

Uma combinação dos algoritmos Isolation Forest e K-Means aplicada às Eleições Brasileiras

Luís Felipe Bueno

Universidade Federal de São Paulo (UNIFESP)
São José dos Campos - SP
lfelipebueno@gmail.com

Mateus Vendramini Polizeli

Universidade Federal de São Paulo (UNIFESP)
São José dos Campos - SP
polizelim@gmail.com

Maurício Vidotti Fernandes

Universidade de Brasília (UnB)
Brasília - DF
mauricio.vidotti@gmail.com

Camila Bertini Martins

Universidade Federal de São Paulo (UNIFESP)
São Paulo - SP
cb.martins@unifesp.br

RESUMO

O objetivo deste artigo é estudar mecanismos de detecção de anomalias associadas ao conjunto de dados das candidaturas eleitorais do Brasil em 2018. A metodologia proposta baseia-se em combinar os algoritmos não supervisionados *K-Means* e *Isolation Forest* visando criar uma ferramenta de apoio à decisão para reguladores direcionarem recursos humanos para investigação. Os resultados indicam que a proposta KM + IF mostra bom desempenho para situações em que dados anômalos não apresentam tendência de agregação. No entanto, esta metodologia pode produzir resultados insatisfatórios quando essa propriedade de aglomeração se mantém. No estudo de caso para o conjunto de candidaturas eleitorais, o resultado geral do algoritmo KM + IF foi inferior ao resultado individual da técnica *Isolation Forest*.

PALAVRAS CHAVE. Detecção de anomalias, dados eleitorais, *Isolation Forest*, métodos de agrupamento.

ABSTRACT

The aim of this paper is to study anomaly detection mechanisms associated with the dataset of the 2018 Brazilian elections. The proposed methodology is based on combining the unsupervised algorithms *K-Means* and *Isolation Forest* to create a decision support tool for regulators to direct human resources for investigation. The results indicate that the KM + IF proposal shows good performance for situations where anomalous data do not present aggregation tendency. However, this methodology can yield unsatisfactory results when such agglomeration property holds. In

the case study for the set of electoral candidates, the performance of the KM + IF algorithm was worst than the individual result of the *Isolation Forest* technique.

KEYWORDS. Anomaly Detection, Electoral Data, Isolation Forest, Clustering methods.

1. Introdução

Há algumas décadas a sociedade brasileira tem apresentado maior desconfiança e falta de esperança nos políticos que lideram o país. Além disso, alguns eventos mais recentes, como operações de investigação da Polícia Federal e análise do Ministério Público mostram grande apoio da população para que atitudes criminosas sejam criteriosamente examinadas e julgadas, a fim de que todos os eventos sejam esclarecidos e os envolvidos punidos conforme determina a lei. Pesquisas realizadas pela *Transparency International* apontam o Brasil com resultados regulares a muito ruins no ranking de países com maior corrupção nos setores públicos [Pfeiffer, 2004], [Saibro, 2006] e [Alley, 2011]. Além disso, a mais recente atualização de seu índice de percepção da corrupção (CPI), de 2018 [Transparency International, 2018], mostra que o Brasil caiu 35 posições no ranking, ou seja, está classificado na 105ª posição de um total de 180 países avaliados, ao lado de nações como El Salvador, Zâmbia, Peru, Timor Leste, muito atrás de países vizinhos como Uruguai (23ª), Chile (27ª) e Argentina (85ª).

Em [Ribeiro et al., 2018] é detalhada de forma muito interessante a estrutura dinâmica das redes de corrupção na política brasileira, capturando informações de mais de 60 casos de corrupção envolvendo mais de 400 indivíduos, nos últimos 27 anos. Uma das conclusões dos autores é que a série temporal do número anual de pessoas envolvidas em casos de corrupção tem um componente periódico (com um intervalo de tempo de 4 anos) que corresponde justamente ao ciclo eleitoral no Brasil, o que os levaram a suspeitar de que as eleições gerais não apenas reformulam a elite política, mas também introduzem novos indivíduos ao poder que logo poderão cair no mesmo ciclo de irregularidades dos antigos parlamentares. Assim, é natural o interesse em mapear e levantar situações atípicas e anômalas em bases eleitorais.

A abordagem aplicada neste trabalho vai ao encontro da análise de detecção de anomalias, por ser um dos métodos mais utilizados no estudo de contrastes e discrepâncias em inspeções em análise de dados. Segundo [Parmar e Patel, 2017], a detecção de anomalias é definida como o problema de encontrar padrões em dados que não estão em conformidade com o comportamento esperado. Este efeito de detecção é amplamente utilizado em uma variedade de áreas como detecção de fraudes em cartões de crédito, seguros ou assistência médica, detecção de intrusão para cibersegurança, detecção de falhas em sistemas industriais, entre outros.

Salienta-se que a detecção de *outliers* ou anomalias nos dados foi estudada na comunidade estatística já no século XIX [Edgeworth, 1887]. Ao longo do tempo, uma variedade de técnicas de detecção de anomalias foi desenvolvida em diferentes comunidades de pesquisa. Muitas dessas técnicas foram criadas especificamente para determinados domínios de aplicação, enquanto outras são mais genéricas, podendo ser aplicadas em uma grande gama de problemas.

Neste contexto, o objetivo deste trabalho é estudar uma combinação dos algoritmos não supervisionados *K-Means* e *Isolation Forest* quando aplicada para detecção de anomalias associadas ao conjunto de dados das candidaturas eleitorais de 2018 do Brasil. Desta forma esperamos estar contribuindo tanto com discussões sobre a técnica proposta como também com a bibliografia escassa sobre ferramentas de apoio à tomada de decisão na aplicação específica. Iniciativas deste tipo podem melhor direcionar os recursos humanos limitados para uma possível investigação mais

eficiente. Visto que o Brasil, na visão de muitos especialistas, provavelmente vai passar por eleições presidenciais conturbadas no ano de 2022, acreditamos que propostas levantam questões sobre estes pontos sejam muito bem vindas no momento.

2. Materiais e métodos

2.1. Banco de dados

Os dados utilizados foram retirados do site do TSE, por meio de uma API (Interface de Programação de Aplicação) desenvolvida na linguagem *Python*. O repositório de dados eleitorais é uma compilação de informações brutas das eleições desde 1945, voltada para pesquisadores, imprensa e pessoas interessadas em analisar os dados de eleitorado, candidaturas, resultados das eleições e prestação de contas. Entre as informações disponíveis, destacam-se: base de candidatos, com informações relacionadas ao perfil dos postulantes nas eleições, declarações de bens e dados sobre os partidos, as coligações e as vagas por cargo e por unidade eleitoral; base de prestação de contas, com arquivos de prestação de contas com todas as receitas e despesas de campanha dos candidatos, de partidos e de comitês; bases processuais, de informações de processos eleitorais, decisões e recursos referentes ao pleito de candidatos; por fim, bases de resultados, com informações dos resultados das eleições, totalizados e por boletins de urna.

2.2. Seleção de público do experimento

A escolha do conjunto inicial de dados contém registros de 8.588 candidaturas de deputados federais por ter abrangência nacional e os tetos de gastos de campanha serem os mesmos e iguais a R\$ 2.5 milhões. Ao observar a existência de candidaturas “vazias”, ou seja, sem votos e sem nenhum valor movimentado na campanha (receitas e despesas) foram aplicados simultaneamente dois filtros de materialidade no conjunto inicial de dados: quantidade total de votos menor do que mil; despesas e receitas inferiores a R\$ 10.000,00. Dessa forma, se um candidato a deputado federal possuir uma quantidade inferior a mil votos e, além disso, apresentar a soma das despesas e receitas inferior à dez mil reais, será excluído da base de treinamento dos algoritmos permitindo assim uma análise mais precisa sobre a movimentação das contas eleitorais. Como consequência, o conjunto de dados final para treinamento dos algoritmos apresenta volumetria igual a 3.066 candidatos a deputado federal.

2.3. K-means

Considere X um conjunto de dados com n elementos e d variáveis. Uma divisão em *clusters* $C = \{C_1, C_2, \dots, C_k\}$ é uma partição de X em subconjuntos C_i (denominados *clusters*) homogêneos, sendo heterogêneos entre si. O algoritmo *K-Means* é uma ferramenta clássica que aplica uma abordagem iterativa para encontrar a divisão em *clusters* que minimiza a soma dos quadrados dos erros (SSE). Esse algoritmo inicia distribuindo aleatoriamente os pontos do conjunto de dados X em k *clusters* e calcula os centroides através da média dos pontos do *cluster* C_i . Em seguida faz-se a atribuição de *clusters*, onde cada ponto $x_i \in X$ é associado ao *cluster* que tem o centroide mais próximo do ponto. Por fim, atualizam-se os centroides através da média de todos os pontos que se encontram no *cluster* C_i . Essas duas etapas são realizadas iterativamente até alcançarem um mínimo local [Forgy, 1965], [Lloyd, 1982] e [Bottou e Bengio, 1994].

2.4. Isolation Forest ou iForest

A técnica *Isolation Forest*, proposta por [Liu et al., 2008], propõe um isolamento explícito das anomalias ao invés de definir perfis normais no conjunto de dados. O uso deste tipo de isolamento permite que este método explore subamostragens, criando um algoritmo com uma baixa complexidade linear e de baixa memória computacional. Além disso, esse algoritmo se mostra muito versátil uma vez que também funciona em problemas de alta dimensão e em situações em

que o conjunto de treinamento não contém anomalias classificadas à priori [Puggini e Mcloone, 2018]. Basicamente, o método utiliza as propriedades quantitativas observadas em anomalias, ou seja, que as anomalias são “poucas e diferentes”, o que as tornam mais suscetíveis ao isolamento do que os pontos normais.

Uma *iTree* é uma árvore binária aleatória com dois tipos de nós: nós externos sem filhos e nós internos com exatamente dois filhos. Seja T um nó interno de uma árvore de isolamento. Um atributo q e um valor de divisão p são selecionados aleatoriamente. O nó T é dividido em dois nós filhos (T_l, T_r) com base no teste $q < p$ até que um nó seja composto por um único elemento. A quantidade de subdivisões para percorrer o caminho do nó inicial até o elemento isolado x é denotada por $h(x)$.

Finalmente, o método constrói um conjunto de *iTrees* (denominado *iForest*) para um determinado conjunto de dados, e em seguida, as anomalias são classificadas como as instâncias que usualmente têm comprimentos de caminho curtos nas *iTrees*. Existem apenas duas variáveis nesse método: o número de árvores a serem construídas e o tamanho das subamostras. Neste trabalho, consideraram-se números de árvores igual a 100 e tamanho mínimo de 256 subamostras. Um fato importante é que o desempenho de detecção do *iForest* converge rapidamente com um número muito pequeno de árvores e exige apenas um pequeno tamanho de subamostras para alcançar um alto desempenho de detecção com alta eficiência [Liu et al., 2012].

As Figuras 1 e 2 ilustram, respectivamente, uma *iTree* e um *iForest*.

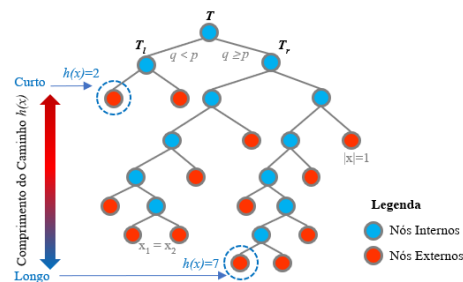


Figura 1: Representação de uma *iTree*. Adaptado de [Wu et al., 2018]. Círculos vermelhos na *iTree* representam valores isolados pelo algoritmo, círculos azuis representam pontos que não foram isolados. Quanto mais curto for o caminho $h(x)$ maior é a chance de o ponto ser uma anomalia.

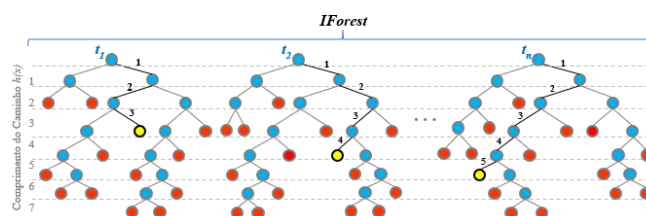


Figura 2: Representação de um *iForest*. Adaptado de [Chen et al., 2016]. O círculo amarelo representa o mesmo ponto isolado em diferentes *iTrees*. Como as partições das árvores são geradas aleatoriamente, cada *iTree* é diferente das demais.

Uma vez aplicado o *iForest*, é possível obter um escore de classificação de anomalias. Um ponto importante é que como as *iTrees* são construídas por subconjuntos de dados divididos por atributos aleatórios, consequentemente os escores de anomalia construídos também são aleatórios.

O escore de anomalia normalizado, proposto em [Liu et al., 2008], contempla informações sobre a frequência de $h(x)$ de forma que maior forem as pontuações de anomalias (próximo de 1) maior serão as chances de serem classificadas como tais.

2.5. Algoritmo combinado *K-Means* e *Isolation Forest* (KM+IF)

A fim de motivar a combinação dos algoritmos *K-Means* e *Isolation Forest*, pense no exemplo lúdico onde haja centenas de gatos pretos misturados a outras centenas de cães brancos. Se no meio destes indivíduos forem acrescentados alguns poucos gatos brancos e cachorros pretos, com certeza seria difícil para um observador humano identificá-los numa rápida visão da cena. Por outro lado, se primeiro fossem separados os cães dos gatos, o observador veria muito mais claramente os animais com cor discrepante ao animal típico do grupo. Intuitivamente, mesmo para uma ferramenta computacional, faz sentido que características anômalas de um grupo, porém comuns a outro, dificultem a identificação de anomalias. Pensando nisto, propõe-se combinar o algoritmo *K-Means* com o *Isolation Forest* para primeiro criar grupos mais homogêneos e então, identificar as anomalias.

Os recentes artigos [Kurnianingsih et al., 2018] e [Gao et al., 2019] já propõem este tipo de abordagem para o problema de identificação de anomalias. Esses artigos mostram que a aplicação conjunta das técnicas traz benefícios para assertividade geral do ajuste. Os autores ainda ressaltam que os resultados experimentais indicam que o algoritmo combinado melhorou a capacidade de identificar anomalias locais e é melhor que os algoritmos tradicionais tanto na precisão quanto na redução de falsos positivos. A presente pesquisa foi realizada de forma independente, tendo sido iniciada antes da publicação destas referências. Algumas das discussões aqui apresentadas não são relatadas nos trabalhos anteriores, sobretudo a respeito de como a distribuição das anomalias pode influenciar o resultado da identificação. Além disto, não foi encontrado nenhum resultado na literatura que abordasse o estudo de anomalias em dados das candidaturas eleitorais, portanto os resultados apresentados constituem uma contribuição original de interesse da sociedade.

Quando o conjunto de dados apresenta uma região relativamente homogênea, com alguns poucos dados discrepantes, é natural que uma série de métodos possam capturar as anomalias com certa facilidade. Entretanto, caso hajam classes disjuntas de dados considerados normais, anomalias que misturam características normais de mais de uma classe são mais dificilmente identificadas. Esta é a situação em que acreditamos existir a vantagem da combinação de um algoritmo de clusterização com um algoritmo de identificação de anomalias. Um agrupamento inicial do conjunto de dados, no nosso caso feito através do *K-Means*, permite que a dimensionalidade do problema seja reduzida, gerando subgrupos homogêneos. Em seguida é possível detectar anomalias de forma mais sensível na região específica de um *cluster*, uma vez que tornamos as anomalias locais em anomalias globais, assim evidenciando o poder do *Isolation Forest*.

Por outro lado, também pode ocorrer uma terceira situação, que não é discutida anteriormente na literatura, que corresponde ao fato das anomalias serem de certa forma semelhantes. Caso o número das anomalias seja pequeno (como deve ser por definição) mas suficientemente significativo em relação a alguma classe associada a uma característica importante de parte dos dados, há a possibilidade que este próprio grupo forme um cluster em uma classificação não supervisionada. Neste contexto pode-se esperar que os resultados do *Isolation Forest* sozinho sejam melhores do que o método combinado, uma vez que a separação de anomalias contidas nesta nuvem de pontos anômalos não é coerente com os propósitos almejados.

3. Estudo de caso com dados de campanhas eleitorais

Neste trabalho aplicamos a técnica proposta em um conjunto de dados referente às campanhas eleitorais ocorridas no Brasil no ano de 2018. A maior motivação para isso é o momento

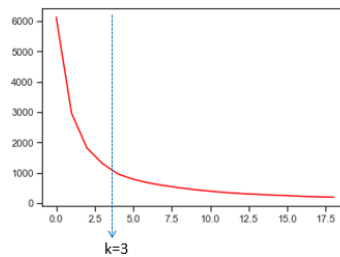


Figura 3: Gráfico do “cotovelo” para definir o número de *clusters*.

de grande instabilidade política ocorrida no país nos últimos anos, a carência de literatura científica relacionada com o tema e a proximidade de novas eleições para cargos federais.

A Lei de Inelegibilidade - Lei Complementar nº 64, de 18 de maio de 1990, determina uma série de situações às quais indivíduos necessitam estar em dia para lançar suas candidaturas, tais como, antecedentes criminais de lavagem de dinheiro, formação de quadrilha, infração do art. 55 da Constituição Federal (situações que podem acarretar na perda do mandato o Deputado ou Senador). Já a Lei nº 9.504, de 30 de setembro de 1997, estabelece regras de arrecadação, coligação, pesquisas eleitorais e, um dos fatores mais importantes, o de prestação de contas após as eleições. Por fim, mais recentemente a Lei nº 13.488, de 6 de outubro de 2017, definiu importantes regras de distribuição das verbas eleitorais, bem como criou tetos de gastos para cada nível de candidatura. Ressalta-se os avanços das leis ao longo dos anos, além de diversas iniciativas de fiscalização do Ministério Público e da Justiça Eleitoral, com consequente punição aos culpados por esses tipos de crimes eleitorais.

A investigação científica realizada neste trabalho tem grande interesse em fornecer ferramentas para mapear e levantar indivíduos que se enquadram em situações atípicas ou anômalas usando apenas dados públicos das campanhas eleitorais. Obviamente que esta contribuição poderia ser expandida e muito mais assertiva caso estivessem disponíveis alguns dados sigilosos dos candidatos, como suas movimentações financeiras e investigações internas dos órgãos competentes.

Vale ressaltar que esse trabalho não visa ser um mecanismo de acusação, apenas um passo na direção do desenvolvimento de instrumentos para que os órgãos responsáveis tenham mais facilidade para identificar prioridades nas investigações. Desta forma não foram apresentados os nomes e partidos relativos ao banco de dados utilizado. O leitor que tenha a curiosidade de consultar estas informações pode facilmente obtê-las acessando o banco de dados do TSE.

3.1. Análise dos resultados: Obtenção dos escores e identificação de variáveis explicativas

Para ajuste do modelo combinado, primeiramente foi aplicado o algoritmo *K-Means*, onde foram utilizadas algumas variáveis que melhor representam as divulgações das contas eleitorais, entre as quais: gastos/despesas, receitas e votos. O número ideal de *clusters*, obtido pelo método heurístico gráfico do “cotovelo” apresentado na Figura 3, foi igual a $k = 3$. Após o agrupamento realizado pelo algoritmo, obteve-se a seguinte volumetria de observações para cada *cluster*: $k_1 = 2.383$, $k_2 = 254$ e $k_3 = 429$. Claramente temos um desbalanceamento entre as classes e, para um conjunto de anomalias da ordem de 10% dos dados, poderemos estar no caso de termos até dois dos *clusters* formados majoritariamente por dados anômalos.

Um fator relevante na construção dos modelos é a definição do ponto de corte para classificação do risco dos indivíduos. Assim, valores acima de um ponto de corte serão considerados anomalias e abaixo serão consideradas observações normais. Utilizou-se como ponto de corte dos algoritmos *Isolation Forest* e *KM+IF* valores do escore normalizado acima do percentil

90% da distribuição dos candidatos. Este ponto de corte resulta, respectivamente, em valores de escore normalizado acima de 0,52 e 0,51.

Os gráficos ilustrados na Figura 4 mostram a associação entre a variável “Despesa Total”, a qual se mostrou uma das mais significativas na análise, e os escores do *Isolation Forest* e do KM+IF. Enquanto é possível ver uma correlação muito clara entre estes atributos para o *Isolation Forest*, um cenário mais complexo acontece em relação ao KM+IF. Dentro do *cluster* k_1 , ainda é possível notar que o algoritmo KM+IF associa fortemente uma anomalia com a “Despesa Total” da candidatura. Já para pontos em k_2 a relação é um pouco menos acentuada e, em k_3 , não existe uma associação muito significativa entre estes atributos.

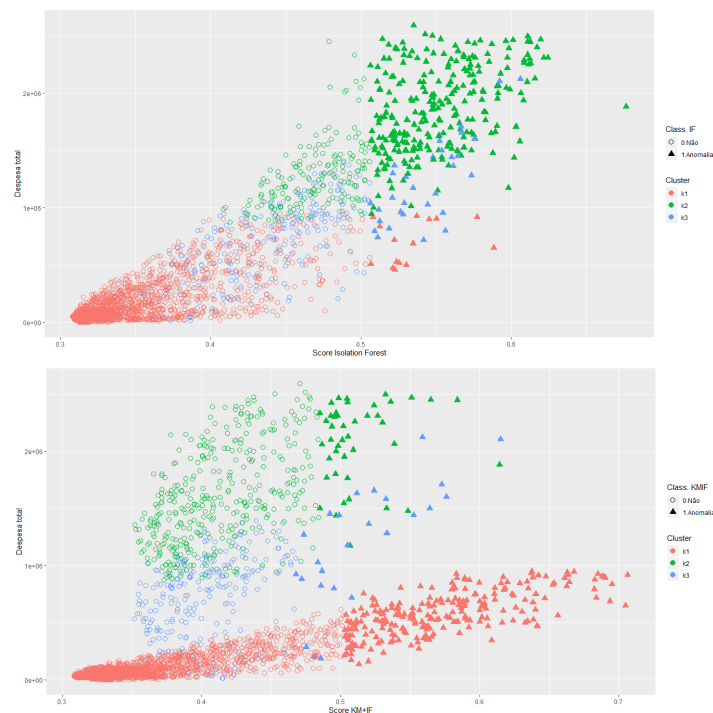


Figura 4: Escore e classificação pelos métodos *Isolation Forest* (superior) e KM+IF (inferior) versus “Despesas Total”. Os círculos representam a classificação dos dados como “normais” e os triângulos dados anômalos em cada método. Os *clusters* k_1 , k_2 e k_3 obtidos pelo método combinado, portanto não utilizados na primeira classificação, são representados, respectivamente, pelas cores vermelha, verde e azul.

3.2. Análise de desempenho dos algoritmos: Construindo uma variável resposta

Para avaliar os resultados obtidos pelo algoritmo, seria muito importante ter a análise de algum órgão regulador para toda a base de candidatos. Como isso não é possível, uma vez que essa informação não está disponível, foi realizada uma avaliação com dados disponíveis pelos tribunais de justiça. Dessa forma, construiu-se uma variável contendo todos processos de crimes eleitorais existentes pelos candidatos. A Tabela 1 mostra a distribuição de candidatos com algum processo de crime na justiça, exibindo separadamente a volumetria do histórico de processos e dos processos ativos em outubro de 2019. Além disso consideram-se duas divisões acerca do tipo de processo: outros – processos oriundos à natureza tributária, trabalhista cível/administrativa; e eleitoral – de crimes exclusivos de processos eleitorais.

Tabela 1: Distribuição de Processos de Crimes Eleitorais dos Candidatos

Possui Processos?	Histórico		Ativos(out/19)	
	Quantidade de candidatos	%	Quantidade de candidatos	Quantidade de processos
Sim, outros	774	25	772	3318
Sim, eleitoral	1212	40	97	404
Não	1080	35	-	-
Total	3066	100	97	3722

Considerando todo o histórico dos dados, 1.212 dos 3.066 candidatos a deputado federal considerados na base de dados utilizada neste trabalho, possuem processos de crimes eleitorais, totalizando mais de 10.000 processos associados a eles. Observa-se também que há 97 candidatos com 404 processos de crimes eleitorais ativos na referência de outubro de 2019, ou seja, uma redução considerável se observarmos o período histórico. A hipótese para que isso ocorra é que processos eleitorais precisam ser tramitados rapidamente para decidir sobre impugnação ou aprovação das candidaturas e da posse após as eleições. Além disso, nota-se que 18% da base de candidatos com algum processo de crime eleitoral (220 candidatos) possuem 10 ou mais processos nessa categoria.

Com estes dados foi construída a variável de interesse para avaliação dos resultados dos modelos. Ela é baseada na combinação dos candidatos que possuem processos de crimes eleitorais ativos com os candidatos cuja quantidade total de processos de crimes eleitorais históricos é significativa (nesse caso a escolha foi 10 ou mais processos). A justificativa para isso é que a recorrência de processos de crimes eleitorais indica um perfil de elevado risco do indivíduo. Outro fator que corrobora com essa decisão é que um processo aberto em 2018 já pode ter sido encerrado e, nesse caso, não estaria ativo no mês de referência de out/19; assim seriam perdidas informações relevantes na análise de performance. Dessa forma, o número total de candidatos classificados com perfil de alto risco para compor a variável de interesse é igual a 270 observações (valores em negrito na Tabela 2 - 50 candidatos apenas com processos de crimes eleitorais ativos em out/19; 173 candidatos com apenas processos de crimes eleitorais históricos; e 47 candidatos em ambas as situações).

Tabela 2: Variável de Interesse: Processos de Crimes Eleitorais Totais (270 candidatos)

Processos Históricos ≥ 10	Processos Ativos(out/19)		
	Não	Sim	Totais
Não	2796	50	2846
Sim	173	47	220
Total	2969	97	3066

Utilizando-se a variável de interesse recém construída, “Processos Eleitorais Totais”, o interesse agora é avaliar os resultados de performance capturados pelos métodos *Isolation Forest* (IF) e *K-Means + Isolation Forest* (KM+IF). A Tabela 3 contém as métricas de performance usuais para avaliação dos algoritmos: Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR), Precisão, Acurácia e F-score [Zhu et al., 2017], aplicadas por exemplo em [Araujo et al., 2019] e [Junior et al., 2016].

Nota-se pela Tabela 3 que os resultados indicam que o método *Isolation Forest* sozinho apresenta indicadores superiores quando comparado ao método combinado (KM+IF).

Analisando as métricas uma a uma, observa-se que no IF a volumetria de casos anômalos

Tabela 3: Performance dos Algoritmos Aplicados nas Contas Eleitorais de 2018

Algoritmo	Medidas de performance(%)				
	TPR	FPR	Precisão	Acurácia	F-score
IF	30.00	8.00	26.50	86.50	28.10
KM+IF	15.90	9.40	14.10	84.00	14.90

capturados é quase duas vezes maior (30,0% contra 15,9%) que KM+IF. Além disso, o algoritmo mostrou-se levemente superior na classificação de casos de não risco como anomalias, apresentou praticamente o dobro da captura das anomalias verdadeiras sobre o total de anomalias classificadas pelos modelos (26,5% contra 14,1%) e acurácia levemente superior no resultado. A acurácia revela o quanto o conjunto de dados está classificado corretamente através dos Verdadeiros Positivos e Verdadeiros Negativos sobre o total de observações. Como a proporção de casos “normais” é alta (conjunto de dados desbalanceados), acaba beneficiando o resultado do KM+IF. Já o resultado do F-score é praticamente o dobro no algoritmo IF. Como ele combina a TPR e a Precisão através de uma média harmônica e esses resultados se mostram superiores, já era esperado que o F-score também fosse melhor. De maneira geral, em termos de qualidade de ajuste do modelo e baseado na variável de interesse construída para embasamento de performance, considera-se o resultado do *Isolation Forest* apenas como razoável. O resultado de captura de eventos de risco (anomalias) é de 30,0% da base total. Se, por exemplo, os órgãos fiscalizadores fossem analisar de forma aleatória a mesma quantidade de contas eleitorais propostas no ponto de corte do modelo (10% dos indivíduos), o resultado de captura seria igual a 8,8%, que é a taxa de classificação de anomalias da base considerando a variável resposta “Processos Totais”. Resumidamente, embora o índice de identificação não seja elevado, o algoritmo é capaz de capturar quase 3,5 vezes mais anomalias nesse público quando comparado à uma escolha aleatória. Já o resultado do algoritmo combinado KM+IF mostrou desempenho insatisfatório, com índices por volta da metade do IF sozinho.

Analizando em detalhes o motivo do método combinado ter tido desempenho inferior ao IF sozinho, há uma hipótese de que essa performance esteja diretamente associada ao nível de acerto do algoritmo *K-Means* na identificação de um *cluster* formado pelas anomalias, conforme nossa discussão inicial. Uma forma de examinar essa suposição é observar a proporção de eventos anômalos (“Processos Totais”) no conjunto de dados e compará-lo à proporção de anomalias de cada *cluster*, disponível na Tabela 4.

Verifica-se que as taxas de anomalias (“Processos Totais”) de cada *cluster* são muito diferentes. Enquanto o *cluster* k_1 apresenta um baixo valor, igual a 4%, os *clusters* k_2 e k_3 apresentam, respectivamente, 28% e 24% de taxa de anomalias. Ou seja, conclui-se que a segmentação via método *K-Means* além de dividir o conjunto de dados em grupos homogêneos entre si, também separou o público com maior concentração de anomalias. Isso acontece porque os eventos de processos de crimes eleitorais devem estar associados a algumas características identificadas na construção não supervisionada dos *clusters*.

Com intuito de avaliar o resultado do método combinado KM+IF dentro de cada *cluster*, na Tabela 5 estão disponibilizados os indicadores de métricas de performance individualmente para k_1 , k_2 e k_3 . Embora os resultados estejam apresentando os índices para cada *cluster* separadamente, eles se referem aos mesmos testes realizados anteriormente, ou seja, envolvendo todos os dados disponíveis. Isso é feito para que se possa observar melhor como foi o desempenho dos algoritmos em cada partição de dados. A proposta é entender a relação entre os resultados do algoritmo e um potencial benefício de seu uso como parte da solução.

Tabela 4: Taxa de Risco em Cada Cluster e no Total

Cluster	Processos Totais	Quantidade	%
k_1	Não	2292	96
	Sim	91	4
	Subtotal	2383	100
k_2	Não	310	72
	Sim	119	28
	Subtotal	429	100
k_3	Não	194	76
	Sim	60	24
	Subtotal	254	100
Total	Não	2796	91
	Sim	270	9
	Total	3066	100

Tabela 5: Performance dos Algoritmos Aplicados Para os Clusters k_1 , k_2 e k_3

Cluster	Algoritmo	Medidas de performance(%)				
		TPR	FPR	Precisão	Acurácia	F-score
k_1	IF	2.20	0.60	13.30	95.70	3.80
	KM+IF	30.80	9.20	11.80	88.50	17.00
k_2	IF	57.10	61.60	26.30	43.60	36.00
	KM+IF	10.10	10.00	27.90	67.80	14.80
k_3	IF	18.30	10.80	34.40	72.40	23.90
	KM+IF	5.00	11.30	12.00	68.90	7.10

Através da Tabela 5 é possível notar que o método combinado KM+IF apresenta melhores resultados para o cluster k_1 . Em contrapartida, para os clusters k_2 e k_3 prevaleceu o melhor ajuste do algoritmo *Isolation Forest*. Ressalta-se que as métricas de performance mudaram consideravelmente quando comparadas individualmente para cada cluster. Por exemplo, em k_1 a TPR é 14 vezes maior no método KM+IF quando comparado ao IF (30,8% contra 2,2%) e, além disso, o F-score é 4,5 vezes maior para o método combinado (17,0% contra 3,8%). Na prática, o método IF não está capturando quase nenhuma anomalia nesse grupo quando aplicado de forma geral no conjunto de dados. Já para os clusters k_2 e k_3 , os resultados do método *Isolation Forest* continuam sendo melhores do que KM+IF, entretanto com uma diferença de performance mais acentuada (4 a 5 vezes maior em termos de TPR e 3 vezes maior em relação ao F-score), ou seja, há um indicativo de que a concentração de anomalias presentes nesses clusters está refletindo nos desempenhos do algoritmo, visto que o próprio *K-Means* identificou partições correlacionadas à variável de interesse “Processos Totais”.

No conjunto de dados identificados pelo cluster k_1 , foram separados os candidatos que formam um grupo grande que não têm perfil de risco elevado, o que é indicado pelo percentual de dados anômalos ser muito baixo. Com isso, as anomalias do conjunto se destacam mais dentro do grupo e, portanto, são mais facilmente identificadas; fazendo com que o método combinado se saia melhor neste grupo. Já para os outros dois clusters a taxa de anomalias é muito alta, 28% e 24%, respectivamente. Isso faz com que o conjunto de dados anômalos (do conjunto completo) seja uma quantidade considerável do subconjunto, não sendo, portanto, anômalos dentro daquele grupo; o

que implica no baixo desempenho do método nos *clusters* k_2 e k_3 . Como existe quase o dobro de anomalias nesses *clusters* (k_2 e k_3) do que em k_1 , o desempenho pior nos dois últimos *clusters* se sobrepõem ao desempenho do *cluster* k_1 o que se reflete em um desempenho geral pior no método KM+IF do que no IF sozinho.

4. Conclusão

Este artigo investiga a viabilidade do uso de uma combinação dos métodos Isolation Forest com o K-Means para identificação de anomalias em bancos de dados. Esse tipo de método híbrido foi proposto recentemente [Kurnianingsih et al., 2018] e [Gao et al., 2019], entretanto suas características ainda não estão totalmente conhecidas. Neste trabalho mostra-se que a combinação dos métodos pode influenciar negativamente os resultados quando algum dos *clusters* é constituído por uma quantidade elevada de dados considerados anômalos. Esta observação não havia sido relatada anteriormente na literatura e apresenta uma contribuição original para este ramo da ciência.

Outro ponto central desse artigo é a identificação de anomalias em dados de campanhas eleitorais. Neste sentido foi exposto que a variável “Despesas Totais” (realizadas pelos candidatos nas eleições de 2018) apresentou a maior contribuição para os escores de anomalia tanto para o *Isolation Forest* aplicado sozinho quanto para o método híbrido KM+IF. Buscando uma possível associação com alguma atividade ilícita dos candidatos, construiu-se uma variável relativa aos processos de crimes eleitorais que esses indivíduos possuem. Os algoritmos utilizados não obtiveram altos níveis de identificação compatível com a variável resposta, embora apresentem uma vantagem significativa frente a um processo aleatório. Estes resultados se justificam pelo fato de a variável resposta não estar fortemente correlacionada às “Despesas Totais” do candidato na campanha.

O algoritmo combinado teve um desempenho pior do que o *Isolation Forest* sozinho, o que pode ser justificado pelo fato do *K-Means* ter construído dois *clusters* com uma taxa elevada de anomalias, corroborando com nossa observação anterior. Infelizmente os autores deste trabalho não conseguiram identificar os fatores que levaram o *K-Means* a construir o *cluster* com poucas anomalias. Perceber isso, ajudaria a ver quais as relações mais propensas a indicar um dado normal e, consequentemente, por exclusão, uma possível atividade irregular. Com estas informações seria possível concentrar esforços onde realmente há mais indícios de atividades suspeitas. Uma observação curiosa é que, embora a variável “Despesas Totais” tenha pouca correlação com a variável resposta, para o *cluster* formado com maior porcentagem de dados normais, a correlação é mais significativa. Uma possibilidade de pesquisa futura seria o uso de técnicas de inteligência artificial tanto para obter um processo de classificação já envolvendo a variável resposta quanto para obter maiores informações sobre como os *clusters* foram formados pelo *K-Means*.

Por fim, um comentário interessante sobre os resultados apresentados neste tipo de trabalho é que, como não existe (e nunca vai existir) uma variável resposta infalivelmente associada a um comportamento impróprio, é impossível uma verificação completamente analítica sobre o desempenho de algoritmos neste tipo de aplicação. Por exemplo, no caso aqui apresentado, talvez os algoritmos estejam apontando uma anomalia associada a uma irregularidade que ainda não foi detectada ou comprovada pelas instâncias jurídicas e/ou investigativas. Outra possibilidade é que de fato estejamos capturando muitas anomalias que não estão associadas a ações criminosas. Uma possibilidade que também deve ser considerada é que, no grupo de candidatos considerados, comportamentos inapropriados podem não ser de fato raros e portanto uma técnica de detecção de anomalias não seria útil. De toda forma, parece ser evidente que novas pesquisas a cerca de métodos quantitativos sobre este tema devem ser considerados para corresponder aos anseios da sociedade.

Referências

- Alley, P. (2011). Corruption: a root cause of deforestation and forest degradation. in Global corruption report 2011: climate change, 6:299–311. [Online]. Available: <https://www.transparency.org/en/>.
- Araujo, B., Almeida, H., e Mello, F. (2019). Computational intelligence methods applied to the fraud detection of electric energy consumers. *IEEE Latin America Transactions*, 17:71–77.
- Bottou, L. e Bengio, Y. (1994). Convergence properties of the k-means algorithms. *Advances in Neural Information Processing System*, p. 585–592.
- Chen, W., Yun, Y. H., Wen, M., Lu, H. M., Zhang, M., e Liang, Y. Z. (2016). Representative subset selection and outlier detection via isolation forest. In *Analytical Methods*, volume 8.
- Edgeworth, F. Y. (1887). Xli. on discordant observations. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 23:364–375.
- Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 21:768–769.
- Gao, R., Zhang, T., Sun, S., e Liu, Z. (2019). Research and improvement of isolation forest in detection of local anomaly points. In *Journal of Physics: Conference Series*, volume 1237.
- Junior, P., Castro, L., e Zambalde, A. (2016). Applying textmining to classify news about supply and demand in the coffee market. Technical report, IEEE Latin America Transactions.
- Kurnianingsih, K., Nugroho, L., Widyawan, W., Lazuardi, L., e Prabuwo, A. S. (2018). Detection of anomalous vital sign of elderly using hybrid k-means clustering and isolation forest. In *IEEE Region 10 Conference*.
- Liu, F. T., Ting, K. M., e Zhou, Z. H. (2008). Isolation forest. In *Eighth IEEE International Conference on Data Mining*, p. 413–422.
- Liu, F. T., Ting, K. M., e Zhou, Z. H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery From Data - TKDD*, 6:1–39.
- Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137.
- Parmar, J. D. e Patel, J. T. (2017). Anomaly detection in data mining: A review. *International Journal of Advanced Research in Computer Science and Software Engineering*, 7:32–40.
- Pfeiffer, S. (2004). Vote buying and its implications for democracy: evidence from latin america. in Global corruption report 2004: political corruption, 5:76–79. [Online]. Available: <https://www.transparency.org/en/>.
- Puggini, F. e Mcloone, S. (2018). An enhanced variable selection and isolation forest based methodology for anomaly detection with oes data. *Engineering Applications of Artificial Intelligence*, 67:126–135.
- Ribeiro, H. V., Martins, A., Lenzi, E., e Perc, M. (2018). The dynamical structure of political corruption networks. *Journal of Complex Networks*, 6:989–1003.

- Saibro, A. L. F. (2006). Brazil. in Global corruption report 2006: corruption and health, 8:133–135 [Online]. Available: <https://www.transparency.org/en/>.
- Transparency International (2018). Corruption perceptions index. [Online]. Available: <https://www.transparency.org/cpi2018/results>.
- Wu, T., Zhang, Y. J., e Tang, X. (2018). Isolation forest based method for low-quality synchrophasor measurements and early events detection. In *IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids*.
- Zhu, X., Tao, H., Wu, Z., Cao, J., Kalish, K., e Kayne, J. (2017). Fraud prevention in online digital advertising. *Springer Briefs in Computer Science*, p. 39–44.