Revista de Informática Aplicada, Volume 14, Número 1, 2018

Extração de informação em bases de dados abertas governamentais através de uma abordagem de mineração descritiva empregando a ferramenta R

Lydia de Castro C. Braga [1], Isabela Neves Drummond [2].

Resumo

O processo de extração de conhecimento de grandes bases de dados envolve diversas técnicas. Cada uma delas é apropriada para um tipo de problema e a análise dos métodos utilizados inclui peculiaridades dos dados e a experiência do analista. Neste contexto, dados abertos governamentais tornam-se essenciais para análise dos diversos componentes presentes no cotidiano populacional de um país, como educação, crescimento econômico e política. Este artigo apresenta uma metodologia de mineração de dados utilizando a ferramenta R como instrumento, buscando verificar tanto o potencial da ferramenta, como da metodologia descrita. Os estudos de casos apresentados utilizam o conjunto de dados Enem 2013, provido pelo Instituto Nacional de Estudos e Pesquisa (INEP), e os Dados das Teses e Dissertações da Pós-Graduação 2012, disponíveis no Banco de Teses e Dissertações da Capes, ambos enquadrados na Lei de Acesso à Informação. Os resultados alcançados demonstram a viabilidade de aplicação da abordagem proposta na construção de aplicativos, tornando possível a difusão de conhecimento em diversas áreas.

Palavras-chave: Extração de conhecimento, ferramenta R, dados abertos governamentais.

Abstract

The process of knowledge discovery from large databases involves several techniques. Each technique is appropriate for one type of problem, and the analysis of the used methods includes peculiarities of the data and the analyst's experience. In this context, open government data becomes essential for the analysis of the various components present in the daily population of a country, such as education, economic growth and politics. This article presents a methodology of data mining using the R tool as instrument, seeking to verify the potential of the tool as well as the methodology. The case studies presented use the data set Enem 2013, provided by the National Institute of Studies and Research (INEP), and the Data of Thesis and Dissertations of Postgraduate 2012, available in the Bank of Thesis and Dissertations of Capes, both framed in the Law on Access to Information. The results show the feasibility of applying the proposed approach in the construction of applications, making it possible to disseminate knowledge in several areas.

Keywords: Knowledge discovery, R tool, open government data.

[1]Universidade Federal de Itajubá, Av. BPS, 1303, Pinheirinho, E-mail: lydiaccbraga@gmail.com
[2]Universidade Federal de Itajubá, Av. BPS, 1303, Pinheirinho, E-mail: isadrummond@unifei.edu.br

1 Introdução

O processo de mineração de dados é uma área da ciência da computação que visa extrair conhecimento de grandes bases de dados. Esse tipo de procedimento é extremamente útil, podendo ser empregado em indústrias, processos administrativos e pesquisas em diversas áreas como forma de auxiliar na tomada de decisão.

As áreas governamentais coletam grandes quantidades de dados todos os dias e, a partir do paradigma de dados abertos governamentais, eles tornam-se disponíveis para a população através do programa de transparência governamental, regido pela Lei de Acesso à Informação nº 12.527/11.

As técnicas de mineração de dados têm fundamentos baseados na estatística e no aprendizado de máquina, utilizando algoritmos complexos no processo de extração de conhecimento, sendo necessárias ferramentas eficientes.

A ferramenta R é uma linguagem de programação e um ambiente integrado de software para estatística computacional e visualização gráfica, desenvolvida como um projeto open-source (RIPLEY, 2001). Diversos algoritmos de mineração de dados são desenvolvidos e publicados todos os anos graças à uma comunidade de desenvolvedores apreciadores da ferramenta R.

Neste contexto, este artigo visa detalhar o modelo de mineração de dados proposto em (BRAGA; DRUMMOND, 2017), estendendo os resultados encontrados com um novo estudo de caso, que abrange o estudo dos dados das Teses e Dissertações da Pós-Graduação 2012 disponível no Banco de Teses e Dissertações da Capes. Os resultados encontrados permitiram validar a metodologia proposta e evidenciar pontos positivos e negativos do processo implementado a partir da ferramenta R. Todo o estudo demonstrado em (BRAGA; DRUMMOND, 2017) que compreende a aplicação da metodologia aos dados do ENEM 2013 é também apresentado, detalhado e comentado neste artigo, buscando evidenciar as características da abordagem proposta.

Este artigo está organizado da seguinte maneira: a Seção 2 apresenta a definição dados abertos governamentais e evidencia as vantagens da abertura de dados no cotidiano populacional. A Seção 3 visa definir o processo de mineração de dados e demonstrar as razões da utilização da ferramenta R neste contexto. A Seção 4 detalha a metodologia de mineração empregada, especificando as técnicas e os pacotes da ferramenta R, bem como a análise proposta. As Seções 5 e 6 apresentam a aplicação do modelo desenvolvido em conjuntos de dados abertos governamentais, como forma de validação. A Seção 7 apresenta as considerações gerais acerca do processo realizado e, por fim, a Seção 8 apresenta as conclusões e possíveis trabalhos futuros.

2 Dados Abertos Governamentais

O paradigma de dados abertos depreende que dados de origem pública, em posse do Governo, devem ser disponibilizados à sociedade na sua forma bruta e em formato aberto, para que a sociedade possa produzir cruzamentos, interpretações e aplicações úteis (DUTRA; LOPES, 2013).

A abertura de dados governamentais objetiva o alcance de três grandes finalidades: transparência, liberação de valor social e comercial e participação governamental (FOUNDATION, 2012). A meta de transparência define que os cidadãos precisam estar cientes sobre as ações e decisões do Governo. Para tanto, eles devem ser capazes de acessar livremente os dados e informações que desejarem e as compartilharem com outros cidadãos. A liberação de valor social e comercial indica que o acesso à informação é um recurso essencial para a abertura de atividades sociais e comerciais. Neste caso, a abertura de dados governamentais pode impulsionar a criação de negócios inovadores e serviços que agregam valor social e comercial. Por último, a participação governamental transcende o princípio de transparência, fazendo com que a sociedade não seja só informada do processo governamental, mas também seja capaz de contribuir com este processo.

Estas razões abrangem motivos sociais e econômicos que beneficiam tanto a população quanto o próprio Governo. Exemplos são a melhoria da transparência e controle democrático, da eficácia dos serviços públicos e da avaliação do impacto da implantação de novas políticas, além do aumento da participação popular e de inovação dos produtos e serviços privados (DIETRICH et al., 2009)

3 Mineração de dados e a Ferramenta R

Mineração de dados é o processo de extração de conhecimento de grandes bases de dados (CORRÊA;

SFERRA, 2003). Sua aplicação tem como objetivo encontrar relacionamentos ocultos entre os dados e sintetizá-los de forma compreensível e útil para seu proprietário (HAND; MANNILA; SMYTH, 2001). O sucesso da mineração de dados depende, principalmente, de uma boa metodologia. Mensurar metas razoáveis, realizar um trabalho interdisciplinar e utilizar um ciclo repetitivo de tarefas e das avaliações de seus resultados são questões essenciais neste processo.

De forma geral, todo processo de extração de conhecimento precisa entender o contexto do problema a ser analisado, definir dados e fazer um pré-processamento, estudar e selecionar modelos de mineração de dados adequados, avaliar os resultados obtidos e, se necessário, recomeçar o processo. Ao atingir resultados apropriados, devem-se estabelecer meios de visualização que facilitem o entendimento dos usuários.

A ferramenta R traz consigo recursos que facilitam este processo. Ela consiste em um conjunto integrado de ferramentas para estatística computacional e visualização gráfica (R Core Team, 2017), podendo ser adquirida diretamente do Comprehensive R Archive Network (CRAN), uma coleção de sites, conhecidos como *mirrors*. Neles estão disponíveis tanto a ferramenta R básica quanto diversos pacotes para funcionalidades específicas. Apesar desta definição, a ferramenta R não está fundamentada apenas em pacotes estatísticos, sendo uma verdadeira linguagem de programação. Usuários desta ferramenta podem utilizar simplesmente linhas de comando ou escrever suas próprias funções e pacotes, utilizando todo o potencial desta linguagem(TEAM, 2000).

Para as tarefas de mineração são definidos pacotes específicos contendo funções que auxiliam as aplicações destas técnicas nos conjuntos de dados trabalhados. Neste trabalho foram empregados os pacotes *Kohonen* (WEHRENS; BUYDENS et al., 2007), *NbClust* (CHARRAD et al., 2014), *party* (HOTHORN; HORNIK; ZEILEIS, 2006), *partykit* (HOTHORN; ZEILEIS, 2015), *wordcloud* (FELLOWS, 2014), *tm* (FEINERER; HORNIK, 2017), *caret* (WING et al., 2017), além do pacote básico *Stats* (R Core Team, 2017).

4 Metodologia Proposta

A metodologia empregada neste trabalho, descrita em (BRAGA; DRUMMOND, 2017), visou à aplicação de técnicas de mineração de dados a dados abertos governamentais por meio da ferramenta R. Com base neste tipo de dados foi elaborado um método próprio capaz de gerar informações descritivas sobre o conjunto, extraindo conhecimento relevante que ajude em um processo de decisão. Para tanto, este processo foi definido de forma a verificar o relacionamento entre as variáveis presentes em um conjunto de dados, focando na abordagem de mineração descritiva. Este processo foi dividido em cinco fases (etapas) principais, conforme fluxograma da Figura 1 e descrito a seguir.

4.1 Pré-Processamento

A fase de pré-processamento engloba as tarefas de coleta de dados, sua exploração e, em caso de inconsistências, limpeza e correção. A ferramenta R disponibiliza estratégias para gerar informações estatísticas sobre os dados trabalhados, sendo possível verificar os principais casos de incoerências, como problemas de completude e inconsistência. O comando *summary*, disponível no pacote *Stats*, é utilizado para realizar a primeira verificação dos dados. Ele retorna informações como média, mediana, valor mínimo e máximo de dados numéricos e a frequência em que ocorrem determinados valores em dados textuais. Também é possível verificar se existem campos vazios nos dados e onde eles ocorrem. Dados lógicos também são verificados, mostrando a frequências de verdadeiros e falsos.

É importante levar em consideração neste tipo de análise o domínio trabalhado, bem como o contexto de onde os dados foram coletados. Informações de unidades, categorias assumidas e faixas de valores consideradas válidas são imprescindíveis ao analisar inconsistências. Caso estes tipos de problemas sejam verificados, o analista precisa decidir entre algumas técnicas de limpeza de dados, como preencher valores em falta, remover campos ou reduzir o conjunto de dados.

4.2 Agrupamento

A fase de agrupamento verifica a relação entre as variáveis do conjunto de dados. O modelo de agrupamento permite identificar e atribuir elementos similares a um mesmo grupo, sem a necessidade dos dados utilizados estarem previamente classificados, utilizando técnicas de aprendizado não supervisionado (CAMILO; SILVA, 2009).

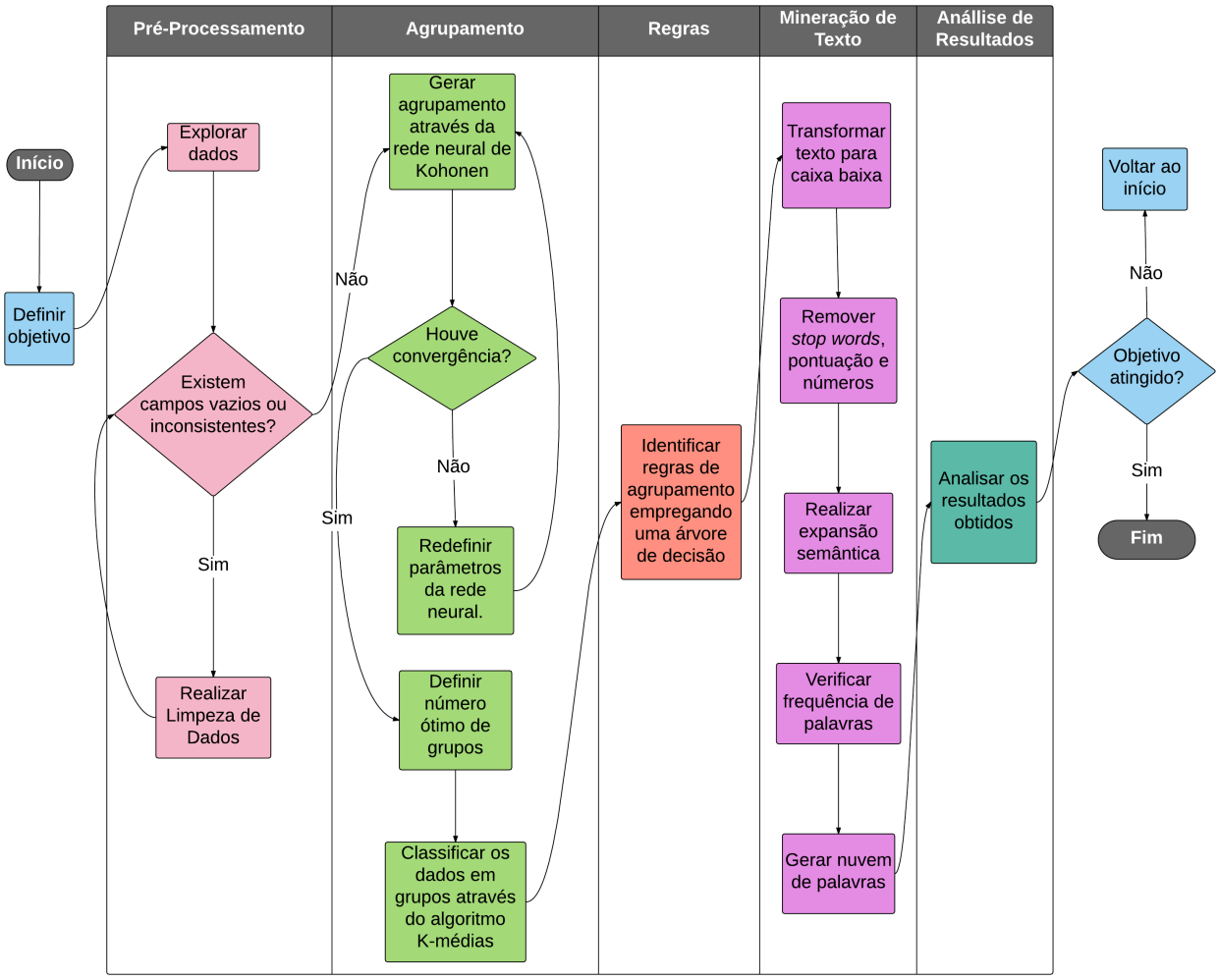

Figura 1: Fluxograma do processo de mineração.

A metodologia desenvolvida para gerar os agrupamentos nos conjuntos de dados trabalhados consistiu no uso de dois algoritmos: a rede neural de Kohonen (KOHONEN, 2013) e o algoritmo K-médias (JAIN, 2010). Como abordagem intermediária entre os dois existem índices de validade de agrupamento que verificam o número de grupos ótimo.

A rede neural de Kohonen (KOHONEN, 2013) organiza, geralmente em um mapa bidimensional, dados complexos em agrupamentos, sendo, por esse motivo, também chamada de mapa auto-organizável. Os pesos finais de cada neurônio da rede determinam um padrão reconhecido, e, por consequência, definem uma classe de dados referente ao que é apresentado na entrada. Segundo KOHONEN, a rede não foi criada para reconhecimento de padrões, mas para agrupamento, visualização e abstração. Ainda assim, pode ser utilizada para reconhecimento e classificação quando utilizada juntamente com um modelo de aprendizado supervisionado. No caso do pacote *Kohonen*, é necessário transformar os dados utilizados em informações numéricas antes do treinamento.

Para definir o número de classes presentes no mapa de Kohonen, índices de validade de agrupamentos podem ser calculados através do pacote *NbClust*. Estes indíces combinam informações sobre a compactação e isolamento dos agrupamentos, assim como propriedades de geometria ou estatística e medidas de dissimilaridade ou similaridade entre os dados. A partir do resultado obtido, as instâncias podem ser classificadas por meio do algoritmo de aprendizado não supervisionado K-médias, que consiste em um método de agrupamento não-hierárquico por reparti-

ção (JAIN, 2010). Ele é um dos algoritmos não-supervisionados mais simples que resolve o problema de agrupamentos onde o número de grupos é conhecido. Para trabalhar com cada uma dessas técnicas na ferramenta R são empregados os pacotes *Kohonen*, *NbClust* e *Stats*.

4.3 Regras

A fase de descoberta das regras tem como objetivo identificar as regras que descrevem cada um dos grupos encontrados através da fase de agrupamento, demonstrando as influências de cada um dos atributos que definem um conjunto de dados. O processo consiste em utilizar uma árvore de decisão nos dados previamente divididos na etapa de agrupamento de forma a gerar regras que descrevem esses conjuntos, permitindo analisar a natureza de cada um dos agrupamentos em relação ao relacionamento das variáveis presentes no conjunto de dados.

Os pacotes *party* e *partykit* estão disponibilizados na ferramenta R e são empregados nesta fase. A metodologia disponível no primeiro pacote utiliza a inferência condicional como método de partição em subconjuntos de forma binária e recursiva (FRIZZARINI; LAURETTO, 2013) e o *partykit* consiste na sua implementação melhorada, provendo a mesma abordagem baseada em uma nova infraestrutura (HOTHORN; ZEILEIS, 2015), descrita em (BREIMAN et al., 1984).

4.4 Mineração de Texto

A fase de mineração de texto visa encontrar novas informações relevantes em campos textuais presentes nos conjuntos de dados. As fases anteriores utilizam técnicas de aprendizado de máquina que trabalham apenas com campos numéricos ou categóricos, de forma que informações importantes podem não ter sido identificadas. Esta etapa do processo é extremamante útil quando se trata de dados abertos governamentais, visto que permite a análise profunda de atas, resumos, contratos, entre outros tipos de documentos.

Os modelos que trabalham com dados textuais diferem dos utilizados em dados numéricos ou categóricos, devido a informação encontrar-se implícita, ou seja, não existem campos que identificam atributos e simplificam as informações representadas. A técnica de mineração de dados deve simular o comportamento de um leitor que, a partir de um texto, naturalmente identifica informações relevantes (ZAMBENEDETTI, 2002). Para tanto é necessário seguir um conjunto de etapas:

- transformar o texto para caixa baixa buscando facilitar os processos de comparação;
- remover pontuação, número, símbolos e *stop words*, que consistem em palavras muito comuns ou sem significado relevante, como artigos e preposições (ZAMBENEDETTI, 2002);
- realizar expansão semântica;
- verificar a frequência de cada palavra;
- definir um método de visualização, como uma nuvem de palavras.

Para mineração de texto é empregado o pacote *tm*, que permite realizar o processo completo, passando por cada uma das etapas listadas. Outro ponto fundamental para a utilização do pacote está no fato de se poder especificar o idioma que se deseja trabalhar, o que é crucial para o resultado deste tipo de processo.

O pacote *wordcloud* permite a visualização em formato de nuvem de palavras. O analista pode definir um número máximo de palavras na visualização e também a frequência mínima de cada uma delas, de forma a melhorar a qualidade do seu resultado. Esse tipo de visualização é uma excelente alternativa para verificar quais termos aparecem mais frequentemente em um documento textual, permitindo induzir quais assuntos se destacaram.

4.5 Análise de Resultados

A análise de resultados visa verificar quais são as vantagens e desvantagens da abordagem de mineração empregando a ferramenta R no âmbito de dados abertos governamentais. Para tanto, é necessário averiguar se os métodos de aprendizado de máquina utilizados possuem validade dentro do domínio do problema e que tipo de informação foi alcançada ao fim do processo de mineração.

Como os dados utilizados para agrupamento não possuem nenhum tipo de classificação prévia não é possível determinar o nível de acurácia da rede neural de Kohonen, mas existem parâmetros que permitem a verificação da sua qualidade. O pacote *Kohonen* disponibiliza diversos tipos de visualização dos resultados obtidos, estando entre eles

um gráfico do processo de treinamento. Este gráfico mede a distância média de um objeto com o vetor de características mais próximo. Em geral, o formato encontrado é uma exponencial decrescente, que indica que os vetores de características dos neurônios da rede encontram-se cada vez mais similares aos objetos contidos no conjunto de dados (WEHRENS; BUYDENS et al., 2007). Gráficos muito distantes deste formato podem indicar que o treinamento da rede não foi eficiente, sendo necessário um novo ajuste dos parâmetros da rede.

As classes geradas pelo algoritmo K-médias já utilizam o número ótimo de grupos definido pelos índices de validade de agrupamento, de forma que já formam partições válidas de acordo com o pacote *NbClust*.

No caso da árvore de decisão, pode-se definir sua acurácia. Por se tratar de um método de aprendizado supervisionado, é possível verificar o número de instâncias corretamente classificadas a partir do método, utilizando como base as classes geradas no processo de agrupamento. Para isso, são utilizadas a matriz de confusão e a medida de acurácia. Ambos os métodos de validação são fornecidos pelo pacote *caret* da ferramenta R, que provê a função *confusionMatrix()* para esse tipo de processo.

A matriz de confusão é utilizada para visualizar o desempenho de um classificador. Ela apresenta o número de predições corretas e incorretas em cada classe do problema, permitindo uma análise do comportamento das classes no espaço, uma vez que é possível identificar, dada a classe, em qual classe está o erro apresentado. O número de acertos para cada uma das classes se localiza na diagonal principal da matriz e os demais elementos representam erros na classificação (FACELI et al., 2011). A medida de acurácia é conhecida como exatidão global e refere-se à estimativa de quanto os resultados representam a classificação correta esperada ou quanto se afastam desta classificação (SUAREZ; CANDEIAS, 2012).

Após a validação da qualidade da árvore de decisão, é preciso extrair suas regras. As regras geradas por esse tipo de modelo são do tipo "se-então", de forma que permitem a visão de quais valores um determinado atributo pode assumir e o que isso implica nos próximos níveis da árvore. Essa informação pode representar a descoberta de padrões nos conjuntos de dados utilizados e, consequentemente, a descoberta de informação relevante.

A análise da fase de mineração de texto consiste em identificar as palavras mais frequentes por meio da nuvem de palavras e verificar, com base no contexto dos dados, os assuntos que mais apareceram dentro dos campos textuais. Esse processo serve como complementação das informações extraídas nas regras derivadas da árvore de decisão, ampliando a visão do analista.

5 Estudo de Caso: Aplicação do processo aos dados do ENEM 2013

A partir da aplicação da abordagem proposta ao conjunto de dados do ENEM 2013 foi possível obter resultados que demonstram o uso da metodologia, em todas suas fases, descrita na Seção 4.

5.1 Conjunto de dados ENEM 2013

O conjunto de dados ENEM 2013 refere-se aos resultados por Escola do Exame Nacional do Ensino Médio aplicado no ano de 2013, podendo ser encontrado no portal do Instituto Nacional de Estudos e Pesquisa (INEP), juntamente com as provas aplicadas e seus gabaritos. Também são disponibilizados documentos descritivos sobre os dados. O conjunto de dados utilizado neste trabalho consiste na Planilha ENEM por Escola, um arquivo que divide o resultado por escola de acordo com as áreas avaliadas na prova do ENEM: Linguagens e Códigos, Redação, Matemática, Ciências Humanas e Ciências da Natureza. Este conjunto possui 14714 instâncias que representam as escolas participantes do exame. Os atributos utilizados na análise referem-se à *unidade da Federação*, *dependência administrativa*, *localização* e *porte da escola*, incluindo ainda o *nível social* dos alunos, o *índice da formação de adequação da formação docente* e as *médias* referentes às provas do ENEM.

5.2 Resultados da Etapa de Pré-Processamento

A primeira etapa da fase de pré-processamento foi definir as colunas apropriadas ao processo de mineração de dados. Para tanto, foram escolhidas colunas que possuíssem informações relevantes sobre as escolas participantes do exame, como *localização* e *nível social*, e as *médias* em cada uma das áreas de conhecimento das provas.

Através do comando *summary*, avaliou-se a qualidade da informação de cada uma das colunas. Os campos referentes à *unidade federativa*, *dependência administrativa*, *localização*, *porte da escola*

e *nível social* foram corretamente definidos como campos categóricos e apresentam a frequência em que cada um destes valores aparece no conjunto de dados. Os campos que representam a *média* das escolas em cada uma das áreas do conhecimento avaliadas estão definidos como campos numéricos e apresentam valores que tendem a 1000, de forma que estão dentro do intervalo esperado como nota da prova. A partir da análise da média das provas em cada uma das áreas de conhecimento, pode-se perceber que a prova de Ciências Naturais possui a menor média e a prova de redação, a maior. Por último, o campo de *formação de docente* assume valores dentro da faixa [0, 100], estando consistente em relação ao seu contexto. Apesar disso, é possível perceber que existem três instâncias do conjunto de dados que estão com este campo vazio, como pode ser visto através do valor NA. Para resolver esta inconsistência foi empregada a abordagem de remoção, eliminando as instâncias que possuíam campos vazios. O resultado pode ser observado na Figura 2.

```
      UF              DEP_ADM       FORMACAO_DOCENTE
SP      :3006   Estadual :7914   Min.    :  0.0
MG      :1688   Federal  : 284   1st Qu.: 48.6
RJ      :1312   Municipal: 115   Median : 60.9
RS      :1071   Privada  :6402   Mean   : 59.4
PR      : 925                    3rd Qu.: 71.9
CE      : 812                    Max.   :100.0
(Other):5901                     NA's   :3

     NIVEL_SOCIAL                  PORTE_ESCOLA
Alto        :3215   De 1 a 30 alunos   :4488
Baixo       : 840   De 31 a 60 alunos  :3603
Medio       :3376   De 61 a 90 alunos  :2100
Medio Alto  :3826   Maior que 90 alunos:4524
Medio Baixo:1916
Muito Alto  :1471        LOCALIZACAO
Muito Baixo:  71         Rural :  627
                         Urbana:14088

    NATURAIS        HUMANAS        LINGUAGENS
Min.   :382.4   Min.   :384.7   Min.   :365.8
1st Qu.:450.8   1st Qu.:491.0   1st Qu.:467.3
Median :474.8   Median :517.7   Median :496.8
Mean   :487.8   Mean   :529.3   Mean   :501.6
3rd Qu.:518.0   3rd Qu.:564.4   3rd Qu.:535.3
Max.   :734.0   Max.   :738.8   Max.   :658.3

   MATEMATICA      REDACAO
Min.   :382.6   Min.   :196.7
1st Qu.:484.3   1st Qu.:482.7
Median :520.7   Median :528.0
Mean   :534.7   Mean   :536.9
3rd Qu.:577.7   3rd Qu.:586.2
Max.   :868.3   Max.   :869.0
```

Figura 2: Pré-processamento do conjunto de dados ENEM 2013.

5.3 Resultados da Etapa de Agrupamento

A rede neural de Kohonen foi aplicada utilizando uma grade de 38x39 neurônios, com vizinhança circular e taxa de aprendizado no intervalo [0.06, 0.01], gerando o mapa de Kohonen. O processo de treinamento da rede pode ser verificado na Figura 3. Observe que a distância média de um objeto com o vetor de características mais próximo tende a diminuir exponencialmente, tendendo a um valor mínimo e se estabilizando a partir da quinquagésima iteração aproximadamente. Este resultado demonstra que os valores escolhidos alcançam um treinamento do tipo adequado e que o número de iterações definido é mais que o suficiente para a convergência da rede neural.

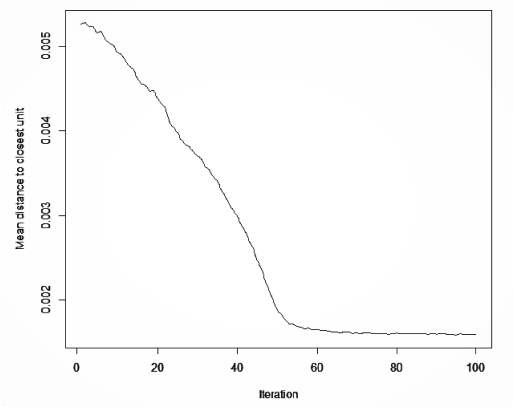

Figura 3: Progresso de treinamento para o conjunto de dados ENEM 2013.

A Figura 4 apresenta os resultados obtidos através da rede; onde são expostos os mapas de calor de cada uma das variáveis do conjunto e, acima à esquerda, o mapa auto-organizável gerado com os agrupamentos. Um mapa de calor consiste na exibição do agrupamento com foco em um único atributo relevante do conjunto de dados. Ele permite a análise do estado e do impacto que um grande número de variáveis causa em um modelo de agrupamento (WILKINSON; FRIENDLY, 2009). A escala de cores utilizada indica que cores mais quentes representam valores mais altos nas instâncias apresentadas em uma região do mapa, enquanto cores mais frias representam o oposto. No mapa referente à *localização*, por exemplo, é possível observar que existem dois grupos, um notavelmente maior que o outro. A comprovação desta característica pode ser conferida no pré-processamento (Figura 2), que mostra que existem 627 escolas na zona rural e 14088 na zona urbana. A mesma análise pode ser feita no mapa de *dependência administrativa*, onde existem dois grandes grupos que

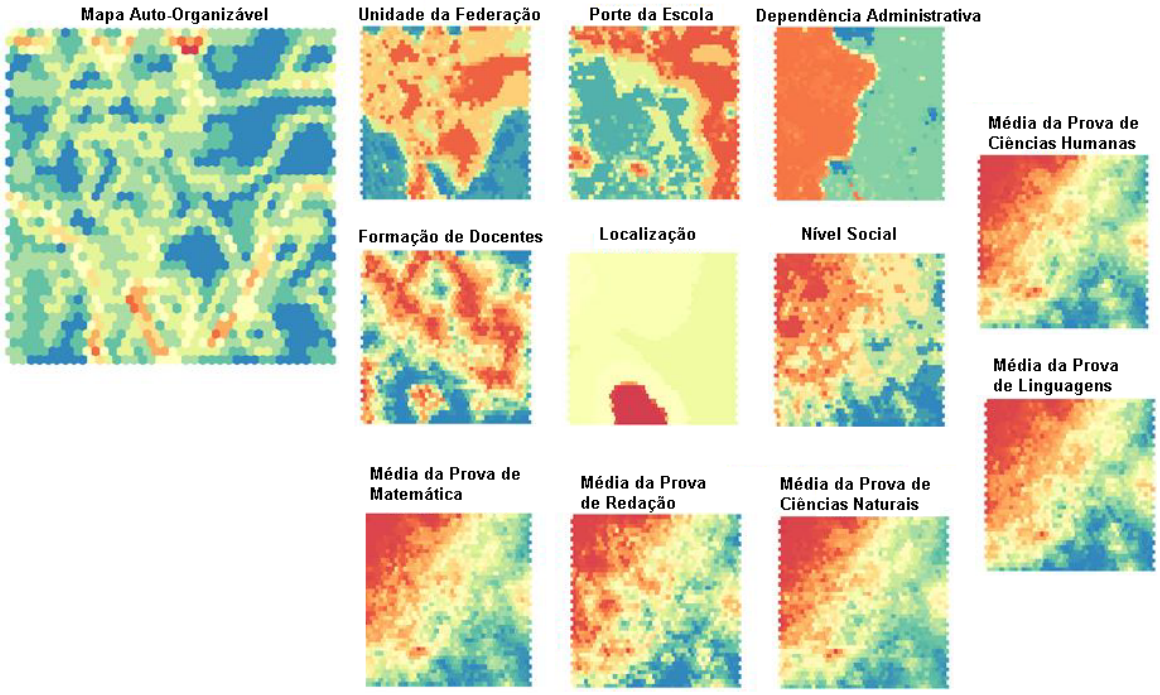

Figura 4: Resultado da Rede de Kohonen para o conjunto de dados ENEM 2013.

representam as categorias estadual e privada, e dois grupos menores referentes aos valores federal e municipal. Também é interessante observar que valores altos para as provas estão localizados na mesma região do mapa, coincidindo também com a região onde o *nível social* tende a ser maior. Posteriormente, foi verificado através do pacote *NbClust* que o número ótimo de grupos deve ser 7, gerando as partições referentes com o algoritmo K-médias.

5.4 Resultados da Etapa de Regras

Para gerar as regras, foi utilizada a árvore de decisão provida pelos pacotes *party* e *partykit*. O conjunto classificado na etapa anterior foi empregado como parâmetro desta abordagem, já que ela consiste em um método de aprendizado supervisionado. A fórmula utilizada como base do processo definiu como variável alvo as classes geradas pelo algoritmo K-médias, mantendo os demais atributos como independentes.

A árvore gerada possuía uma estrutura de 1011 nós, onde 505 são nós internos e 506 são folhas. Essa quantidade de nós pode indicar que muitas arestas e sub-árvores refletem ruídos ou erros, problema conhecido como sobreajuste e que reflete um aprendizado muito específico do conjunto de treinamento, não permitindo a generalização do modelo. Um processo de poda foi aplicado para facilitar a leitura dos resultados através do pacote *partykit*. Ele permite que sejam realizadas podas após o treinamento da árvore através da função *nodeprune*. Para isso, é preciso definir quais nós serão podados. A escolha desses nós foi realizada por meio da imposição de significância aos nós, definindo um valor de 10^{-5} para o teste de hipótese de independência entre uma variável de entrada e sua resposta. Após este processo, a árvore sofreu uma diminuição de 80%.

Tabela 1: Resultados de acurácia para o conjunto ENEM 2013.

Classe	Acurácia
1	0,45
2	0,64
3	0,37
4	0,36
5	0,53
6	0,26
7	0,70
Acurária Global	0,556

Utilizando a matriz de confusão como base de cálculo, a acurácia do modelo atingiu 55,6%. Verificando a acurácia individual de cada uma das parti-

ções, medida também chama de sensibilidade dentro da estatística, pode-se perceber um resultado de 36% a 70% de acurácia. Isso pode indicar que as regras geradas não representam seu comportamento em totalidade. A Tabela 1 sumariza os valores de acurácia obtidos para a árvore.

Mesmo com a poda, a quantidade de nós obtidos impossibilita uma análise por meio da visualização gráfica, de forma que é necessário extrair as regras através da função disponível no pacote *partykit*. As regras geradas permitem a visão das escolas que apresentam notas abaixo da média em cada uma das áreas de conhecimento do ENEM. A Figura 5 apresenta a porcentagem de regras que indicam escolas abaixo da média geral dos inscritos do ano de 2013, separando-as por área de conhecimento. Com base neste gráfico, fica evidente que 54% das regras apresentaram notas abaixo da média, sendo que 36% relacionavam-se com a prova de ciências da natureza.

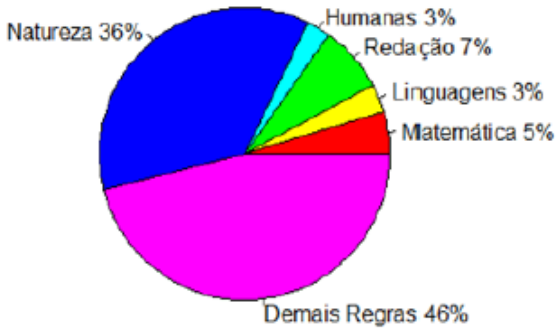

Figura 5: Gráfico da porcentagem de Regras com escolas abaixo da média por área de conhecimento.

5.5 Resultados da etapa Mineração de Texto

A etapa de mineração de texto foi aplicada sobre as provas do exame do ano de 2013, trazendo resultados voltados ao conteúdo escolhido como foco das áreas de conhecimento do ENEM, como sendo as matérias mais cobradas. A apreciação desta visualização neste tipo de contexto depende do analista e seu conhecimento nas áreas estudadas, de forma a relacionar termos com os conteúdos do ensino médio. Esse tipo de informação relaciona-se com as notas apresentadas nas regras obtidas, visto que especifica os conteúdos onde os alunos apresentaram maiores dificuldades. A Figura 6 demonstra as nuvens geradas.

Em relação à prova de matemática, pode-se inferir que seu foco foi em questões relacionadas a conversões de tempo, como indica as palavras "tempo", "horas", "dia"e "anos", e geometria plana, espacial e analítica através, por exemplo, das palavras "raio", "centro", "sistema"e "vértices".

Sobre a prova de redação, vê-se com clareza o tema do ano de 2013: a lei seca. A prova de português destaca questões sobre gêneros textuais, como demonstra os termos "carta"e "crítica". E também podem ser observados temas voltados à literatura, com base nas palavras "arte", "poema", "obra"e "fragmento". Outro ponto a ser observado são as palavras "negro"e "indígenas". Em 2003, a Lei 10.639 incluiu à Lei de Diretrizes e Bases da Educação Nacional (LDB) o ensino obrigatório de história e cultura afro-brasileira durante o ensino médio e, em 2008, a Lei 11.645 acrescentou as populações indígenas, dessa forma, a presença destas duas palavras indica a cobrança de questões sobre a cultura destes dois excluídos históricos.

Em relação a Ciências Humanas, percebe-se que as palavras características deste contexto são naturalmente mais difíceis de enquadrar em conteúdos específicos do ensino médio, podendo-se inferir com base nas palavras "brasil", "rei", "poder", "guerra"e "filosofia"que os conteúdos cobrados giravam em torna de história do brasil, monarquias, conflitos históricos e conteúdos filosóficos.

Por fim, em relação à prova de Ciências da Natureza, vê-se a presença de termos de química como "carbono"e "moléculas". Também percebe-se termos típicos de física mecânica e eletrodinâmica, como "movimento", "força", "elétrico"e "circuito". Os termos "dna", "pai"e "mãe"podem ser enquadrados em questões de genética e as palavras "água", "árvores"e "aquecimento"indicam matérias ligadas a ecologia.

Através desse resultado, é possível estabelecer uma relação entre as regras geradas e as matérias cobradas na prova, concluindo-se que no ano de 2013 os inscritos na prova do Enem tiveram mais dificuldade em matérias relacionadas a química, física e biologia, sugerindo uma possível deficiência no ensino destas matérias.

Figura 6: Nuvens de Palavra para as provas do Enem 2013.

6 Estudo de Caso: Aplicação do processo aos dados das Teses e Dissertações da Pós-Graduação 2012

Os resultados obtidos utilizando os dados das Teses e Dissertações da Pós-Graduação 2012 tem como objetivo espandir a análise realizada com o conjunto de dados ENEM, buscando evidenciar os pontos de sucesso e de falha da metodologia descrita na Seção 4.

6.1 Dados Empregados

O conjunto de dados das Teses e Dissertações da Pós-Graduação 2012 pode ser encontrado no Banco de Teses e Dissertações (BTD) da Capes (BANCO, 2013), permitindo a consulta de todos os trabalhos defendidos na pós-graduação brasileira no referido ano. As informações contidas no BTD são fornecidas diretamente à Capes pelos programas de pós-graduação, sendo estes responsáveis pela veracidade dos dados. O conjunto utilizado possui 61.054 registros que representam teses e dissertações com 61.050 autores distintos. Ele engloba informações sobre o autor da tese, data de defesa, instituição a qual o autor está vinculado e área de conhecimento. Juntamente com os dados, é possível encontrar no BTD seu dicionário, onde é descrito o significado de cada um dos campos do conjunto de dados, bem como as categorias por eles assumidas, seu tipo e tamanho máximo.

6.2 Resultados da etapa Pré-Processamento

Os dados das Teses e Dissertações do ano 2012 possuem ótima documentação e a equipe de análise responsável mantém a consistência dos dados disponibilizados, por isso a etapa de pré-processamento consistiu em tarefas simples. A primeira tarefa foi a seleção das colunas com informações relevantes para o processo de mineração, de forma que, mantiveram-se informações relacionadas a *região*, *institutos* a qual as teses estão vinculadas, *grande área de conhecimento*, *grau da tese* e *idioma* no qual ela foi escrita. Através do comando *summary*, as instâncias que possuíam alguma informação faltante foram eliminadas, resultando em um total de 60545 teses no conjunto de dados. O tratamento necessário consistiu em identificar cada tipo de dado corretamente para as etapas seguintes do processo. Por exemplo, informações de *região*

e *institutos* são dados categóricos e a *grande área de conhecimento* é identificada por valores numéricos. Os resultados obtidos a partir do comando *summary* podem ser observado na Figura 7. Desta forma, é possível confirmar os tipos de dados e verificar a frequência de cada uma das categorias que um atributo pode assumir.

```
          REGIAO        UNIVERSIDADE
CENTRO-OESTE: 3967   USP     : 4113
NORDESTE     :10025  UFRJ    : 2415
NORTE        : 2291  UFRGS   : 2362
SUDESTE      :31962  UFMG    : 1941
SUL          :12300  UNICAMP: 1929
                     UFPE    : 1557
                     (Other):46228

   GRANDE_AREA            IDIOMA
 40000001:10022   Alemao   :    2
 70000000: 9694   Diversos :    2
 60000007: 8050   Espanhol :   14
 50000004: 6992   Frances  :   10
 30000009: 6934   Ingles   :  271
 90000005: 5270   Outro    :   54
 (Other) :13583   Portugues:60192

                  NIVEL
      Doutorado          :13833
      Mestrado           :42514
      Profissionalizante: 4198
```

Figura 7: Pré-processamento do conjunto de Teses e Dissertações do ano 2012.

6.3 Resultados da etapa Agrupamento

A rede neural de Kohonen foi aplicada utilizando uma grade de 25x26 neurônios, com vizinhança circular e taxa de aprendizado no intervalo [0.06, 0.01], gerando o mapa auto-organizável. O processo de treinamento da rede pode ser visualizado na figura 8. Segundo o gráfico, a distância média de um objeto com o vetor de características mais próximo sempre diminui, assemelhando-se a uma exponencial negativa, demonstrando que os valores selecionados alcançam um treinamento adequado.

Figura 8: Progresso de treinamento para o conjunto de Teses e Dissertações do ano 2012.

A Figura 9 exibe os resultados desta rede. O mapa auto-organizável de Kohonen, do lado esquerdo, exibe os agrupamentos gerados ao fim do processo. Também estão definidos os mapas de calor de cada uma das variáveis utilizadas na análise, permitindo verificar como cada um dos atributos se distribui pelo mapa. A escala de cores utilizada é a mesma apresentada na análise dos agrupamentos dos dados de Enem 2013. No mapa de *grau da tese*, por exemplo, existem claramente 3 grupos, representando os valores de doutorado, mestrado e mestrado profissionalizante encontrados dentro desta categoria. O mapa para a variável *idioma* apresenta um pequeno grupo em vermelho, podendo representar as teses redigidas em idiomas diferentes do português. Em relação ao mapa de *região*, observa-se tons de cores que representam as regiões do país. É importante notar como os agrupamentos referentes a *região* e a *área de conhecimento* não coincidem, demonstrando que existe produção de conhecimento em diferentes áreas em cada região do país.

Com estes resultados, os dados foram classificados através do algoritmo K-médias, onde foram gerados 3 grupos.

6.4 Resultados da etapa Regras

As regras contidas neste conjunto de dados, assim como no estudo de caso da Seção 5, foram geradas através do pacote *party* e *partykit* da ferramenta R. A árvore de decisão foi produzida a partir do conjunto classificado empregando o algoritmo K-médias, com a variável alvo definida como as classes encontradas na etapa de agrupamento.

Assim como aconteceu nos resultados desta etapa do conjuntos de dados ENEM, a árvore foi gerada com um grande número de nós, sendo 99 internos e 100 folhas. Esta quantidade de nós indica que a árvore enfrentou dificuldades de generalização. Para tentar generalizar os resultados, o mesmo processo de poda do conjunto anterior foi aplicado, mas ainda assim restaram 90 folhas, indicando que a árvore realizou um aprendizado muito específico

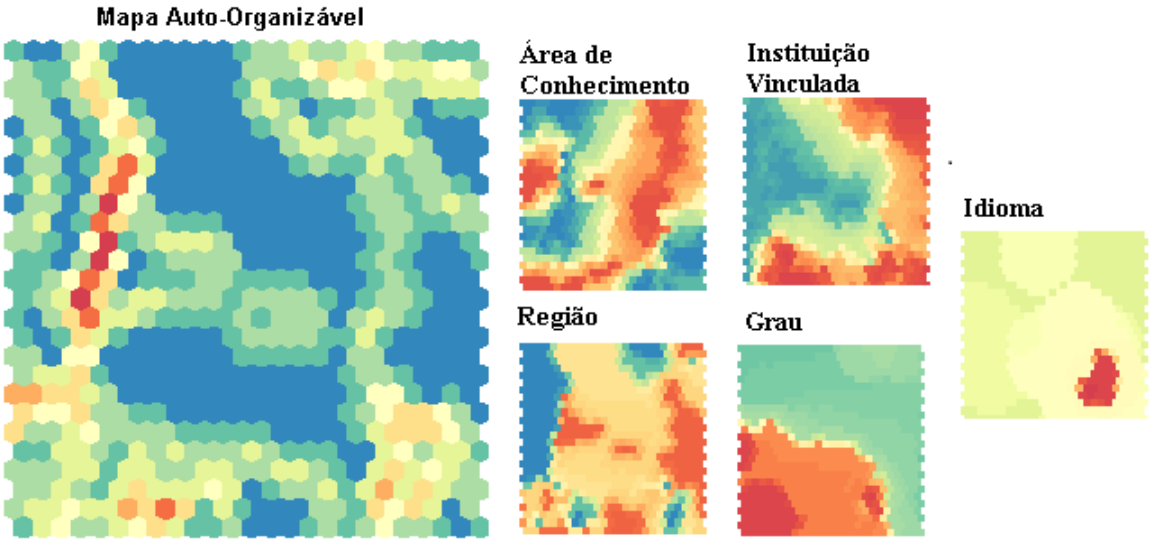

Figura 9: Resultado da Rede de Kohonen para o conjunto de Teses e Dissertações do ano 2012.

para o conjunto de dados.

A acurácia do modelo foi calculada com base na matriz de confusão do resultado da árvore de decisão, atingindo um valor de 80,56%. A sensibilidade de cada partição também foi calculada, encontrando o resultado de 30,95% para a classe 1, 62,34% para a classe 2 e 92,48% para a classe 3, evidenciando a dificuldade de predição para a primeira e segunda classe do modelo. Estes resultados estão concentrados na Tabela 2.

Tabela 2: Resultados de acurácia para o conjunto de Teses e Dissertações do ano 2012.

Classe	Acurácia
1	0,3095
2	0,6234
3	0,9248
Acurácia Global	0,8056

As regras foram extraídas através do pacote *partykit*, totalizando 90 regras para a descrição do conjunto. Numa análise manual, ou seja, sem a aplicação de modelos computacionais que pudessem restringir o tamanho do conjunto de regras, ou combiná-las de maneira a agrupá-las facilitando a interpretação, não foi possível identificar padrões pudessem gerar informações novas e/ou relevantes sobre os registros de teses e dissertações em estudo.

Como exemplo das informações geradas, pode-de citar que 2261 das Teses e Dissertações são referentes aos trabalhos de mestrado na área de Ciências Agrárias, estando vinculadas a institutos da região sudeste. Outra informação consiste em dizer que 4329 trabalhos de mestrado sobre Ciências Sociais Aplicadas estão vinculados as regiões sul e sudeste do país. Estas duas regras classificam, aproximadamente, apenas 4% e 7% do total de teses consideradas no conjunto na etapa de pré-processamento. O conjunto de regras é composto por muitas regras que se aplicam a poucos elementos, evidenciando a dificuldade de generalização do modelo na descrição dos dados.

6.5 Resultados da etapa Mineração de Texto

A etapa de mineração de texto foi aplicada sobre as teses e dissertações redigidas no idioma português, que representam aproximadamente 99% do conjunto de dados. O processo foi realizado para cada uma das grandes áreas de conhecimento da Capes, totalizando nove nuvens de palavras, que permitem a visualização dos termos mais utilizados nas teses classificadas dentro de uma área de conhecimento. Esses termos podem ajudar a inferir os temas mais estudados no ano de 2012.

Figura 10: Nuvem de palavras de Ciências Agrárias.

Na nuvem gerada para Ciências Agrárias (Figura 10) são observadas as palavras "animais", "alimentos", "manejo"e "solos"em destaque, podendo indicar que as teses relacionavam-se a cultivos e criação de animais. Também é possível verificar palavras "ambiente"e "sustentabilidade", apontando pesquisas sobre esses temas.

Para a nuvem de Ciências Biológicas (Figura 11), pode-se inferir que as teses possuem temas que incluem genética, biodiversidade e pesquisas moleculares. Também percebem-se temas relacionados a infectologia e diagnóstico.

Figura 11: Nuvem de palavras de Ciências Biológicas.

A visualização de Ciências da Saúde (Figura 12) possui poucas palavras em destaque, denunciando um grande número de termos. As palavras "doenças", "tratamento", "enfermagem", "materiais", "clínicos"e "diagnósticos"se sobressaem e indicam trabalhos relacionados a estudos clínicos médicos. A palavra "odontologia"também aparece, apontando para trabalhos desta área. Uma visualização no formato nuvem, que aponta maior ocorrência de termos pelo tamanho das palavras que compõe a nuvem, só aponta contribuição de informação quando o conjunto de dados tem maior ocorrência e palavras, caso contrário, a nuvem é composta de uma grande quantidade de termos de mesmo tamanho, dificultando a interpretação.

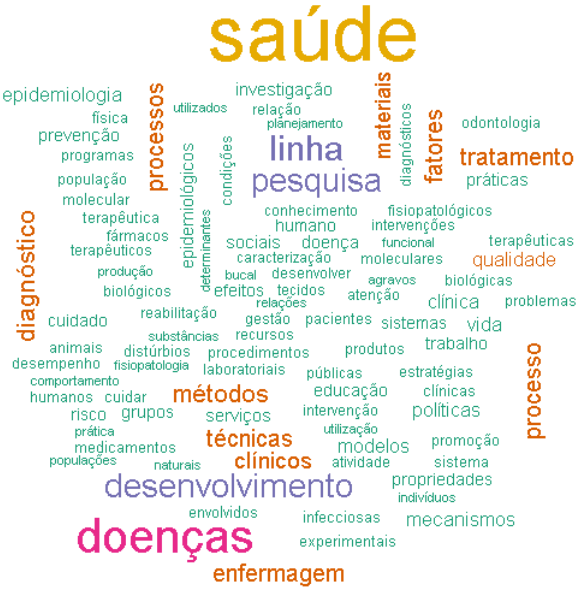

Figura 12: Nuvem de palavras de Ciências da Saúde.

No caso da nuvem de palavras da grande área de conhecimento Multidisciplinar (Figura 13, pode-se identificar palavras de diversos campos. Palavras como "educação", "metodologias"e "aprendizagem"indicam teses relacionadas a ensino, podendo englobar diversos conteúdos. Também existem palavras relacionadas a biologia e saúde como "biodiversidade", "natureza"e "doenças". Além destas, pode-se observar as palavras "tecnologia", "sistemas"e "informação"que estão relacionadas às tecnologias de informação e comunicação, entre outras possíveis análises. A nuvem de palavras é um tipo de visualização que envolve o contexto, e as possíveis inferências não são exatas. Esta observação na área multidisciplinar é ainda mais evidente, já que a área abrange teses que combinam diversas outras áreas de conhecimento.

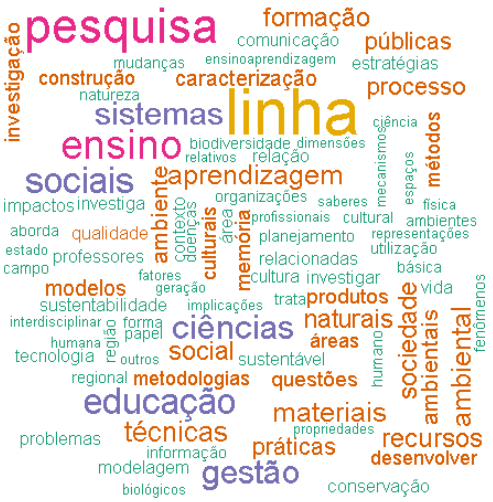

Figura 13: Nuvem de palavras de Multidisciplinar.

Para as engenharias, a nuvem de palavras demonstrada na Figura 14 contém palavras como "problemas", "materiais", "simulação", "otimização", "processos", "energia", "qualidade", "construção", "software"e "programação". Essas palavras relacionam-se com diversas especialidades na engenharia, tais como engenharia civil, de produção, de computação e de energia.

Figura 14: Nuvem de palavras de Engenharias.

A nuvem de Linguistica, Letras e Artes (Figura 15) possui palavras que indicam gêneros textuais como "discursos", "poéticas"e "crítica". Também existem palavras relacionadas a literatura como "obra"e "literárias". Palavras como "sala", "aprendizagem"e "metodologias"indicam teses que discutem o ensino de literatura e português.

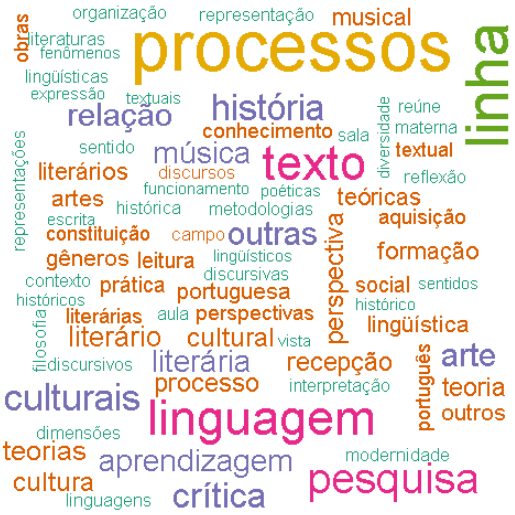

Figura 15: Nuvem de palavras de Linguistica, Letras e Artes.

Na nuvem de Ciências Exatas e da Terra (Figura 16) existem muitos termos de ciência da computação, como "software", "redes", "dados"e "algoritmos". Essas palavras evidenciam trabalhos em várias áreas da tecnologia de informação como desenvolvimento de software, arquitetura de redes e banco de dados. Outros termos como "geometria", "química"e "biológica"indicam estudos em áreas, como matemática e química. Apesar disso, pelo volume de termos relacionados a computação, percebe-se que uma grande parcela das teses relacionavam-se a estudos computacionais de diferentes tipos.

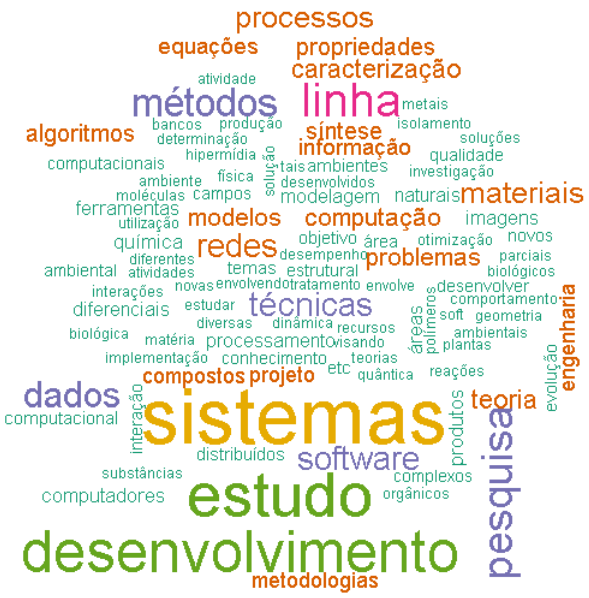

Figura 16: Nuvem de palavras de Ciências Exatas e da Terra.

A nuvem de palavras para Ciências Sociais pode ser vista na Figura 17. Nela as palavras "políticas", "públicas", "estado"e "sociedade"aparecem denotando temas governamentais, sociais e políticos. As palavras "direito", "constitucional"e "tutela"são termos comuns da área de direito. Além disso, pode-se verificar palavras relacionadas a organizações e empreendedorismo, como os termos "organização", "empresa", "inovação", "competividade"e "currículo". É possível inferir pelas palavras "docente", "educativos", "professores"e "escolar"temas voltados a pesquisas educativas e psicopedagogia. As palavras "marketing", "design"e "produtos"também aparecem, sugerindo temas voltados a publicidade e propaganda.

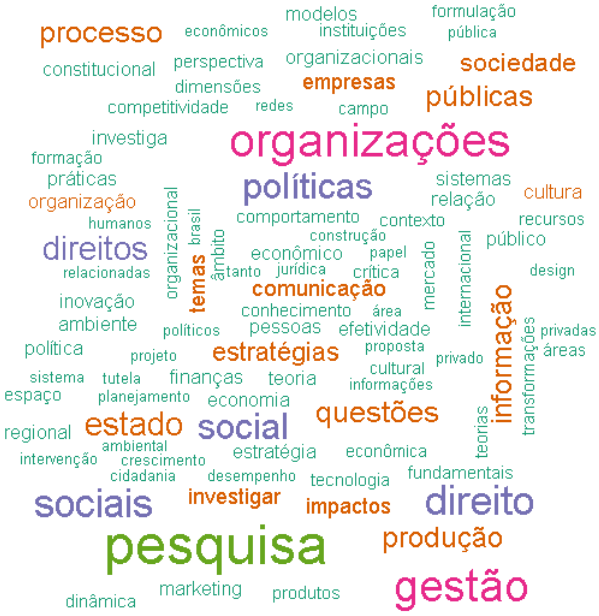

Figura 17: Nuvem de palavras de Ciências Sociais.

A nuvem de Ciências Humanas pode ser vista na (Figura 18) e apresenta as palavras "educação", "formação", "educativos"e "professores", "aprendizagem", "ensino", "escolar", "docente"e "educacional", apontando trabalhos relacionados a temas voltados a pesquisas educativas e aprendizado. As palavras "sociais", "gestão", "político"e "estado"indicam trabalhos relacionados às questões governamentais, sociais e políticas, assim como ocorre na nuvem de Ciências Sociais. Palavras como "filosofia", "pensamento", "subjetividade", "ética", "poder, "movimentos", "sociais", "culturais"e "psicologia"sugerem teses na área de filosofia, sociologia e psicologia.

Figura 18: Nuvem de palavras de Ciências Humanas.

7 Considerações Finais

Os conjuntos de dados empregados para validação da metodologia descrita permitiram observar que o processo desenvolvido incluiu etapas capazes de buscar diversos tipos de informações ocultas em conjuntos de dados. As etapas vão desde a procura por semelhanças entre os dados, utilizando processos de agrupamentos não-supervisionado, relacionamento entre cada uma das instâncias, procurando por regras do tipo "se-então"que indiquem padrões, até a procura de informações implícitas em colunas ou documentos textuais relacionados ao domínio do conjunto. Foi possível demonstrar uma abordagem completa e versátil, que pode ser aplicada a conjuntos não previamente classificados e em diversos tipos de dados. Na sequência são ressaltadas as conclusões obtidas a partir da aplicação de cada etapa do processo.

A etapa de pré-processamento mostrou-se eficiente na descoberta de problemas de completude e inconsistência de dados, permitindo a visualização dos valores assumidos em cada um dos atributos trabalhados e dos tipos de dados. Esta fase do processo também proporciona o cálculo de informações estatísticas sobre os dados, de modo que o analista é capaz de verificar se os campos possuem valores contidos dentro da faixa válida. No conjunto de dados ENEM o pré-processamento foi um

passo extremamente importante, dado que existiam muitas instâncias com dados incompletos ou inconsistentes. No caso do conjunto de Teses e Dissertações da Capes, esta etapa serviu mais como uma comprovação da qualidade do dados disponibilizados.

O agrupamento de dados demonstrou ser capaz de organizar os dados nos dois estudos de caso, gerando boas visualizações. A procura por semelhanças entre dados é importante na análise descritiva de um conjunto de dados e a abordagem selecionada incluiu técnicas de visualização que ajudam a sumarizar a informação através de mapas, facilitando a interpretação e análise. Os mapas de calor criados são uma ótima forma de se observar a distribuição dos atributos pelo agrupamento, trazendo informações mais completas sobre o mapa auto-organizável, revelando-se muito útil na análise dos dois conjuntos de dados. A combinação dos algoritmos de Kohonen e K-médias se apresentou como uma alternativa adequada para problemas onde o dado não possui classificação, sendo uma abordagem já empregada na literatura, como em (GODIN; HUGUET; GAERTNER, 2005) e (CHI; YANG, 2006) que realizam o estudo de um agrupamento em dois estágios utilizando estes dois métodos.

O processo de construção da árvore de decisão revelou-se um desafio em ambos os conjuntos de dados. Apesar da ferramenta R disponibilizar diversos pacotes para a execução deste tipo de tarefa, seu resultado apresentou um problema de generalização: o sobreajuste. Isso pode ter acontecido devido a uma grande quantidade de atributos utilizados na árvore de decisão, o que a tornou esparsa, gerando muitas regras para o conjunto. Algumas técnicas podem ser consideradas para atenuar este problema como agrupamentos hierárquicos e a utilização de modelos de seleção de atributos, que buscam extrair de um conjunto de dados as características mais relevantes, além de identificar atributos altamente correlacionados que não contribuem com o processo de classificação. Algoritmos de árvores de decisão costumam ser afetados pela presença de atributos irrelevantes no conjunto de dados, de modo que seu resultado pode não ser o esperado. O emprego de técnicas de seleção de atributos permite a melhoria do desempenho preditivo e acelera o aprendizado, gerando uma árvore mais compacta e generalizada (TANG; ALELYANI; LIU, 2014).

A etapa de mineração de texto foi de grande utilidade nos dois conjuntos de dados. Realizar a análise de campos e documentos textuais é um processo longo e cansativo se realizado manualmente, de modo que automatizar este processo pode ser muito útil na descoberta de conhecimento. Outro ponto importante desta fase é o fato de se procurar informações implícitas nos dados, podendo servir de complementação nas demais etapas do processo. Isso pode ser exemplificado na mineração do conjunto de dados ENEM, onde a informação encontrada na etapa de mineração de texto foi agregada à informação gerada nas demais etapas do processo para identificar os conteúdos onde os alunos encontraram mais dificuldades no ano de 2013.

Em relação à ferramenta R, notou-se que existe uma variedade muito grande de pacotes implementando técnicas de mineração de dados disponíveis nos *mirrors* da linguagem. A própria estrutura da linguagem facilita a manipulação de dados e permite que eles sejam armazenados em objetos que simulam tabelas e facilitam a visualização das transformações realizadas. Ela também está preparada para realizar operações com matrizes e vetores de forma ágil e simplificada, facilitando o desenvolvimento de funções próprias necessárias ao processo desenvolvido. Um dos pontos fortes desta ferramenta é a sua capacidade de gerar diferentes tipos de visualização gráfica, o que é uma etapa essencial para o processo de descoberta de informações. Como nota-se nos estudos de caso descritos nas Seções 5 e 6, as visualizações foram primordiais na análise. Mesmo quando a visualização gráfica não era possível, como ocorreu com as árvores de decisão geradas nos dois estudos realizados, foi possível extrair as regras de forma automatizada através do pacote *partykit*, agilizando o estudo das informações encontradas nas árvores.

8 Conclusões e Trabalhos Futuros

As pesquisas e estudos realizados neste trabalho demonstram que a área de mineração de dados está presente em diversos campos cotidianos, podendo ser considerada essencial na indústria, na economia e, como o foco deste trabalho indica, no governo.

O processo de mineração desenvolvido teve como base o ciclo de análise de dados proposto na literatura e demonstrou ser eficaz para a realização de mineração descritiva. Este processo teve como motivação o fato dos dados disponibilizados

em sites governamentais brasileiros não serem previamente classificados, sendo necessária a elaboração de um procedimento que utilizasse aprendizado não supervisionado para definir as partições presentes nos dados para posterior uso de algoritmos supervisionados para extração de regras. A etapa de mineração de texto também se mostrou adequada para a descoberta de novas informações sobre o conjunto, já que muitos dados relevantes encontram-se camuflados em textos.

A ferramenta R se apresentou como um recurso adequado para as tarefas mineração de dados, facilitando o processo desde a leitura do dado até a aplicação de técnicas de aprendizado de máquina complexas. Os pacotes relacionados às tarefas de mineração estão disponíveis em grande número, sendo que todos eles são muito bem documentados e mantidos atualizados.

Apesar de em crescente desenvolvimento, a mineração de dados traz consigo desafios. As técnicas utilizadas dependem substancialmente do domínio do problema trabalhado, bem como dos dados disponíveis para análise. As escolhas dos algoritmos de aprendizado de máquina muitas vezes só podem ser feitas após vários testes com os dados de treinamento, pois seus resultados estão intrinsicamente ligados com as propriedades de cada um deles.

Os trabalhos futuros pretendem seguir duas vertentes. A primeira está voltada para novos estudos de casos empregando a metodologia proposta, com o objetivo de gerar conhecimento a partir de outras bases de dados governamentais. Neste contexto espera-se contribuir com informação que auxilie nos processos administrativos do governo, e no dia a dia da sociedade, uma vez que a informação gerada pode fazer parte de uma nova plataforma ou um novo aplicativo a ser disponibilizado para o cidadão. O segundo foco de trabalho está na implementação da metodologia proposta empregando outras ferramentas de mineração de dados, como Weka ou Python. O Weka (HALL et al., 2009) é uma coleção de algoritmos de aprendizado de máquina voltada a mineração de dados, implementada em Java; e Python (ROSSUM; DRAKE, 2011), que não pode deixar de ser mencionada, é uma linguagem que tem se mostrado muito eficiente nesta área, com diversas bibliotecas de aprendizado de máquina disponíveis. A utilização de outras ferramentas permite gerar estudos comparativos quanto ao tempo de execução, dificuldade ou facilidade de implementação e qualidade da informação gerada a partir de uma base de dados.

Referências

BANCO, D. T. E. D. D. *CAPES*. 2013.

BRAGA, L. C.; DRUMMOND, I. N. Uma abordagem de mineração descritiva aplicada a dados abertos governamentais empregando a ferramenta r. *Anais do Computer on the Beach*, p. 051–060, 2017.

BREIMAN, L. et al. *Classification and regression trees*. [S.l.]: CRC press, 1984.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, p. 1–29, 2009.

CHARRAD, M. et al. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, v. 61, n. 6, p. 1–36, 2014. Disponível em: <http://www.jstatsoft.org/v61/i06/>.

CHI, S.-C.; YANG, C. C. Integration of ant colony som and k-means for clustering analysis. In: SPRINGER. *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. [S.l.], 2006. p. 1–8.

CORRÊA, Â. M. J.; SFERRA, H. Conceitos e aplicações de data mining. *Revista de ciência & tecnologia*, v. 11, p. 19–34, 2003.

DIETRICH, D. et al. *Open data handbook*. 2009.

DUTRA, C. C.; LOPES, K. M. G. Dados abertos: Uma forma inovadora de transparência. 2013.

FACELI, K. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, v. 2, p. 192, 2011.

FEINERER, I.; HORNIK, K. *tm: Text Mining Package*. [S.l.], 2017. R package version 0.7-1. Disponível em: <https://CRAN.R-project.org/package=tm>.

FELLOWS, I. *wordcloud: Word Clouds*. [S.l.], 2014. R package version 2.5. Disponível em: <https://CRAN.R-project.org/package=wordcloud>.

FOUNDATION, O. K. (Ed.). *The Open Data Handbook*. [S.l.], 2012. Disponível em: <http://opendatahandbook.org/>.

FRIZZARINI, C.; LAURETTO, M. S. Proposta de um algoritmo para induçao de árvores de classificaçao para dados desbalanceados. *Annais do X Simpósio Brasileiro de Sistemas de Informação*, p. 722–733, 2013.

GODIN, N.; HUGUET, S.; GAERTNER, R. Integration of the kohonen's self-organising map and k-means algorithm for the segmentation of the ae data collected during tensile tests on cross-ply composites. *NDT & E International*, Elsevier, v. 38, n. 4, p. 299–309, 2005.

HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009.

HAND, D. J.; MANNILA, H.; SMYTH, P. *Principles of data mining*. [S.l.]: MIT press, 2001.

HOTHORN, T.; HORNIK, K.; ZEILEIS, A. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, v. 15, n. 3, p. 651–674, 2006.

HOTHORN, T.; ZEILEIS, A. partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, v. 16, p. 3905–3909, 2015. Disponível em: <http://jmlr.org/papers/v16/hothorn15a.html>.

JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, Elsevier, v. 31, n. 8, p. 651–666, 2010.

KOHONEN, T. Essentials of the self-organizing map. *Neural networks*, Elsevier, v. 37, p. 52–65, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>.

RIPLEY, B. D. The r project in statistical computing. *MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network*, v. 1, n. 1, p. 23–25, 2001.

ROSSUM, G. V.; DRAKE, F. L. *The python language reference manual*. [S.l.]: Network Theory Ltd., 2011.

SUAREZ, A. F.; CANDEIAS, A. Avaliação de acurácia da classificação de dados de sensoriamento remoto para o município de maragogipe. *IV Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação, DeCart-UFPE, Recife*, 2012.

TANG, J.; ALELYANI, S.; LIU, H. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, CRC Press, p. 37, 2014.

TEAM, R. C. R language definition. *Vienna, Austria: R foundation for statistical computing*, 2000.

WEHRENS, R.; BUYDENS, L. M. et al. Self-and super-organizing maps in r: the kohonen package. *J Stat Softw*, v. 21, n. 5, p. 1–19, 2007. Disponível em: <http://www.jstatsoft.org/v21/i05>.

WILKINSON, L.; FRIENDLY, M. The history of the cluster heat map. *The American Statistician*, Taylor & Francis, v. 63, n. 2, p. 179–184, 2009.

WING, M. K. C. from J. et al. *caret: Classification and Regression Training*. [S.l.], 2017. R package version 6.0-77. Disponível em: <https://CRAN.R-project.org/package=caret>.

ZAMBENEDETTI, C. Extração de informação sobre bases de dados textuais. 2002.