontre ambiguidade, então esse ramo é encerrado e o valor obtido é atribuído a ele (RAVICHANDRAN, SRINIVASAN E RAMASAMY, 2012).

3.2.3 Regras de associação

As regras de associação tem papel importante na tarefa de mineração de dados, o objetivo delas é encontrar associações, relações ou correlações interessantes em um grande conjunto de itens de dados, em resumo, as regras de associação mostram as condições de ocorrência de valor dos atributos em um determinado conjunto de dados (KUMAR E CHADHA, 2012).

Em diferentes áreas, segundo Lee *et al.* (2013), as regras de associação ganharam grande importância, entre elas, comércio, telecomunicações, seguros e bioinformática, auxiliando no desenvolvimento de várias corporações.

Regra de associação é definida por Shweta e Garg (2013) como as declarações que fazem relações entre dados de um determinado banco de dados. Segundo o autor, uma regra de associação tem duas partes, aquela denominada antecedente, e a seguinte, definida por consequente, por exemplo, {ovo} => {leite} (nesse caso o ovo é o antecedente e o leite é o consequente), no caso, a ocorrência do primeiro item depende do item seguinte da regra.

Conforme cita Angeline e James (2012), as regras de associações foram propostas por Agrawal *et al.* (1993), desde então muitos algoritmos para a geração de regras de associação foram criados, entre os mais populares estão Apriori, Eclat e FP-Growth.

3.2.4 Algoritmo Apriori

O algoritmo Apriori foi proposto por R. Agarwal e R. Srikant em 1994 com objetivo de minerar conjuntos de itens frequentes em banco de dados, utilizando regras de associação booleanas (KUMAR E CHADHA, 2012).

De acordo com Shweta e Garg (2013), Apriori é uma palavra de origem latina e seu significado é "com o que vem antes". Segundo o autor este tipo de algoritmo utiliza estratégia de baixo para cima; é considerado o mais famoso e clássico algoritmo para mineração de padrões frequentes.

3.3 Ferramentas de mineração de dados

A tarefa de mineração de dados se torna facilitada se tiver o auxílio de uma ferramenta apropriada. Entre os diversos softwares disponíveis foram escolhidas dois, para serem analisados e conhecer melhor as suas finalidades, entre eles o software Weka (utilizado no presente trabalho) e o Rapidminer.

3.3.1 Rapidminer

Rapidminer, segundo Costa *et al.* (2012), é um software livre e com código aberto, distribuído de forma independente para análise de dados, permitindo a integração com outros produtos desenvolvidos pelo mesmo projeto. Algumas características destacadas pelo autor são:

- Disponibilidade gratuita da ferramenta
- Funciona na maioria das plataformas
- Interface gráfica intuitiva
- Integração com diferentes fontes de dados

3.3.2 WEKA

O software WEKA é descrito por Nassif (2013) como uma ferramenta livre, sob a licença GPL *(Genaral Public License)*, desenvolvida na Universidade de Waikato. De acordo com o autor, ela foi desenvolvida em Java e contém uma interface gráfica para interagir com arquivos de dados e produzir resultados visuais (árvores, curvas e tabelas).

De acordo com Hall *et al.* (2009) a ferramenta WEKA tem várias interfaces gráficas que possibilitam o fácil acesso as suas funcionalidades. O autor relata também que a ferramenta possui, como interface principal, a “Explorer”, onde cada painel representa uma tarefa de mineração de dados diferente, além de um painel denominado “Pré-processo” (*preprocess* em inglês), onde as fontes de dados são carregadas e os atributos são selecionados.

4 DESENVOLVIMENTO

Neste capítulo serão exibidos os detalhes do processo de mineração de dados eleitorais, conforme a abordagem proposta. Este processo tem como objetivo encontrar regras de classificação, descobrindo padrões nos dados apresentados, utilizando algoritmos de *Data Mining*, através do software WEKA. Foi definido como escopo apenas os dados de candidatos a prefeito, permitindo assim uma analise de todos os municípios de cada estado (conforme citado na seção de introdução), sem utilizar uma base de dados tão extensa, se comparado aos dados de candidatos a vereador.

4.1 Procedimentos

Nesta etapa serão exibidos os procedimentos realizados para o desenvolvimento do trabalho proposto. Será apresentada a origem dos dados utilizados e em qual formato eles estão disponíveis, assim como as ferramentas escolhidas para realização das tarefas que foram idealizadas.

4.1.1 Dados utilizados

Os dados utilizados no presente trabalho são fornecidos, de forma livre e gratuita, pelo Tribunal Superior Eleitoral através do portal "Repositório de dados eleitorais" (http://www.tse.jus.br/eleicoes/repositorio-de-dados-eleitorais), estando os mesmos disponíveis para *download*. Estes encontram-se organizados cronologicamente, permitindo os usuários acessarem dados das eleições municipais e nacionais do ano de 1994 até o ano de 2012.

Estes dados são disponibilizados em diretórios compactados, contendo arquivos de texto (no formato de texto plano) para cada unidade federativa. Estes arquivos contém 1 (um) registro por linha e dados dos candidatos separados por ponto e vírgula (;), fornecendo, como por exemplo, estado civil, data de nascimento, partido e resultado, no caso, se foi ou não eleito. Além de dados de candidatos de cada eleição, pode-se ter acesso a relatórios de dados do eleitorado.

Em relação à proposta de mineração de dados do presente trabalho, foram selecionados os dados dos candidatos a prefeito dos estados do Rio Grande do Sul, São Paulo, Bahia, Goiás e Pará, podendo assim avaliar a situação de todas as regiões do país, cada uma

com suas características sociais, culturais e financeiras. Abaixo encontra-se uma tabela que exibe a quantidade de candidatos por estado e o total de dados selecionados.

Tabela 1 - Dados utilizados

Estado	Total de candidatos	Candidatos a prefeito
Rio Grande do Sul	29052	1216
São Paulo	81622	2119
Bahia	36345	1206
Goiás	20587	690
Pará	18780	504
Total	**186386**	**5735**

Fonte: Arquivo pessoal

Além das informações pessoais dos candidatos, torna-se interessante avaliar a importância que outros dados, que se encontram disponíveis no repositório citado, teriam na tarefa de descoberta de conhecimento proposta, mesmo que estes não estejam no conjunto de dados citado anteriormente.

Entre estes dados que não foram citados, aqueles referentes aos bens dos candidatos também estão disponíveis. Os mesmos estão disponibilizados da mesma maneira que os dados cadastrais dos candidatos, ou seja, armazenados em arquivos de texto, sendo um arquivo para cada unidade federativa, contendo informações dos bens dos candidatos que foram declarados antes da eleição. Entre estes valores estão o valor estimado, tipo, descrição e, principalmente, um número que identifica o candidato proprietário do bem.

A partir destes dados é possível relacionar os mesmos com os dados cadastrais dos candidatos, gerando, por exemplo, o total de bens de cada indivíduo, exibindo assim um novo atributo que pode auxiliar na tarefa de mineração de dados pretendida, tornando possível avaliar o quanto que esta característica pode influenciar no resultado da eleição.

4.1.2 Ferramentas utilizadas

Para processamento e filtragem dos dados foi utilizada a planilha eletrônica Microsoft Office Excel 2007, pois permite a manipulação dos dados de maneira facilitada com a utilização de mecanismos de filtragem.

A realização do processo de mineração de dados, proposto neste trabalho, depende da utilização de algoritmos apropriados. Para dar suporte a esta tarefa foi utilizada a ferramenta

WEKA (*Waikato Environment for Knowledge Analysis)*, que encontra-se disponibilizada no site http://sourceforge.net/projects/weka/?source=dlp. Outras ferramentas

4.2 Etapas do desenvolvimento

O processo de mineração de dados proposto compõe-se das seguintes etapas:

- Definição do escopo
- Seleção de dados
- Definição de atributos
- Transformação de dados
- Mineração dos dados
- Interpretação dos resultados

4.2.1Definição do escopo

Após analisar os dados que são disponibilizados no site do TSE, definiu-se qual escopo deveria ser adotado para realização da tarefa proposta. Foram escolhidos, portanto, os dados das eleições municipais do ano de 2012, referentes aos estados do Rio Grande do Sul, São Paulo, Bahia, Goiás e Pará.

Após analisar os arquivos disponíveis, foram verificados que os mesmos continham dados de candidatos a prefeito, vice-prefeito e vereador, possuindo aproximadamente 30 mil registros (no caso, os dados referentes aos candidatos do estado do Rio Grande do Sul). Como o objetivo é analisar apenas um tipo de candidatura, foram filtrados apenas os dados referentes a candidatos a prefeito, o que reduz o total de registros e, consequentemente, abrangendo todos os municípios dos estados escolhidos, conhecendo a realidade das pequenas cidades até as capitais dos estados.

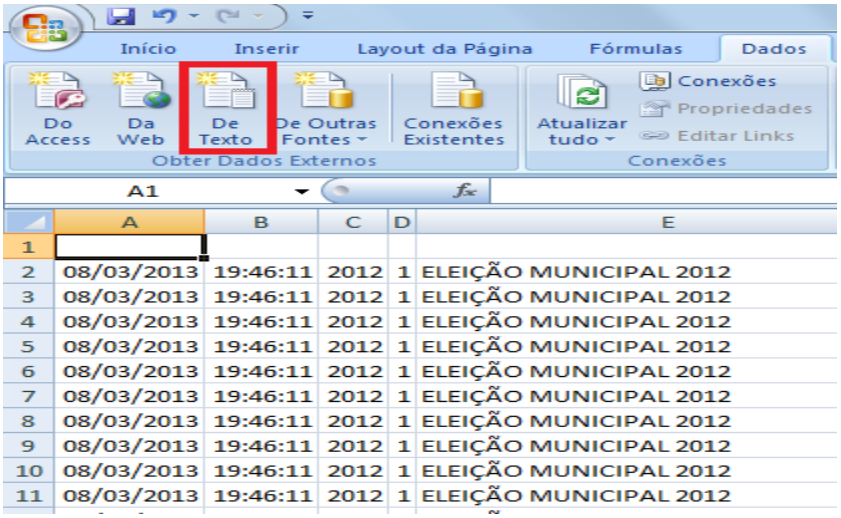
4.2.2 Seleção de dados

Após a definição do escopo a ser estudado, a próxima tarefa refere-se à filtragem e seleção dos dados, necessária, pois é preciso utilizar apenas as informações referentes aos candidatos a prefeito.

Considerando que os dados encontram-se armazenados em arquivos de texto, é preciso importar os mesmos para alguma ferramenta capaz de realizar a filtragem pretendida. Para

isso foi escolhida a planilha eletrônica *Excel* 2007, da *Microsoft*, utilizando a opção de importação de dados (conforme exibido na figura 3).

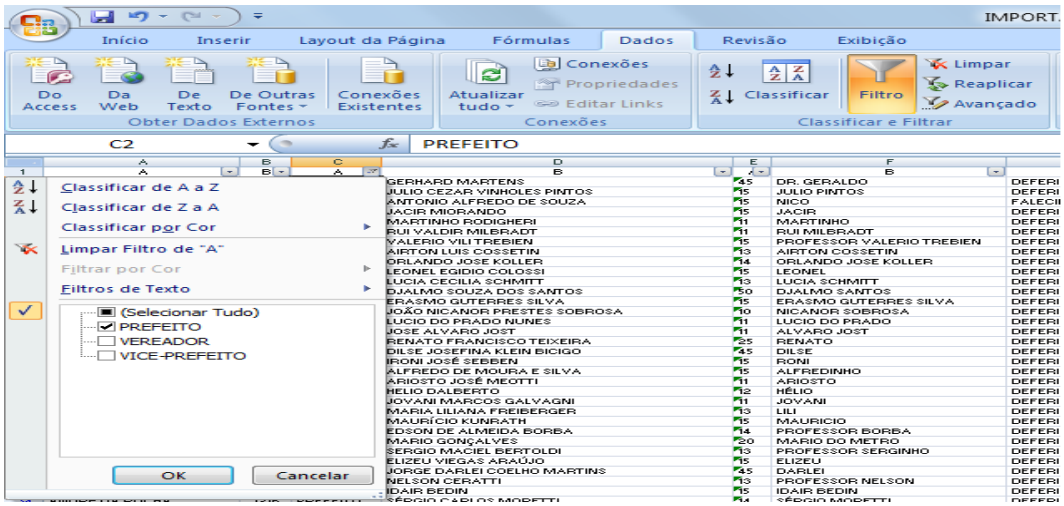
Figura 3 - Importação de dados



Fonte: Arquivo pessoal

Após a importação dos dados, a tarefa de filtragem de dados se faz necessária, considerando que o arquivo importado possui informações de candidatos a prefeito, vereador e vice-prefeito. Para realizar esta tarefa, foi utilizado o recurso "filtro" (conforme exibido na figura 4), selecionando apenas o valor "prefeito".

Figura 4 - Filtragem de registros



Fonte: Arquivo pessoal

4.2.3 Definição de atributos

Após realizar a seleção de dados foi necessário definir quais atributos seriam utilizados e quais poderiam ser úteis para alcançar o objetivo proposto. O arquivo disponível pelo TSE contém 28 atributos, sendo muito deles, apenas de informação como, partido, coligação, nome do candidato, entre outros, que não foram considerados importantes para o que foi proposto, por não serem ligados às características pessoais dos candidatos.

Dentre os atributos disponíveis foram selecionados:

- Data de nascimento
    - Atributo referente a data de nascimento do candidato, indicado no formato "dia/mês/ano".
- Profissão
    - Profissão do candidato, informada antes da eleição, descrita por extenso (professor, dentista, veterinário, etc.)
- Sexo
    - Gênero do candidato (masculino ou feminino)
- EnsSuperior
    - Atributo que indica se o candidato possui ou não ensino superior.
- Estado Civil
    - Condição civil do candidato antes das eleições (casado, solteiro, divorciado, etc).
- Despesa de Campanha
    - Total de investimento do candidato na campanha eleitoral, representado por um valor numérico.
- Cidade Nascimento
    - Cidade onde candidato nasceu ou foi registrado.
- Eleito/Não eleito
    - Situação final do candidato no pleito, indicando se o mesmo foi ou não eleito.

Na escolha destes atributos é importante destacar que o objetivo é selecionar apenas aqueles que podem trazer algum resultado, ou seja, que podem indicar alguma tendência ou padrão. Portanto, aqueles atributos relacionados a partidos ou coligações não foram

considerados relevantes, pois o foco principal do trabalho está nas características dos candidatos.

4.2.4 Transformação de dados

Durante os primeiros testes realizados na ferramenta WEKA, com os dados que foram selecionados, foi observado que os resultados não eram claros e a ferramenta não permitia a execução de alguns algoritmos, principalmente as árvores de decisão e regras de associação. O motivo principal é que os dados não estavam formatados e organizados devidamente. Por exemplo, o atributo que exibe a data de nascimento possui uma grande variação de valores, aspecto que dificulta o processamento dos dados e o entendimento dos resultados.

Portanto, para maior aproveitamento da ferramenta WEKA, alguns atributos foram alterados, com objetivo de reduzir a variação dos valores, classificando os dados em níveis. Foi necessário então atribuir uma escala para cada um. Abaixo estão as alterações que foram feitas nos atributos citados.

- Data de nascimento:
    - Neste atributo foi preciso alterar a representação dos valores, alterando para idade, ao invés da data de nascimento. Para isso foi realizado um cálculo da idade do candidato, tendo como referência a data da eleição.
    - Após esta alteração os dados foram classificados da seguinte maneira:
        - mAlta(71-80)
        - Alta(61-70)
        - Media(41-60)
        - Baixa(31-40)
        - mBaixa(18-30)
- Despesa
    - Seguindo o princípio do atributo anterior, os dados referentes aos gastos em campanha foram classificados, com objetivo de facilitar a tarefa de mineração de dados. A classificação ficou na seguinte forma:
        - mAlta(>1Milhao)
        - Alta(100mil-1mi)
        - Média(10-100mil)
        - Baixa(<10mil) }

Além disso, outros atributos foram padronizados, permitindo apenas os valores “sim” e “não”, seguindo o esquema abaixo.

- Profissão => Politico { não,sim }
    - Devido a este atributo possuir uma grande variação de valores, foi decidido atribuir apenas os valores “sim” ou “não”, identificando se o candidato exerce algum cargo político (prefeito, vereador, deputado, por exemplo) no período da campanha.
- Estado Civil => Casado { sim,não }
    - Este atributo foi padronizado para exibir apenas se o estado civil do candidato é igual a “casado”. Permitindo os valores “sim” e “não”.
- Cidade Nascimento = > CidNasc { não,sim }
    - Já este atributo foi definido como “CidNasc”, indicando se o candidato nasceu ou não na cidade de candidatura. Permitindo também os valores “sim” e “não”. No caso, foi feito uma fórmula para verificar se o atributo que indica a cidade de nascimento é igual ao atributo que representa a cidade de candidatura.
- EnsSuperior
    - Este atributo foi alterado para exibir apenas se o candidato possui ou não nível superior.

Após as alterações, a tabela com os dados finais ficou com a seguinte composição, conforme é exibido na imagem abaixo (ver figura 5).

Figura 5 - Dados finais

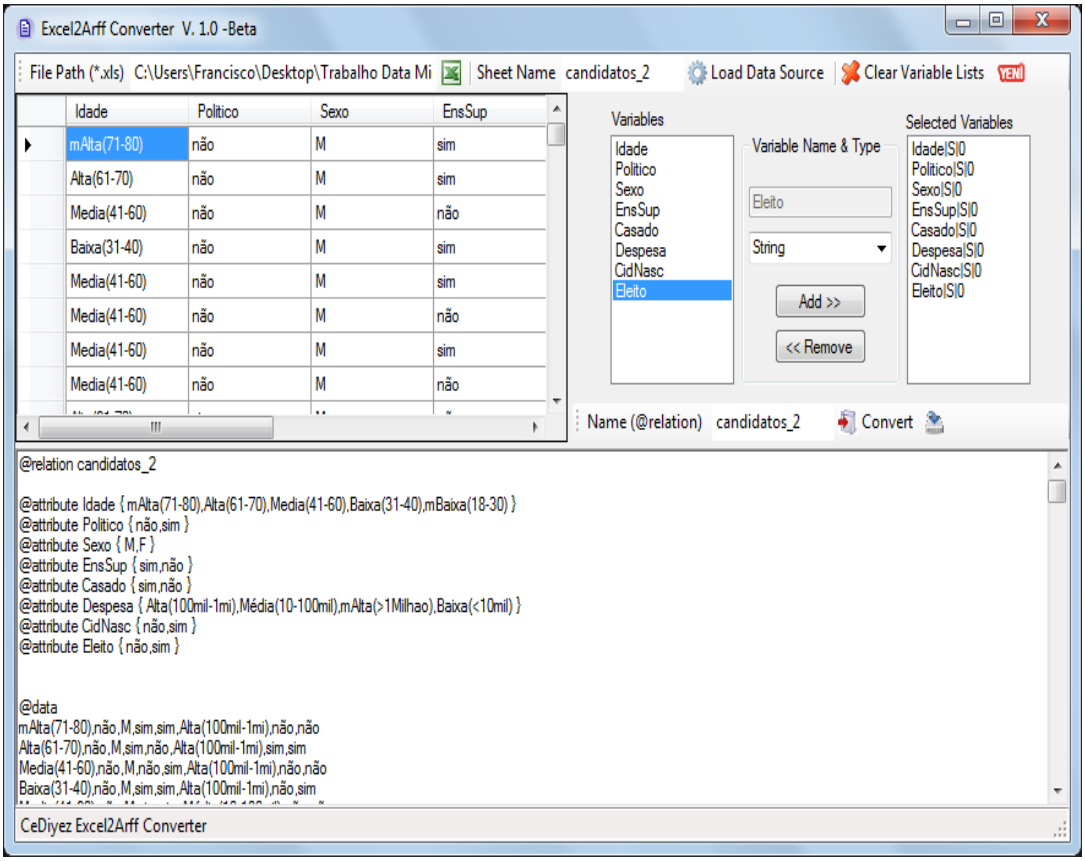
Político	Faixa de idade	Sexo	GrauInst	EnsSuperior	Casado	CidNasc	Total de bens	Despesa	Eleito
x	x	x	x		x	x		x	x
não	Alta(61-70)	M	7-SC	sim	sim	não	Medio(500mil-1mi)	Alta(100mil-1mi)	não
não	Alta(61-70)	M	7-SC	sim	não	sim	Baixo(100mil-500mil)	Alta(100mil-1mi)	sim
não	Media(41-60)	M	5-EMC	não	sim	não	Alto(1mi-50mi)	Alta(100mil-1mi)	não
não	Baixa(31-40)	M	7-SC	sim	sim	não	Alto(1mi-50mi)	Alta(100mil-1mi)	sim
não	Media(41-60)	M	7-SC	sim	sim	não	Baixo(100mil-500mil)	Média(10-100mil)	não
não	Media(41-60)	M	6-SI	não	sim	sim	Baixo(100mil-500mil)	Alta(100mil-1mi)	não
não	Media(41-60)	M	7-SC	sim	não	sim	mBaixo(<100mil)	Alta(100mil-1mi)	sim
não	Media(41-60)	M	6-SI	não	sim	não	Baixo(100mil-500mil)	Média(10-100mil)	sim
sim	Alta(61-70)	M	6-SI	não	sim	sim	Baixo(100mil-500mil)	Alta(100mil-1mi)	não
não	Media(41-60)	M	6-SI	não	sim	sim	Alto(1mi-50mi)	Média(10-100mil)	sim
não	Media(41-60)	F	7-SC	sim	não	sim	Baixo(100mil-500mil)	Média(10-100mil)	não
não	Media(41-60)	M	7-SC	sim	sim	não	Baixo(100mil-500mil)	Alta(100mil-1mi)	não
não	Media(41-60)	M	7-SC	sim	não	sim	Medio(500mil-1mi)	Alta(100mil-1mi)	sim
não	Alta(61-70)	M	7-SC	sim	sim	não	Baixo(100mil-500mil)	Alta(100mil-1mi)	não
não	mBaixa(18-30)	M	6-SI	não	sim	sim	Baixo(100mil-500mil)	Alta(100mil-1mi)	não
não	Media(41-60)	M	7-SC	sim	sim	não	Medio(500mil-1mi)	Alta(100mil-1mi)	não
não	Media(41-60)	M	3-EFC	não	sim	não	Baixo(100mil-500mil)	Alta(100mil-1mi)	sim
sim	Media(41-60)	F	7-SC	sim	sim	não	Baixo(100mil-500mil)	Alta(100mil-1mi)	não
sim	Media(41-60)	M	2-EFI	não	sim	não	mBaixo(<100mil)	Média(10-100mil)	sim
não	Alta(61-70)	M	5-EMC	não	sim	não	Baixo(100mil-500mil)	Média(10-100mil)	sim
não	Media(41-60)	M	7-SC	sim	sim	sim	mBaixo(<100mil)	Alta(100mil-1mi)	não

Fonte: Arquivo pessoal

Realizado esta etapa, foi criado um arquivo XLS (arquivo no formato da planilha Excel 2003) apenas com os dados que serão utilizados para realizar o processo de mineração de dados. Após a criação deste arquivo foi realizada a conversão do mesmo para o formato ARFF (Attribute-Relation File Format), utilizado pela ferramenta WEKA.

Para realizar esta conversão foi utilizado a ferramenta "Excel2Arff Converter" (conforme exibido na figura 6).

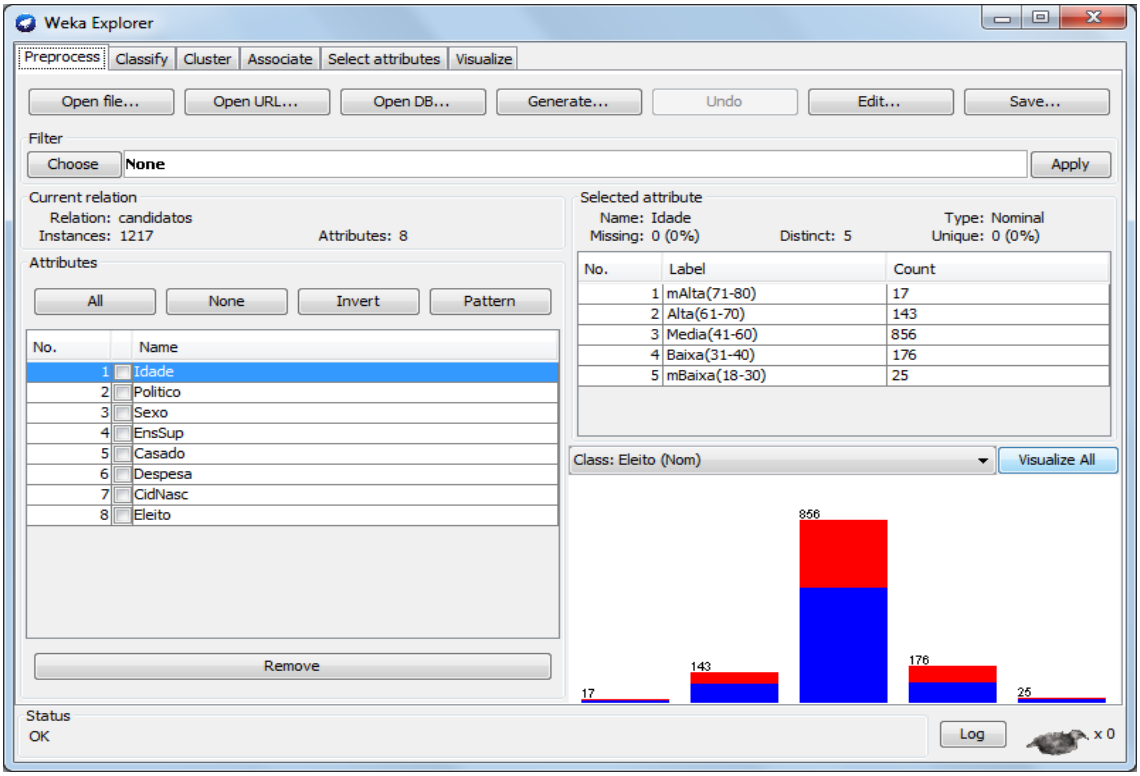
Figura 6 - Ferramenta Excel2Arff Converter



Fonte: Arquivo pessoal

Realizada a conversão, foi preciso processar o arquivo criado na ferramenta WEKA, a fim de obter os resultados pretendidos. Abaixo está a tela principal do software WEKA, onde foi carregado o arquivo ARFF e exibido as informações preliminares (ver figura 7).

Figura 7 - Ferramenta WEKA



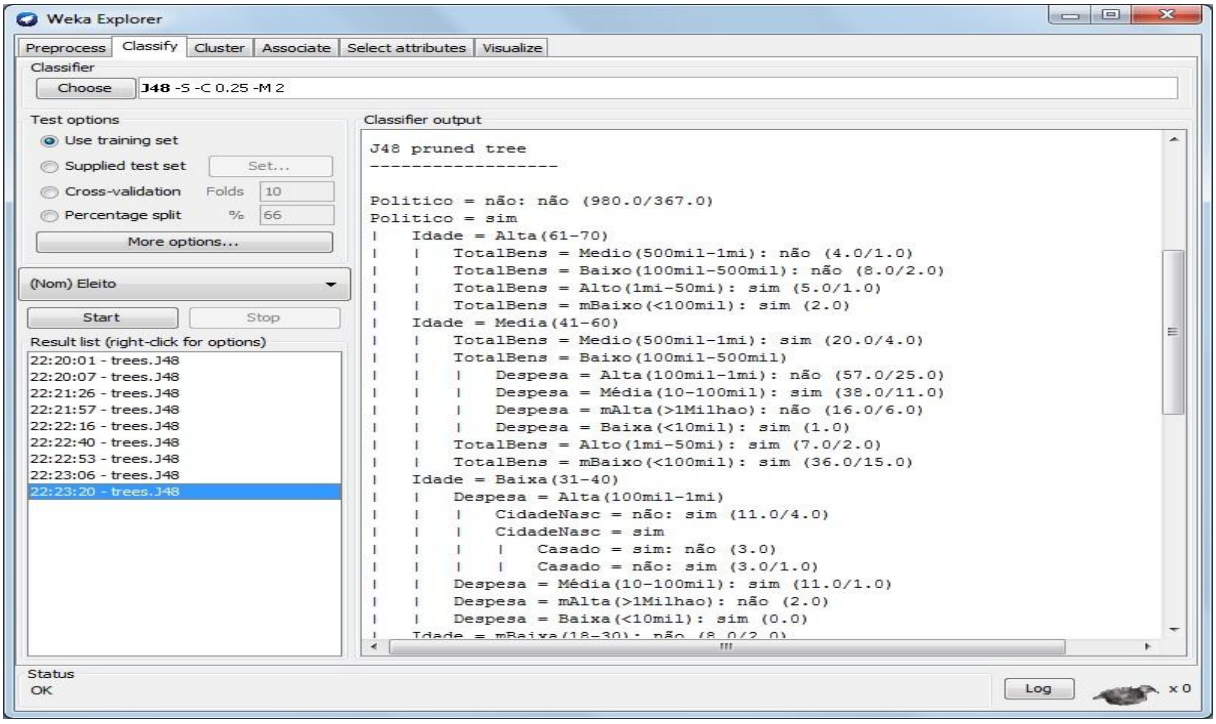
Fonte: Arquivo pessoal

4.2.5 Mineração dos dados

Após a filtragem e seleção dos dados foram carregados os mesmos para ferramenta WEKA, para serem submetidos aos algoritmos de mineração de dados. Entre os diversos tipos de algoritmos disponíveis na ferramenta, foi escolhido o algoritmo J48, do tipo classificação, pois as árvores de decisão (geradas a partir desse algoritmo) são de fácil entendimento. Os resultados da mineração dos dados referentes às eleições municipais do estado do Rio Grande do Sul serão comparados aos resultados obtidos a partir dos dados das eleições municipais de São Paulo, Bahia, Goiás e Pará, identificando possíveis tendências ou diferenças.

Após as etapas de seleção e filtragem de dados, se faz necessário a utilização dos algoritmos escolhidos, através da ferramenta WEKA, para gerar os resultados pretendidos. Um dos algoritmos escolhidos é o J48, do tipo árvore de decisão. Na figura 8 é exibida a tela onde é definido o algoritmo e suas devidas configurações, assim como, é onde a mineração de dados é executada e o resultado é exibido.

Figura 8 - J48 (árvore de decisão)



Fonte: Arquivo pessoal

Após carregar os dados e selecionar o algoritmo J48, foi realizada a execução do algoritmo J48, utilizando a seguinte esquema de execução (*schema*):

- *weka.classifiers.trees.J48 -S -C 0.25 -M 2*

Entre os atributos disponíveis, foram selecionados aqueles considerados mais relevantes para obtenção do conhecimento pretendido, ou seja, aqueles que retornaram um resultado mais significante (indicado pela maior porcentagem de instancias classificadas corretamente, exibida na figura 11).

- Politico
- Idade
- EnsSuperior
- Casado
- CidadeNasc
- TotalBens
- Despesa
- Eleito

4.2.6 Resultados obtidos

Para cumprir o objetivo central do trabalho, a mineração de dados proposta necessita ter uma grande abrangência, podendo assim coletar possíveis tendências de regiões diferentes do Brasil, com suas características culturais, sociais e políticas. Para isso, foi realizado processo de mineração de dados selecionando apenas uma unidade federativa de cada região do Brasil, tornando o processo mais reduzido e facilitado. O critério selecionado para seleção destes Estados foi a quantidade de habitantes, seguindo a classificação publicada no Portal G1 (2012), que menciona dados do IBGE do mesmo ano (ver tabela 2).

Tabela 2 - População dos estados brasileiros

ESTADO	POPULAÇÃO
Região Sudeste	
São Paulo	41.901.219
Minas Gerais	19.855.332
Rio de Janeiro	16.231.365
Espírito Santo	3.578.067
Região Nordeste	
Bahia	14.175.341
Pernambuco	8.931.028
Ceará	8.606.005
Maranhão	6.714.314
Paraíba	3.815.171
Rio Grande do Norte	3.228.198
Alagoas	3.165.472
Piauí	3.160.748
Sergipe	2.110.867
Região Sul	
Rio Grande do Sul	10.770.603
Paraná	10.577.755
Santa Catarina	6.383.286
Região Norte	
Pará	7.792.561
Amazonas	3.590.985
Rondônia	1.590.011
Tocantins	1.417.694
Acre	758.786
Amapá	698.602
Roraima	469.524
Região Centro-Oeste	
Goiás	6.154.996
Mato Grosso	3.115.336
Distrito Federal	2.648.532
Mato Grosso do Sul	2.505.088

Fonte: (Portal G1, 2012)

Sendo assim, as seguintes unidades federativas foram selecionadas, abrangendo as cinco regiões do Brasil:

- Região Sul
  - Rio Grande do Sul
- Sudeste
  - São Paulo
- Nordeste
  - Bahia
- Centro-oeste
  - Goiás
- Norte
  - Pará

Rio Grande do Sul

Portanto, para iniciar o processo idealizado, foram selecionados os dados do Estado do Rio Grande do Sul. Após a carregar estes dados para ferramenta e a escolha dos atributos envolvidos no processo, foi realizado a execução do algoritmo. Abaixo encontra-se o cabeçalho do resultado apresentado (ver figura 9).

Figura 9 - Informações de execução (Rio Grande do Sul)

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -S -C 0.25 -M 2
Relation:     dados-weka.filters.unsupervised.attribute.Remove-R3-4
Instances:    1216
Attributes:   8
              Politico
              Idade
              EnsSuperior
              Casado
              CidadeNasc
              TotalBens
              Despesa
              Eleito
Test mode:evaluate on training data
```

Fonte: Arquivo pessoal

Neste trecho do resultado é exibido um resumo do processo de mineração realizado incluindo:

- Esquema utilizado: *weka.classifiers.trees.J48 -S -C 0.25 -M 2*

- Número de atributos utilizados: 8
- Quantidade de registros: 1216
- Modo de teste: *evaluate on training data*

Após o cabeçalho do resultado, é exibida a árvore de decisão, onde está a relação que cada atributo possuiu no processo realizado.

Figura 10 - Árvore de decisão (Rio Grande do Sul)

```
Politico = não: não (980.0/367.0)
Politico = sim
|   Idade = Alta(61-70)
|   |   TotalBens = Medio(500mil-1mi): não (4.0/1.0)
|   |   TotalBens = Baixo(100mil-500mil): não (8.0/2.0)
|   |   TotalBens = Alto(1mi-50mi): sim (5.0/1.0)
|   |   TotalBens = mBaixo(<100mil): sim (2.0)
|   Idade = Media(41-60)
|   |   TotalBens = Medio(500mil-1mi): sim (20.0/4.0)
|   |   TotalBens = Baixo(100mil-500mil)
|   |   |   Despesa = Alta(100mil-1mi): não (57.0/25.0)
|   |   |   Despesa = Média(10-100mil): sim (38.0/11.0)
|   |   |   Despesa = mAlta(>1Milhao): não (16.0/6.0)
|   |   |   Despesa = Baixa(<10mil): sim (1.0)
|   |   TotalBens = Alto(1mi-50mi): sim (7.0/2.0)
|   |   TotalBens = mBaixo(<100mil): sim (36.0/15.0)
|   Idade = Baixa(31-40)
|   |   Despesa = Alta(100mil-1mi)
|   |   |   CidadeNasc = não: sim (11.0/4.0)
|   |   |   CidadeNasc = sim
|   |   |   |   Casado = sim: não (3.0)
|   |   |   |   Casado = não: sim (3.0/1.0)
|   |   Despesa = Média(10-100mil): sim (11.0/1.0)
|   |   Despesa = mAlta(>1Milhao): não (2.0)
|   |   Despesa = Baixa(<10mil): sim (0.0)
|   Idade = mBaixa(18-30): não (8.0/2.0)
|   Idade = mAlta(71-80): não (4.0/1.0)

Number of Leaves  :     20

Size of the tree :      28
```

Fonte: Arquivo pessoal

Pode-se observar que a estrutura lembra uma árvore, onde cada valor representa uma relação com o valor seguinte e cada linha uma regra. No caso, a primeira linha, ou regra, há a expressão *"Politico = não: não (980.0/367.0)*". Essa linha indica que em 980 registros, quando o atributo político foi igual a "não", o atributo "eleito" resultou em "não" e em 367 registros essa regra não se aplica, ou seja, em 980 casos onde o candidato não possuía um cargo eletivo antes da eleição o mesmo não foi eleito.

Nota-se também que há uma diferença nos valores, em relação ao primeiro nível, o que indica que as regras ficam mais específicas conforme o nível da árvore muda. Além disso, quanto mais níveis são avançados, mais restrições a regra irá possuir. Por exemplo, a regra

"*Despesa = Alta(100mil-1mi): não (57.0/25.0)*" não possui apenas este elemento, e sim é formada da seguinte maneira:

Politico = sim
Idade = Media(41-60)
TotalBens = Baixo(100mil-500mil)
Despesa = Alta(100mil-1mi): não (57.0/25.0)

Nesta regra é indicado que os candidatos que possuíam cargo político, tinham idade entre 41 e 60 anos, declararam um total de bens entre 100 e 500 mil reais, além de ter investido entre 100 mil e 1 milhão de reais na campanha, 57 deles não foram eleitos. Já em 25 registros essa regra não foi confirmada, ou seja, não foi aplicada corretamente.

Após a classificação dos dados, outros resultados são apresentados (figuras 11, 12 e 13), referentes ao nível de precisão da classificação das instancias.

Figura 11 - Sumário (Rio Grande do Sul)

```
=== Summary ===

Correctly Classified Instances         773               63.5691 %
Incorrectly Classified Instances       443               36.4309 %
Kappa statistic                          0.153
Mean absolute error                      0.4572
Root mean squared error                  0.4781
Relative absolute error                 94.4663 %
Root relative squared error             97.1964 %
Total Number of Instances             1216
```

Fonte: Arquivo pessoal

No resultado apresentado (figura 11), o item "*correctly classified instances"* indica o número de instancias corretamente classificadas. Já o item "*Incorrectly classified instances"* indica o número de instancias classificadas incorretamente.

Figura 12 - Precisão detalhada por classe (Rio Grande do Sul)

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.946    0.81       0.627     0.946     0.754      0.582    não
                 0.19     0.054      0.709     0.19      0.3        0.582    sim
Weighted Avg.    0.636    0.5        0.66      0.636     0.568      0.582
```

Fonte: Arquivo pessoal

Abaixo destes resultados, outros também são exibidos. No caso exibido acima (figura 12) é exibido a acurácia ou exatidão. Segundo Schwerz *et al.* (2013), acurácia pode ser definida como a qualidade da classificação dos valores, independente da classe a qual pertencem, quanto mais a acurácia se aproximar a 1 (um), mais os resultados estarão próximos aos que eram conhecidos previamente. Nos resultados que foram obtidos, pode-se perceber que a exatidão dos valores "sim" e "não" para o atributo "eleito" são, respectivamente, 0,709 e 0,627 (conforme exibido na figura 12).

Outra parte importante do resultado é a matriz de confusão. De acordo com Rovedder (2007), a matriz de confusão ou erro é usada para avaliar o resultado de uma classificação, exibindo através de uma matriz, a quantidade de instancias classificadas corretamente e incorretamente. No conjunto de dados analisados, obteve-se o seguinte resultado, apresentado na figura 13.

Figura 13 - Matriz de confusão (Rio Grande do Sul)

```
=== Confusion Matrix ===

   a   b   <-- classified as
 678  39 |   a = não
 404  95 |   b = sim
```

Fonte: Arquivo pessoal

Na diagonal da matriz encontra-se o total de instâncias corretamente classificadas para cada item (a e b), ou seja, 678 e 95, totalizando 773 registros. Os valores restantes representam as instâncias classificadas incorretamente, sendo 404 e 39, totalizando 443 registros.

São Paulo

Após a mineração dos dados eleitorais de candidatos do Rio Grande do Sul, foi realizado o mesmo processo com dados dos candidatos do estado de São Paulo, utilizando os mesmos parâmetros citados anteriormente e selecionando os mesmos atributos, resultando na seguinte árvore de decisão, exibida na figura 14.

Figura 14 - Árvore de decisão (São Paulo)

```
Politico = não: não (1757.0/492.0)
Politico = sim
|   Despesa = Alta(100mil-1mi)
|   |   TotalBens = mBaixo(<100mil): não (163.0/62.0)
|   |   TotalBens = Baixo(100mil-500mil): sim (43.0/21.0)
|   |   TotalBens = Alto(1mi-50mi): sim (5.0/1.0)
|   |   TotalBens = Medio(500mil-1mi)
|   |   |   Casado = sim
|   |   |   |   EnsSuperior = sim: não (3.0/1.0)
|   |   |   |   EnsSuperior = não: sim (7.0)
|   |   |   Casado = não: não (2.0)
|   Despesa = Média(10-100mil)
|   |   TotalBens = mBaixo(<100mil)
|   |   |   Idade = Baixa(31-40): não (2.0/1.0)
|   |   |   Idade = Media(41-60): sim (21.0/5.0)
|   |   |   Idade = mAlta(71-80): sim (0.0)
|   |   |   Idade = Alta(61-70): não (4.0/1.0)
|   |   |   Idade = mBaixa(18-30): sim (0.0)
|   |   TotalBens = Baixo(100mil-500mil): não (9.0/3.0)
|   |   TotalBens = Alto(1mi-50mi): sim (1.0)
|   |   TotalBens = Medio(500mil-1mi): sim (1.0)
|   Despesa = Baixa(<10mil)
|   |   Idade = Baixa(31-40): sim (2.0)
|   |   Idade = Media(41-60): não (3.0/1.0)
|   |   Idade = mAlta(71-80): sim (0.0)
|   |   Idade = Alta(61-70): sim (0.0)
|   |   Idade = mBaixa(18-30): sim (0.0)
|   Despesa = mAlta(>1Milhao): não (96.0/29.0)

Number of Leaves  :     21

Size of the tree :      29
```

Fonte: Arquivo pessoal

Pode-se observar que a árvore resultante apresenta algumas diferenças em relação à árvore anterior, referente aos dados eleitorais do Rio Grande do Sul. Em compensação uma similaridade pode ser observada, o atributo "*político"* também está no topo da árvore. Conforme o resultado exibe, em 1757 registros, quando o atributo foi igual a "não", o atributo eleito possuía o valor "não", ou seja, na maioria dos casos essa regra foi aplicada corretamente, confirmando em mais um cenário a importância desse fator na escolha do candidato.

Outro atributo em destaque é "despesa", que indica o quanto que cada candidato investiu na campanha política. No resultado obtido é possível observar que este atributo esta logo abaixo do atributo "político", ficando no segundo nível, o que indica sua importância na classificação de dados. Além da árvore de decisão, o sumário foi gerado, onde exibido um resumo das informações (ver figura 15).

Figura 15 - Sumário (São Paulo)

```
=== Summary ===

Correctly Classified Instances        1502               70.8825 %
Incorrectly Classified Instances       617               29.1175 %
Kappa statistic                          0.0852
Mean absolute error                      0.4085
Root mean squared error                  0.4519
Relative absolute error                 96.6204 %
Root relative squared error             98.3041 %
Total Number of Instances             2119
```

Fonte: Arquivo pessoal

Neste trecho do resultado pode-se observar que aproximadamente 70% das instancias foram corretamente classificadas e 29% foram incorretamente classificadas. Outra parte importante do resultado é a tabela onde é exibido o fator de precisão do algoritmo utilizado, conforme é exibido na figura 16.

Figura 16 - Precisão detalhada por classe (São Paulo)

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.982    0.918      0.711     0.982     0.824      0.552    não
                 0.082    0.018      0.663     0.082     0.147      0.552    sim
Weighted Avg.    0.709    0.645      0.696     0.709     0.619      0.552
```

Fonte: Arquivo pessoal

Nota-se que a precisão do algoritmo para o valor “não” do atributo “eleito” foi igual a 0,711. Já para o valor sim foi de 0,663.

Outra parte importante do resultado é a matriz de confusão. Na presente análise, foi gerada a seguinte matriz, representada na figura 17.

Figura 17 - Matriz de confusão (São Paulo)

```
=== Confusion Matrix ===

   a    b   <-- classified as
1449   27 |    a = não
 590   53 |    b = sim
```

Fonte: Arquivo pessoal

Pode-se dizer que o total de instancias classificadas corretamente foi 1502 (soma dos valores 1449 e 53), enquanto que o total de instancias classificadas de maneira erronia foi 617 (soma dos valores restantes).

Bahia

Após realizar o processo de mineração nos dados de dois estados, um da região sul e outro da sudeste, torna-se interessante analisar os dados de outras regiões do Brasil. A escolhida nesse caso é a região nordeste, representada pelo estado da Bahia, o qual possui uma maior quantidade de habitantes na região.

Depois de carregar os dados para a ferramenta WEKA, foi selecionado o algoritmo J48, definindo seus parâmetros e realizando a execução do mesmo. Após a execução do algoritmo, foi gerada uma árvore de decisão, que possui o seguinte cabeçalho, de acordo com figura abaixo.

Figura 18 - Informações de execução (Bahia)

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -S -C 0.4 -M 2
Relation:     dados-weka.filters.unsupervised.attribute.Remove-R3-4
Instances:    1206
Attributes:   8
              Politico
              Idade
              EnsSuperior
              Casado
              CidadeNasc
              TotalBens
              Despesa
              Eleito
Test mode:evaluate on training data

=== Classifier model (full training set) ===
```

Fonte: Arquivo pessoal

O cabeçalho exibe, entre as diversas informações, os seguintes itens principais:

- Esquema: weka.classifiers.trees.J48 -S -C 0.4 -M 2
    - Diferentemente das análises anteriores, neste caso foi aumentado o fator de confiança, passando de 0,25 (valor padrão da ferramenta) para 0,4. Esta alteração foi feita devido às restrições do esquema proposto anteriormente. Com o valor fixado em 0,25, o algoritmo executou o processo de poda de maneira mais drástica, sobre os dados analisados, não retornando a árvore de decisão. O aumento do valor de confiança para 0,4 fez com que a árvore fosse

exibida. De acordo com Silva (2007), níveis de confiança menores tendem a penalizar a classificação que envolve um número menor de registros.

- Número de instancias: 1206
    - Total de registros do *dataset* utilizado que corresponde ao número de candidatos.
- Nº de atributos utilizados: 8
- Modelo de classificação selecionado: *full training set*

Após o cabeçalho do resultado, é exibida a árvore propriamente dita, mostrando as regras que foram geradas. A mesma é composta da seguinte estrutura, conforme é exibido na figura 19.

Figura 19 - Árvore de decisão (Bahia)

```
Despesa = Alta(100mil-1mi)
|   Politico = não: não (836.0/290.0)
|   Politico = sim
|   |   TotalBens = Medio(500mil-1mi): sim (18.0/7.0)
|   |   TotalBens = Alto(1mi-50mi)
|   |   |   Casado = sim: não (13.0/2.0)
|   |   |   Casado = não: sim (5.0/2.0)
|   |   TotalBens = mBaixo(<100mil)
|   |   |   EnsSuperior = sim: não (17.0/7.0)
|   |   |   EnsSuperior = não
|   |   |   |   Casado = sim: sim (23.0/8.0)
|   |   |   |   Casado = não: não (15.0/6.0)
|   |   TotalBens = Baixo(100mil-500mil)
|   |   |   Idade = Media(41-60)
|   |   |   |   CidadeNasc = não: não (34.0/13.0)
|   |   |   |   CidadeNasc = sim
|   |   |   |   |   Casado = sim: sim (23.0/10.0)
|   |   |   |   |   Casado = não: não (4.0/1.0)
|   |   |   Idade = Baixa(31-40)
|   |   |   |   CidadeNasc = não: sim (6.0/2.0)
|   |   |   |   CidadeNasc = sim: não (6.0/2.0)
|   |   |   Idade = mBaixa(18-30): sim (1.0)
|   |   |   Idade = Alta(61-70)
|   |   |   |   EnsSuperior = sim: sim (7.0/2.0)
|   |   |   |   EnsSuperior = não: não (3.0/1.0)
|   |   |   Idade = mAlta(71-80): sim (1.0)
Despesa = mAlta(>1Milhao)
|   TotalBens = Medio(500mil-1mi): não (20.0/4.0)
|   TotalBens = Alto(1mi-50mi)
|   |   EnsSuperior = sim: não (18.0/4.0)
|   |   EnsSuperior = não: sim (7.0/2.0)
|   TotalBens = mBaixo(<100mil): não (18.0/1.0)
|   TotalBens = Baixo(100mil-500mil)
|   |   Politico = não
|   |   |   CidadeNasc = não: sim (13.0/5.0)
|   |   |   CidadeNasc = sim: não (9.0/1.0)
|   |   Politico = sim: sim (9.0/3.0)
Despesa = Média(10-100mil): não (86.0/11.0)
Despesa = Baixa(<10mil): não (14.0/3.0)

Number of Leaves  :     25

Size of the tree :      40
```

Fonte: Arquivo pessoal

A principal diferença que nota-se, em relação às árvores anteriores, refere-se à posição dos atributos na árvore. Por exemplo, o atributo “despesa” encontra-se no topo da árvore e no primeiro nível, diferente dos casos anteriores, onde o atributo “político” se encontrava nesta posição.

Em relação ao atributo “despesa”, os valores “Alta(100mil-1mi)” e “mAlta(>1Milhao)” possuíram um maior número de registros do que os demais valores da série, indicando uma possível influência do valor do investimento em campanha no resultado final das eleições municipais do estado da Bahia.

Já no sumário do resultado foram exibidas as seguintes informações, conforme a figura 20 apresenta.

Figura 20 - Sumário (Bahia)

```
=== Summary ===

Correctly Classified Instances         819               67.9104 %
Incorrectly Classified Instances       387               32.0896 %
Kappa statistic                          0.1451
Mean absolute error                      0.4234
Root mean squared error                  0.4601
Relative absolute error                 93.4592 %
Root relative squared error             96.6826 %
Total Number of Instances             1206
```

Fonte: Arquivo pessoal

Neste sumário pode-se destacar que as instâncias corretamente classificadas correspondem a 819 (67,91%), e as instâncias incorretamente classificadas somam 387 (32,08%). Após o sumário é exibido uma análise da precisão do procedimento (ver figura 21).

Figura 21 - Precisão detalhada por classe (Bahia)

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.948    0.828      0.683     0.948      0.794      0.617    não
                 0.172    0.052      0.637     0.172      0.271      0.617    sim
Weighted Avg.    0.679    0.559      0.667     0.679      0.613      0.617
```

Fonte: Arquivo pessoal

A precisão em relação ao valor “não” do atributo eleito resultou em 0.683. Já a precisão referente ao valor “sim“ foi igual a 0.637. Em relação a matriz de confusão, o resultado foi aproximado em relação aos outros casos.

Figura 22 - Matriz de confusão (Bahia)

```
=== Confusion Matrix ===

   a   b   <-- classified as
 747  41 |   a = não
 346  72 |   b = sim
```

Fonte: Arquivo pessoal

Pode-se observar que a as instancias classificadas corretamente equivalem a 747 e 72, enquanto aquelas classificadas equivocadamente são 41 e 346.

Goiás

Seguindo o objetivo do trabalho proposto, e após analisar os dados dos estados já citados, é preciso analisar um estado da região norte e um da região centro-oeste, abrangendo assim as 5 (cinco) regiões do Brasil. O próximo estado analisado possui um menor número de candidatos, o que pode gerar resultados diferentes dos anteriores, tornando interessante realizar uma comparação entre os resultados já apresentados.

Portanto, após carregar o arquivo contendo os dados para a ferramenta WEKA e executar o algoritmo J48, foi gerado o resultado, com o seguinte cabeçalho (ver figura 23).

Figura 23 - Informações de execução (Goiás)

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -S -C 0.4 -M 2
Relation:     dados-weka.filters.unsupervised.attribute.Remove-R3-4
Instances:    690
Attributes:   8
              Politico
              Idade
              EnsSuperior
              Casado
              CidadeNasc
              TotalBens
              Despesa
              Eleito
Test mode:evaluate on training data
```

Fonte: Arquivo pessoal

Este cabeçalho é formado pelas seguintes informações:

- Schema: weka.classifiers.trees.J48 -S -C 0.4 -M 2
    - Para este conjunto de dados foi utilizado o mesmo esquema utilizado nos dados do estado da Bahia.
- Instâncias: 690
- Atributos utilizados: 8
- Modo de teste: *Training data*

Após o cabeçalho segue a árvore de decisão resultante do processo, que possui a seguinte estrutura, conforme a figura abaixo demonstra (figura 24).

Figura 24 - Árvore de execução (Goiás)

```
Despesa = mAlta(>1Milhao)
|   Idade = Media(41-60): não (61.0/19.0)
|   Idade = Alta(61-70): sim (13.0/4.0)
|   Idade = Baixa(31-40): não (23.0/5.0)
|   Idade = mBaixa(18-30): não (3.0/1.0)
|   Idade = mAlta(71-80): não (0.0)
Despesa = Média(10-100mil): não (26.0/5.0)
Despesa = Alta(100mil-1mi)
|   Casado = sim
|   |   TotalBens = Baixo(100mil-500mil)
|   |   |   Idade = Media(41-60)
|   |   |   |   Politico = sim
|   |   |   |   |   EnsSuperior = sim
|   |   |   |   |   |   CidadeNasc = não: sim (6.0/1.0)
|   |   |   |   |   |   CidadeNasc = sim: não (2.0)
|   |   |   |   |   EnsSuperior = não: não (17.0/5.0)
|   |   |   |   Politico = não: não (92.0/34.0)
|   |   |   Idade = Alta(61-70): não (21.0/7.0)
|   |   |   Idade = Baixa(31-40): não (34.0/16.0)
|   |   |   Idade = mBaixa(18-30): sim (3.0)
|   |   |   Idade = mAlta(71-80): não (0.0)
|   |   TotalBens = Alto(1mi-50mi): não (97.0/38.0)
|   |   TotalBens = Medio(500mil-1mi)
|   |   |   Idade = Media(41-60)
|   |   |   |   EnsSuperior = sim: sim (22.0/8.0)
|   |   |   |   EnsSuperior = não
|   |   |   |   |   Politico = sim: sim (7.0/3.0)
|   |   |   |   |   Politico = não: não (29.0/13.0)
|   |   |   Idade = Alta(61-70): não (4.0)
|   |   |   Idade = Baixa(31-40): sim (4.0/1.0)
|   |   |   Idade = mBaixa(18-30): sim (0.0)
|   |   |   Idade = mAlta(71-80): não (1.0)
|   |   TotalBens = mBaixo(<100mil)
|   |   |   Politico = sim
|   |   |   |   Idade = Media(41-60): não (11.0/4.0)
|   |   |   |   Idade = Alta(61-70): sim (1.0)
|   |   |   |   Idade = Baixa(31-40): sim (4.0/1.0)
|   |   |   |   Idade = mBaixa(18-30): sim (0.0)
|   |   |   |   Idade = mAlta(71-80): sim (0.0)
|   |   |   Politico = não: não (68.0/19.0)
|   |   TotalBens = mAlto(>50milhoes): não (0.0)
|   Casado = não: não (132.0/37.0)
Despesa = Baixa(<10mil): não (9.0)

Number of Leaves  :     31

Size of the tree :      44
```

Fonte: Arquivo pessoal

Assim como nos resultados referentes ao Estado da Bahia, o atributo despesa também possui destaque na árvore de decisão, nesse caso, ficando acima de atributos como “Idade”, “TotalBens” entre outros. Outro fator observado é que o atributo “político”, que possuiu grande importância nos resultados apresentados a partir dos dados dos Estados de São Paulo e Rio Grande do Sul, no estado de Goiás não ficou no topo da árvore, ficando abaixo de atributos como “Idade” e “EnsSuperior”. Após analisar a árvore, é preciso verificar o sumário do resultado, que é exibido na figura 25.

Figura 25 - Sumário (Goiás)

```
=== Summary ===

Correctly Classified Instances         424               61.4493 %
Incorrectly Classified Instances       266               38.5507 %
Kappa statistic                         -0.0038
Mean absolute error                      0.4521
Root mean squared error                  0.4909
Relative absolute error                 98.6871 %
Root relative squared error            102.5757 %
Total Number of Instances              690
```

Fonte: Arquivo pessoal

Pode-se verificar que de 690 instâncias, 424 foram classificas corretamente e 266 incorretamente, resultando nas porcentagens 61,4% e 38,5%, respectivamente.

Em relação à precisão referente aos valores do atributo eleito, o resultado apontou uma precisão de 0,348 para o valor “sim” e 0,644 para o valor “não”, conforme pode ser visto na figura 26.

Figura 26 - Precisão detalhada por classe (Goiás)

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.098     0.101      0.348     0.098     0.153      0.528    sim
                 0.899     0.902      0.644     0.899     0.75       0.528    não
Weighted Avg.    0.614     0.618      0.539     0.614     0.538      0.528
```

Fonte: Arquivo pessoal

Além disso, a matriz de confusão possuiu uma distribuição diferente, se comparada aos resultados anteriores, conforme pode ser visto na figura 27.

Figura 27 - Matriz de confusão (Goiás)

```
=== Confusion Matrix ===

   a   b   <-- classified as
  24 221 |   a = sim
  45 400 |   b = não
```

Fonte: Arquivo pessoal

No caso, os valores mais altos ficaram na coluna b, ou seja, referente ao valor “não”, enquanto que na coluna “b” os valores foram menores. Em relação aos registros classificados corretamente, os mesmos são 24 e 400 (totalizando 424) e os classificados incorretamente são 45 e 221 (somando 266).

Pará

Após minerar os dados eleitorais dos Estados do Rio Grande do Sul, São Paulo, Bahia e Goiás, das regiões sul, sudeste, nordeste e centro-oeste, respectivamente, restou apenas minerar os dados de um Estado da região norte do Brasil, completando o que foi proposto inicialmente. Foi selecionado então o Estado do Pará como alvo. Portanto, após carregar os dados selecionados para a ferramenta WEKA, foi executado o algoritmo J48, tendo os seguintes parâmetros exibidos na figura abaixo.

Figura 28 - Informações de execução (Pará)

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -S -C 0.4 -M 2
Relation:     dados-weka.filters.unsupervised.attribute.Remove-R3-4
Instances:    504
Attributes:   8
              Politico
              Idade
              EnsSuperior
              Casado
              CidadeNasc
              TotalBens
              Despesa
              Eleito
Test mode:evaluate on training data
```

Fonte: Arquivo pessoal

Entre os parâmetros exibidos pode-se destacar:

- Esquema: weka.classifiers.trees.J48 -S -C 0.4 -M 2
    - Foi definido o mesmo esquema utilizado para os Estados da Bahia e Goiás, devido à característica do conjunto de dados.
- Instancias: 504
    - Total de registros utilizados, ou seja, total de candidatos a prefeito.
- Atributos utilizados: 8
    - Definido os mesmos atributos utilizados nos casos anteriores.
- Modo de teste: *evaluate on training data*

Baseando-se nesses parâmetros, o algoritmo é executado, gerando a seguinte árvore de decisão, conforme é exibido na imagem abaixo.

Figura 29 - Árvore de decisão (Pará)

```
J48 pruned tree
------------------

Politico = sim
|   Idade = Media(41-60)
|   |   EnsSuperior = sim
|   |   |   TotalBens = mBaixo(<100mil): sim (3.0)
|   |   |   TotalBens = Baixo(100mil-500mil): não (9.0/3.0)
|   |   |   TotalBens = Medio(500mil-1mi): sim (5.0/2.0)
|   |   |   TotalBens = Alto(1mi-50mi)
|   |   |   |   Despesa = mAlta(>1Milhao): não (2.0)
|   |   |   |   Despesa = Alta(100mil-1mi): sim (3.0/1.0)
|   |   |   |   Despesa = Média(10-100mil): não (0.0)
|   |   |   |   Despesa = Baixa(<10mil): não (0.0)
|   |   |   TotalBens = mAlto(>50milhoes): sim (0.0)
|   |   EnsSuperior = não: não (50.0/16.0)
|   Idade = Alta(61-70): sim (7.0/2.0)
|   Idade = Baixa(31-40): não (18.0/5.0)
|   Idade = mBaixa(18-30): sim (2.0)
|   Idade = mAlta(71-80): não (1.0)
Politico = não: não (404.0/103.0)
```

Fonte: Arquivo pessoal

A primeira característica que nota-se nessa árvore é sua estrutura, que é bem menor, se comparar com as demais árvores geradas no presente trabalho. A posição dos atributos é outro aspecto importante a ser destacado. Assim como na árvore gerada a partir dos dados do Rio Grande do Sul, os atributos “politico” e “idade” encontram-se acima dos demais atributos.

Outro atributo que pode ser destacado é “EnsSuperior”. Este se encontra logo abaixo do atributo idade, tornando-se importante na classificação dos dados.

Em relação ao nível de acerto na classificação dos dados, a proporção foi aproximada em relação aos casos anteriores, ficando em 73,8 %, as instancias classificadas corretamente e 26,19% as classificadas incorretamente, mas ficando um pouco acima das demais, conforme pode ser observado na figura 30.

Figura 30 - Sumário (Pará)

```
=== Summary ===

Correctly Classified Instances         372               73.8095 %
Incorrectly Classified Instances       132               26.1905 %
Kappa statistic                          0.1243
Mean absolute error                      0.383
Root mean squared error                  0.4376
Relative absolute error                 94.5532 %
Root relative squared error             97.2835 %
Total Number of Instances              504
```

Fonte: Arquivo pessoal

Referente à precisão do algoritmo para os valores do atributo “eleito” ocorreu certa aproximação. Para o valor “sim” a precisão foi de 0,75 e para o valor “não” foi de 0,738, como pode ser confirmado na figura 31.

Figura 31 - Precisão detalhada por classe (Pará)

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.106     0.014      0.75      0.106     0.185      0.568    sim
                 0.986     0.894      0.738     0.986     0.844      0.568    não
Weighted Avg.    0.738     0.646      0.741     0.738     0.658      0.568
```

Fonte: Arquivo pessoal

Já na matriz de confusão, como no caso anterior, ocorreu certa diferença na distribuição dos valores, ficando os valores mais altos na coluna B. Em relação aos registros classificados corretamente, foi exibido os valores de 15 e 357, conforme é exibido na figura 32.

Figura 32 - Matriz de confusão (Pará)

```
=== Confusion Matrix ===

   a   b   <-- classified as
  15 127 |   a = sim
   5 357 |   b = não
```

Fonte: Arquivo pessoal

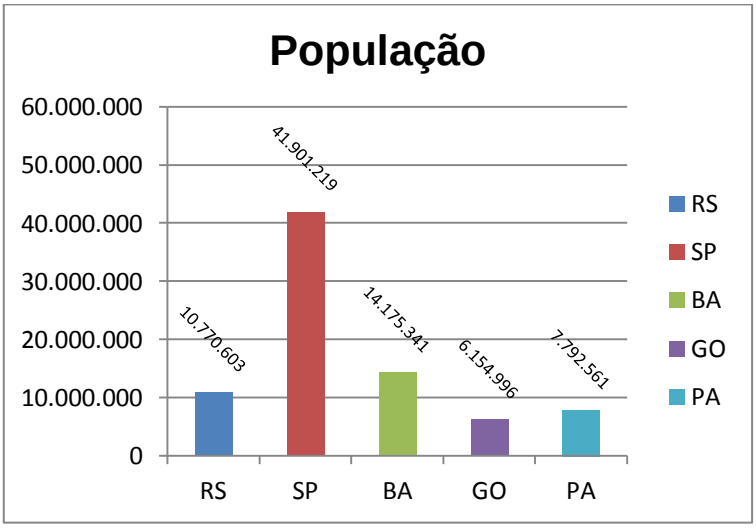
4.2.7 Interpretação dos resultados

Foi possível observar, a partir dos resultados gerados, vários aspectos importantes, tanto referentes às regras geradas, como também aos números e percentuais resultantes.

Em relação aos números gerados, podem-se observar alguns aspectos que são úteis para analisar, de certa forma, a eficiência do algoritmo J48, assim como, possíveis relações entre esses resultados.

Ao analisar os resultados apresentados e o total de habitantes de cada estado (figura 33) é possível fazer uma comparação entre esses dados.

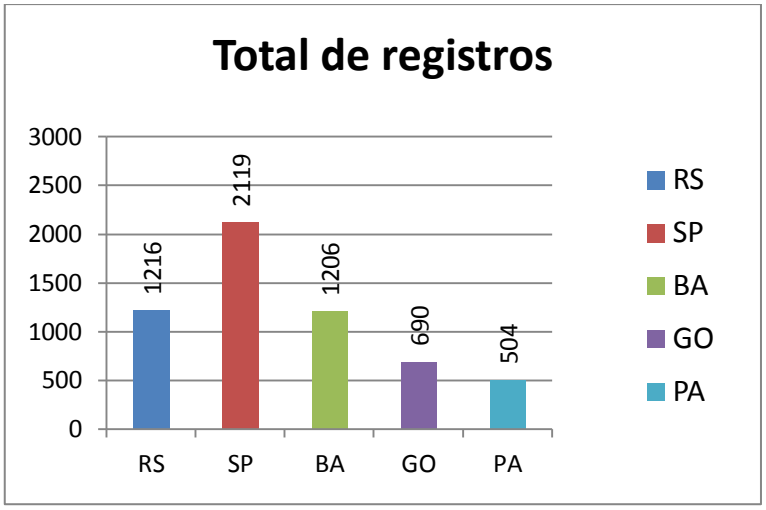
Figura 33 - População dos Estados analisados



Fonte: Arquivo pessoal

Na figura 33 pode-se observar o total de habitantes por estado, baseando-se nos dados do IBGE do ano de 2012 (ver seção 3.2.2, página 32), divulgados pelo portal G1, (2012). E abaixo observa-se, na figura 34, o total de instancias ou registros de cada estado selecionado.

Figura 34 - Total de registros (candidatos)

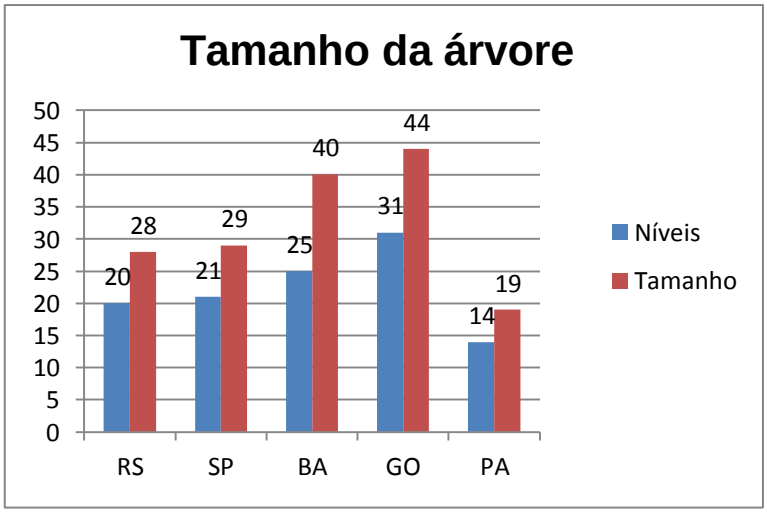


Fonte: Arquivo pessoal

Visualizando estes gráficos pode-se afirmar que ocorreu certa similaridade em relação à população de cada Estado (figura 33) com o número de instancias (figura 34), ou seja, com o total de candidatos a prefeito. Por exemplo, o Estado de São Paulo, o mais populoso, é aquele que possuía mais instancias a serem analisadas.

Já em relação ao tamanho da árvore gerada (ver figura 35), não surgiu qualquer relação ao tamanho da população, pois como pode ser visto para os dados do Estado de Goiás, o menos populoso, a ferramenta gerou uma árvore de 31 níveis e 44 linhas, a maior gerada no estudo realizado.

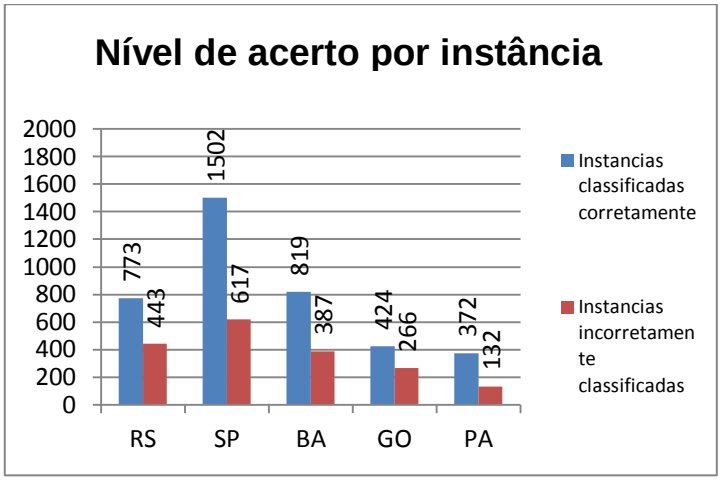
Figura 35 - Tamanho da árvore



Fonte: Arquivo pessoal

Em relação à precisão do algoritmo, outros aspectos podem ser verificados, conforme é observado nos gráficos abaixo (figura 36 e 37).

Figura 36 - Acerto por instância

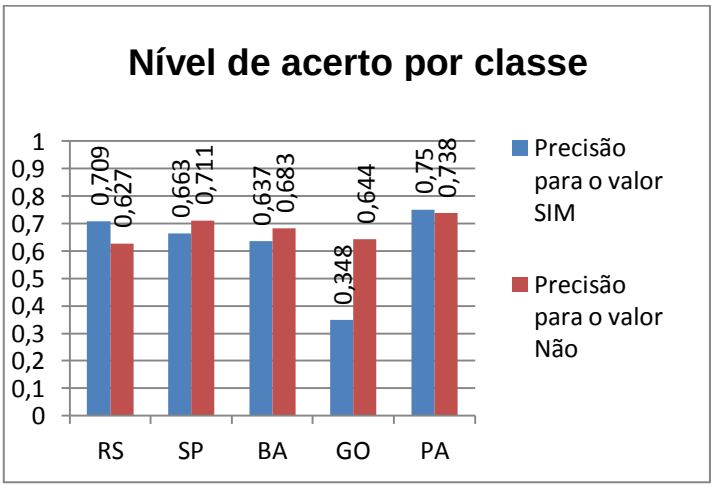


Fonte: Arquivo pessoal

Nesse gráfico (figura 36) observa-se que as instancias classificadas corretamente estão sempre acima das instâncias classificadas incorretamente. Em alguns casos, o total de classificações corretas está bem acima das incorretas, como é o caso dos Estados de São Paulo e Bahia, por exemplo.

Já na figura 37 pode-se observar o nível de exatidão ou acerto por classe, de cada estado escolhido.

Figura 37 - Nível de acerto por classe



Fonte: Arquivo pessoal

Observa-se que a precisão do algoritmo para o valor “sim” (eleito) e “não” (não eleito) na maioria dos casos foi similar (conforme pode ser visto na figura 37), resultando em pouca diferença. Exceto para o Estado de Goiás, onde a precisão para o valor “sim” foi bem menor a precisão para o valor “não”.

Em relação às regras geradas, alvo central do trabalho, foi possível observar muitas similaridades entre os resultados. Alguns atributos tiveram maior importância na classificação dos dados, tornando-se importantes para identificar tendências ou padrões entre os Estados.

Entre esses atributos, o mais importante, foi o “Politico”, que indica se o candidato possuiu ou exercia cargo público eletivo antes da eleição de 2012. Entre os cinco Estados analisados, em três deles esse atributo esteve no topo da árvore, indicando a importância desse fator na escolha do candidato, ou seja, a experiência política influência diretamente na escolha do eleitor nas urnas. Um exemplo é a árvore gerada a partir dos dados do Estado do Rio Grande do Sul, onde a primeira regra exibe: “Politico=não: não (980.0/ 367.0)”. Essa regra indica que, em 980 registros, os candidatos que não possuíam cargo político não foram eleitos. Já em 367 registros essa regra não se aplica corretamente.

O atributo “Despesa” também pode ser considerado outro destaque nos resultados. Em duas árvores geradas, referente aos dados de Goiás e Bahia, o mesmo ficou no topo, e na árvore gerada a partir dos dados do Estado de São Paulo, o atributo ficou logo abaixo do atributo “politico”, indicando sua importância para a classificação dos dados.

Referente às regras exibidas para este atributo, aquelas que indicam um maior gasto na campanha eleitoral resultaram em uma maior quantidade de registros classificados corretamente, favoráveis à eleição dos candidatos. Já os valores que indicam pouco gasto em campanha, indicaram a tendência negativa, em relação à eleição dos candidatos. Por exemplo, nas árvores geradas a partir de dados dos Estados Goiás e Bahia, a regra para o valor “Baixa(<10mil)”, que indica que o candidato investiu menos de 10 mil reais na campanha, não indicou probabilidade do candidato ser eleito.

Pode-se afirmar também que a idade do candidato é outro fator importante observado nos resultados obtidos. Em três Estados, sendo eles, Bahia, Goiás e Rio Grande do Sul, o atributo “Idade” esteve logo abaixo do primeiro atributo na árvore de decisão. Analisando os resultados é possível observar uma tendência a não eleição de candidatos com baixa idade (entre 18 e 30 anos) e idade avançada (entre 71 e 80 anos).

Além dos atributos já citados, outros também tiveram influência no resultado final, entre eles está o atributo “TotalBens”, que indica o total de bens do candidato. Já outros atributos não tiveram tanta influência no resultado. Entre eles, “EnsSuperior”, “Casado” e “CidNasc”.

5 METODOLOGIA

Para construção do presente trabalho foi preciso passar por algumas etapas, estas consideradas importantes para se alcançar o que foi proposto. O primeiro passo, já tendo em mente o tema do central do trabalho, foi pesquisar o referencial teórico, necessário para possuir embasamento para desenvolver as próximas etapas.

Entre os tópicos centrais pesquisados está a mineração de dados, incluindo as técnicas de classificação. Além deste tópico foi pesquisado também sobre descoberta de conhecimento em banco de dados e também um breve referencial sobre as principais ferramentas de mineração de dados, incluindo WEKA.

Além do que foi citado, outro tópico importante é a situação política do Brasil, cenário onde o trabalho se baseia. Foram citados aspectos importantes da política atual brasileira, incluindo também um breve histórico. Outro fator relevante para o presente trabalho é identificar quais os objetivos e expectativas dos eleitores no momento do voto, ou seja, qual o perfil de cada eleitor e o que motiva a escolha de seus candidatos.

Após adquirir embasamento teórico foram realizados os primeiros passos do desenvolvimento do trabalho. Entre esses passos esta a definição do escopo estudado, ou seja, qual fonte de dados a ser utilizada para realizar o processo de mineração de dados. Para isso, foi feito uma breve seleção e filtragem dos dados disponíveis, selecionando apenas aqueles considerados úteis para o trabalho proposto.

Para executar esse passo importante do desenvolvimento, foi preciso definir qual ferramenta responsável por extrair o conhecimento pretendido. Além disso, foram determinados quais métodos e algoritmos seriam utilizados nesse procedimento.

Seguida da realização da mineração dos dados, os resultados foram gerados, sendo possível analisar os mesmos a fim de descobrir prováveis padrões e tendências que possam ser observadas.

6 CONCLUSÃO

Quando o eleitor vivencia um período eleitoral, muitos pensamentos vêm a tona, muitas informações são fornecidas e muitas ideias são formadas sobre os candidatos. Essas ideias são influenciadas de diversas maneiras, tanto pelo grupo onde se vive, pelas pessoas que participam do cotidiano do cidadão e, cada vez mais, pela mídia, determinante para levar ao conhecimento do povo as características de cada candidato. O presente trabalho proporcionou saber o quanto que essas características são determinantes no resultado de uma eleição. Os resultados apontaram que, principalmente, fatores como experiência política, despesa em campanha e idade foram importantes na classificação dos dados. Até alcançar a etapa final do trabalho, muitos desafios foram encontrados. Entre eles, a tarefa de seleção e filtragem de dados, principalmente pela quantidade e características dos dados disponíveis, foi o principal desafio, pois foi preciso encontrar a formatação adequada dos dados, a fim de obter um melhor aproveitamento da ferramenta de mineração de dados. Tendo em vista as dificuldades apresentadas e os resultados que foram obtidos, pode-se afirmar que os objetivos foram alcançados e foi possível obter o conhecimento idealizado. Para trabalho futuro, torna-se interessante avaliar os padrões encontrados comparando com outros períodos eleitorais, podendo comparar com os dados das eleições seguintes ou com pleitos anteriores.

REFERÊNCIAS

AGRAWAL, R.; FALOUTSOS, C.; SWAMI, A. Efficient similarity search in sequence databases. **Proceedings of the 4th International Conference of Foundations of Data Organization and Algorithms (FODO)**, 1993.

ANGELINE, D. M. D.; JAMES, I. S. P. Association Rule Generation Using Apriori Mend Algorithm for Student ' s Placement. **Int. J. Emerg. Sci.**, v. 2, n. March, p. 78-86, 2012.

BACARDIT, J.; LLORÀ, X. Large-scale data mining using genetics-based machine learning. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 3, n. 1, p. 37-61, 9 jan. 2013.

BARADWAJ, B. K.; PAL, S. Mining Educational Data to Analyze Students ‟ Performance. **International Journal of Advanced Computer Science and Applications**, v. 2, n. 6, p. 63-69, 2011.

COSTA, E. *et al.* Mineração de Dados Educacionais: Conceitos, Téc- nicas, Ferramentas e Aplicações. v. d, p. 1-29, 2012.

DIGERATI, E. **Hardware para profissionais**. [s.l.] Digerati Books, 2009. p. 128

FELISBINO, R. DE A.; BERNABEL, R. T.; KERBAUY, M. T. M. Somente um deve vencer: as bases de recrutamento dos candidatos à prefeitura das capitais brasileiras em 2008. **Revista de Sociologia e Política**, v. 21, n. 41, p. 219-234, 2012.

FIGUEIRA, P. R. Efeitos da campanha virtual no universo das mídias sociais: o comportamento do eleitor no Twitter nas Eleições 2010. **Revista Compolítica**, v. 1, n. 3, p. 8-26, 2013.

G1, P. **Brasil tem 193.946.886 habitantes, aponta estimativa do IBGE**. Disponível em: <http://g1.globo.com/brasil/noticia/2012/08/brasil-tem-193946886-habitantes-aponta-estimativa-do-ibge.html>. Acesso em: 15 jul. 2013.

GAXIE, D. As lógicas do recrutamento político. **Revista Brasileira de Ciência Política**, n. 8, p. 165-208, 2012.

GRÖGER, C.; NIEDERMANN, F.; MITSCHANG, B. Data Mining-driven Manufacturing Process Optimization. **Proceedings of the World Congress on Engineering**, v. III, p. 0-6, 2012.

GUPTA, S.; KUMAR, D.; SHARMA, A. DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS. **Journal of Computer Science**, v. 2, n. 2, p. 188-195, 2011.

HAGOOD, J. A Brief Introduction to Data Mining Projects in the Humanities. **Bulletin of the American Society for Information Science and Technology**, v. 38, n. 4, p. 20-23, 2012.

HALL, M. *et al.* The WEKA Data Mining Software : An Update. **SIGKDD Explorations**, v. 11, n. 1, p. 10-18, 2009.

JOSEPH, M. V; SADATH, L.; RAJAN, V. Data Mining : A Comparative Study on Various Techniques and Methods. **International Journal of Advanced Research in Computer Science and Software Engineering**, v. 3, n. 2, p. 106-113, 2013.

KAUR, K.; MOHAN, N.; SANDHU, P. S. Reusability of Software Components using J48 Decision Tree. **International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012)**, p. 15-17, 2012.

KUMAR, V.; CHADHA, A. Mining Association Rules in Student ' s Assessment Data. **International Journal of Computer Science**, v. 9, n. 5, p. 211-216, 2012.

LEE, I.; YONG, H.; ENGINEERING, C. S. Common Sense Knowledge Discovery for Association Rule. **IST**, v. 23, p. 188-191, 2013.

LIMA, D. DE. A ditadura militar, a redemocratização e a democracia representativa no Brasil. **Revista Jurídica - CCJ**, v. 16, n. 31, p. 75-92, 2012.

LUIZ, O. *et al*. Aplicação de algoritmos de aprendizagem de máquina para mineração de dados sobre beneficiários de planos de saúde suplementar. **Journal of Health Informatics**, v. 4, n. 2, p. 43-49, 2012.

NASSIF, D. R. J. **Uma ferramenta para mineração de dados de projetos de software livre e criação de redes sócio-técnicas**. [s.l.] Universidade Tecnológica Federal do Paraná, 2013.

OLIVEIRA, A. O estado da arte dos determinantes do voto no Brasil e as lacunas existentes. **Sociedade e Cultura**, v. 15, n. 1, p. 193-206, 17 out. 2012.

OLIVEIRA, I. C. DE. **Aplicação de Data Mining na Busca de um Modelo de Prevenção da Mortalidade Infantil**. [s.l.] Universidade Federal de Santa Maria, 2001.

PADHY, N.; MISHRA, P.; PANIGRAHI, R. The Survey of Data Mining Applications And Feature Scope. **International Journal of Computer Science, Engineering and Information Technology**, v. 2, n. 3, p. 43-58, 2012.

PATIL, T. R.; SHEREKAR, S. S. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. **International Journal Of Computer Science And Applications**, v. 6, n. 2, 2013.

RAVICHANDRAN, S.; SRINIVASAN, V. B.; RAMASAMY, C. Comparative Study on Decision Tree Techniques for Mobile Call Detail Record. **Journal of Communication and Computer**, v. 9, p. 1331-1335, 2012.

RODRIGUES, M. O que determina o número de candidatos em uma eleição? **Economia & Tecnologia**, v. 25, 2011.

ROVEDDER, J. **Validação da classificação orientada a objetos em imagens de satélite Ikonos II e elaboração de indicadores ambientais georreferenciados no município de torres, planície costeira do Rio Grande do Sul, Brasil**. [s.l.] UFRGS – Universidade Federal do Rio Grande do Sul, 2007.

SCHWERZ, A. L. *et al.* Predizendo a Participação de Desenvolvedores em Discussões em Projetos de Software Livre. 2013.

SHWETA, M.; GARG, K. Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms. **International Journal of Advanced Research in Computer Science and Software Engineering**, v. 3, n. 6, p. 306-312, 2013.

SILVA, I. A. F. **Descoberta de conhecimento em base de dados de monitoramento ambiental para avaliação da qualidade da água**. [s.l.] Universidade Federal de Mato Grosso, 2007.

SUNDAR, N. A.; LATHA, P. P.; CHANDRA, M. R. PERFORMANCE ANALYSIS OF CLASSIFICATION DATA MINING TECHNIQUES OVER HEART DISEASE DATA BASE. **International Journal of Engineering Science & Advanced Technology**, v. 2, n. 3, p. 470-478, 2012.

WITTEN, I. H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques, Second Edition**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.