Detecção de anomalia nas Eleições de 2018 com *Isolation Forest*

Outlier detection in the 2018 Elections with Isolation Forest

José Edson de Albuquerque Filho [1] orcid.org/0000-0001-9340-0086
Camila Luisa Farias de Lima [2] orcid.org/0000-0001-5727-2427
Larissa Maria Costa Rêgo Perboire [2] orcid.org/0000-0003-1724-3394
Paula Beserra Pithon [2] orcid.org/0000-0002-9424-2750

[1] Secretaria de Tecnologia da Informação, Ministério Público de Pernambuco, Recife, Brasil.
[2] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil.

E-mail do autor principal: Edson Filho jeaf@ecomp.poli.br

Resumo

Mineração de exceção é o ramo da ciência de dados que busca descobrir anomalias no conjunto de dados e pode ser um caminho para auxiliar a detecção de distorções relevantes em uma eventual auditoria. Esse artigo busca identificar possíveis candidaturas de fachada através de ferramentas de mineração de dados. Verificamos correlações estatísticas e a existência de características dependentes. Buscamos o estado da arte da detecção de anomalia e selecionamos os algoritmos mais adequados ao perfil do problema. O *Isolation Forest* destacou os candidatos mais destoantes e uma árvore de decisão escolheu o perfil de candidatos com distorções relevantes nos dados da campanha.

Palavras-Chave: Mineração; Política; Eleições; Detecção de anomalias.

Abstract

Exception mining is the branch of data science that detects outliers in the data set and can be a path to figure out relevant distortions in accountability. This paper seeks to identify possible fake applications through data mining tools. We verified statistical correlations in the database and the existence of dependent characteristics. We search the state of the art of anomaly detection and select the most appropriate algorithms for the problem profile. Isolation Forest highlighted the most deviant candidates and decision tree chose the profile of candidates with relevant distortions in the campaign data.

Key-words: *Data Mining; Politics; Elections; Outlier Detection.*

DOI: 1025286/repa.v5i1.1308

1 INTRODUÇÃO

Mineração de exceção é a descoberta de estruturas interessantes, inesperadas ou valiosas em grandes conjuntos de dados. Como tal, tem dois aspectos bastante diferentes. Um deles diz respeito às estruturas em larga escala e o objetivo é modelar as formas ou características. O outro diz respeito a estruturas locais. O objetivo é detectar anomalias e decidir se elas são reais ou ocorrências fortuitas **[1]**.

O programa Jornal Nacional da Tv Globo fez uma reportagem sobre o uso de "laranjas" para se ter mais verba vindas do dinheiro público **[2]**. O jornal analisou os dados de 24.765 candidatos a deputados estaduais e federais e fez a identificação de possíveis laranjas, através do cruzamento dos dados a respeito do quanto os candidatos receberam dos fundos públicos e o número de votos. O jornal fez o comparativo com todos os candidatos eleitos. Foi identificado que destes 51 candidatos, 45 são mulheres, o jornal afirma que essa informação é relevante pois o Tribunal Superior Eleitoral decidiu que ao menos 30% dos recursos do fundo eleitoral devem ser destinados a candidaturas femininas **[3]**.

Nesse contexto, surge a mineração de exceções, conhecida como o processo de encontrar anomalias, padrões e correlações em grandes conjuntos de dados para prever resultados. Através de uma variedade de técnicas, é possível detectar distorções e indícios de fraude.

O intuito deste artigo consiste em verificar anomalias nos gastos de verba empregada nas campanhas eleitorais a cargos públicos, visando identificar distorções relevantes.

Para atingir esse objetivo, serão utilizados dados obtidos no repositório de dados eleitorais do Tribunal Superior Eleitoral (TSE), nele podemos encontrar informações como nome do candidato, cargo almejado, valor recebido para investir na campanha e quantidades de votos recebidos.

2 FUNDAMENTAÇÃO TEÓRICA

O Brasil permite que uma campanha seja financiada com verba pública. Quando um cidadão se inscreve sem a intenção real de concorrer, com objetivos irregulares, dizemos que é um Candidato laranja **[2]**. Isso pode ser modelado como um problema de detecção de anomalias e classificação.

Há várias definições para os dados divergentes (anomalias). Em 2017, Ayadi *et al*. **[4]** deram doze interpretações diferentes a partir de diversos autores. Isso demonstra como é complexo fornecer uma definição precisa de uma anomalia.

Entretanto uma definição clássica é a de Aggarwal **[5]**: um Outlier é um dado significativamente diferente dos demais. Na mesma linha de pensamento, Hawkins **[6]** disse que uma anomalia é uma observação tão diferente das outras que parece que foi gerada por um mecanismo diferente.

Uma definição prática de anomalias para computação inteligente é: São os elementos que, quando retirados do conjunto de treinamento, o desempenho de um algoritmo de aprendizagem da máquina melhora **[7]**.

Outliers, anomalias, anormalidades, discordantes, extremidades, discrepâncias, desvios são sinônimos da literatura quando falamos em mineração de exceção **[6]**.

Nesse artigo vamos combinar uma técnica de detecção de anomalia **[8]** (*Isolation Forest*), análise de correlação **[9]** e árvore de decisão **[10]** para encontrar indícios de candidaturas fora do padrão conforme detalharemos a seguir. A suposta fraude pode ser considerada uma distorção relevante na prestação de contas. A tarefa é identificar se uma candidatura é legítima ou supostamente fraudulenta.

2.1 Isolation Forest

Isolation Forest (IF) é uma combinação de *Isolation trees* (IT), semelhante às *decision trees e random forest* **[11]**. Uma *isolation tree* é construída a partir de uma matriz X, como descrito na Figura 1. Uma *Isolation Forest* é então definida por inúmeras árvores de isolamento **[12]**.

$$IF = \{ IT_1, \ldots, IT_n \}$$

Para cada árvore t, é possível calcular o número de iterações $h_t(x)$ necessárias para isolar uma amostra x. O número médio de etapas necessárias para isolar uma amostra x em uma floresta é então

$$h(x) = \frac{1}{T}\sum h_t(x)$$

Quanto menos passos para isolar uma anomalia melhor. O número de etapas necessárias para isolar uma observação x é influenciado pelo número de amostras n nos dados. Para explicar isso, um escore de anomalia normalizado $s(x,n)$ é definido como:

$$s(x,n) = 2^{-\frac{h(x)}{c(n)}}$$

onde $c(n)$ é:

$$c(n) = \begin{cases} 2H(n-1) - \frac{2(n-1)}{n}, & se\ n > 2 \\ 1, & se\ n = 2 \\ 0, & se\ n < 2 \end{cases}$$

e $H(i)$ é o número harmônico estimado:

$$H(i) \approx ln(i) + 0.577216649$$

Pode-se provar que *c(n)* é o número médio de etapas necessárias para isolar uma amostra das outras n amostras. Assim, há um fator de normalização que torna o valor *s* independentemente do número de amostras *(n)* **[8]**.

Podemos simplificar o IF com o pseudocódigo a seguir:

1. Escolha um ponto aleatoriamente.
2. Para cada dimensão, defina o intervalo entre o mínimo e o máximo.
3. Escolha uma dimensão aleatória
4. Escolha um valor que esteja no intervalo (novamente aleatoriamente):
    a. Se o valor escolhido mantiver o ponto acima, mude o mínimo do intervalo do recurso para o valor.
    b. Se o valor escolhido mantiver o ponto abaixo, mude o máximo do intervalo do recurso para o valor.
5. Repita as etapas 3 e 4 até o ponto ser isolado. Ou seja, até que o ponto seja o único dentro do intervalo para todos os recursos
6. Conte quantas vezes você repetiu as etapas 3 e 4. Chamamos essa quantidade de número de isolamento.

O algoritmo IF afirma que um ponto é uma anomalia se não precisar repetir as etapas 3 e 4 muitas vezes **[8]**.

2.2 Análise de correlação

É o estudo estatístico das relações eventualmente existentes entre variáveis. A investigação de correlação compreende a análise de dados amostrais para saber se e como variáveis distintas estão relacionadas numa população. Coloquialmente, correlação pode ser interpretada como sinônimo de dependência. De uma maneira mais técnica, podemos afirmar que variáveis são dependentes se não satisfizerem a propriedade matemática da independência probabilística **[9]**.

2.3 Árvore de decisão

Uma árvore de decisão **[10]** é um mapa de possibilidades de uma série de escolhas relacionadas. Ela permite comparação de ações com base em seus custos, probabilidades e benefícios. As árvores podem ser usadas para mapear a melhor escolha. Iniciam com um único nó, que se divide em alternativas. Cada ramo leva a nós adicionais, que se desdobram em outras possibilidades.

2.4 Fraudes eleitorais

O Ministério Público Estadual, comandado pelo Procurador Geral de Justiça, existe para fazer com que empresas, pessoas e governos cumpram as leis. Nas eleições, por exemplo, este é responsável por fiscalizar o andamento do processo eleitoral, assim como investigar a existência de candidatos-laranjas ou candidato de fachada: aquele que entra na eleição sem a intenção de concorrer de fato, com objetivos supostamente irregulares, tais como desviar dinheiro do fundo eleitoral. Um levantamento feito pelo Jornal Nacional identificou 51 casos de candidatos supostamente laranjas nas eleições de 2018. Eles concorreram por 18 partidos, em 18 estados **[2]**.

O objetivo desse artigo é analisar, através de técnicas de mineração de dados, os candidatos a deputado federal e estadual de todos os estados do Brasil no ano de 2018 e gerar uma árvore de decisão que auxilie a identificar potenciais candidatos laranja nas próximas eleições.

2.5 Detecção de anomalias

Em seu livro *Outlier Analysis*, Aggarwal **[5]** classifica métodos de detecção de anomalias em dados uni variados. Modelos probabilísticos e estatísticos que determinam instâncias improváveis de um modelo probabilístico dos dados. Modelos que usam correlações lineares. Modelos baseados em proximidade determinada pela análise de cluster, densidade ou vizinhança. Métodos que pesquisam subespaços para valores discrepantes **[5]**.

Existem pesquisas que investigam a análise de dados a respeito do assunto abordado. O artigo Mulheres e Política no Brasil: trajetos e perspectivas sobre a Lei de Cotas de Gênero analisa, com base em dados empíricos, informações importantes para o desenvolvimento desse projeto fazendo uma

DOI: 1025286/repa.v5i1.1308

investigação de como se deu a distribuição dos recursos financeiros entre mulheres e homens em diferentes estados, nas competições proporcionais das eleições de 2014. O artigo também declara que mesmo que não seja fator único, o financiamento se confirmou como peça-chave para uma real e maior inclusão feminina na política **[13]**.

Outro Artigo relevante para a pesquisa é o Mulheres em campanha: uma análise da distribuição de recursos financeiros nos estados brasileiros e o desempenho eleitoral das mulheres nas eleições de 2014 que faz uma análise da distribuição financeira dos recursos envolvidos em eleições e a relação entre o valor investido em campanhas e sua relevância nos resultados da eleição. Constatando que a maioria desses recursos estavam concentrados entre os candidatos homens, porém nos locais onde mulheres tinham um investimento equivalente, elas tiveram resultados positivos. Sendo todas estas, constatações semelhantes às do artigo anterior, reforçando a relevância dessas informações **[3]**.

3 MATERIAIS E MÉTODOS

Um dos grandes problemas da mineração é o tratamento inicial dos dados. Realizamos várias transformações para adequar os dados para o trabalho dos algoritmos.

Inicialmente analisamos e transformamos os dados disponíveis no repositório de dados eleitorais do formato texto para um banco relacional. Com os dados no banco, pudemos realizar várias análises e empiricamente decidir qual seria a amostra ideal para nosso estudo.

Para o projeto proposto, além de realizarmos análise e limpeza dos dados, utilizamos técnicas como o *Isolation Forest* **[8]**, análise de correlação **[9]** e árvore de decisão **[10]** para montarmos um modelo de classificação de candidatos que devem ser investigados por apresentarem indícios de anomalias na sua prestação de contas da campanha.

Fazendo uma análise da coerência **[9]** desse sistema, foi identificado a relevância que altos investimentos em campanhas políticas leva a mais chances de sucesso nas eleições, também a diferença entre a participação entre candidatos do sexo masculino e feminino. A partir dessas informações foi identificada inconsistência do uso das cotas partidárias. Alguns candidatos receberam doações elevadas e tiveram um número de votos extremamente baixo. Isso sugere a existência de candidatura de "laranjas" para desvio de verba dos cofres públicos. O alto número de mulheres inaptas (1/4 das candidaturas femininas) é um dos fatos que possibilita essa conclusão.

Para descobrir os outliers usamos o *Isolation Forest* **[8]**, um algoritmo não supervisionado. poderíamos complementar com um algoritmo de clusterização como o *k-means* **[14]**. A Figura 1 a apresenta a relação entre número de votos e valor investido na campanha dos candidatos.

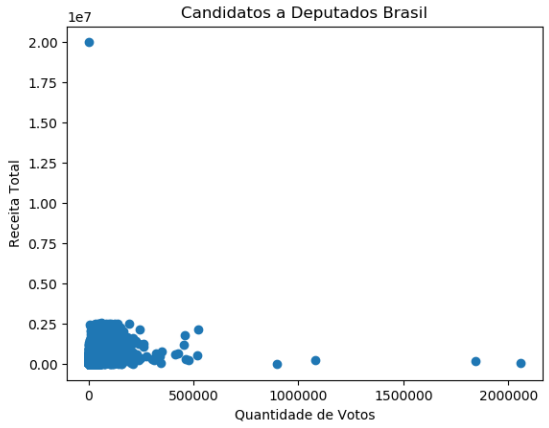

Figura 1: Receita Total x Quantidade de votos dos Candidatos a Deputados.
Fonte: O Autor.

O algoritmo *Isolation Forest* **[8]** então se mostrou um método relevante a ser aplicado, justamente pela sua capacidade de apresentar os candidatos que possuem características destoantes dos demais.

Para manter os outliers que, possuíam características de candidatos de fachada, de acordo com os artigos mencionados, extraímos dos resultados obtidos pelo algoritmo aqueles candidatos com maiores gastos em suas campanhas.

Em seguida realizamos uma análise de correlação **[9]** entre os atributos, que julgamos mais relevantes dos candidatos, e o resultado do *Isolation Forest* **[8]**, com isso conseguimos encontrar algumas informações importantes como 95% dos possíveis laranjas serem mulheres e a 60% dos suspeitos possuírem mais de R$100.000 em despesas contratadas. Como pode ser percebido nas tabelas 1 e 2.

Tabela 1: Quantitativo de anomalias por "Sexo".

Sexo	Não é outlier	É outlier	Total Geral
FEMININO	7.601	19	7.620
MASCULINO	15.830	1	15.831
Total Geral	**23.431**	**20**	**23.451**

Tabela 2: Quantitativo de anomalias por "Despesa Contratada".

Despesa Contratada	Não é outlier	É outlier	Total Geral
(vazio)	7.514	3	7.517
x<=1.000	1.106	0	1.106
10000<x<=50000	1.511	3	1.514
100000< x	9.786	12	9.798
50000<x<=100000	331	1	332
1000>x<=10000	3.183	1	3.184
Total Geral	**23.431**	**20**	**23.451**

Além da análise de correlação, utilizamos os atributos selecionados e o resultado do *Isolation Forest* **[8]** como entrada para gerar o modelo da Árvore de Decisão que irá determinar quais candidatos precisam ser investigados por apresentar fortes indícios de ter realizado uma candidatura e fachada.

Para isso, criamos um script em Python que transforma o .CSV com os dados em um arquivo ARFF que inserido no software Weka para gerar o modelo da Árvore de Decisão.

3.1 Base de Dados

O Repositório de dados eleitorais é uma compilação de informações brutas das eleições, desde as de 1945, voltada para pesquisadores, imprensa e pessoas interessadas em analisar os dados de eleitorado, candidaturas, resultados e prestação de contas. Todos os arquivos fornecidos estão em formato TXT e podem ser importados para qualquer programa estatístico, base de dados ou planilha eletrônica. Consultas, filtros e cruzamentos são de responsabilidade do pesquisador. É importante ler o arquivo de instruções e atentar à data de geração do arquivo, para fazer as importações e as consultas de forma adequada.

Utilizamos um subconjunto das bases disponíveis sobre as candidaturas de 2018. Nesta seção, constam arquivos com informações acerca do perfil dos candidatos nas eleições, declarações de bens e dados sobre os partidos, as coligações e as vagas por cargo e por unidade eleitoral. Após análise de correlação dos dados selecionamos os seguintes atributos para treinar a nossa Árvore de Decisão:

1- DS_GENERO: gênero do candidato;

2- DS_GRAU_INSTRUCAO: grau de instrução do candidato;

3- SG_PARTIDO: sigla do partido do candidato;

4-DS_OCUPACAO: descrição da ocupação do candidato;

5-DESPESA_CONTRATADA: valor total da despesa contratada pelo candidato;

6-VR_TOTAL_BEM_CANDIDATO: valor total dos bens declarados pelo candidato.

4 RESULTADOS OBTIDOS

Em nossos experimentos pudemos observar a discrepância no custo por voto de alguns candidatos na eleição de 2018 em Pernambuco, conforme mostra a Tabela 3.

Tabela 3: Discrepância no custo por voto.

Candidato	Receita	Votos	Custo
Candidato 1	R$ 200.000,00	37	R$ 5.405,41
Candidato 2	R$ 35.600,00	20	R$ 1.780,00
Candidato 3	R$ 12.013,00	7	R$ 1.716,22
Candidato 4	R$ 400.000,00	274	R$ 1.459,85
Candidato 5	R$ 150.000,00	147	R$ 1.020,41

Não podemos afirmar que houve fraude nem qualquer tipo de irregularidade na campanha dos candidatos. Podemos apenas dizer que os algoritmos utilizados indicam que os alguns números da campanha estão em desarmonia em relação à maioria e merecem um olhar mais criterioso das autoridades competentes.

Vale salientar que o candidato eleito com o voto mais caro em Pernambuco recebeu para campanha aproximadamente R$ 361.792,00 e teve 65.750 votos, ou seja, um custo por voto de R$ 30,48. No Brasil, o candidato eleito com o voto mais caro foi de Roraima. A receita da sua campanha foi de R$ 2.342.528,00 para obter 12.129 votos a um custo unitário de R$ 193,13.

Ficamos atentos aos candidatos que apresentaram grandes discrepâncias desses valores e usamos o *Isolation Forest* para encontrá-los. Em todo Brasil o candidato mais discrepante recebeu cerca de R$ 20.000.00,00 e teve apenas 119 votos. outros 162 candidatos gastaram mais de R$ 1.000,00 por voto e não foram eleitos.

Com a identificação desses outliers e com os atributos descritos anteriormente, treinamos uma árvore de decisão para reconhecer esse perfil de candidato e poder prever nas próximas eleições com antecedência perfis relevantes para investigação. A Figura 2 mostra o modelo da Árvore de Decisão gerado no Weka.

DOI: 1025286/repa.v5i1.1308

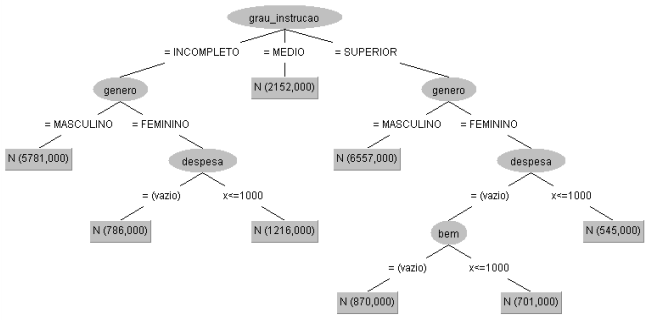

Figura 2: Árvore de Decisão para possível detecção de anomalias.
Fonte: O Autor.

Como trabalho futuro, poderíamos usar uma técnica de agrupamento (clusterização) tal como o *k-means* [5] para melhorar a predição e servir de parâmetro de comparação ou para combinar o resultado (Ensemble).

REFERÊNCIAS

[1] HAND, D. J.; BACH, F. **Principles of Data Mining**. MIT Press, 2001.

[2] BARBOSA, A., AMORIM, T. **MPE detecta indícios de candidaturas 'laranja' pelo PSL nas eleições de 2018 no Maranhão**. Em: https://g1.globo.com/ma/maranhao/noticia/2019/04/04/mpe-detecta-indicios-de-candidaturas-laranja-pelo-psl-nas-eleicoes-de-2018-no-maranhao.ghtml.
Último acesso: 14/04/2020.

[3] EDUARDO, M. C. **Mulheres em campanha: uma análise da distribuição de recursos financeiros nos estados brasileiros e o desempenho eleitoral das mulheres nas eleições de 2014**. *Guaju* – Revista Brasileira de Desenvolvimento Territorial Sustentável. v.4(2), p.187-208, 2018.

[4] AYADI, A., *et al*. **Outlier detection approaches for wireless sensor networks: A survey**. Computer Networks v. 129(1), p.319-333, 2017.

[5] AGGARWAL, C. C. **An introduction to outlier analysis**. In Outlier analysis, second ed. Springer, 2017. ISBN: 978-3-319-47577-6, 978-3-319-47578-3.

[6] HAWKINS, D. M. **Identification of Outliers**, 1 ed. Monographs on Applied Probability and Statistics. Springer Netherlands, 1980. ISBN 978-94-015-3996-8,978-94-015-3994-4.

[7] LIU, H.; SHAH, S.; JIANG, W. **On-line outlier detection and data cleaning**. Computers & Chemical Enggineering. Vol. 28 (9) p.1635–1647, 2004.

[8] LIU, F. T., TING, K. M., ZHOU, Z. H., **Isolation forest**. In: Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, pp. 413– 422.

[9] BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. 6ª ed. Editora Saraiva, 2010.

[10] FRIEDL, M. A., BRODLEY, C.E. **Decision tree classification of land cover from remotely sensed data**. Remote Sensing of Environment v.61(3), p.399-409, 1997.

[11] BREIMAN, L. **Random forests**. Machine learning 45.1 (2001): 5-32.

[12] MURTHY, S. K. **Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey**. Data Mining and Knowledge Discovery 2, 345–389, 1998.

[13] DE ALMEIDA NETO, Antônio Lopes; FORTUNATO, Caio Emanuel Brasil; DA SILVA CARDOSO, Fernando. **Mulheres e política no Brasil: trajetos e perspectivas sobre a lei de cotas de gênero**. Caderno Espaço Feminino, v. 30, n. 2, 2018.

[14] MACQUEEN, J. B. (1967). **Some Methods for classification and Analysis of Multivariate Observations**. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Consultado em 14 de novembro de 2012.