Centro Universitário Estácio de Brasília

DAYANE MARTINS MOURA

DATA MINING APLICADO AOS DADOS ORÇAMENTÁRIOS ABERTOS DO SENADO FEDERAL

Taguatinga – DF
2016

Centro Universitário Estácio de Brasília

DAYANE MARTINS MOURA

DATA MINING APLICADO AOS DADOS ORÇAMENTÁRIOS ABERTOS DO SENADO FEDERAL

Projeto apresentado à banca examinadora do Centro Universitário Estácio de Brasília, como exigência parcial para a obtenção do grau de Bacharel em Sistemas de Informação, sob orientação do Prof. João Paulo Pimentel.

Taguatinga – DF
2016

Centro Universitário Estácio de Brasília

Dayane Martins Moura

DATA MINING APLICADO AOS DADOS ORÇAMENTÁRIOS ABERTOS DO SENADO FEDERAL

Projeto aprovado como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação pelo Centro Universitário Estácio de Brasília. Banca examinadora:

Taguatinga, DF 08/11/2016

_____

Prof. Esp. João Paulo Pimentel (Orientador)

Presidente

_____

Prof. MSc. Paulo Roberto Lobão Lima

1º Examinador

_____

Prof. Esp. Raphael Alves Bruce

2º Examinador

DEDICATÓRIA

“Dedico a minha mãe, graças a seus esforços estou concluindo mais esta etapa. Obrigada por sempre me incentivar a estudar e nunca desistir de mim”.

Dayane Martins Moura

AGRADECIMENTOS

“Agradeço à Deus por todo discernimento e sabedoria, e a minha namorada Carolina Alencar por toda paciência e dedicação”.

Dayane Martins Moura

A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê.

(Arthur Schopenhauer)

RESUMO

A Mineração de Dados ou *Data Mining*, é uma funcionalidade diretamente ligada à tecnologia da informação, que tem como principal função o agregamento e organização de grandes massas de dados. É uma prática considerada recente, surgiu com a necessidade que muitas organizações tiveram de encontrar técnicas de associação em grandes bases de dados. Uma vez que estes dados são minerados, podem ser transformados em informação útil para a organização, auxiliando na tomada de decisão e até mesmo na reestruturação de alguns setores. Atualmente, a mineração de dados tem sido amplamente utilizada em áreas de educação, engenharia, ciência e medicina em geral. Este trabalho teve como objetivo esclarecer os conceitos e técnicas utilizados pelo *Data Mining* aplicadas ao Orçamento Público e Programas Sociais. Os dados utilizados foram previamente escolhidos e minerados, com a finalidade de apresentar análises aplicadas a atributos de educação, saúde, inclusão social e segurança pública. Com todos os dados coletados, foi possível obter informações importantes. Mesmo com todos os investimentos na saúde, o Sistema único de Saúde ainda carece de recursos; por outro lado, é possível notar que o governo não amplia os investimentos de acordo com as necessidades básicas do cidadão comum. A técnica de *Data Mining* é uma forte aliada para as organizações público ou privadas, que precisam utilizar estratégias e se adequar as necessidades das partes interessadas.

Palavras-chave: *Data Mining*, Orçamento Público, Programas Sociais.

ABSTRACT

Data Mining is a functionality directly linked to the information of technology, whose main functions are the aggregation and organization of large masses of data. It is a recent practice that arose from the needing that organizations have to find techniques in large databases. Once this data is mined, it can be transformed into useful information for the organization, in decision for making and even rebild some sectors. Currently, data mining has been widely used in areas of education, engineering, science and general medicine. This work aimed to clarify the concepts and techniques used by Data Mining applied to the Public Budget and Social Programs. The data used were previously chosen and mined, with the purpose of presenting analyzes applied to attributes of education, health, social inclusion and public safety. With all the data collected, it was possible to obtain important information. Even with all investments in health, the Single Health System still lacks resources; On the other hand, it is possible to notice that the government does not extend the investments according to the basic necessities of the common citizen. The Data Mining technique is a strong ally for public or private organizations that needing to use strategies and tailor of stakeholders.

Key words: data mining, public budget and social programs.

LISTA DE ILUSTRAÇÕES

Figura 1 - FONTE: Adaptada de Data Mining: Técnicas e Aplicações para o Marketing Direto, Berkeley, 2001. 19

Figura 2 - FONTE: DevMedia – Mineração de Dados Tarefas e Técnicas 22

Figura 3 - Fonte: Adaptada de AWAD, E. M; GHAZIRI, H. Knowlwdge management. Pearson: Englewood Cliffs: Prentice Hall, 2004. 24

Figura 4 - FONTE: Escola Superior de Tecnologia – Instituto Politécnico de Castelo Branco 25

Figura 5 - FONTE: Rezende, 2003. P.318 28

Figura 6 - FONTE: Ermilov, 2015 – An Open – Source Data Management Solution For Open Data 29

Figura 7 - FONTE: Portal Siga Brasil – Despesa Fiscal e Seguridade – Pesquisa de Campo 2016 31

Figura 8 – FONTE: Pesquisa de Campo 2016 32

Figura 9 – FONTE: Pesquisa de Campo 2016 33

Figura 10 – FONTE: Pesquisa de Campo 2016 35

Figura 11 - FONTE: Pesquisa de Campo 2016 36

Figura 12 - FONTE: Pesquisa de Campo 2016 37

Figura 13 – FONTE: Pesquisa de Campo 2016 38

Figura 14 – FONTE: Pesquisa de Campo 2016 38

Figura 15 – FONTE: Pesquisa de Campo 2016 39

Figura 16 – FONTE: Pesquisa de Campo 2016 40

Figura 17 – FONTE: Pesquisa de Campo 2016 40

Figura 18 – FONTE: Pesquisa de Campo 2016 47

Figura 19 – FONTE: Pesquisa de Campo 2016 48

Figura 20 – FONTE: Pesquisa de Campo 2016 48

Figura 21 – FONTE: Pesquisa de Campo 2016 49

Figura 22 – FONTE: Pesquisa de Campo 2016 49

LISTA DE ABREVIATURAS E SIGLAS

DM – Data Mining

CKAN – Comprehensive Knowledge Archive Network

SGBD – Sistema Gerenciador de Banco de Dados

KDD – Knowledge Discovery in Database

OWL – Web Ontology Language

XML – Extensible Markup Language

PDF – Portable Document Format

HTML – Hyper Text Markup Language

LOA – Lei Orçamentária Anual

SUS – Sistema Único de Saúde

HIV - Human Immunodeficiency Virus

API – Applications Programming Interface

JSON – JavaScript Object Notation

URL – Uniform Resource Locator

CSV – Comma-separated Values

JDK – Java Development Kit

YAML – Aint’t Another Markup Language

Sumário

1 – Introdução 14
1.1 – Formulação do Problema 14
1.2 – Justificativa 14
1.3 – Objetivos 15
1.3.1 – Objetivo Geral 15
1.3.2 - Objetivos Específicos 15
1.4 – Metodologia 15
1.5 – Estrutura do Trabalho 16
2 – REFERENCIAL TEÓRICO 18
2.1 – Introduzindo Conceitos 18
2.2 – Banco de Dados Relacionais 19
2.3 – KDD e Mineração de Dados 19
2.3.1 – Desenvolvimento Tecnológico 20
2.3.2 – Execução de KDD 20
2.3.3 – Aplicação de Resultados 20
2.3 – Técnicas de Data Mining 24
2.3.1 – Descoberta de Associações 25
2.3.2 – Classificação 25
2.3.3 - Regressão 26
2.3.4 - Agrupamento (Clusterização) 26
2.3.5 – Sumarização 27
2.3.6 – Detecção de Desvios 27
2.3.7 – Descoberta de Sequências 27
2.3.8 – Previsão de Séries Temporais 27
2.3.9 - MBR – Memory Based Reasoning 27
2.4 – CKAN 28
3 – METODOLOGIA 30
3.1 – OWL 31
3.2 – Exemplificando o Processo de Pesquisa 32
3.3 – Elaboração do processo de pesquisa 33
3.3.1 – Escolher o Processo de Pesquisa 35
3.4 – Técnicas 36

3.4.1 – Descoberta de Associações 37
3.4.2 – Classificação 37
3.4.3 – Agrupamento 38
3.4.4 – Sumarização 38
3.4.5 – Descoberta de Sequências 39
3.4.6 – MBR – Memory Based Reasoning 40
3.5 – Aplicando o CKAN 41
3.5.1 – Objetivos 43
3.5.2 – Estrutura 43
3.5.3 – Ferramenta de Pesquisa 44
3.5.4 – Utilização 44
3.5.5 – Workflow 44
3.5.6 – Estruturando Dados 45
3.5.6.1 – Utilizando o JSON 45
3.5.7 – XML 45
4 – APRESENTAÇÃO E ANÁLISE DOS RESULTADOS 46
5 – CONCLUSÃO 51
REFERÊNCIAS 53
GLOSSARIO 55

1 – Introdução

O *Data Mining* ou Mineração de Dados consiste em um processo de análise projetado para extrair dados específicos de grandes bases, usando métodos e cálculos para que se possam apresentar as informações solicitadas. Para que o processo de extração seja possível, é preciso que o projeto seja estruturado com base em determinadas variáveis e padrões previamente definidos, em suma, a técnica permite filtrar o que é repetitivo e caótico, utilizando técnicas de exploração para se encontrar inconsistências. Os meios mais comuns da utilização de DM estão relacionados com o mercado de negócios, envolvendo grandes empresas que necessitam de ferramentas capazes de fornecer buscas baseados em padrões consistentes e relacionamentos sistemáticos.

1.1 – Formulação do Problema

O *Data Mining* é uma técnica considerada recente no universo da Tecnologia da Informação, a prática utiliza meios modernos de análise a grandes quantidades de dados. Entretanto, até alcançar a simplicidade na análise de dados, muitas técnicas devem ser combinadas. Como as técnicas utilizadas pelo *Data Mining* são elaboradas? Como elas funcionam na prática? Como saber se utilizamos os melhores caminhos ?

1.2 – Justificativa

Com o advento das melhores práticas contidas no mundo da computação, surgiram com elas grandes necessidades. Uma delas é a adoção a práticas de análise e mineração de grandes bases de dados.

Muitas são as vantagens adquiridas por uma companhia ao se programar um sistema de *Data Mining*. Sejam elas;

- ✓ As informações solicitadas podem ser filtradas, simplificando o retorno de dados;
- ✓ O caminho utilizado pelo *Data Mining* na base de dados pode ser selecionado, limitando ou não o campo de busca;
- ✓ Desvendar tendências e preferências do mercado;
- ✓ Prevenir e detectar riscos de fraudes.

1.3 – Objetivos

1.3.1 – Objetivo Geral

Fornecer um estudo sistemático referente às técnicas de utilização *Data Mining*, funcionalidades e onde aplicá-las.

1.3.2 - Objetivos Específicos

- Detalhar a utilização correta do *Data Mining*, desde o desenvolvimento de técnicas e coleta de informações, até a conclusão e retorno de dados para o usuário final;
- Demonstrar de forma clara e coesa, os passos necessários para melhor utilização das técnicas de mineração de dados, de acordo com cada necessidade previamente estabelecida;
- Descrever com dados comprobatórios, os benefícios propostos pelo *Data Mining* após a sua implementação na companhia.

1.4 – Metodologia

Para a concretização dos objetivos inicialmente propostos por este projeto, serão utilizadas pesquisas bibliográficas, tendo em vista as técnicas que o método abrange: leitura, análise e interpretação de arquivos já publicados. Segundo Vergara (2005; p.48), a pesquisa bibliográfica é o estudo sistematizado desenvolvido com base em material publicado em livros,

revistas, jornais, redes eletrônicas, isto é, material acessível ao público em geral.

Será realizado um trabalho baseado em dados legislativos do Senado Federal, bem como o orçamento aberto e o portal da transparência, disponibilizados pela ferramenta CKAN – *Comprehensive Knowledge Archive Network*.

1.5 – Estrutura do Trabalho

No capítulo 2, será abordado o referencial teórico, que é basicamente a ideologia de alguns autores em relação ao *Data Mining*. Como surgiu, porque surgiu, como deve ser utilizado para o alcance do objetivo final. Serão exemplificadas algumas técnicas muito utilizadas, e por fim, as vantagens de serem utilizadas.

No capítulo 3, será abordada a concepção do problema proposto. Será utilizada uma rica fonte de dados a respeito do orçamento público referente a despesas e relatórios de gestão fiscal do Senado Federal. O portal da transparência, onde podemos colher informações do que diz respeito aos recursos utilizados pelos senadores, estrutura administrativa do Senado Federal, modalidades de licitações e contratos firmados pelo órgão, sobre servidores ativos, aposentados e colaboradores de um modo geral. Depois de finalizada a pesquisa, será possível o entendimento das partes interessadas de como são aplicados os recursos financeiros que cabem ao poder legislativo.

O capítulo 4 será estruturado de acordo com as conclusões obtidas com a ferramenta de *Data Mining* do Portal da Transparência. Será representado tudo que foi possível extrair, registrar e filtrar sobre os colaboradores que prestam serviço ao Senado Federal. De modo geral, será feita uma avaliação do que foi possível extrair da ferramenta disponível, se é acessível para todo o público, se esclarece aspectos importantes como o que é feito com o orçamento público, e se de fato traz resultados rápidos e informações concisas.

No capítulo 5 faremos o encerramento do projeto. Será mostrado tudo que foi possível extrair, os obstáculos encontrados, as vantagens e avanços possíveis para a companhia que implantar a ferramenta de mineração

de dados, e se houver os pontos negativos de se trabalhar com a aplicação de *Data Mining*.

2 – REFERENCIAL TEÓRICO

2.1 – Introduzindo Conceitos

Com o passar do tempo, vem crescendo a necessidade de se tirar proveitos cada vez maiores dos dados. Para tanto, surge o triangulo dado, informação e conhecimento. O dado é matéria bruta, é a essência da informação. Informação é o dado trabalhado, bem contextualizado e definido. O conhecimento é a aplicação correta da informação, é a informação utilizada de forma inteligente. Por fim, a boa utilização da informação é o alicerce do conhecimento.

Para Silva, Peres e Boscarioli (2006), é exorbitante a quantidade de dados gerados atualmente, o que acaba ultrapassando os limites da capacidade humana de interpretação. A necessidade de armazenamento de determinadas informações, era um desejo antigo de grandes empresas. Sanado este problema, surgiu outra questão importante: Como essa imensa quantidade de dados pode ser analisada? Com este desafio em especial, surgiu o interesse em elaborar planos estratégicos, tendo como foco principal a descoberta do conhecimento em grandes bases de dados, com o objetivo de elaborar novos planos estratégicos que auxiliassem o aumento das vendas, a definição de perfis e a seleção de produtos específicos.

A descoberta do conhecimento é composta por uma série de processos que em etapas previamente definidas, são capazes de auxiliar e desmistificar complexas analises de dados. A primeira etapa tem como função o pré-processamento de dados na base de dados, com o objetivo de entregar dados limpos, preparados e selecionados a fase seguinte. A segunda etapa, que apesar da ordem é a mais importante, consiste na própria Mineração de Dados.

Na fase de Mineração de Dados, são executados algoritmos em linguagens de máquina ou de redes neurais artificiais sobre os dados, com o objetivo de criar modelos que auxiliem em tarefas básicas, como classificação, agrupamento e associação de dados. Como terceira e última etapa, os resultados obtidos com a Mineração de Dados são interpretados e analisados qualitativamente e quantitativamente. É possível notar que o *Data Mining* ou Mineração de Dados, é uma área interdisciplinar e possui como pré-requisito a

vivência do leitor em: banco de dados, álgebra linear, matemática básica, estrutura de dados e algoritmos.

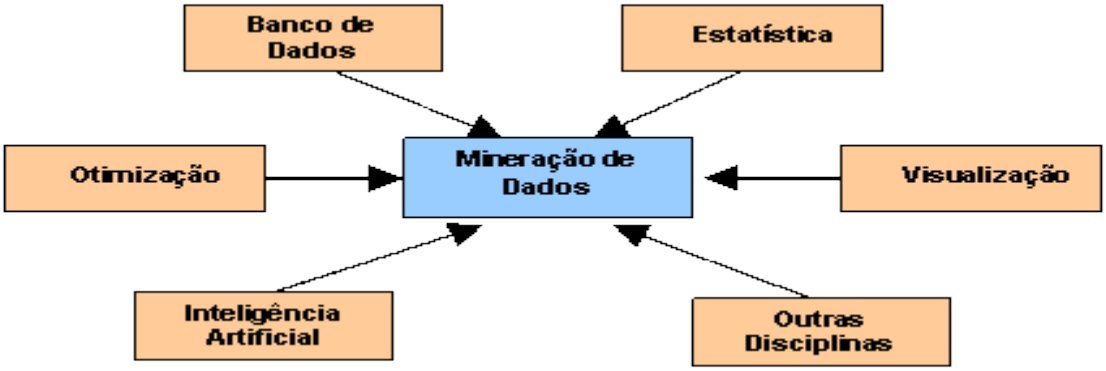


Figura 1 - FONTE: Adaptada de Data Mining: Técnicas e Aplicações para o Marketing Direto, Berkeley, 2001.

2.2 – Banco de Dados Relacionais

Para Amaral (2001), como principal função, são responsáveis por armazenar e recuperar dados de forma eficiente. Bancos de Dados relacionais devem ser projetos de forma eficiente para que possam ser utilizados corretamente, ou seja, para que torne possível a extração de todo tipo de informação nele contido. Mas de que forma esse processo é executado?

É mais simples do que aparenta, o primeiro passo é a elaboração do problema, em seguida é realizado um mapeamento para a linguagem de consulta, esta consulta é submetida ao SGBD (Sistema Gerenciador de Banco de Dados). Este processo é criado para solucionar problemas que devem ser necessariamente definidos, em suma, as informações retiradas são respostas a uma consulta previamente definida. Porém, os dados salvos em uma base de dados, podem conter vários tipos de padrões e comportamentos relevantes que não podem ser descobertos no primeiro momento. Dentro desta contextualidade está inserida a aplicabilidade do *Data Mining*.

2.3 – KDD e Mineração de Dados

Em diversas fontes de pesquisa como livros e publicações, o processo de mineração de dados é visto como parte de um processo maior, denominado

KDD – *Knowledge Discovery in Database,* que significa basicamente a descoberta do conhecimento em bases de dados.

Segundo Goldschmidt e Passos (2005), o conceito de KDD foi implementado no ano de 1989, em referência ao enorme conceito de descobrir conhecimento a partir de bases de dados.

As atividades executadas na área de KDD podem ser organizadas em três grandes grupos: Atividades voltadas ao desenvolvimento tecnológico, atividades de execução de processos de KDD e atividades envolvendo a aplicação de resultados obtidos em processos de KDD. Sejam:

2.3.1 – Desenvolvimento Tecnológico

Esta etapa concentra todos os propósitos de concepção, aprimoramento e desenvolvimento de algoritmos, ferramentas e tecnologias de apoio que podem e ser utilizadas na busca por conhecimento em grandes bases de dados.

2.3.2 – Execução de KDD

Esta etapa diz respeito às atividades concentradas à busca efetiva de conhecimento na base de dados. É onde as ferramentas desenvolvidas na etapa de “Desenvolvimento Tecnológico” serão aplicadas.

2.3.3 – Aplicação de Resultados

Para finalizar, consolidados os modelos de conhecimento que podem ser aproveitados, as atividades se concentram à aplicação dos resultados no contexto em que foi realizado o processo de KDD.

O processo de KDD é subdividido em seis tarefas e envolve duas fases: A preparação dos dados e a mineração propriamente dita.

Este processo é inicializado pelo entendimento do domínio da aplicação e a definição dos objetivos que devem ser concluídos. Nesta etapa, as principais questões que justificam a mineração são apontadas. Definidos os objetivos, tendo o problema a ser solucionado, é hora de escolher a técnica que deve ser

aplicada. Os objetivos apontados e a técnica a ser utilizada, definem por onde a mineração será iniciada.

O próximo passo serve para realizar a limpeza de um pré-processamento de dados. É uma fase importante, serve para eliminar inconsistências, redundâncias e eventuais problemas de tipos.

Feita a limpeza, é o momento de passar os dados pré-processados por uma transformação, com o objetivo de simplificar o manuseio dos dados pelas técnicas de mineração previamente definidas.

Em seguida, é chegada a fase de mineração, onde é iniciada com a seleção dos algoritmos a serem utilizados. Essa escolha envolve o objetivo definido no processo de KDD. No final do processo, o sistema que executou a mineração de dados, poderá apresentar os relatórios das descobertas de dados, que passam por prévias interpretações pelos desenvolvedores ou administradores de dados. Podendo assim transformar as informações coletadas em conhecimento.

Segundo os autores Goldschmidt e Passos (2015), a área de Tecnologia da Informação vem trazendo inovações com seus avanços frequentes, como consequência, os dados são classificados com bastante heterogeneidade, ou seja, não possuem uma forma ou natureza definida.

O *Data Mining* é um subconjunto do *Big Data*, sendo este uma referência a um grande volume de dados. O *DM* se distingue de acordo com as informações que se precisa extrair.

Segundo Polito (1997), o Data Mining pode ser caracterizado como a análise de grandes bases de dados que possuem informações ocultas. Para o autor, a empresa que adquire um sistema de Data Mining, estará sempre à frente de outras que não possuem aplicação igual ou semelhante.

Para Figueira (1998), com o advento das técnicas de Data Mining, a informação passou a ser valorizada como nunca havia acontecido desde a criação da Informática.

Ávila (1998) ressalta que o Data Mining é basicamente a busca por padrões simplificados de dados contidos em grandes massas, é a simplificação do KDD – Knowledge Discovery Database, que se resume a extração de dados de modo complexo e analítico.

Para Ronaldo Goldschmidt e Emanuel Passos (2005), o uso do KDD numa visão geral, tem proporcionado à descoberta de dados em diversas áreas

do conhecimento: comercial, administrativa, governamental e cientifica, áreas estas que analisam grandes fontes de dados. Nesta etapa, procura-se obter um melhor entendimento do problema para se usar a ferramenta adequada. Tais descobertas tem como ponto de partida a divisão do problema, a diferença entre conhecimento, informação e dado. No KDD são tratados os dados para um melhor entendimento do problema, para assim se aplicar o Data Mining. Quando identificado, escolhe o padrão para sua interpretação. No KDD esta descoberta de dados é multidisciplinar, ou seja, onde se conhece diversas áreas do conhecimento, estatística, inteligência computacional, reconhecimento de padrões e banco de dados. Na figura abaixo os processos do KDD estão muito bem ilustrados:

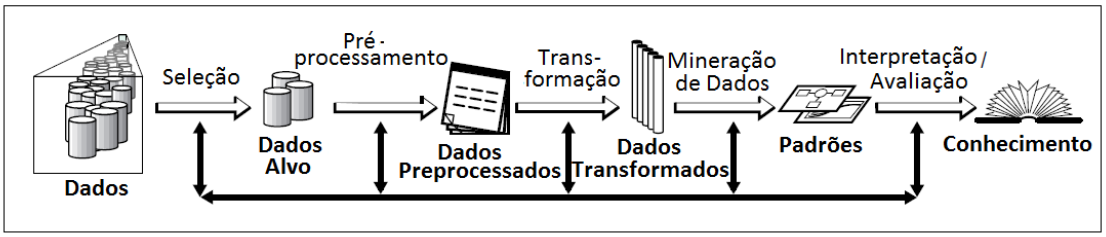


Figura 2 - FONTE: DevMedia – Mineração de Dados Tarefas e Técnicas

Para um melhor entendimento do KDD no Data Mining, é necessário apresentar os elementos envolvidos na área de aplicação. Basicamente, o KDD é composto por três componentes: o problema, o recurso e os resultados possíveis da solução. O problema em que vai ser usado o KDD é dividido em observação dos dados e seus aspectos, especialistas no domínio da aplicação e os objetivos da aplicação.

- Observação dos dados – É feita em um aspecto intencional e extensional, onde o intencional refere-se à estrutura e o extensional aos registros ou casos que o compõem.
- Especialista no domínio da aplicação - É o profissional ou grupo, que domina o ambiente e conhece o assunto onde será aplicado o KDD.
- Objetivos da aplicação – São as características dos resultados encontrados ao final do processo, usando uma precisão mínima de modelo de conhecimento.
  Os recursos são as ferramentas KDD, os especialistas a plataforma disponível.
- Especialista, profissional ou grupo com experiência no uso do KDD para sua melhor aplicação.
- Ferramentas, tipo de recurso computacional a ser utilizado.

- Plataforma computacional indica os recursos de hardware disponíveis para uso e aplicação.

Os resultados englobam todo o esperado com a aplicação do que se tem disponível.

Como exemplo de fundamentação das técnicas de KDD, podemos citar a teoria de redes neurais, o método Rule Evolver e o método NFHB-Class.

Segundo Tamio Shimizu (1988), a busca de novos conhecimentos sobre banco de dados, relatórios e organização, é um conceito de inteligência organizacional, onde as organizações tradicionais tomam decisões com base em experiências passadas. As decisões são tomadas a partir resultados positivos ou negativos, onde os positivos são reutilizados e os negativos reestruturados para obterem resultados positivos.

Achar um equilíbrio ótimo entre a exploração e o refinamento é difícil, é uma técnica que envolve um entendimento profundo entre pessoas e os níveis de sistemas. Um ótimo equilíbrio pode variar de profissional para profissional, ou de organização para organização, existem limites rígidos entre um conceito e outro. O conhecimento de um especialista é mais amplo ou profundo, depende do nível e da qualidade da informação a qual ele tem acesso.

Na próxima página, será possível visualizar uma tabela comparativa entre dados, informação e conhecimento.

DADOS	INFORMAÇÃO	CONHECIMENTO
Elementos transacionais sobre fatos de interesse da organização.	Agregação de dados processados para certa finalidade.	Conjunto de informações individuais ou organizacionais baseado na percepção, habilidade, treinamento e experiência.
Exemplos: - Dados sobre vendas ou itens produzidos	Exemplo: - Relatório de vendas; - Livro contábil; - Arquivo de inventário.	Exemplo: - Conhecimento de especialistas, analistas; - Decisão do administrador

Figura 3 - Fonte: Adaptada de AWAD, E. M; GHAZIRI, H. Knowlwdge management. Pearson: Englewood Cliffs: Prentice Hall, 2004.

Data Mining é basicamente o processo utilizado para extrair informações validas, utilizáveis, desconhecidas e abrangentes, existentes em um data warehouse.

2.3 – Técnicas de Data Mining

O surgimento das técnicas de Data Mining não é muito recente como muitos imaginam, teve início em meados dos anos 1980 quando as pesquisas de Inteligência Artificial foram aprofundadas. O que faz muita gente pensar que são técnicas recentes, é fato de elas terem tido uma adoção recente por parte dos grandes sistemas de Bancos de Dados, Figueira (1998). Para o autor, existem três características que se destacam:

- A expansão e difusão de sistemas transacionais volumosos;
- A informação como vantagem competitiva;
- A difusão de tecnologia de informação escalável.

Cada técnica de Dara Mining possui uma característica única, pontos específicos onde podem e devem ser aplicadas, dependendo de casa objetivo em particular. Na modelagem a seguir, podemos observar um breve resumo do que consiste o processo de Data Mining:

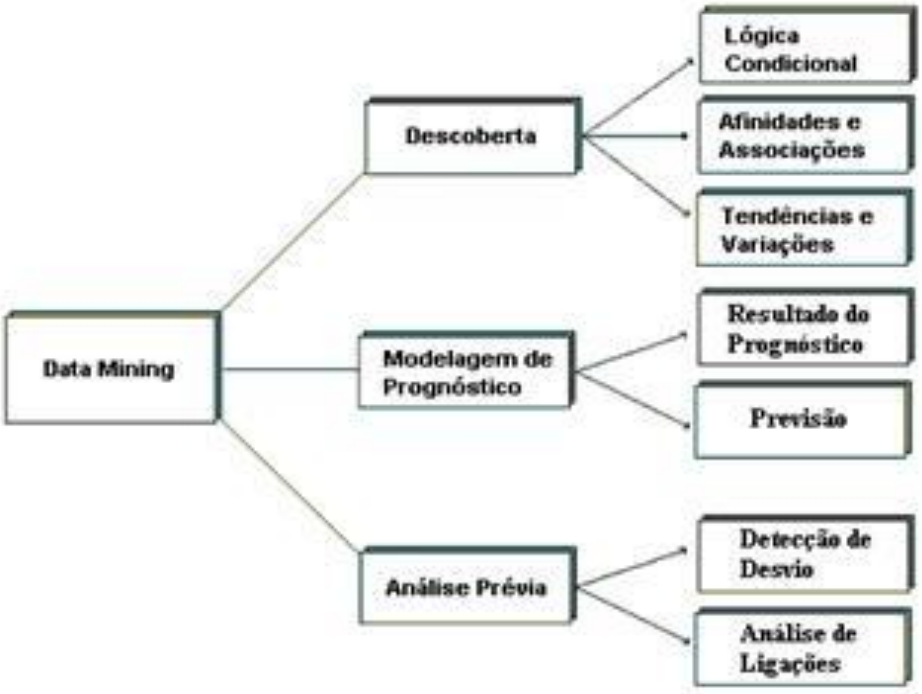


Figura 4 - FONTE: Escola Superior de Tecnologia – Instituto Politécnico de Castelo Branco

2.3.1 – Descoberta de Associações

Para Goldschmidt e Passos (2015), nesta etapa são definidos alguns conceitos importantes. Os conjuntos de dados possuem registros, e estes registros são conhecidos como transações – cada registro é uma transação. As transações também possuem composições, e estas composições são conhecidas como itens. Funciona basicamente da seguinte forma: Uma busca é efetuada na base de dados, a procura de itens que ocorrem de forma simultânea com outras transações. É uma técnica considerada por muitos autores como descritiva, ou seja, é utilizada para identificar padrões em dados históricos.

2.3.2 – Classificação

Para Goldschmidt e Passos (2015), na classificação, ocorre à divisão dos atributos do conjunto de dados em dois tipos, o primeiro deles é o previsor e o segundo é o alvo. Para cada atributo alvo existe uma classe que normalmente é vinculado a um rótulo que pertence a um conjunto previamente definido. Em

suma, a classificação tem como função mapear os registros em classes. Quando desvendada, tal função tem o objetivo de aplicar o mapeamento em novos registros de forma a prever a classe em que os registros se encaixam.

2.3.3 - Regressão

Assim como a classificação, a regressão também consiste em um mapeamento de registro de uma determinada fonte de dados em um espaçamento de valores numéricos reais. Goldschmidt e Passos (2015).
Tem o objetivo de definir um valor numérico de alguma variável desconhecida a partir dos valores de variáveis já conhecidas.

A técnica de regressão é considerada preditiva.

É considerada a técnica mais simples de ser utilizada, mas com certeza não é a mais poderosa, na verdade é a menos poderosa. O modelo é considerado fácil, pois só possui duas variáveis, uma de entrada e outra de saída.

2.3.4 - Agrupamento (Clusterização)

De acordo com Harrison (1998), o agrupamento possui a função de separar os registros da fonte de dados em subconjuntos, conhecidos como clusters, o procedimento é feito de tal maneira que os itens dos subconjuntos são capazes de compartilhar as propriedades simples que os diferenciem dos demais itens dispostos no subconjunto. O objetivo desta técnica é aumentar a igualdade intracluster e diminuir a igualdade intercluster.

As informações podem ser divididas em classes de elementos parecidos. Neste caso, nenhuma informação é transmitida ao sistema no que diz respeito às classes existentes. O algoritmo utilizado desvenda as classes a partir das opções disponibilizadas na base de dados, agrupando assim um conjunto de objetos em classes de objetos que possuem alguma semelhança entre si. A técnica de agrupamento é considerada uma tarefa descritiva.

2.3.5 – Sumarização

Pode ser considerada uma das técnicas mais simples, como o próprio nome já diz, a sumarização tem como objetivo apontar a semelhança entre os registros da fonte de dados. Goldschmidt e Passos (2015).

2.3.6 – Detecção de Desvios

Para Goldschmidt e Passos (2015), a detecção de desvios tem o objetivo de mencionar itens da fonte de dados cujas identificações dos que é considerado a normalização do contexto em análise. Estes itens são conhecidos como valores anormais ou outliers.

2.3.7 – Descoberta de Sequências

Segundo os autores Goldschmidt e Passos (2015), a Descoberta de Sequências possui uma grande similaridade com a técnica Descoberta de Associações, a tarefa é apontar registros comuns levando em consideração um espaço de tempo previamente definido.

2.3.8 – Previsão de Séries Temporais

Resume-se a uma série de apontamentos (necessariamente numéricas) organizada no tempo. Goldschmidt e Passos (2015).

2.3.9 - MBR – Memory Based Reasoning

Raciocínio Baseado em Memória é uma técnica de Data Mining direcionada que usa padrões e casos concretos como modelos para facilitar suposições e exemplificar fatos desconhecidos. Em suma, combinada em pares os dados conhecidos para prever resultados futuros. A principal vantagem desta técnica é a capacidade de se obter novas habilidades pelo

simples gesto de introduzir novos exemplos a base de dados, Berry e Kremer (1997).

Para finalizar, será apresentada uma figura de sequências da utilização de técnicas de Data Mining:

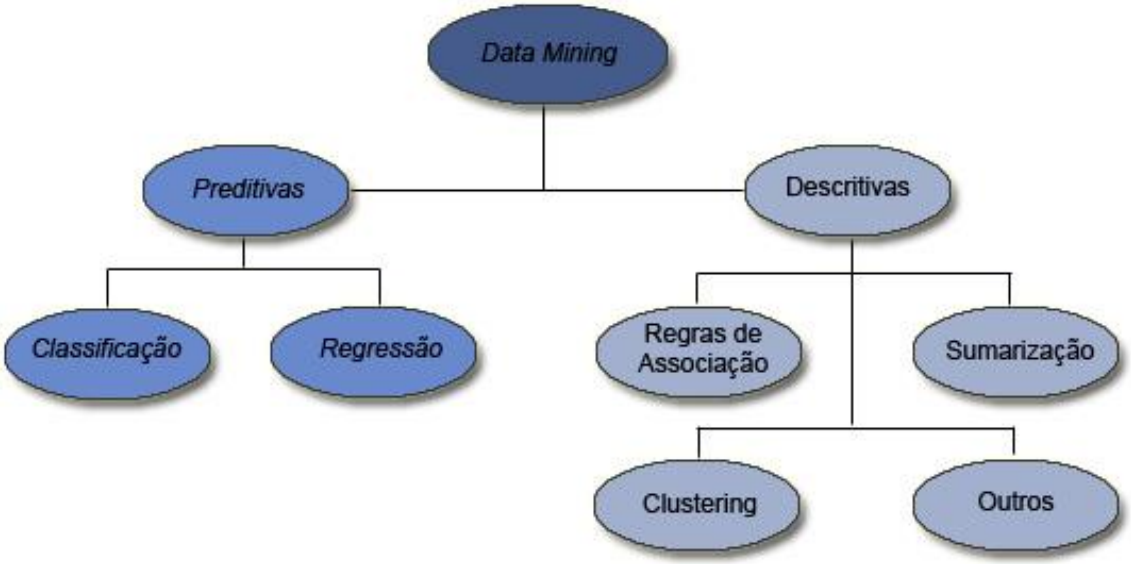


Figura 5 - FONTE: Rezende, 2003. P.318

2.4 – CKAN

Umas das maiores ferramentas de catalogação de dados *Open Source* desenvolvidas até hoje, possui uma interface web intuitiva que permite o fácil manuseio da aplicação.

Conforme Santarem (2013), o CKAN é um ambiente digital para o controle e manuseio de dados abertos, que disponibiliza ferramentas para acelerar o compartilhamento e utilização de dados públicos.

Segundo Melo (2015), o CKAN é um sistema fortemente elaborado e desenvolvido nas linguagens Python e Javascript, para o gerenciamento de dados. Elaborado pela Open Knowledge Foudation, este sistema foi projetado para disponibilizar dados de forma prática e descomplicada. Tem como público alvo as empresas e governos que desejam mapear dados e transparece-los para as partes interessadas.

Segue um exemplo da arquitetura simplista utilizada pela ferramenta CKAN:

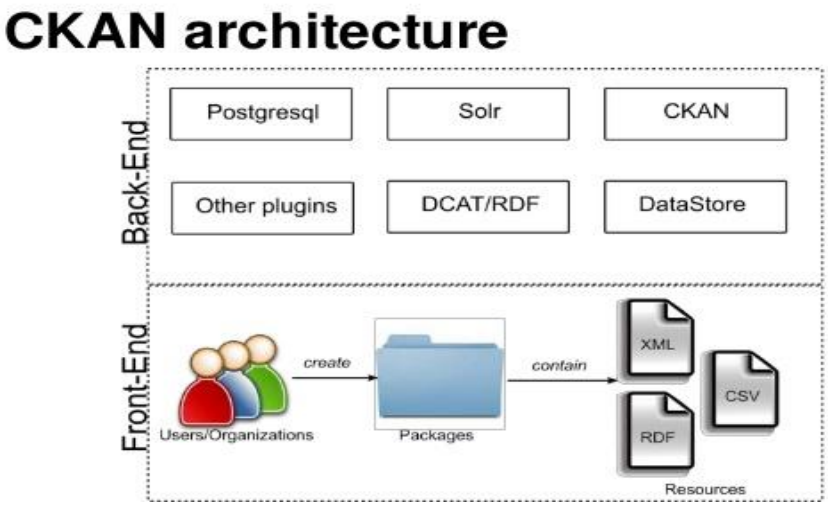


Figura 6 - FONTE: Ermilov, 2015 – An Open – Source Data Management Solution For Open Data

3 – METODOLOGIA

De acordo com o Senado (2016), o Portal da Transparência assim como o Orçamento Aberto, foi desenvolvido antes da promulgação da Lei de Acesso a Informação Pública 12.527/2011, que regula o acesso a dados e informações detidas pelo Governo Federal. Esta lei compõe um limite para a acessibilização da informação pública, e prioriza, entre inúmeras condições técnicas, que a informação solicitada pelo cidadão deve seguir à risca requisitos ordenados com as chamadas três leis dos dados abertos propostas pelo ativista David Eaves, são elas:

1- Se o dado não pode ser encontrado e indexado na Web, ele não existe;
2- Se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser reaproveitado; e
3- Se algum dispositivo legal não permitir sua replicação, ele não é útil.

Baseado neste relato foi criado o Portal Brasileiro de Dados Abertos, que foi desenvolvido com o intuito de disponibilizar simplificadamente os dados e informações públicas.

Este portal nada mais é do que a própria ferramenta CKAN, corretamente adequada ao Senado Federal e devidamente formatada para melhor atender o propósito do órgão em questão.

A escolha dos conceitos de *Data Mining* para a elaboração deste projeto, se deu devido à facilidade e acessibilidade que está ferramenta proporciona. O objetivo é tornar acessível às informações dos gastos e funções executadas pelo poder legislativo, realizando uma análise do Orçamento, gastos de senadores e pautas de sessões, obtendo informações de extrema importância para o cidadão comum.

Os conjuntos de dados contidos neste trabalho estão disponíveis para consulta e *download* no Portal Brasileiro de Dados Abertos e no Portal da Transparência do Senado Federal.

Segundo o Portal da Transparência, os registros no CKAN são vistos como acervos de dados, este acervo é composto por várias informações, como descrições, origens, autoridade competente, tipo de licença e recursos.

No portal Dados Abertos e por meio do Siga Brasil, é possível encontrar conjuntos de dados diversos. Informações sobre dados orçamentários abertos, gastos do Senado Federal, despesas dos senadores, recursos humanos,

contratos e licitações. As consultas podem ser prontas ou elaboradas, e podem ser salvas em diversos formatos (XML, HTML e TEXT).

No que diz respeito às pesquisas prontas, ao acessar o portal é possível visualizar ou baixar vários conjuntos de dados prontos, como reuniões de comissões, diários do senado e do congresso, emendas, medidas provisórias, pronunciamento de senadores, vetos e votações normais, e projetos sociais.

Para melhor entendimento, será apresentada abaixo uma tabela extraída do Siga Brasil, disponível no Portal da Transparência, em formato XML, a tabela contém dados referentes aos gastos com a saúde pública em relação a alguns programas específicos que serão descritos na imagem:

LOA 2016 - Execução Orçamentáriapor Programa

R$ 1,00

Programa (Cod/Desc)	Dotação Inicial	Autorizado	Empenhado	Liquidado	Pago	RP Pago
0002 - PREVENÇÃO E CONTROLE DE DOENÇAS TRANSMITIDAS POR VETORES	0	0	0	0	0	0
0004 - QUALIDADE E EFICIÊNCIA DO SUS	0	0	0	0	0	0
0005 - ASSISTÊNCIA FARMACÊUTICA	0	0	0	0	0	0
0011 - PREVENÇÃO E CONTROLE DO CÂNCER E ASSISTÊNCIA ONCOLÓGICA	0	0	0	0	0	0
0016 - GESTÃO DA POLÍTICA DE SAÚDE	0	0	0	0	0	282.758

Figura 7 - FONTE: Portal Siga Brasil – Despesa Fiscal e Seguridade – Pesquisa de Campo 2016

3.1 – OWL

A *web ontology language* é um método de desenvolvimento que possui o objetivo de conceituar e instanciar modelos de dados para informações na web. É usada em aplicações que precisam processar informações contidas em documentos web, não sendo apresentada em formato legível apenas por humanos. É uma recomendação do W3C.

A OWL oferece recursos para elaboração de vocabulários que permitem que a web seja mais semântica, oferecendo significados para serem utilizados

em softwares. Com ela, facilita-se a interpretação de dados por máquinas, utilizando estruturas como XML, RDF e RDFSs.

3.2 – Exemplificando o Processo de Pesquisa

Na imagem abaixo, podemos observar a página de pesquisa dos conjuntos de dados disponíveis no Portal Dados Abertos. Note que o item um lista uma série de pacotes de dados, este item diz respeito aos conjuntos prontos.

Já no campo pesquisar, é possível elaborar o próprio filtro de dados e elaborar o seu próprio conjunto.

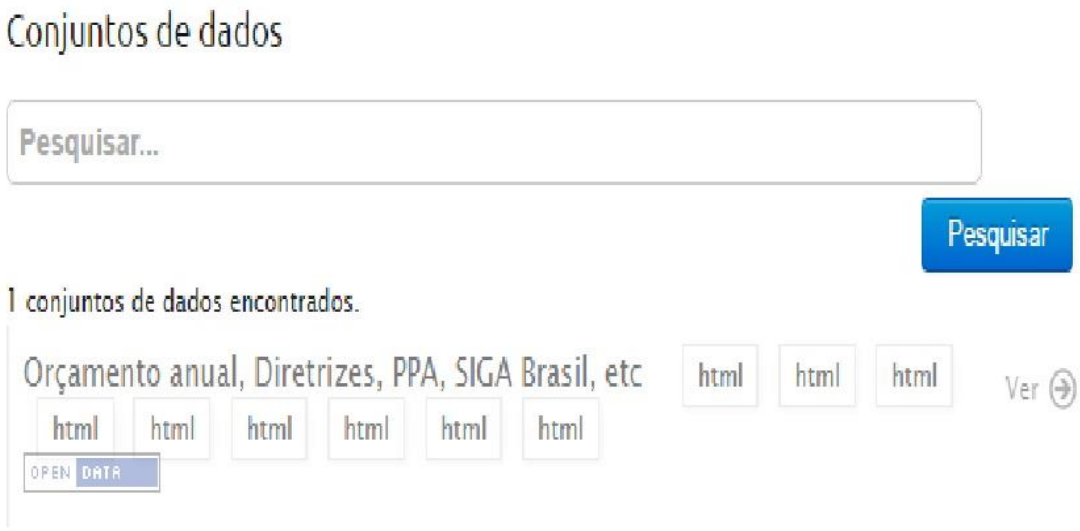


Figura 8 – FONTE: Pesquisa de Campo 2016

Informações referentes ao Orçamento Público são regulamentadas pela Lei Orçamentária Anual (LOA), que calcula as receitas que o governo federal tem a pretensão de recolher durante o ano, baseado nesse preceito fixa os gastos a serem realizados com tais recursos.

Os pacotes de dados prontos estão dispostos pela ferramenta em formatos de diagramas. A pesquisa pode ser realizada clicando no ano, escolhendo entre elaboração e execução e, após seguir estes passos, é possível navegar pelo diagrama. Segue o exemplo:

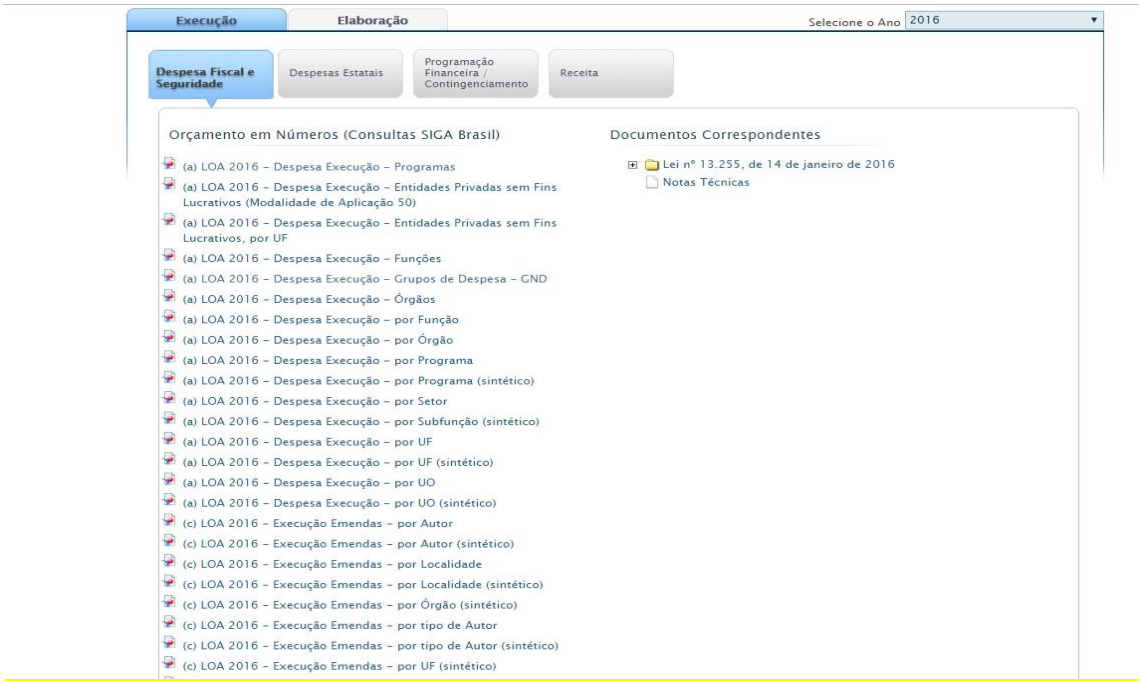


Figura 9 – FONTE: Pesquisa de Campo 2016

3.3 – Elaboração do processo de pesquisa

No Portal Dados Abertos, é possível encontrar informações sobre todo o processo orçamentário, elaboração e execução das leis orçamentárias, emendas parlamentares, transferências para estados, municípios e entidades privadas. Todo este processo é feito em paralelo com o portal Siga Brasil, que é responsável pelo fornecimento de informações do orçamento público que alimentam o Portal Dados Abertos.

As informações encontradas no Portal Dados Abertos, podem ser pacotes prontos ou elaborados, os pacotes elaborados estão dispostos para download ou visualização, no formato HTML.

Como uma das definições principais de Mineração de Dados é a transformação de Dados e Informações em Conhecimento, é exatamente está linha que será trabalhada neste projeto. Com a imensa quantidade de dados disponibilizados pelo sistema, faremos um filtro do que é ou não interessante. Por exemplo, no que diz respeito à saúde, encontramos todo e qualquer tipo de dado e informação; Valores Gastos com o Sistema Único de Saúde (SUS), Implementação de Melhorias, Tratamentos de Doenças como o Câncer e HIV. O que será feito é basicamente transformar estas informações em

conhecimento, informação útil, mostrar como o sistema interpreta comandos específicos e retorna valores.

Para o desenvolvimento deste estudo, serão tratados os dados do ano de 2016. Devido à grande quantidade de dados e sua imensa variação ao longo dos anos, caso não houvesse a exclusividade do ano de 2016, o trabalho ficaria extenso e dificilmente alcançaria um de seus principais objetivos que é demonstrar de forma clara e objetiva dados de interesse público. Tendo está conclusão como ponto de partida, foram realizadas análises com base na ferramenta e no interesse público. Os conjuntos de dados que serão abordados neste projeto, dizem respeito ao Orçamento Anual Público para os Projetos Sociais. Como existem muitos programas sociais e muitos deles inclusive nem possuem recursos, um filtro foi realizado para a escolha dos programas que irão compor este projeto, o critério escolhido foi o interesse da sociedade por determinados assuntos de interesse comum, como Segurança, Saúde, Educação e Inclusão Social.

Na próxima página é possível observarmos um quadro com informações sucintas ao que diz respeito aos Dados Minerados e Tratados neste trabalho. Mais informações poderão ser coletadas no decorrer do projeto.

	PROGRAMAS SOCIAIS
LOA 2016 – LEI ORCAMENTÁRIA ANUAL	APRIMORAMENTO DA EXECUÇÃO PENAL
	SEGURANÇA E EDUCAÇÃO DE TRÂNSITO: DIREITO E RESPONSABILIDADE DE TODOS
	BRASIL ESCOLARIZADO
	QUALIDADE DOS SERVIÇOS DE TELECOMUNICAÇÕES
	CONTROLE INTERNO, PREVENÇÃO E COMBATE A CORRUPÇÃO

	ATENÇÃO BÁSICA EM SAÚDE
	QUALIDADE NA ESCOLA
	QUALIDADE DOS SERVIÇOS DE TRANSPORTE
	INCLUSÃO SOCIAL POR MEIO DO BOLSA FAMÍLIA, DO CADASTRO ÚNICO E DA ARTICULAÇÃO DE POLÍTICAS SOCIAIS

Figura 10 – FONTE: Pesquisa de Campo 2016

3.3.1 – Escolher o Processo de Pesquisa

O Portal de Dados Abertos assim como o Portal da Transparência, possuem uma infinidade de dados referentes ao Orçamento Público, à definição do que seria abordado se deu devido às necessidades das partes interessadas.

As análises serão feitas a partir de Programas Sociais criados pela LOA (Lei Orçamentária Anual).

Todas as informações colhidas serão cruzadas com os propósitos dos Programas Sociais. O objetivo é compreender os critérios usados pelo Governo Federal para a disponibilização de recursos e se as verbas distribuídas pelo Órgão são suficientes para alcançar os interesses das partes interessadas.

Com a escolha dos Programas Sociais, foram criados Conjuntos de Dados com campos e dados necessários e como eles serão pré-processados para uma análise devida.

O Primeiro Conjunto de Dados é chamado de Execução Orçamentária por Programa, ele será composto por itens de avaliação necessárias:

- Programa – O nome, código e uma breve explicação do programa escolhido;
- Dotação Inicial – Refere-se ao valor inicialmente disponibilizado para o programa;
- Autorizado – Após a dotação inicial, o valor deve ser autorizado pela Comissão do Senado Federal;
- Empenhado – Valor que estão à disposição do Programa;

- Liquidado – Valor utilizado pelo Programa;
- Pago – É a soma de todos os valores utilizados pelo Programa;
- RP Pago – Refere-se ao método de pagamento utilizado.

LOA 2016 - Execução Orçamentáriapor Programa

R$ 1,00

Programa (Cod/Desc)	Dotação Inicial	Autorizado	Empenhado	Liquidado	Pago	RP Pago
0004 - QUALIDADE E EFICIÊNCIA DO SUS	0	0	0	0	0	0
0016 - GESTÃO DA POLÍTICA DE SAÚDE	0	0	0	0	0	282.758
0225 - GESTÃO DA POLÍTICA DOS TRANSPORTES	0	0	0	0	0	7.956
0660 - SEGURANÇA E EDUCAÇÃO DE TRÂNSITO: DIREITO E RESPONSABILIDADE DE TODOS	0	0	0	0	0	0
0661 - APRIMORAMENTO DA EXECUÇÃO PENAL	0	0	0	0	0	53.262
0662 - PREVENÇÃO E REPRESSÃO À CRIMINALIDADE	0	0	0	0	0	0
0665 - GESTÃO DA POLÍTICA NACIONAL SOBRE DROGAS	0	0	0	0	0	459.416
1061 - BRASIL ESCOLARIZADO	0	0	0	0	0	0
1157 - QUALIDADE DOS SERVIÇOS DE TELECOMUNICAÇÕES	0	0	0	0	0	0
1173 - CONTROLE INTERNO, PREVENÇÃO E COMBATE À CORRUPÇÃO	0	0	0	0	0	0
1184 - SEGURANÇA E SAÚDE NO TRABALHO	0	0	0	0	0	193.600.809
1214 - ATENÇÃO BÁSICA EM SAÚDE	0	0	0	0	0	1.407.806
1448 - QUALIDADE NA ESCOLA	0	0	0	0	0	31.613.095
1463 - QUALIDADE DOS SERVIÇOS DE TRANSPORTE	0	0	0	0	0	0
2015 - APERFEIÇOAMENTO DO SISTEMA ÚNICO DE SAÚDE (SUS)	9.122.421.859	9.074.963.667	7.153.904.100	5.418.404.152	5.196.251.178	0
2015 - FORTALECIMENTO DO SISTEMA ÚNICO DE SAÚDE (SUS)	86.364.379.731	86.635.562.091	74.502.071.594	55.747.634.527	55.106.422.234	6.684.810
2019 - BOLSA FAMÍLIA	13.348.037.830	13.283.682.440	13.272.817.441	13.013.874.113	13.011.373.072	293.889
2019 - INCLUSÃO SOCIAL POR MEIO DO BOLSA FAMÍLIA, DO CADASTRO ÚNICO E DA ARTICULAÇÃO DE POLÍTICAS SOCIAIS	15.338.403.538	15.638.403.538	15.338.403.538	7.584.664.718	6.835.903.538	306.033

Figura 11 - FONTE: Pesquisa de Campo 2016

O Segundo Conjunto de Dados é composto pelo código de identificação do programa e por uma breve descrição de onde os gastos foram aplicados de acordo com cada programa.

O Terceiro Conjunto de Dados será composto pelo código de identificação do programa e pelo nome do Órgão responsável.

3.4 – Técnicas

Como citado anteriormente, são muitas as técnicas de *Data Mining,* para este projeto as principais foram cuidadosamente escolhidas para um melhor entendimento e utilização da ferramenta escolhida CKAN. Para demonstração das técnicas serão utilizados painéis extraídos da fonte de dados.

3.4.1 – Descoberta de Associações

Conforme referenciado anteriormente, o conceito de descobertas de associações está relacionado a padrões de relacionamentos entre itens localizados na mesma base de dados.

Neste projeto de pesquisa, a as regras de associação podem ser aplicadas na análise de relacionamentos entre programas. Quando o SuS é referenciado, por exemplo, é feita uma referência a vários programas que estão relacionados à área de Saúde, com o uso de “tags” é possível retornar valores e elaborar conjuntos de dados com todos os programas relacionados ao assunto, vejam o exemplo:

LOA 2016 - Execução Orçamentáriapor Programa

R$ 1,00

Programa (Cod/Desc)	Dotação Inicial	Autorizado	Empenhado	Liquidado	Pago	RP Pago
0004 - QUALIDADE E EFICIÊNCIA DO SUS	0	0	0	0	0	0
0016 - GESTÃO DA POLÍTICA DE SAÚDE	0	0	0	0	0	282.758
1214 - ATENÇÃO BÁSICA EM SAÚDE	0	0	0	0	0	1.407.806
2015 - APERFEIÇOAMENTO DO SISTEMA ÚNICO DE SAÚDE (SUS)	9.122.421.859	9.074.963.667	7.153.904.100	5.418.404.152	5.196.251.178	0
2015 - FORTALECIMENTO DO SISTEMA ÚNICO DE SAÚDE (SUS)	86.364.379.731	86.635.562.091	74.502.071.594	55.747.634.527	55.106.422.234	6.684.810

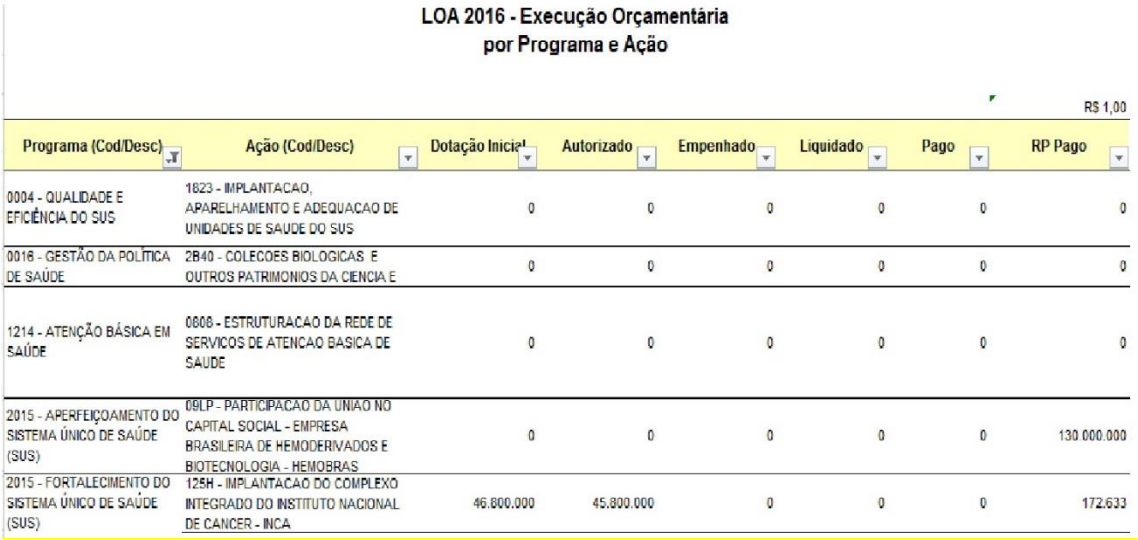
Figura 12 - FONTE: Pesquisa de Campo 2016

Na imagem a cima, é possível visualizar todos os programas do conjunto de dados criado, que estão relacionados ao SUS. Apenas com a aplicação de regras de associação.

3.4.2 – Classificação

A Classificação associa objetos a determinadas classes, ela procura prever uma classe de um novo dado automaticamente.

Continuando com o exemplo do SUS, em um novo conjunto de dados é possível observar várias informações classificadas de acordo com a previsão da técnica de classificação.



LOA 2016 - Execução Orçamentária
por Programa e Ação

R$ 1,00

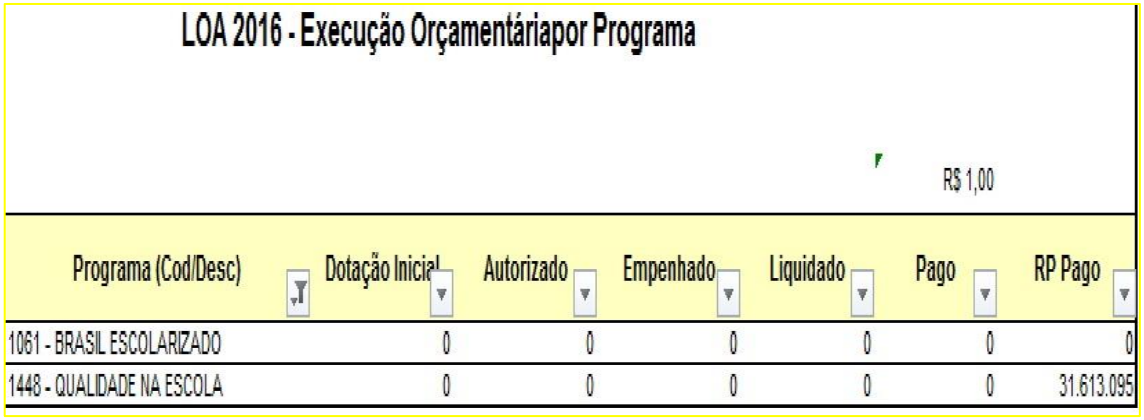
Programa (Cod/Desc)	Ação (Cod/Desc)	Dotação Inicial	Autorizado	Empenhado	Liquidado	Pago	RP Pago
0004 - QUALIDADE E EFICIÊNCIA DO SUS	1823 - IMPLANTACAO, APARELHAMENTO E ADEQUACAO DE UNIDADES DE SAUDE DO SUS	0	0	0	0	0	0
0016 - GESTÃO DA POLÍTICA DE SAÚDE	2B40 - COLECOES BIOLOGICAS E OUTROS PATRIMONIOS DA CIENCIA E	0	0	0	0	0	0
1214 - ATENÇÃO BÁSICA EM SAÚDE	0808 - ESTRUTURACAO DA REDE DE SERVICOS DE ATENCAO BASICA DE SAUDE	0	0	0	0	0	0
2015 - APERFEIÇOAMENTO DO SISTEMA ÚNICO DE SAÚDE (SUS)	09LP - PARTICIPACAO DA UNIAO NO CAPITAL SOCIAL - EMPRESA BRASILEIRA DE HEMODERIVADOS E BIOTECNOLOGIA - HEMOBRAS	0	0	0	0	0	130.000.000
2015 - FORTALECIMENTO DO SISTEMA ÚNICO DE SAÚDE (SUS)	125H - IMPLANTACAO DO COMPLEXO INTEGRADO DO INSTITUTO NACIONAL DE CANCER - INCA	46.800.000	45.800.000	0	0	0	172.633

Figura 13 – FONTE: Pesquisa de Campo 2016

3.4.3 – Agrupamento

Com o objetivo de dividir os registros da fonte de dados em subconjuntos, os itens que compões os subconjuntos compartilham propriedades simples que os diferenciam dos demais itens do grupo.

Para a demonstração desta técnica, será usada a análise de dados referentes à educação:



LOA 2016 - Execução Orçamentáriapor Programa

R$ 1,00

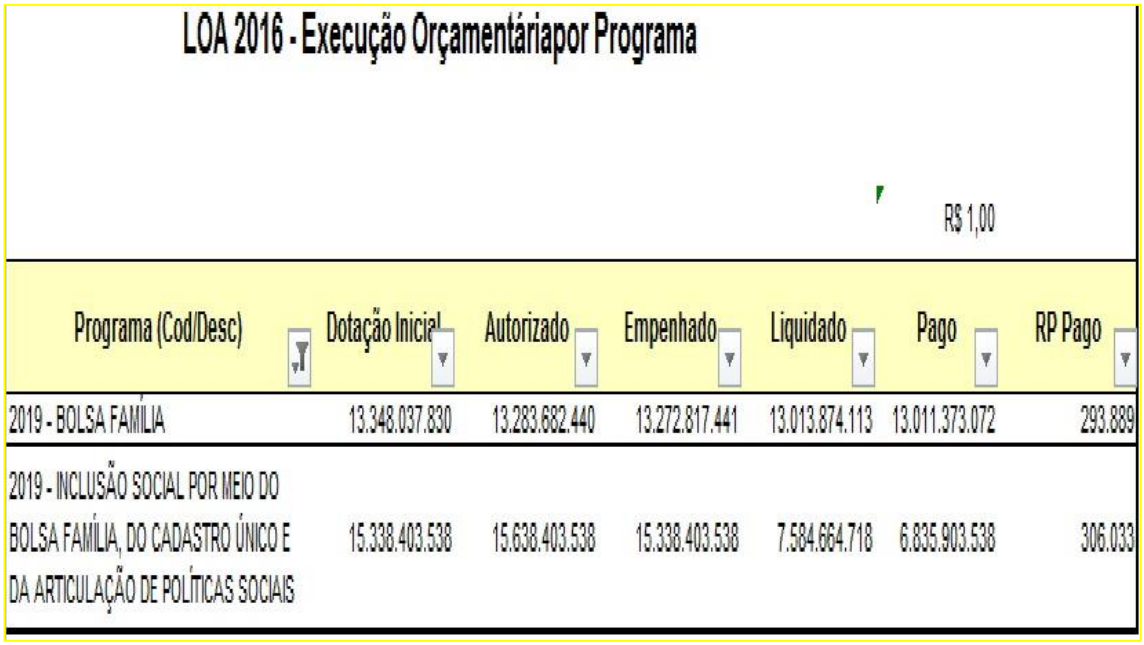
Programa (Cod/Desc)	Dotação Inicial	Autorizado	Empenhado	Liquidado	Pago	RP Pago
1061 - BRASIL ESCOLARIZADO	0	0	0	0	0	0
1448 - QUALIDADE NA ESCOLA	0	0	0	0	0	31.613.095

Figura 14 – FONTE: Pesquisa de Campo 2016

Note que ao aplicar a técnica de agrupamento, o sistema retorna informações que diferenciam um registro do outro.

3.4.4 – Sumarização

Tem o objetivo de apresentar as semelhanças entre os registros da fonte de dados. Para demonstração, serão usados os dados referentes à Inclusão Social:



LOA 2016 - Execução Orçamentáriapor Programa

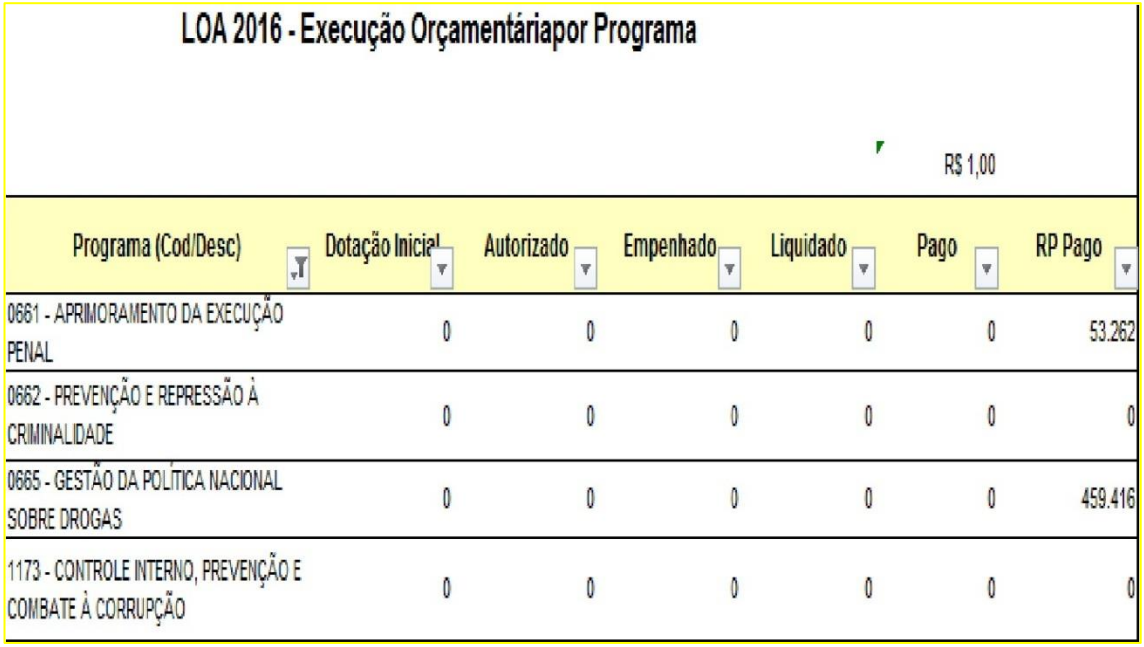
R$ 1,00

Programa (Cod/Desc)	Dotação Inicial	Autorizado	Empenhado	Liquidado	Pago	RP Pago
2019 - BOLSA FAMÍLIA	13.348.037.830	13.283.682.440	13.272.817.441	13.013.874.113	13.011.373.072	293.889
2019 - INCLUSÃO SOCIAL POR MEIO DO BOLSA FAMÍLIA, DO CADASTRO ÚNICO E DA ARTICULAÇÃO DE POLÍTICAS SOCIAIS	15.338.403.538	15.638.403.538	15.338.403.538	7.584.664.718	6.835.903.538	306.033

Figura 15 – FONTE: Pesquisa de Campo 2016

3.4.5 – Descoberta de Sequências

O propósito é demonstrar registros comuns levando em conta um espaço de tempo previamente definido. Para a demonstração da técnica, será levado em conta o Orçamento direcionado aos programas das ações penais definidos e apresentados anteriormente, o espaço de tempo definido nesta análise foram os investimentos do governo entre um programa e outro:



LOA 2016 - Execução Orçamentáriapor Programa

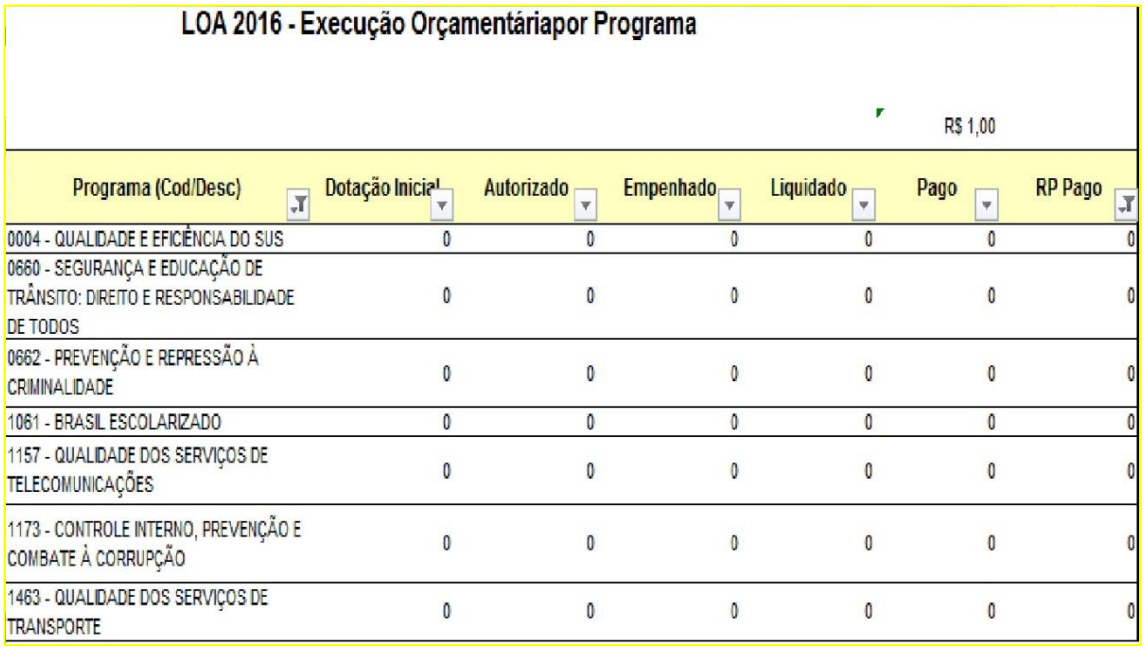
R$ 1,00

Programa (Cod/Desc)	Dotação Inicial	Autorizado	Empenhado	Liquidado	Pago	RP Pago
0661 - APRIMORAMENTO DA EXECUÇÃO PENAL	0	0	0	0	0	53.262
0662 - PREVENÇÃO E REPRESSÃO À CRIMINALIDADE	0	0	0	0	0	0
0665 - GESTÃO DA POLÍTICA NACIONAL SOBRE DROGAS	0	0	0	0	0	459.416
1173 - CONTROLE INTERNO, PREVENÇÃO E COMBATE À CORRUPÇÃO	0	0	0	0	0	0

Figura 16 – FONTE: Pesquisa de Campo 2016

3.4.6 – MBR – Memory Based Reasoning

É uma técnica direcionada a padrões definidos pelas partes interessadas, combina pares de dados para prever resultados futuros. Como demonstração da técnica, serão utilizados programas sociais que não possuem nenhuma aplicação de verbas públicas e que possivelmente não existe valor reservado para este fim:



LOA 2016 - Execução Orçamentáriapor Programa

R$ 1,00

Programa (Cod/Desc)	Dotação Inicial	Autorizado	Empenhado	Liquidado	Pago	RP Pago
0004 - QUALIDADE E EFICIÊNCIA DO SUS	0	0	0	0	0	0
0660 - SEGURANÇA E EDUCAÇÃO DE TRÂNSITO: DIREITO E RESPONSABILIDADE DE TODOS	0	0	0	0	0	0
0662 - PREVENÇÃO E REPRESSÃO À CRIMINALIDADE	0	0	0	0	0	0
1061 - BRASIL ESCOLARIZADO	0	0	0	0	0	0
1157 - QUALIDADE DOS SERVIÇOS DE TELECOMUNICAÇÕES	0	0	0	0	0	0
1173 - CONTROLE INTERNO, PREVENÇÃO E COMBATE À CORRUPÇÃO	0	0	0	0	0	0
1463 - QUALIDADE DOS SERVIÇOS DE TRANSPORTE	0	0	0	0	0	0

Figura 17 – FONTE: Pesquisa de Campo 2016

3.5 – Aplicando o CKAN

Como o CKAN é uma ferramenta livre, ela também é disponibilizada por sua API (*Application Programming Interface*), que fornece um acesso simplista para o sistema CKAN. A API é muito influente e possui uma interface prática, que permite ao desenvolvedor ou ao simples usuário, fazer tudo que é possível através da interface web. Por exemplo:

- Acessar todo e qualquer bit de informação no CKAN;
- Se previamente autorizado, permite alterar qualquer informação no CKAN;
- Caso seja da vontade do desenvolvedor ou do usuário, permite à criação de um *front-end web* inteiro para o CKAN; e
- A API CKAN segue o estilo RESTful (Transferência de Estado Representativo) e usa JSON por padrão.

Sintaxe do CKAN aplicado aos dados abertos do Senado Federal:

```
/
*
  * Included as inline javascript in layout_base.html. Simpler than
  * trying to trick the translation system into reading a js file.
  */
  var CKAN = CKAN || {};
  CKAN.Strings = CKAN.Strings || {};
  /*
  * Used in application.js.
  */
  CKAN.Strings.checking = "Verificando...";
  CKAN.Strings.urlIsTooShort = "Digite pelo menos dois caracteres...";
  CKAN.Strings.urlIsAvailable = "Esta URL está disponível!";
  CKAN.Strings.urlIsNotAvailable = "Esta URL já é utilizada, por favor use uma
  diferente.";
  CKAN.Strings.bracketsNone = "(Nenhum)";
  CKAN.Strings.failedToSave = "Falha ao salvar, possivelmente devido a
  dados inválidos";
  CKAN.Strings.addDataset = "Adicionar Conjunto de Dados";
  CKAN.Strings.addGroup = "Adicionar Grupo";
  CKAN.Strings.youHaveUnsavedChanges = "Você tem alterações não salvas.
  Certifique-se de ter clicado "Salvar Alterações" abaixo antes de sair desta
  página.";
  CKAN.Strings.loading = "Carregando ...";
  CKAN.Strings.noNameBrackets = "(nenhum nome)";
  CKAN.Strings.deleteThisResourceQuestion = "Excluir o recurso '%name%'?"
```

```
/*
* Used in templates.js.
*/
CKAN.Strings.fileUrl = "URL do arquivo";
CKAN.Strings.apiUrl = "URL da API";
CKAN.Strings.add = "Adicionar";
CKAN.Strings.cancel = "Cancelar";
CKAN.Strings.file = "Arquivo";
CKAN.Strings.name = "Nome";
CKAN.Strings.description = "Descrição";
CKAN.Strings.url = "Url";
CKAN.Strings.format = "Formato";
CKAN.Strings.resourceType = "Tipo de Recurso";
CKAN.Strings.sizeBracketsBytes = "Tamanho (Bytes)";
CKAN.Strings.mimetype = "Mimetype";
CKAN.Strings.lastModified = "Modificada pela última vez";
CKAN.Strings.mimetypeInner = "Mimetype (Interno)";
CKAN.Strings.hash = "Resumo criptográfico";
CKAN.Strings.id = "ID";
CKAN.Strings.doneEditing = "Pronto";
CKAN.Strings.resourceHasUnsavedChanges = "Este recurso tem alterações
não salvas.";
</script>
<!-- finally our application js that sets everything up-->
<script type="text/javascript"
src="/scripts/application.js?lang=${c.locale}"></script>
<script type="text/javascript" src="/scripts/templates.js"></script>
<script src="/scripts/vendor/modernizr/1.7/modernizr.min.js"></script>
<script type="text/javascript">
CKAN.plugins = [
// Declare js array from Python string
",
];
CKAN.plugins.push('storage');
CKAN.SITE_URL = '/';
// later use will add offsets with leading '/' so ensure no trailing slash
CKAN.SITE_URL = CKAN.SITE_URL.replace(/\/$/, '');
$(document).ready(function() {
var ckan_user = '';
if (ckan_user) {
$(".ckan-logged-out").hide();
$(".ckan-logged-in").show();
```

```
}
$('input[placeholder], textarea[placeholder]').placeholder();
});
</script>
```

3.5.1 – Objetivos

- Disponibilizar e facilitar a pesquisa de dados, por meio da utilização de filtros ou tags;
- Conservar dados brutos e também metadados;
- Permitir a visualização de dados em formatos estruturados como planilhas XML, gráficos, textos, páginas da web e mapas;
- Ter uma ligação direta com os gerenciadores de conteúdo como Drupal e Joomla;
- Possibilita a criação de um sistema de *harvester* (aplicativo de interoperabilidade através de um processo de coleta de informações) para interoperabilidade com outros portais de dados abertos;
- Como se trata de uma ferramenta *open source permite* o uso e também a customização.

3.5.2 – Estrutura

A ferramenta CKAN utiliza uma estrutura particular que é composta pelos seguintes elementos:

- Título;
- Identificador próprio (também conhecido como URL);
- Descrição;
- Histórico de Revisão;
- Visualização de Dados (aglomerado de informações obrigatoriamente em formato CSV);
- Campos Extras (tornam possível a inclusão de qualquer informação adicional);
- Licença (informações referentes ao tipo de licença em que os dados estão publicados);
- Tags (conjunto de rótulos para os dados publicados);

- Grupos (categorias dos dados);
- Múltiplos Formatos.

3.5.3 – Ferramenta de Pesquisa

- Possui uma interface de pesquisa Google-Style;
- A pesquisa pode ser realizada em todos os atributos do conjunto de dados;
- Fornece uma pesquisa de texto completa nos campos;
- Aplica a lógica *fuzzy* em modelos de pesquisa, que significa procurar por termos que combinam em vez de correspondências exatas.

3.5.4 – Utilização

Como a grande maioria das ferramentas, o CKAN possui pré-requisitos para que possa ser corretamente utilizado, a título de desenvolvimento, podemos citar os principais requisitos:

- Deve ser instalado em plataformas Linux, Debian ou Ubuntu;
- Ferramentas:
  - Linguagem de Programação Python;
  - Banco de Dados PostgreSQL;
  - Apache Solr – Plataforma de Pesquisa;
  - OpenJDK 6 JDK – Kit de Desenvolvedor Java.
- Opção de imagem disponível através de uma Máquina Virtual para Oracle.

3.5.5 – Workflow

Os Conjuntos de dados disponibilizados pelo CKAN podem ser públicos ou privados. Se eles são privados são visíveis apenas aos membros autorizados. Os administradores de dados são capazes de aprovar conjuntos de dados elaborados pelo público em geral, através da ferramenta de edição em massa que permite pesquisar e escolher conjuntos de dados para tornar lós públicos ou privado.

3.5.6 – Estruturando Dados

3.5.6.1 – Utilizando o JSON

O formato JSON é muito prático. É estruturado de forma simples e facilita o entendimento, ele utiliza convenções comuns em muitas linguagens de programação.

A sintaxe padrão do JSON é ainda mais simples, observe o exemplo:

use JSON;
my $json = to_json($saude_sus);
Agora, é possível "renderizar" a variável $json, lembrando-se de definir o Content-Type da página para "application/json".
O resultado será:
{"Melhorias":6,"Qualidade":62,"Gestão":72,"Atenção Básica":30}

3.5.7 – XML

O formato XML foi um dos formatos padronizados mais populares para serialização de dados via rede. Atualmente, muitos desenvolvedores evitam seu uso em *web services REST*, optando por alternativas mais simples, como JSON ou YAML. Mas o XML ainda é uma boa escolha, especialmente quando se quer validações mais robustas, usando os chamados XML-schemas.

Se o arquivo contendo o schema do XML estiver em "dados.xsd",
Pode-se gerar um XML da seguinte forma:

use XML::Compile::Schema;
use XML::LibXML;
my $doc = XML::LibXML->createDocument('1.0', 'UTF-8');
my $schema = XML::Compile::Schema->new('dados.xsd');
my $xml = $schema->compile(WRITER => {http://dadosabertos.senado.leg.br/api}
dados')
->($programas_sociais, $hash)
->toString;

Há pouca vantagem em usar XML sem validação, mas isso também pode ser obtido com relativa facilidade. Não se pode esquecer de definir o Content-Type da página gerada para “text/xml”.

4 – APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Para que se chegue à usabilidade apresentada aos usuários finais, a aplicação das técnicas de Mineração de Dados mostrou-se bastante complexa devido à exigência do prévio conhecimento e estudo de vários conceitos.

A utilização de tais técnicas demonstrou ser de grande importância quando aplicadas ao Portal de Dados Abertos, pois é capaz de proporcionar profundos conhecimentos de assuntos de interesse comum. É possível extrair informações de onde e como os impostos são aplicados e se são aplicados como se deve, bem como os programas sociais para os quais as verbas são destinadas.

Com a correta aplicação das técnicas, é possível identificar padrões nas fontes de dados.

Analisando as verbas destinadas à saúde pública, é possível notar que apesar do grande investimento, o Sistema Único de Saúde ainda se mostra ineficaz no atendimento às necessidades básicas do cidadão comum.

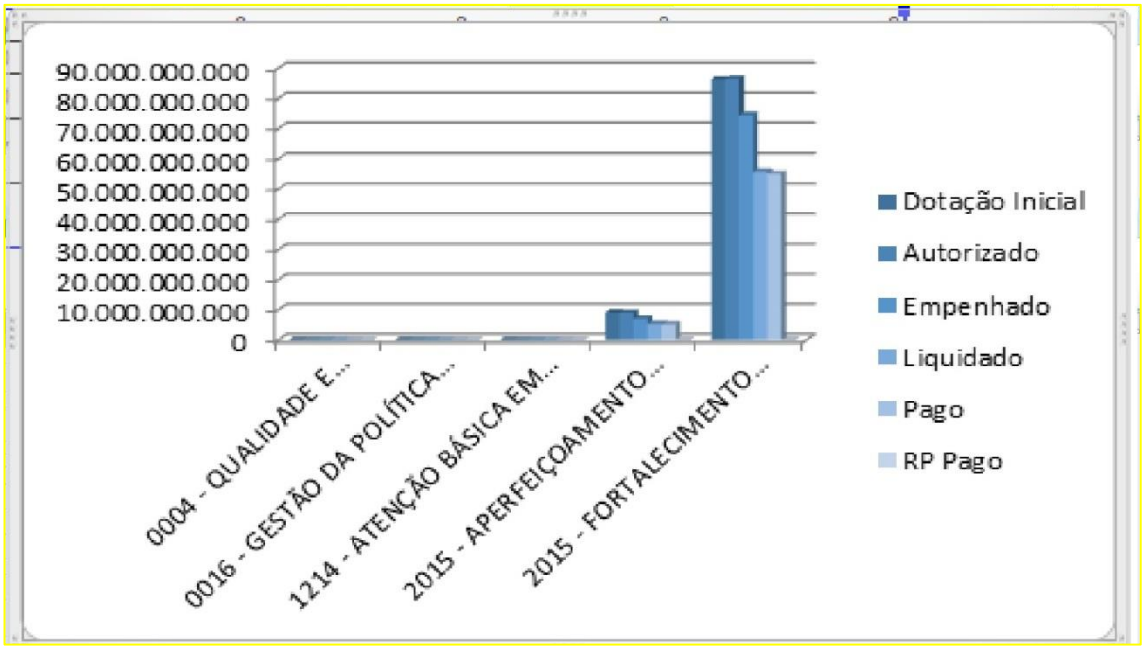


Figura 18 – FONTE: Pesquisa de Campo 2016

No gráfico acima, é possível observar os valores investidos no Sistema único de Saúde no ano de 2016.

Dois programas sociais referentes à educação foram selecionados para a mineração e análise de dados, nenhum valor foi retornado para ambos os programas.

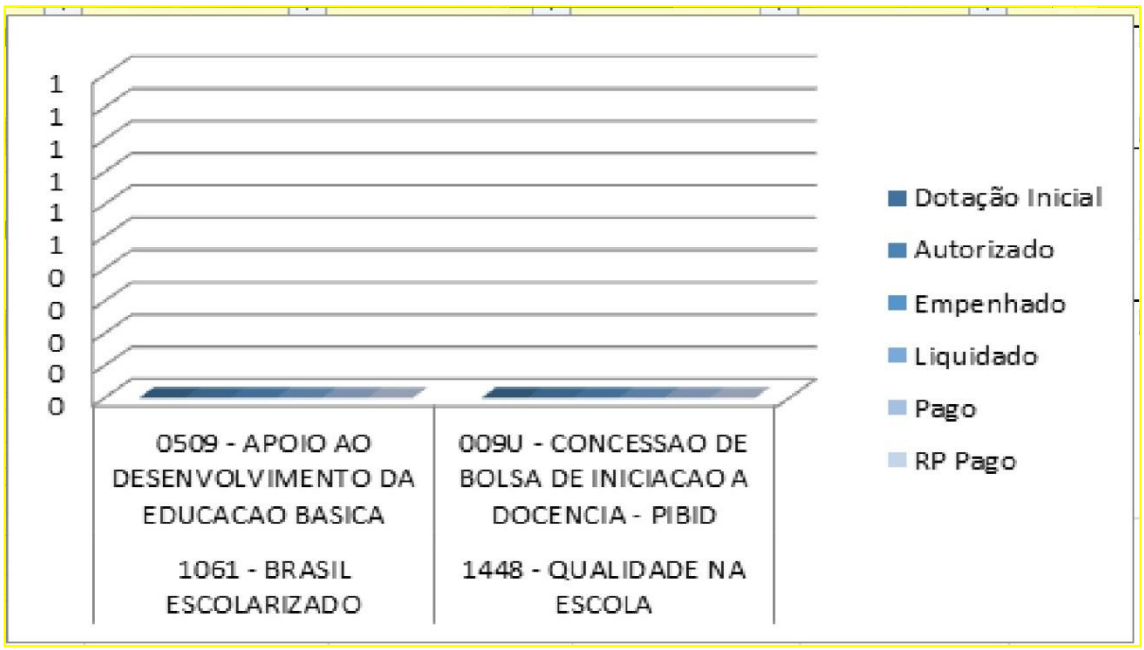


Figura 19 – FONTE: Pesquisa de Campo 2016

É possível notar que apesar de existirem propósitos para melhorias na Educação, nada foi orçado no ano de 2016.

O mesmo pode ser observado em Programas Sociais voltados para ações penais, apesar da grande precariedade que o Sistema Penitenciário vem demonstrado ter, nenhum valor foi reservado para o abastecimento de Políticas Voltadas para a melhoria do mesmo.

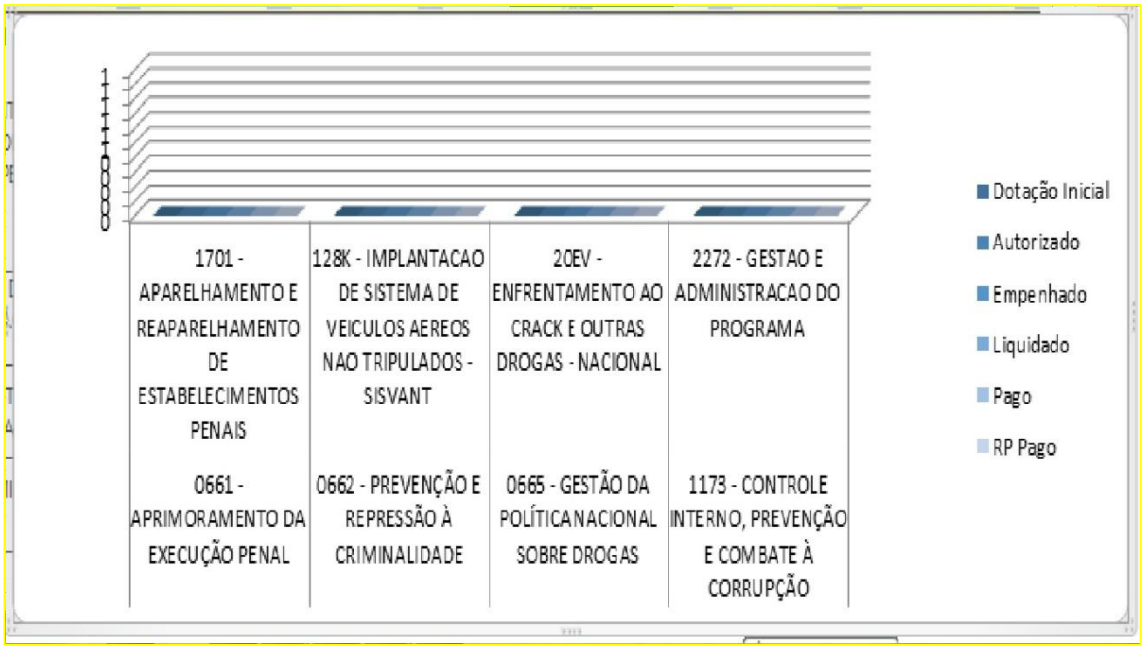


Figura 20 – FONTE: Pesquisa de Campo 2016

Em contra partida, pode ser observado que os investimentos do governo com a Inclusão Social com base no Programa Bolsa Família são bem altos.

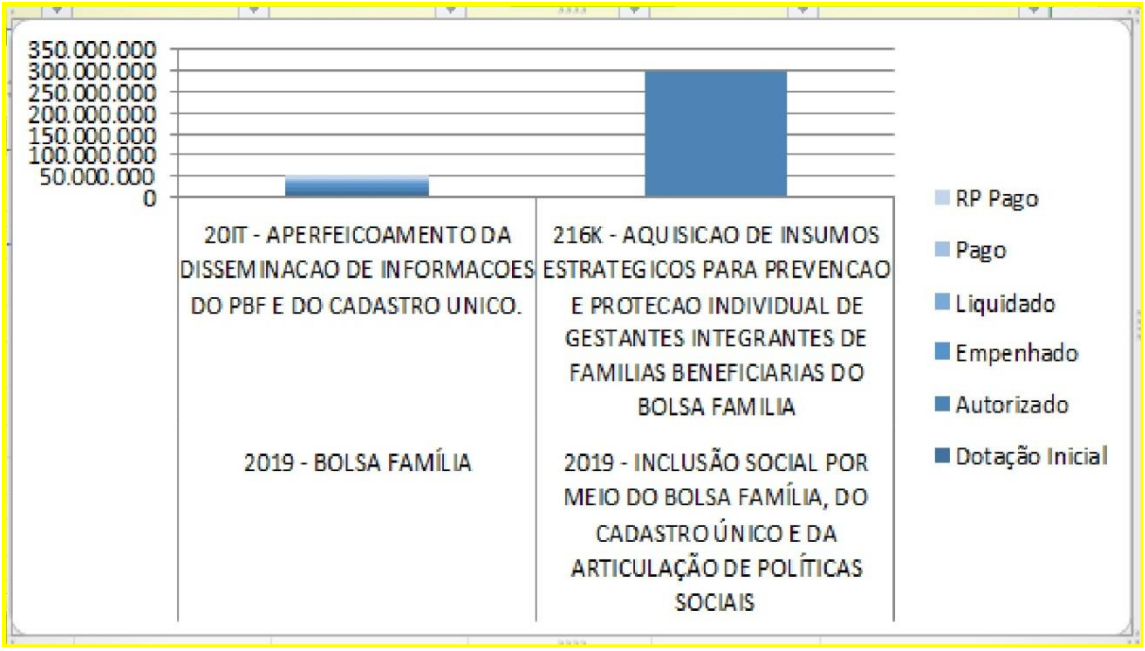


Figura 21 – FONTE: Pesquisa de Campo 2016

É possível observar que o orçamento público não é corretamente aplicado às necessidades básicas do cidadão comum. De forma geral, será apresentado um gráfico demonstrativo dos investimentos do governo em Programas Sociais essencias.

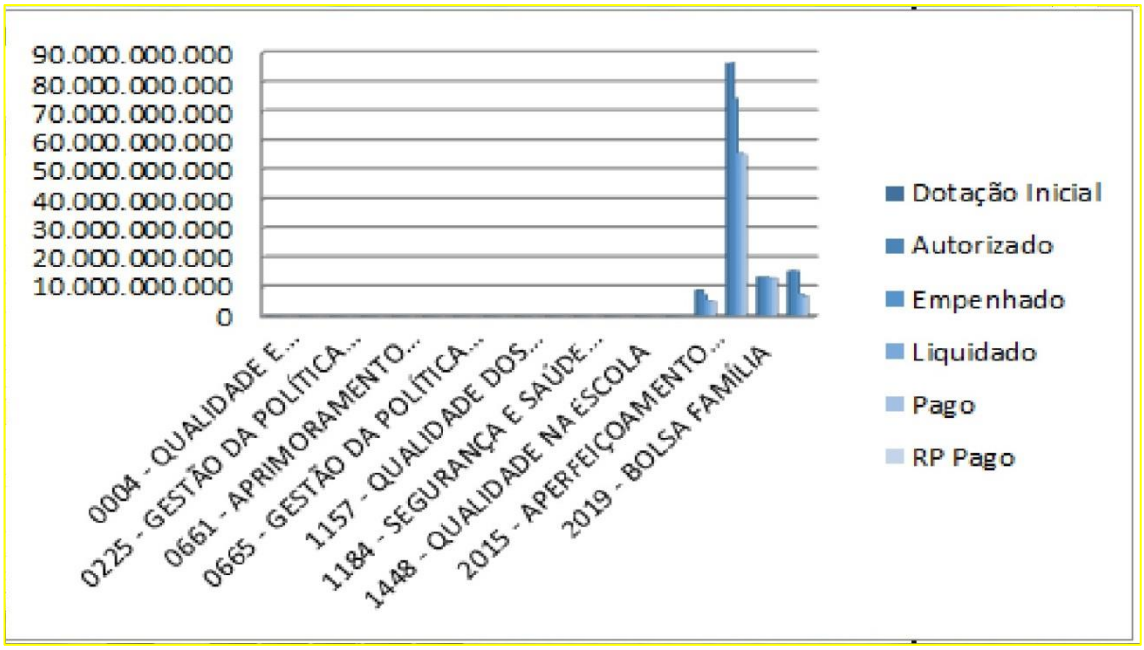


Figura 22 – FONTE: Pesquisa de Campo 2016

Com base nos dados apresentados, é possível observar que a ferramenta CKAN em paralelo a utilização de outros aplicativos, mostrou ser uma poderosa Base de Dados capaz de armazenar e disponibilizar de forma descomplicada, grandes massas de informações úteis.

Além dos conjuntos de dados criados para compor este projeto, é possível encontrar diversas informações referentes aos Gastos Públicos; Pautas votadas pelo quadro de Senadores do Órgão em questão; Funções e Atribuições do quadro de funcionários do Senado Federal; Documentos Relacionados ao Poder Legislativo; Normas Jurídicas Federais desde o Império até os dias de hoje; entre outros assuntos.

Com um cenário bem estruturado, as partes interessadas podem acrescentar ou encontrar conjuntos de dados sem grandes dificuldades.

O CKAN foi implantado pela maioria dos órgãos governamentais no Brasil e no mundo, de forma a facilitar o acesso a informações de interesse público.

5 – CONCLUSÃO

Data Mining ou Mineração de Dados é o ato de selecionar dados, dos quais podem ser extraídas informações relevantes, para que posteriormente possam sem transformados em conhecimento útil.

A composição do projeto manteve como principal objetivo o estudo e aplicação das técnicas de *Data Mining* nos dados obtidos através do Portal da Transparência do Senado Federal.

Foi evidenciada a importância dos processos de KDD em especial sua principal etapa, e núcleo desse processo, denominada *Data Mining* ou Mineração de Dados para a análise dos dados de forma significativa em nível de detalhes através das técnicas de Classificação, *Memory Based Reasoning*, previsão de séries temporais, descoberta de sequências, detecção de desvios, sumarização, agrupamento, regressão e descoberta de associações, que foram utilizadas especificamente neste projeto.

Inferiu-se do Estudo de Caso realizado no capítulo 3 e na descrição do capítulo 4, que os resultados foram considerados satisfatórios e o objetivo alcançado. As técnicas utilizadas de *Data Mining* e Mineração de Dados por muito vistas na área de mercado financeiro e publicidade, demonstrara-se amplamente eficaz na análise detalhista dos dados do Portal da Transparência do Senado Federal, o que trouxe em sua essência registros de gastos efetuados pelos senadores durante o exercício do seu mandato no período de 2016.

A utilização de *Data Mining* é justificada pelo fato de que os analistas de mineração de dados podem usar as informações obtidas e com um grau elevado de acerto tornar orçamentos do governo federal públicos, de forma que essa análise sirva para tomada de decisões no que refere ao aprumo financeiro, trazendo maior entendimento das informações para avaliação dos gastos efetuados e que serão efetuados.

Por fim, entende-se que a análise dos gastos públicos por parte dos senadores utilizando as técnicas descritas acima, trouxe maiores detalhes dos valores gastos favorecendo uma nova forma de pensar sobre as verbas publicas.

A informação produzida na conjuntura orçamentária pode servir de base para estudos que venham ajudar a promover uma política de controle,

equilíbrio e economia nos gastos dos senadores. No futuro propõe-se desenvolver aplicações que disponibilize as informações apresentadas, tanto para o(s) órgão(s) avaliado(s), quanto de forma aberta para o publico, visando maior controle e transparência dos gastos realizados pelos parlamentares. A exposição das informações de forma lúcida permite o acesso facilitado e o entendimento claro do total de gastos efetuados por toda a sociedade, do qual, poderão ser elaboradas políticas de melhor uso da verba pública.

REFERÊNCIAS

ABERNETHY Michael. Mineração de Dados com WEKA, Parte 1: Introdução e Regressão. Disponível em: <http://www.ibm.com/developerworks/br/opensource/library/os-weka1/>. Acessado em: 21 out. 2016.

AMARAL, Fernando. Fundamentos de Programação em R. 1° Ed. Amazon; São Paulo; 2001.

AMARAL, Fernando. *Data Mining*: Técnicas e Aplicações para o Marketing Direto. 1° Ed. Berkeley; São Paulo; 2001.

BRITO, Daniel. Aspectos Teóricos da Mineração de Dados e Aplicação das Regras de Classificação para Apoiar o Comércio. Disponível em: <http://www.devmedia.com.br/aspectos-teoricos-da-mineracao-de-dados-e-aplicacao-das-regras-de-classificacao-para-apoiar-o-comercio/25429>. Acessado em: 29 set. 2016.

CASTRO, Leandro; FERRARI, Daniel. Introdução a Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações. 1° Ed. Saraiva, 2016.

CKAN. *Data Explorer Examples.* Disponível em: <http://demo.ckan.org/pt_PT/>*.* Acessado em: 28 out. 2016.

CKAN. Criar um Conjunto de Dados. Disponível em: http://ckan01.myersmediagroup.com/pt_BR/dataset/edit/43recr2ft3f-scdft4fwrexewredcw. Acessado em: 13 set. 2016.

Comitê Gestor da Internet no Brasil. Manual dos Dados Abertos: Desenvolvedores. 1° Ed. Laboratório Brasileiro de Cultura Digital; São Paulo; 2011.

DADOS ABERTOS. Busca por conjuntos de Dados prontos ou elaboração de conjuntos. Disponível em: <http://dadosabertos.senado.leg.br/>. Acessado em: 31 ago. 2016.

DEVMEDIA. Mineração de Dados: Tarefas e Técnicas. Disponível em: <http://www.devmedia.com.br/mineracao-de-dados-tarefas-e-tecnicas/30919>. Acessado em: 06 set. 2016.

FARIA, Tuane. Explorando Técnicas e Recursos do Gerenciador de Dados Abertos CKAN. Disponível em: <http://docplayer.com.br/4765804-Explorando-tecnicas-e-recursos-do-gerenciador-de-dados-abertos-ckan-tuanefaria-usp-tuanefaria-yahoo-com-br.html>. Acessado em 19 out. 2016.

FIVE ACTS. O que é Mineração de Dados. Disponível em: <http://fiveacts.com.br/o-que-e-mineracao-de-dados/>. Acessado em: 21 out. 2016.

GALVÃO, Noemi; MARIN, Heimar. Técnica de Mineração de Dados: Uma Visão da Literatura. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-21002009000500014>. Acessado em: 17 set. 2016.

GITHUB. Scripts para fazer Operações Úteis no Portal Dados Abertos. Disponível em: <https://github.com/dadosgovbr/scripts-dadosgovbr>. Acessado em: 24 out. 2016.

GOLDSCHMIDT, Ronaldo; PASSOS, Emanuel. *Data Mining*: Um guia prático. 1° Ed. Campus, 2005.

GONÇALVES, Eduardo. Data Mining de Regras de Associação. Disponível em: http://www.devmedia.com.br/data-mining-de-regras-de-associacao-parte-1/6533. Acessado em: 28 set. 2016.

MICROSOFT. Conceitos de Mineração de Dados. Disponível em: <https://msdn.microsoft.com/pt-br/library/ms174949.aspx>. Acessado em: 23 ago. 2016.

PULCINELLI, Márcio. As Técnicas do *Data Mining.* Disponível em: <http://www.tiespecialistas.com.br/2013/08/as-tecnicas-do-datamining/>. Acessado em: 19 out. 2016.

SAS. Definição de Mineração de Dados. Disponível em: <http://www.sas.com/pt_br/insights/analytics/mineracao-de-dados.html>. Acessado em: 23 ago. 2016.

SIGNIFICADOS. Significado de Data Mining. Disponível em: <https://www.significados.com.br/data-mining/>. Acessado em: 11 nov. 2016.

SILVA, Leandro; PERES, Sarajane; BOSCARIOLI, CLODIS. Introdução a Mineração de Dados. 1° Ed. Elsevier, 2016.

SHIMIZU, Tamio. Processamento de Dados: Conceitos Básicos. 1° Ed. Atlas, 1988.

VIANA, Reinaldo. Introdução ao *Data Mining.* Disponível em: <http://www.devmedia.com.br/artigo-sql-magazine-10-introducao-ao-data-mining/7820>. Acessado em 06 set. 2016.

GLOSSARIO

Bancos de Dados Relacionais – Armazenamento de dados em tabelas.

Sistema Único de Saúde – É o sistema público de saúde no Brasil.

Informação Útil – Confiável, que auxilia na tomada de decisão.

Partes Interessadas – Pessoa ou grupo que tem influência direta no resultado do projeto.

Conjunto de Dados – É uma coletânea de dados organizados em tabelas, cada coluna da tabela representa uma variável particular.

Necessidades Básicas – Todas aquelas a que o ser humano deve ter acesso para sobreviver com o mínimo de dignidade.

Open Source – Diz respeito ao código fonte de um determinado programa de computador, o que o denomina *Open Source* é o fato de se tratar de um código aberto, livre e gratuito, podendo ser alterado a qualquer tempo.

Usabilidade – Facilidade de utilização de uma ferramenta ou objeto com a finalidade de executar uma tarefa específica.