Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - IFPB
Programa de Pós-Graduação em Tecnologia da Informação
Mestrado Profissional em Tecnologia da Informação

Relatório Técnico

Uma análise de possíveis anomalias em dados da administração para gastos públicos

Flavio Henrique de Sousa Silva, IFPB - Campus João Pessoa

Damires Souza, IFPB-PPGTI-Campus João Pessoa

João Pessoa, 16/03/2021

Resumo

Este relatório apresenta os resultados de um estudo realizado com dados abertos governamentais que são empenhados mensalmente, visando à identificação de anomalias na administração desses gastos. Para isso, foi utilizada a técnica de agrupamento com o intuito de identificar grupos baseados em algumas informações de cada gasto público. O objetivo geral foi a identificação de registros fora dos grupos gerados e o estudo mais aprofundado sobre estes itens.
Palavras-chave: Dados abertos. Agrupamento. Anomalias.

Abstract

This report presents results obtained from a study with open government data. These data are committed monthly. This works aims to identify possible anomalies in the management of those expenses. To this end, the data mining clustering technique was used in order to create groups based on similar information from each public expenditure. The general goal of this work was to identify outliers in the data and to further study these specific objects.
Keywords: Open data. Clustering. Outliers.

Sumário

1. Introdução

Com o crescimento tecnológico e a facilidade de utilização da internet por grande parte da população, o interesse em acompanhar as ações dos seus representantes políticos tem crescido. Uma das formas de realizar este acompanhamento é por meio da análise de dados abertos governamentais, disponibilizados em diversos portais (SILVA e GALVÃO, 2018). Porém, muitas vezes, a tentativa de analisar e compreender os conjuntos de dados publicados é frustrada devido aos formatos ou distribuições em que eles estão disponibilizados (e.g., CSV, JSON). Além disso, o volume dos conjuntos de dados torna praticamente impossível analisá-los a "olho nu", sem ferramentas que facilitem seu entendimento e visualização.

Nesse panorama, o desenvolvimento de modelos de aprendizado de máquina pode ajudar na análise dos dados, de modo a buscar padrões que não seriam visualizados diretamente no próprio conjunto de dados disponibilizado (Fayyad, Piatetsky-Shapiro e Smyth, 1996). O aprendizado de máquina pode auxiliar também a identificar possíveis anomalias nos dados (Alpaydin 2010).

Esse relatório utiliza métodos de aprendizado de máquina não supervisionados na busca pela identificação de anomalias no contexto de empenhos governamentais. Para isso foram usados métodos de agrupamento.

Neste relatório, é relatado o estudo feito na utilização de um *dataset* criado a partir de dados abertos governamentais disponibilizados no Portal da Transparência[1], relacionados aos empenhos realizados pelo governo federal. O objetivo foi o agrupamento dos empenhos, de forma a ser possível identificar possíveis anomalias, ou seja, registros não pertencentes a nenhum grupo e que se distanciam dos padrões encontrados (*outliers*).

2. Fundamentação teórica e trabalhos relacionados

Esta seção introduz o referencial teórico que foi utilizado para a pesquisa, descrevendo os principais conceitos necessários. Além disso, também são citados alguns trabalhos relacionados que também utilizam a mineração de dados para descoberta de padrões em determinados conjuntos de dados abertos.

2.1 Dados abertos governamentais

Segundo a *Open Knowledge Internacional*, dados abertos caracterizam-se por serem de livre acesso a qualquer pessoa, permitindo sua utilização, modificação e compartilhamento com a finalidade desejada, estando sujeitos apenas a vigências que preservem sua origem e abertura (Open Knowledge Internacional, 2018).

David Eaves, Professor da Harvard Kennedy School of Government, propôs três leis sobre dados abertos (Eaves, 2009). A primeira lei indica que se um dado não pode ser encontrado na Web, ele não existe. Na segunda lei, ele define que se o dado não estiver disponível em um formato que possa ser interpretado por máquina, ele não pode ser

[1] http://www.portaltransparencia.gov.br

reutilizado. Finalmente, na terceira lei, ele diz que se um dispositivo legal não permitir que este dado seja replicado, então ele não é considerado útil.

No Brasil, existem dois principais portais onde os dados abertos governamentais federais são disponibilizados, que são: O portal Brasileiro de Dados Abertos[2] e o Portal da Transparência[3].

2.2 Tarefa de Agrupamento

Kaufman e Rousseeuw (1990) definem a tarefa de agrupamento como a arte de encontrar grupos para os dados. Trata-se de uma tarefa de mineração de dados onde o objetivo é a execução de algoritmos para o agrupamento de um conjunto de dados (*dataset*) de forma que seja possível identificar e visualizar os grupos formados a partir de características comuns.

Com a formação dos grupos, abre-se a possibilidade também de identificação de anomalias (outliers). Hawkins (1980), citado por Santoyo (2017), afirma que anomalias são itens observados que se desviam muito de outras observações, gerando a dúvida se estes podem ter sido gerados a partir de outros mecanismos.

Alguns algoritmos utilizados na tarefa de agrupamento são os K-Means e o DBScan (Kaufman e Rousseeuw, 1990).

O K-means é um algoritmo de aprendizado não supervisionado, ou seja, é utilizado em conjuntos de dados sem nenhum tipo de rótulo, visando à extração de características similares entre os registros (Garbade, 2018).

O DBScan (*Density-based spatial clustering of applications with noise*) trata-se de um algoritmo comumente usado na mineração de dados, que funciona da seguinte forma: baseando-se em um conjunto de pontos, o algoritmo agrupa todos os pontos que estiverem próximos, geralmente medindo a distância euclidiana para realizar o agrupamento (PRADO, 2017).

2.3 Trabalhos relacionados

A seguir, alguns trabalhos relacionados ao tipo da pesquisa em questão são descritos.

2.3.1 Mineração de Dados Abertos sobre o ENEM

Embora não sejam dados abertos governamentais do mesmo tipo que os propostos neste trabalho, este trabalho apresenta um estudo de caso utilizando mineração de dados abertos relacionados ao Exame Nacional do Ensino Médio (ENEM) de 2011. O objetivo do trabalho é identificar o perfil de cada participante baseado nos dados obtidos. Para isso, foi utilizado o algoritmo K-means para o agrupamento.

Após sua realização, foi possível perceber que características como renda familiar, escolaridade dos pais, escola do inscrito, lugar de residência e o fato se o candidato possui internet ou não, impactam na nota final do participante. A limitação que pode

[2] http://dados.gov.br

[3] http://www.portaltransparencia.gov.br

ser citada é a falta de descrição detalhada da tarefa de mineração de dados, onde foi dado foco apenas no resultado final.

2.3.2 Mineração de Dados Abertos: Uma análise do uso de *bots* em pregões eletrônicos

O pregão eletrônico é uma modalidade licitatória utilizada pelo governo brasileiro para contratar bens e serviços, independentemente do valor estimado, facilitando a aquisição de compras governamentais, bens e serviços considerados comuns. Este trabalho demonstra a importância de uma modalidade licitatória como esta e relata as dificuldades de se gerenciar estes eventos pela internet, além de tentar identificar possíveis atuações de bots durante as licitações.

Foram utilizadas técnicas de agrupamento e associação através da ferramenta RapidMiner para aplicar os algoritmos de mineração de dados. Como conclusões, foram constatados comportamentos anômalos, considerados como sendo de *bots,* devido à alta frequência de lances emitidos por itens, habitualidade no registro de lances com intervalos similares de tempo, entre outros fatores. Como limitação, pode ser citada a falta de detalhamento nos algoritmos de mineração utilizados, já que só é dito que o RapidMiner contempla vários.

3. Metodologia

A metodologia para desenvolvimento deste trabalho é baseada no *Cross Industry Standard Process for Data Mining* (CRISP-DM) (Chapman, 1999). Este modelo de processo possui seis etapas, que seriam o ciclo de vida de um projeto de ciência de dados (Data Science, 2021). As etapas a serem seguidas são: (i) Entendimento do negócio; (ii) Entendimento dos dados; (iii) Preparação dos dados; (iv) Modelagem; (v) Avaliação e (vi) Implantação do modelo. As etapas são descritas ao longo do presente relatório, conforme a execução realizada neste trabalho.

3.1 Entendimento do negócio

A etapa de entendimento do negócio promoveu a compreensão da necessidade da população poder acompanhar de forma mais específica o desenrolar dos gastos realizados pelo governo federal. Isto pode ser feito através de dados abertos, pois os mesmos são publicados e estão disponíveis para qualquer cidadão verificá-los. Com essas informações, é possível buscar identificar se algum gasto ou algum empenho aconteceu de forma indevida, ou se ele está muito distante em relação aos demais.

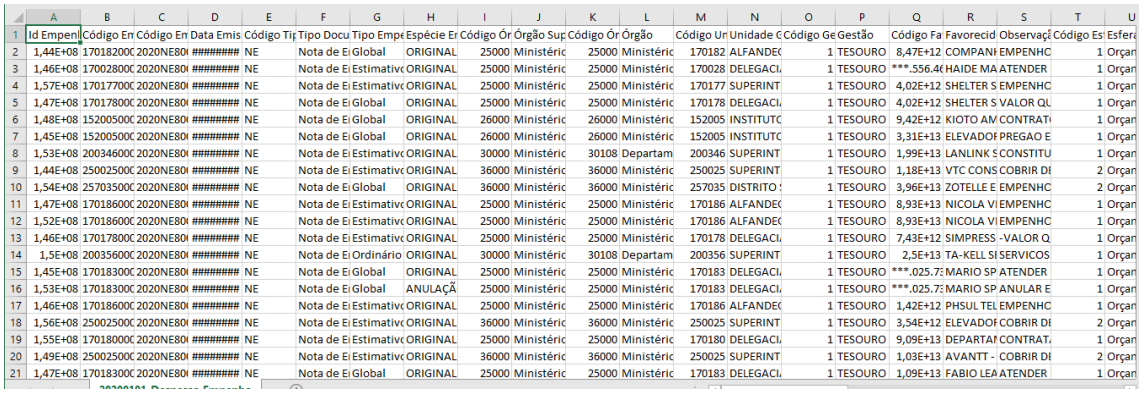
3.2 Entendimento dos dados

Na etapa de entendimento dos dados, busca-se a coleta e exploração do conjunto de dados a ser usado. Neste trabalho, o objetivo principal é a geração de grupos identificados a partir do dataset com dados abertos e a identificação de *outliers*. O dataset utilizado neste trabalho encontra-se no portal da transparência, sendo os dados dos empenhos governamentais realizados pelo governo federal. Empenho é a etapa em que o governo reserva o dinheiro que será pago quando o bem comprado for entregue ou quando o serviço contratado for concluído. Isso ajuda o governo a organizar os gastos

pelas diferentes áreas do governo, evitando que se gaste mais do que foi planejado (Portal da Transparência, 2018).

Os dados dos empenhos governamentais são disponibilizados mensalmente, em arquivos de formato aberto .csv. O dataset original dos empenhos contém 35 campos. A Figura 1 demonstra um fragmento do arquivo .csv recuperado do Portal da Transparência.

Figura 1 – Fragmento do *dataset*



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Id Empen	Código Em	Código Em	Data Emis	Código Tip	Tipo Docu	Tipo Empe	Espécie Er	Código Ór	Órgão Sup	Código Ór	Órgão	Código Un	Unidade G	Código Ge	Gestão	Código Fa	Favorecid	Observaçã	Código Es	Esfera
2	1,44E+08	17018200(	2020NE80(	########	NE	Nota de Ei	Global	ORIGINAL	25000	Ministério	25000	Ministério	170182	ALFANDE(	1	TESOURO	8,47E+12	COMPANI	EMPENHO	1	Orçan
3	1,46E+08	17002800(	2020NE80(	########	NE	Nota de Ei	Estimativo	ORIGINAL	25000	Ministério	25000	Ministério	170028	DELEGACI/	1	TESOURO	***.556.4(	HAIDE MA	ATENDER	1	Orçan
4	1,57E+08	17017700(	2020NE80(	########	NE	Nota de Ei	Estimativo	ORIGINAL	25000	Ministério	25000	Ministério	170177	SUPERINT	1	TESOURO	4,02E+12	SHELTER S	EMPENHO	1	Orçan
5	1,47E+08	17017800(	2020NE80(	########	NE	Nota de Ei	Global	ORIGINAL	25000	Ministério	25000	Ministério	170178	DELEGACI/	1	TESOURO	4,02E+12	SHELTER S	VALOR QU	1	Orçan
6	1,48E+08	15200500(	2020NE80(	########	NE	Nota de Ei	Global	ORIGINAL	26000	Ministério	26000	Ministério	152005	INSTITUTO	1	TESOURO	9,42E+12	KIOTO AM	CONTRAT	1	Orçan
7	1,45E+08	15200500(	2020NE80(	########	NE	Nota de Ei	Global	ORIGINAL	26000	Ministério	26000	Ministério	152005	INSTITUTO	1	TESOURO	3,31E+13	ELEVADOF	PREGAO E	1	Orçan
8	1,53E+08	20034600(	2020NE80(	########	NE	Nota de Ei	Estimativo	ORIGINAL	30000	Ministério	30108	Departam	200346	SUPERINT	1	TESOURO	1,99E+13	LANLINK S	CONSTITU	1	Orçan
9	1,44E+08	25002500(	2020NE80(	########	NE	Nota de Ei	Estimativo	ORIGINAL	36000	Ministério	36000	Ministério	250025	SUPERINT	1	TESOURO	1,18E+13	VTC CONS	COBRIR DI	2	Orçan
10	1,54E+08	25703500(	2020NE80(	########	NE	Nota de Ei	Global	ORIGINAL	36000	Ministério	36000	Ministério	257035	DISTRITO	1	TESOURO	3,96E+13	ZOTELLE E	EMPENHO	2	Orçan
11	1,47E+08	17018600(	2020NE80(	########	NE	Nota de Ei	Estimativo	ORIGINAL	25000	Ministério	25000	Ministério	170186	ALFANDE(	1	TESOURO	8,93E+13	NICOLA VI	EMPENHO	1	Orçan
12	1,52E+08	17018600(	2020NE80(	########	NE	Nota de Ei	Estimativo	ORIGINAL	25000	Ministério	25000	Ministério	170186	ALFANDE(	1	TESOURO	8,93E+13	NICOLA VI	EMPENHO	1	Orçan
13	1,46E+08	17017800(	2020NE80(	########	NE	Nota de Ei	Estimativo	ORIGINAL	25000	Ministério	25000	Ministério	170178	DELEGACI/	1	TESOURO	7,43E+12	SIMPRESS	-VALOR Q	1	Orçan
14	1,5E+08	20035600(	2020NE80(	########	NE	Nota de Ei	Ordinário	ORIGINAL	30000	Ministério	30108	Departam	200356	SUPERINT	1	TESOURO	2,5E+13	TA-KELL SI	SERVICOS	1	Orçan
15	1,45E+08	17018300(	2020NE80(	########	NE	Nota de Ei	Global	ORIGINAL	25000	Ministério	25000	Ministério	170183	DELEGACI/	1	TESOURO	***.025.7(	MARIO SP	ATENDER	1	Orçan
16	1,53E+08	17018300(	2020NE80(	########	NE	Nota de Ei	Global	ANULAÇÃ	25000	Ministério	25000	Ministério	170183	DELEGACI/	1	TESOURO	***.025.7(	MARIO SP	ANULAR E	1	Orçan
17	1,46E+08	17018600(	2020NE80(	########	NE	Nota de Ei	Estimativo	ORIGINAL	25000	Ministério	25000	Ministério	170186	ALFANDE(	1	TESOURO	1,42E+12	PHSUL TEL	EMPENHO	1	Orçan
18	1,56E+08	25002500(	2020NE80(	########	NE	Nota de Ei	Estimativo	ORIGINAL	36000	Ministério	36000	Ministério	250025	SUPERINT	1	TESOURO	3,54E+12	ELEVADOF	COBRIR DI	2	Orçan
19	1,55E+08	17018000(	2020NE80(	########	NE	Nota de Ei	Estimativo	ORIGINAL	25000	Ministério	25000	Ministério	170180	DELEGACI/	1	TESOURO	9,09E+13	DEPARTAI	CONTRAT/	1	Orçan
20	1,49E+08	25002500(	2020NE80(	########	NE	Nota de Ei	Estimativo	ORIGINAL	36000	Ministério	36000	Ministério	250025	SUPERINT	1	TESOURO	1,03E+13	AVANTT -	COBRIR DI	2	Orçan
21	1,47E+08	17018300(	2020NE80(	########	NE	Nota de Ei	Global	ORIGINAL	25000	Ministério	25000	Ministério	170183	DELEGACI/	1	TESOURO	1,09E+13	FABIO LEA	ATENDER	1	Orçan

3.3 Preparação dos dados

Após a análise inicial do conjunto de dados e seus metadados, foram definidos quatro campos a serem considerados para geração do modelo de agrupamento e para análise. A Tabela 1 mostra a descrição dos campos selecionados para o dataset.

Tabela 1 – Descrição dos campos do dataset

CAMPO	**DESCRIÇÃO**
FUNCAO	Código da função (Ex: 5,6,7).
ACAO	Código da ação, que é relacionado a função (Ex: 2,6,7).
EMPENHO	Valor empenhado na despesa (Ex: 25.000,00).
PAGO	Valor gasto realmente na execução da despesa (Ex: 25.500,00).

Descrevendo brevemente, a função é uma forma de categorização das despesas públicas, onde existem por exemplo: Saúde, Educação, Economia. Já a ação é uma ramificação de cada função. Por exemplo, pode-se ter um registro de uma despesa de função “Educação” e ação “Funcionamento das instituições da rede federal”.

Como os dados são disponibilizados mensalmente, foi definido durante esta etapa que seriam utilizados dados dos meses de Janeiro a Abril de 2020. Após a extração dos campos necessários e montagem de um único dataset dos 4 meses, o resultado final foi um conjunto de dados contendo 60.416 registros.

3.4 Modelagem

Após a preparação do conjunto de dados, foram selecionados os algoritmos para execução do aprendizado de máquina. Durante a execução do projeto, foram utilizados dois algoritmos para análise dos resultados que foram o K-means e o DBScan.

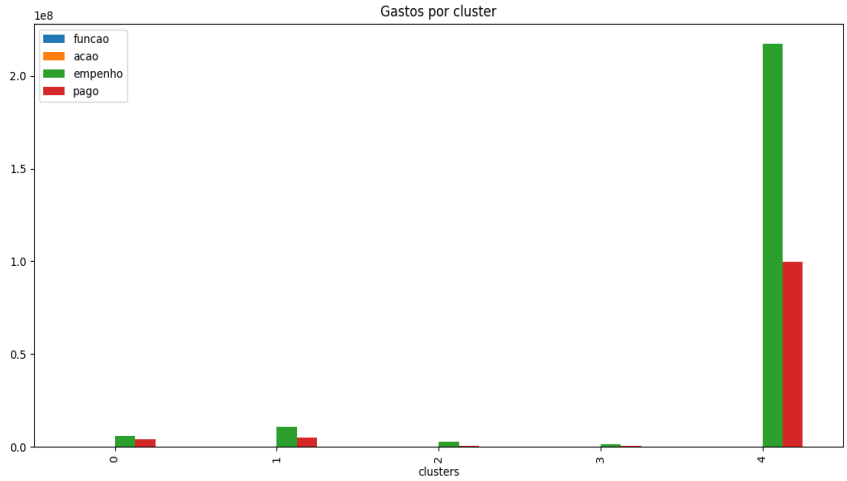
3.4.1 K-means

O algoritmo K-means foi o primeiro a ser selecionado para a execução dos testes devido à sua popularidade e à facilidade do aprendizado para sua utilização.

Este algoritmo determina que deve ser definido um valor para k, que será o número de grupos (clusters) esperados após a execução do algoritmo. Neste caso, foram realizados testes com vários valores para a variável k, variando de 5 a 10. Após a execução, foi definido que o valor final de k seria 10, pois desta forma os grupos acabam sendo mais específicos em suas similaridades.

As Figuras 2 e 3 mostram os grupos que foram formados em um gráfico de barras, onde a diferença entre o tamanho das barras se dá devido aos valores das colunas EMPENHO e PAGO, que são valores monetários.

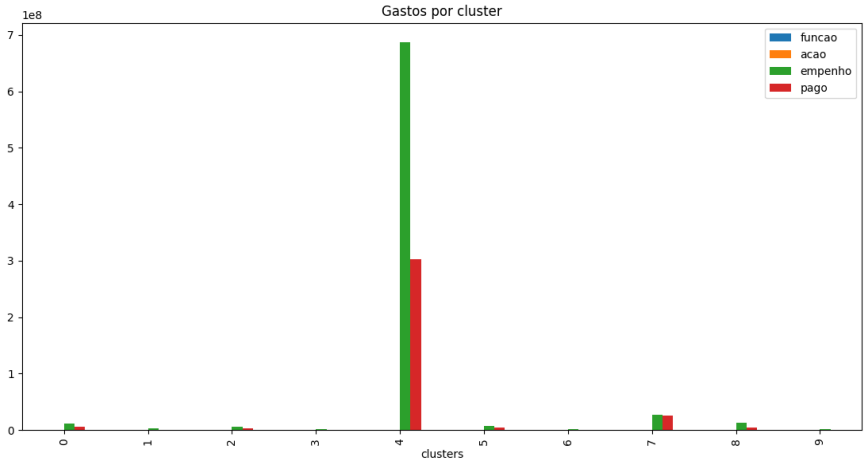
Figura 2 - Grupos formados com 5 clusters



Na Figura 2 é possível verificar que o último grupo (grupo 5) é formado por registros onde os valores de empenhados e pagos são consideravelmente maiores que os demais, porém contém diferença entre o valor que foi empenhado inicialmente e o que realmente foi pago.

Na Figura 3, pode-se ver que o mesmo padrão aparece quando são gerados 10 clusters. O grupo 4 contém as mesmas características do grupo 5 da Figura 2. Além disso, o grupo 4 é o grupo com menos registros vinculados, contendo 2.103 registros.

Figura 3 - Grupos formados com 10 clusters



A Tabela 2 indica alguns exemplos de registros retirados do grupo 4 gerado a partir da execução com k = 10, onde pode-se ver a diferença entre os valores que foram pagos do que foi realmente empenhado. É possível verificar que nem sempre a diferença é simplesmente o valor pago igual a zero, em muitos casos se pôde ver que a diferença é menor.

Tabela 2 – Exemplos de valores do cluster com menos registros

VALOR EMPENHADO	**VALOR PAGO**
20.536.555,91	5.399.813,92
13.095.500,00	8.694.522,72
270.000,00	5.097,70
973,50	4.733,60

3.4.2 DBScan

A razão pela qual este algoritmo foi escolhido é a facilidade que ele tem para marcação de pontos como anomalias, sendo pontos em regiões de baixa densidade (PRADO, 2017). Como o objetivo deste estudo é a identificação de possíveis pontos anômalos nos gastos públicos, o algoritmo foi utilizado.

Diferente dos testes com K-means, o número de grupos gerados pelo algoritmo é gerado pelo próprio algoritmo. No caso do dataset estudado, o algoritmo chegou a gerar 38 clusters. Após a execução do algoritmo, foi possível verificar o grupo de label = -1, que seria o grupo com registros que não se encaixaram em nenhum outro cluster, sendo considerados ruídos ou outliers. Ao analisar esses registros, foi possível enxergar o mesmo padrão dos grupos estudados após a execução do K-means, que são registros com valores empenhados diferentes do que realmente foi pago. A Tabela 3 demonstra alguns dos registros encontrados neste grupo.

Tabela 3 – Exemplos de valores dos outliers da execução do DBScan

VALOR EMPENHADO	VALOR PAGO
267.285.000,00	6.540.106,12
730.587.400.00	141.900.000,00
411.050,27	15,12
92.481,60	255.584,11

4. Discussão

Após a execução dos algoritmos, foi possível constatar que os grupos formados com menos registros aparentam ter um padrão, que seria a diferença entre os valores dos pagamentos e dos empenhos iniciais.

Não foi possível atestar com toda a certeza que estes valores são anomalias pois, para isso, deveria ser feito um estudo aprofundado de cada um dos registros, pois pode haver explicações plausíveis para esta diferença de valores. Contudo, esta base de dados contendo os registros identificados como possíveis *outliers* pode ser usada para um futuro estudo, caso seja necessário, servindo como base para definir quais registros podem ser estudados mais a fundo a fim de encontrar realmente uma anomalia na execução da despesa.

5. Considerações e trabalhos futuros

Após este estudo, pode-se concluir que os dois algoritmos usados contém eficácia semelhante no agrupamento dos registros, porém o algoritmo DBScan oferece mais opções na identificação de *outliers* que, para este caso, foi mais útil. Com os resultados, foi possível verificar que existem registros que podem ser considerados ruídos ou registros fora da curva, permitindo utilizar esta base de dados para futuros estudos mais aprofundados.

Como trabalhos futuros, podem ser citados a execução de mais um algoritmo para análise dos resultados, como o K-medoids, além de aumentar o escopo do dataset, adicionando mais meses para a análise. Além disso, os *clusters* formados devem ser analisados de maneira mais específica, de forma a encontrar possíveis outros padrões que não são apenas baseados nos valores monetários de empenhos e pagamento.

6. Referências

Alpaydin, E. (2010). Introduction to machine learning. MIT press.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (1999). The CRISP-DM user guide. In 4th CRISP-DM SIG Workshop in Brussels in March (Vol. 1999).

Data Science. What is CRISP DM? Disponível em: www.datascience-pm.com/crisp-dm-2. Acesso em: 10 Jan 2021.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37.

GARBADE, M. Understanding K-means clustering in Machine Learning. Disponível em:https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. Acesso em: 24 Jan 2021.

HAWKINS, D. (1980). Identification of Outliers.1 ed. New York: Chapman and Hall.

JUNIOR, R; JUNIOR, J; SILVA, T; SILVA, T; MAGALHAES, R. Mineração de Dados Abertos. 1 ed. Parnaíba, Faculdade Maurício de Nassau.

KAUFMAN, L; ROUSSEEUW, P. (1990). Finding groups in data: An introduction to cluster analysis. 1 ed. New Jersey: John Wiley & Sons, Inc.

Open Knowledge International. What is Open? Disponível em: https://okfn.org/opendata/. Acesso em: 21 Jan. 2021.

Portal da Transparência. Execução da despesa pública. Disponível em: http://www.portaltransparencia.gov.br/entenda-a-gestao-publica/execucao-despesa-publica. Acesso em: 02 Jan. 2021.

PRADO, K. How DBScan works and why should we use it? Disponível em: https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80. Acesso em: 24 Jan. 2021.

SANTOYO, Sergio. A brief overview of Outlier Detection Techniques. Disponível em: A Brief Overview of Outlier Detection Techniques | by Sergio Santoyo | Towards Data Science. Acesso em: 05 Jan.2021.

SILVA, Ângela; GALVÃO, Maria. A importância do uso de dados abertos pelo poder público para o fortalecimento da governança pública. UFRN, 2018. Acesso em: 10 Mar. 2021.

SOUTO, H. Mineração de Dados Abertos: Uma análise do uso de bots em pregões eletrônicos. Disponível em: https://sig-arq.ufpb.br/arquivos/2019071230f6981803056bc243c9a4b41/Dissertao_-_Hugo_Medeiros_Souto_-_Minerao_de_Dados_Abertos_2.pdf. Acesso em: 10 Dez. 2020.