Dissertação apresentada à Pró-Reitoria de Pós-Graduação e Pesquisa do Instituto Tecnológico de Aeronáutica e da Universidade Federal de São Paulo, como parte dos requisitos para obtenção do título de Mestre em Ciências no Programa de Pós-Graduação em Engenharia de Produção, Área de Pesquisa Operacional.

Mateus Vendramini Polizeli

APLICAÇÃO DE ALGORITMOS NÃO SUPERVISIONADOS EM DADOS ELEITORAIS

Dissertação aprovada em sua versão final pelos abaixo assinados:

Prof. Dr. Luís Felipe C. da R. Bueno

Orientador

São José dos Campos, SP - Brasil
2019

Dados Internacionais de Catalogação-na-Publicação (CIP)
Divisão de Informação e Documentação

Polizeli, Mateus Vendramini
Aplicação de algoritmos não supervisionados em dados eleitorais / Mateus Vendramini Polizeli. São José dos Campos, 2019.
95f.

Dissertação de mestrado – Programa de Pós-Graduação de Engenharia de Produção. Área de Pesquisa Operacional – Instituto Tecnológico de Aeronáutica e Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo, 2019. Orientador: Luís Felipe Cesar da Rocha Bueno.

1. Processamento de dados. 2. Anomalias. 3. Eleição. 4. Detecção. I. Instituto Tecnológico de Aeronáutica. II. Universidade Federal de São Paulo. III. Título.

REFERÊNCIA BIBLIOGRÁFICA

POLIZELI, Mateus Vendramini. **Aplicação de algoritmos não supervisionados em dados eleitorais**. 2019. 95f. Dissertação de mestrado em Engenharia de Produção, Área de Pesquisa Operacional – Instituto Tecnológico de Aeronáutica e Universidade Federal de São Paulo, São José dos Campos.

CESSÃO DE DIREITOS

NOME DO AUTOR: Mateus Vendramini Polizeli
TÍTULO DO TRABALHO: Aplicação de algoritmos não supervisionados em dados eleitorais
TIPO DO TRABALHO/ANO: Dissertação / 2019

_____

Mateus Vendramini Polizeli
Rua João Álvares Correia, 111
04.115-030 – São Paulo - SP

APLICAÇÃO DE ALGORITMOS NÃO SUPERVISIONADOS EM DADOS ELEITORAIS

Mateus Vendramini Polizeli

Composição da Banca Examinadora:

Profa. Dra.	Ana Carolina Lorena	Presidente	-	ITA
Prof. Dr.	Luís Felipe Cesar da Rocha Bueno	Orientador	-	UNIFESP
Prof. Dr.	Rafael Izbicki	Membro Externo	-	UFSCar
Prof. Dr.	Weldon Alexander Lodwick	Membro Externo	-	UNIFESP

ITA/UNIFESP

Mateus Vendramini Polizeli

APLICAÇÃO DE ALGORITMOS NÃO SUPERVISIONADOS EM DADOS ELEITORAIS

Dissertação apresentada à Universidade Federal São Paulo como requisito parcial para obtenção do título de Mestre em Ciências.

Área de Concentração: Engenharia de Produção, Área de Pesquisa Operacional.

Aprovada em 10 de dezembro de 2019.

Presidente da Banca:

Prof. Dr. Luís Felipe C. da R. Bueno 

Banca Examinadora:

Prof. Dr. Weldon Alexander Lodwick

Prof. Dr. Rafael Izbicki

Profa. Dra. Ana Carolina Lorena

"Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world." — Albert Einstein

Resumo

Diante da busca incessante da sociedade por clareza nos gastos públicos, eficiência na gestão e transparência com uso da máquina pública, torna-se relevante a estruturação de trabalhos que possibilitem uma apuração aprofundada para acompanhamento eficiente dessas ações. A partir de um estudo inicial na literatura, verificou-se a existência de uma série de controles e divulgação de prestação de contas de setores e órgãos públicos. Contudo, apesar de iniciativas como essas, ainda há poucos trabalhos considerando uma investigação mais aprofundada para capturar possíveis irregularidades do meio político. Dessa forma, o objetivo deste projeto é estudar alguns mecanismos de detecção de anomalias associados ao conjunto de dados das candidaturas eleitorais de 2018. As metodologias propostas são baseadas nos algoritmos não supervisionados *K-Means* e *Isolation Forest* como tentativa de criar uma ferramenta de apoio à tomada de decisão para os reguladores, visando direcionar os recursos humanos para investigação. É sugerida também uma combinação desses algoritmos, denominado aqui como *KM+IF*, com intuito de melhorar a acurácia e diminuir as taxas de erro associadas aos modelos. Os resultados observados neste projeto indicam que a proposta *KM+IF* mostra bom desempenho para situações onde estão disponíveis as variáveis de interesse. Entretanto, pode apresentar resultados insatisfatórios quando tais não estão disponíveis. No estudo de caso realizado para o conjunto de candidaturas eleitorais, o resultado geral do algoritmo *KM+IF* foi inferior ao resultado individual das técnicas *K-Means* e *Isolation Forest*.

Abstract

Given the incessant search of society for clarity in government spending, management efficiency and transparency using the public agency, the structuring of works that allow a thorough investigation to efficiently monitor these actions becomes relevant. From an initial study in the literature, it was verified the existence of a series of controls and disclosure of accountability of sectors and public agencies. However, despite initiatives such as these, there is still little work considering further investigation to capture possible irregularities in the policy instrument. Thus, the objective of this project is to study some mechanisms for detecting anomalies associated with the 2018 electoral candidate data set. The proposed methodologies are based on unsupervised algorithms *K-Means* and *Isolation Forest* in an attempt to create a decision support tool for regulators to direct human resources for research. A combination of these algorithms, referred to here as *KM+IF*, is also suggested in order to improve accuracy and decrease the error rates associated with the models. The results observed in this project indicate that the proposal *KM+IF* shows good performance for situations where the variables of interest are available. However, it may yield unsatisfactory results when they are not available. In the case study for the set of electoral candidates, the overall result of the *KM+IF* algorithm was lower than the individual result of the *K-Means* and *Isolation Forest* techniques.

Lista de Figuras

FIGURA 2.1 – Anomalia Pontual. Fonte: (CHANDOLA *et al.*, 2009) 24

FIGURA 2.2 – Exemplo de Anomalia Contextual. Fonte: (CHANDOLA *et al.*, 2009). 28

FIGURA 2.3 – Exemplo de Anomalia Coletiva. Contração atrial prematura. Fonte: (CHANDOLA *et al.*, 2009). 29

FIGURA 3.1 – Partições necessárias para separarem um ponto normal x_i (à esquerda) e um ponto anômalo x_0 (à direita). Fonte:(LIU *et al.*, 2008). 45

FIGURA 3.2 – Comprimentos médios de convergência para x_i e x_0 quando o número de árvores aumenta. Fonte: (LIU *et al.*, 2008). 46

FIGURA 3.3 – Representação de uma *iTree*. 47

FIGURA 3.4 – Exemplo ilustrativo de construção de uma *iTree*. 48

FIGURA 3.5 – Representação de um *iForest*. 48

FIGURA 3.6 – Relação do comprimento esperado do caminho $E(h(x))$ e pontuação s da anomalia. Fonte: (LIU *et al.*, 2008). 50

FIGURA 3.7 – Contagem de anomalias do *iForest* para uma distribuição normal de 64 observações. Fronterias de contorno para $s = 0,5; 0,6; 0,7$ são ilustradas. Fonte: (LIU *et al.*, 2008). 51

FIGURA 3.8 – Pseudo algoritmo *iForest*. Fonte: (LIU *et al.*, 2008) 52

FIGURA 3.9 – Pseudo algoritmo *iTree*. Fonte: (LIU *et al.*, 2008) 52

FIGURA 3.10 –Pseudo algoritmo do comprimento do caminho. Fonte: (LIU *et al.*, 2008) 53

FIGURA 3.11 –Desafios e complexidade na detecção de anomalias. Fonte: (CHANDOLA *et al.*, 2009). 53

FIGURA 4.1 – Imagem das três espécies de Iris (Setosa, Virgínica e Versicolor). . . 57

FIGURA 4.2 – Box-plot do conjunto de dados Iris. Comprimento e Largura das Sépalas e Pétalas. Há presença de 4 *ouliers* na largura das sépalas. . 57

FIGURA 4.3 – Comparação entre classes originais e após aplicação do método *K-Means* com k=3. 58

FIGURA 4.4 – *Outliers* simulados para aplicação do modelo *K-Means*. 58

FIGURA 4.5 – Classificação de Anomalias após ajuste do modelo *K-Means*. 59

FIGURA 4.6 – Classificação de Anomalias após ajuste do modelo *Isolation Forest*. . 59

FIGURA 4.7 – Classificação de Anomalias após ajuste do modelo combinado *KM+IF*. 60

FIGURA 4.8 – Gráfico do "cotovelo" para definir o número de *clusters* (à esquerda) e Volumetria de cada *cluster* gerado (à direita). 65

FIGURA 4.9 – Distribuição do Score dos Candidatos e marcação do percentil 90%. 66

FIGURA 4.10 –Distribuição do score *Isolation Forest* e variáveis: bens de alto risco, custo por voto e votos total (à esquerda); bens total, inadimplência e despesa total (à direita) 66

FIGURA 4.11 –Distribuição do score do modelo combinado *K-Means + Isolation Forest* e variáveis: bens de alto risco, custo por voto e votos total (à esquerda); bens total, inadimplência e despesa total (à direita) . . . 67

FIGURA 4.12 –Score do modelo *Isolation Forest* x Despesas Total 73

FIGURA 4.13 –Score do modelo *Isolation Forest* x Votos Total 74

FIGURA 4.14 –Score do modelo *Isolation Forest* x Custo por Voto 74

FIGURA 4.15 –Score do modelo combinado *KM+IF* x Despesas Total 75

FIGURA 4.16 –Score do modelo combinado *KM+IF* x Votos Total 75

FIGURA 4.17 –Score do modelo combinado *KM+IF* x Custo por Voto 76

FIGURA 4.18 –Score e Classificação do modelo *Isolation Forest* x Despesas Total . 76

FIGURA 4.19 –Score e Classificação do modelo *Isolation Forest* x Votos Total . . . 77

FIGURA 4.20 –Score e Classificação do modelo *Isolation Forest* x Custo por Voto . 77

FIGURA 4.21 –Score e Classificação do modelo *KM+IF* x Despesas Total 78

FIGURA 4.22 –Score e Classificação do modelo *KM+IF* x Votos Total 78

FIGURA 4.23 –Score e Classificação do modelo *KM+IF* x Custo por Voto 79

FIGURA 4.24 –Comparação de Processos Totais e Classificação do modelo *IF* x Despesas Total 79

FIGURA 4.25 –Comparação de Processos Totais e Classificação do modelo *KM+IF* x Despesas Total 80

FIGURA 4.26 –Comparação de Processos Totais e Classificação do modelo *IF* x Votos Total 80

FIGURA 4.27 –Comparação de Processos Totais e Classificação do modelo *KM+IF* x Votos Total 81

FIGURA 4.28 –Comparação de Processos Totais e Classificação do modelo *IF* x Custo por Voto 81

FIGURA 4.29 –Comparação de Processos Totais e Classificação do modelo *KM+IF* x Custo por Voto 82

FIGURA 4.30 –Comparação entre os modelos *IF e KM+IF* x Classificação IF . . . 82

FIGURA 4.31 –Comparação entre os modelos *IF e KM+IF* x Classificação KMIF . 83

FIGURA A.1 –Pseudo algoritmo do método *K-Means*. Fonte: [Nunes 2016] 95

Lista de Tabelas

TABELA 3.1 – Matriz de Confusão. 54

TABELA 4.1 – Distribuição de variáveis de Despesa (R\$), Receita (R\$) e Votos dos Candidatos. 64

TABELA 4.2 – Distribuição de Candidatos e Processos Judiciais. 68

TABELA 4.3 – Percentuais dos Candidatos com Processos de Crimes Eleitorais. . . 68

TABELA 4.4 – Quantidade de Processos de Crimes Eleitorais Históricos dos Candidatos. 68

TABELA 4.5 – Variável de interesse: Processos Eleitorais Totais (270 candidatos). . 69

TABELA 4.6 – Performance dos Algoritmos aplicados às Contas Eleitorais. 69

TABELA 4.7 – Taxa de Risco em cada *Cluster* e no Conjunto de Dados Total. . . . 71

TABELA 4.8 – Performance dos Algoritmos aplicados para o *Cluster* k_1. 71

TABELA 4.9 – Performance dos Algoritmos aplicados para o *Cluster* k_2. 71

TABELA 4.10 –Performance dos Algoritmos aplicados para o *Cluster* k_3. 72

Lista de Abreviaturas e Siglas

API	*Application Programming Interface*
AUC	*Area Under the Curve*
BIRCH	*Balanced Iterative Reducing and Clustering Using Hierarchies*
BST	*Binary Search Tree*
COAF	Conselho de Controle de Atividades Financeiras
CPI	*Corruption Perceptions Index*
CURE	*Clustering Using Representatives*
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
DBCLASD	*Distribution-Based Clustering of Large Spatial Databases*
FPR	*False Positive Ratio*
GMM	*Gaussian Mixture Models*
IFOREST	*Isolation Forest*
ITREE	*Isolation Tree*
KNN	*K-Nearest Neighbors*
OPTICS	*Ordering Points to Identify the Clustering Structure*
RIPPER	*Repeated Incremental Pruning to pProduce Error Reduction*
RNA	Redes Neurais Artificiais
SSE	*Sum of Squared Errors*
SVM	*Support Vector Machine*
TM	*Text Mining*
TPR	*True Positive Ratio*
TSE	Tribunal Superior Eleitoral
UIF	Unidade de Inteligência Financeira

Lista de Símbolos

C	um conjunto de *clusters* ou partição de um conjunto de dados
D	um conjunto de dados com n pontos num espaço de dimensão d
z	ponto que representa o centroide
k	o número de *clusters*
x	uma observação (ou um ponto de dados)
X	um conjunto de dados de n instâncias
X'	um conjunto de dados selecionado aleatoramente sem reposição de X
n	número de observações em um conjunto de dados, $n = \|X\|$
m	índice das observações x_m, $m \in 0, ..., n-1$
Q	um conjunto de atributos
d	número de atributos, $d = \|Q\|$
q	um atributo
T	uma árvore ou um nó
T_l	um nó filho l
T_r	um nó filho r
t	número de árvores
$h(x)$	comprimento do caminho de x
$hlim$	limite de altura de avaliação
ψ	tamanho da subamostragem
l	um caminho possível
s	uma pontuação ou score de anomalia

Sumário

1 Introdução 17

1.1 Objetivo 19

1.2 Metodologia 19

1.2.1 Levantamento Bibliográfico 20

1.2.2 Coleta de Dados 20

1.2.3 Combinação de Técnicas 21

1.3 Organização do trabalho 22

2 Definição e Abordagem de Detecção de Anomalias . . 23

2.1 Desafios na Detecção de Anomalias 24

2.2 Aspectos do problema de detecção de anomalias 26

2.2.1 Natureza dos dados de entrada 26

2.2.2 Tipos de Anomalia 27

2.2.3 Variável de Interesse 30

2.2.4 Resultado da detecção de anomalia 32

2.3 Aplicações de detecção de anomalia 32

2.3.1 Detecção de invasão 33

2.3.2 Detecção de fraude 33

2.3.3 Detecção de anomalias médicas e de saúde pública 35

2.3.4 Detecção de danos industriais 36

3 Métodos Utilizados para Detecção de Anomalias . . . 37

3.1 Métodos de Classificação 38

3.2 Métodos de Agrupamento 40

3.2.1 Algoritmo *K-Means* 41
3.3 Método *Isolation Forest* 44
3.4 Método Combinado *K-Means+Isolation Forest* 53
3.4.1 Métricas de Performance 54
4 Estudo de Caso 56
4.1 Experimento Inicial 56
4.2 Estudo de Caso Principal 60
4.2.1 Descrição do Problema 61
4.2.2 Fontes de Informação 62
4.2.3 Estrutura do Conjunto de Dados 63
4.2.4 Análise Descritiva 64
4.2.5 Construção do Modelo 64
4.2.6 Construção de uma Variável de Interesse 67
4.2.7 Resultados 69
5 Conclusão e Perspectivas Futuras 84
Referências 86
Apêndice A – Pseudocódigo dos Algoritmos 95

1 Introdução

Há algumas décadas, a sociedade tem vivido com maior desconfiança e falta de esperança nos políticos que os representam e lideram o país. Em (LAZZARI, 2016) é fornecida uma explicação para o quadro de ampla desconfiança em partidos políticos no Brasil. Além disso, alguns eventos mais recentes, como operações de investigação da Polícia Federal e análise do Ministério Público mostra-se 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. Em (RIBEIRO *et al.*, 2018) foi construída uma estrutura dinâmica das redes de corrupção política no país, onde mostram que elas podem ser usadas para prever com êxito futuros escândalos políticos.

Alinhado a isso, algumas pesquisas realizadas pela *Transparency International*, (TI, 2004), (TI, 2006) e (TI, 2011) apontam o Brasil com resultados regulares a muito ruins no *ranking* de países com maior corrupção nos setores públicos (político, saúde e mudança do clima). Além disso, a mais recente atualização de seu índice de percepção da corrupção (CPI), de 2018 (TI, 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). São vários os fatores que contribuem para essa avaliação, como baixo índice de desenvolvimento humano, escândalos políticos, mão de obra escrava e infantil, tráfico de pessoas e animais silvestres, destruição ambiental, aumento da ineficiência do sistema político e econômico, etc.

Por conta desses fatores, é cada vez mais importante que a população tenha acesso a um maior número de informações e ferramentas como fontes de pesquisa para cobrança daqueles que a representam no meio político e, obviamente, para fundamentar sua decisão nas urnas a cada ano eleitoral. Além disso, através de mais referências e notícias, a sociedade tem a possibilidade de acentuar e estender sua visão crítica para o meio privado, como indústrias, instituições financeiras ou empresas de serviços que estejam diretamente ligados a episódios de corrupção ou desvios de conduta junto ao meio político.

Neste sentido, este projeto visa estudar uma ferramenta que contemple uma abordagem de riscos, ou seja, um método que auxilie o julgamento de órgãos públicos que apuram

irregularidades no meio. Isso será feito através de algoritmos computacionais que buscam indicar fatores mais relevantes em processos com padrão fora da normalidade e onde devem ser alocados os recursos para uma possível investigação. Diante do fato desse processo depender de pessoas extremamente competentes, considerando sua natureza delicada e a necessidade de sua condução ser realizada com cautela e precisão, faz-se necessário este tipo de instrumento de apoio à tomada de decisão, aumentando a eficiência no uso dos recursos disponíveis.

A proposta deste trabalho é estudar uma maneira de construir um dispositivo de alertas para investigação, e, de forma alguma, o intuito será gerar acusações para os riscos levantados. Sendo assim, por meio de informações públicas disponibilizadas pelo próprio governo federal, mais especificamente do TSE (Tribunal Superior Eleitoral), serão investigadas incompatibilidades nos bancos de dados referentes às prestações de contas realizadas pelos candidatos políticos nas eleições de 2018.

A abordagem escolhida 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; 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.

A importância da detecção de anomalias deve-se ao fato delas estarem associadas a informações relevantes e significativas e, muitas vezes, críticas, em uma ampla variedade de domínios de aplicação. Em (CHANDOLA *et al.*, 2009) são apresentados alguns exemplos destas situações. Um deles indica que um padrão de tráfego anômalo em uma rede de computadores significa que um computador hackeado está enviando dados confidenciais para um destino não autorizado. Já uma imagem de ressonância magnética anômala pode indicar a presença de tumores (SPENCE *et al.*, 2001). Por outro lado, anomalias nos dados de transações com cartão de crédito podem indicar roubo ou clonagem dos dados do cartão (ALESKEROV *et al.*, 1997), ou ainda leituras anômalas de uma nave espacial poderiam indicar uma falha em algum componente da espaçonave (FUJIMAKI *et al.*, 2005). Acreditamos que anomalias em bases de candidaturas eleitorais possam estar associadas a algum desvio de conduta e portanto estudar técnicas de detecção das mesmas é algo fundamental e importante nesse contexto.

Neste trabalho serão analisadas as técnicas presentes na literatura com enfoque principal nas abordagens não supervisionadas, as quais têm vasta aplicabilidade em problemas práticos e cujo objetivo é encontrar soluções sem a necessidade de uma variável resposta pré definida.

1.1 Objetivo

O objetivo do presente estudo é propor uma ferramenta que auxilie na análise e tomada de decisão de agências reguladoras e, ao mesmo tempo, compartilhar este instrumento com indivíduos que queiram pesquisar de forma mais detalhada o perfil de gastos de seus candidatos políticos. Para tal, a proposta é aplicar, por meio de métodos estatísticos e computacionais, algumas metodologias capazes de avaliar a existência de anomalias no comportamento dos dados das divulgações de contas de candidatos e partidos políticos nas eleições de 2018. Com isso pretende-se contribuir com um processo que futuramente reduza o risco de situações, como desvio do dinheiro público ou mesmo a geração de situações com benefícios e vantagens por meio de atos considerados imorais e/ou ilícitos.

Ressalta-se que a análise que será realizada não possui uma informação a priori de risco dos candidatos políticos. Ou seja, nesta situação não é possível validar com exatidão a eficiência obtida pelos algoritmos utilizados. Isso é completamente diferente de uma situação onde previamente observa-se algum fenômeno já ocorrido, por exemplo, se existe uma informação de fraude comprovada no cartão de crédito ou se um cliente bancário deixou de pagar seu empréstimo ou ainda se um paciente apresenta uma doença. Para todas essas situações, é possível encontrar as relações entre variáveis explicativas e alguma variável de interesse. Voltando ao problema inicial, estão disponibilizadas apenas informações preditoras que podem ser exploradas em um contexto de decisão, e por esta razão não é possível fazer afirmações concretas quanto aos resultados estimados, sendo necessário um parecer humano para tal.

Para que o objetivo geral seja alcançado, alguns objetivos específicos são explorados:

(i) Realizar um levantamento das principais técnicas de captura de anomalias existentes na literatura, com ênfase nos métodos estatísticos e de aprendizado de máquina;

(ii) Estudar o desafio de se encontrar anomalias em conjunto de dados, relacionados às informações públicas das campanhas eleitorais; e

(iii) Analisar os resultados obtidos e avaliar a correlação desses com alguma variável de risco adicional (por exemplo: se houve algum processo/crime dos candidatos após o período eleitoral).

1.2 Metodologia

A metodologia utilizada nesta dissertação foi dividida em três partes. A primeira é constituída de um levantamento bibliográfico relacionado à investigação de anomalias, incluindo os principais métodos empregados bem como algumas aplicações em diversas áreas

de pesquisa. A segunda contém detalhes da obtenção e extração dos dados das campanhas eleitorais do ano de 2018. A última parte refere-se à construção de um modelo combinando duas técnicas de investigação de anomalias, uma já consagrada de agrupamento conhecida por *K-Means* (LLOYD, 1957; FORGY, 1965) e outra mais recente chamada de *Isolation Forest* (LIU *et al.*, 2008).

1.2.1 Levantamento Bibliográfico

A pesquisa bibliográfica foi conduzida na literatura com o objetivo de identificar os principais métodos de investigação de anomalias associadas a quaisquer tipos de informação.

Ressalta-se que a detecção de anomalias é um importante problema existente em diversas áreas do conhecimento e que há uma grande gama de técnicas de detecção destes eventos desenvolvidas para domínios específicos de aplicações. O levantamento realizado tenta fornecer uma abordagem estruturada e uma visão abrangente da pesquisa sobre detecção de anomalias. Foram coletadas técnicas existentes em diferentes esferas com base na abordagem de uso. Ou seja, para cada área foram avaliadas as principais técnicas utilizadas na identificação de perfis de comportamento normal e anômalo.

Um fato relevante sobre a pesquisa realizada é que não foram encontrados registros de aplicações de algoritmos de detecção de anomalias em bases com perfis de campanhas eleitorais. Este detalhe reforça que este estudo pode contribuir com o avanço e controle na gestão de recursos federais, além de motivar outras pesquisas na mesma direção.

1.2.2 Coleta de Dados

Toda a informação utilizada neste projeto será condicionada a dados públicos e os principais insumos foram coletados por meio da internet, utilizando métodos de *web crawler* para elaborar um banco de dados com o maior número de variáveis e eventos possíveis. Salienta-se que o acesso à informação está previsto na Constituição Federal e na Declaração Universal dos Direitos Humanos. O direito do cidadão em obter informações de órgãos públicos, previsto em diversos tratados internacionais, já constava do texto original da Constituição de 1988 no inciso XXXIII do art. 5º. Desde então, o acesso à informação passou a ser previsto em leis específicas sobre temas, como licitações ou finanças públicas. No entanto, somente em 18 de novembro de 2011 foi sancionada a lei 12.527 de acesso à informação pública, que regula o acesso de dados e informações detidas pelo governo.

Uma fonte preciosa de informações pode ser obtida em www.dados.gov.br. Neste projeto, os principais insumos foram coletados através do site do Tribunal Superior Eleitoral

(TSE), relacionados às campanhas eleitorais de 2018. Mais detalhes serão apresentados no Capítulo 4 sobre o estudo de caso.

1.2.3 Combinação de Técnicas

Na construção deste projeto, enquanto eram estudados alguns algoritmos em situações simuladas e reais como no conjunto de dados de eleições levantado, foi discutida a possibilidade de agregar mais de uma técnica para combinação dos resultados na tentativa de melhorar a acurácia final do modelo. A partir disso, e entendendo a complementaridade de algumas técnicas, foi estabelecido um método, que aqui será chamado de *K-Means + Isolation Forest* ou simplesmente de *KM+IF*, combinando uma técnica de agrupamento, mais especificamente *K-Means*, e uma técnica que visa isolar informações anômalas, chamada de *Isolation Forest*. Os detalhes dessas técnicas são mostrados no Capítulo 3.

Em (MUNIYANDI *et al.*, 2012) os autores propõem algo semelhante com *K-Means* + Algoritmo de classificação (Árvore de Decisão). O método consiste em duas etapas: na primeira são criados *clusters* utilizando a distância euclidiana e na sequência aplicam árvore de decisão em cada partição, refinando os limites de decisão e avaliando a existência de anomalias. Recentemente, (Kurnianingsih *et al.*, 2018) e (GAO *et al.*, 2019) discutiram o mesmo método sugerido neste trabalho. A justificativa de (GAO *et al.*, 2019) é que o método *Isolation Forest* apresenta o desafio de identificar pontos de anomalias locais e por isso propõem um algoritmo aprimorado. Já (Kurnianingsih *et al.*, 2018) defendem essa metodologia híbrida visando eficiência computacional, reduzindo o custo de cálculo de distâncias do conjunto de dados. Ambos os autores aplicam seus algoritmos em dados rotulados da literatura e afirmam que eles reduziram a taxa de erros, além de aumentar sua acurácia.

Uma vez aplicados os algoritmos, os resultados de performance comparativa serão apresentados através de cinco medidas muito utilizadas na literatura:

1) Taxa de verdadeiros positivos ou *True Positive Ratio* ou ainda *Recall* (TPR);
2) Taxa de falsos positivos ou *False Positive Ratio* (FPR);
3) Precisão;
4) Acurácia;
5) F1-score ou F-*measure*.

A descrição dessas medidas, bem como os resultados obtidos nos testes simulados são encontrados no Capítulo 3. Os resultados em bases de candidaturas eleitorais serão apresentados no Capítulo 4.

1.3 Organização do trabalho

A dissertação está estruturada em cinco capítulos.

O Capítulo 1 contém a introdução do trabalho, descrevendo a importância de estudos relacionados a informações públicas e como podem se transformar em ferramentas para tomada de decisão. Também é apresentado o objetivo, contemplando detalhes esperados do projeto. Por fim, o capítulo menciona a metodologia que será aplicada, neste caso através do levantamento bibliográfico, coleta de dados e combinação de técnicas com uma sugestão de uso de algoritmos na captura de anomalias.

O Capítulo 2 apresenta de forma resumida um extenso levantamento sobre algumas aplicações das técnicas de detecção de anomalias, através de uma revisão bibliográfica dos últimos anos nas áreas acadêmica e aplicada. Além disso, aborda também a definição de anomalia, os desafios na sua detecção, sua tipologia, como são classificadas, abordagens de detecção e formas de representação.

O Capítulo 3 retrata os principais métodos utilizados em detecções de anomalias. Primeiramente, são citados os modelos supervisionados de classificação, como os baseados em redes neurais, redes bayesianas, *support vector machine* e baseadas em regras. Em seguida, são abordados os métodos não supervisionados de agrupamento com um detalhamento do método *K-Means* que será aplicado neste trabalho. Além disso, é apresentado, de forma aprofundada, um método particular baseado no isolamento de anomalias, conhecido como *Isolation forest*. Por fim, é discutido o método combinado *K-Means + Isolation Forest* e as métricas de performance que serão utilizadas para verificar acurácia dos modelos.

O Capítulo 4 evidencia duas aplicações em dados reais dos métodos descritos no capítulo 3, contemplando a descrição do problema, uma análise descritiva e resultados de performance dos modelos. A primeira aplicação é relativa ao estudo bastante conhecido na literatura sobre o conjunto de dados de flores Íris (FISHER, 1936). A segunda aplicação está relacionada com o propósito principal deste projeto: a detecção de anomalias em conjunto de dados de candidaturas eleitorais.

Por fim, o Capítulo 5 apresenta as conclusões do projeto, com um resumo dos principais resultados obtidos, considerações finais sobre uso dos métodos e perspectivas de trabalhos futuros.

2 Definição e Abordagem de Detecção de Anomalias

Neste capítulo serão discutidos conceitos fundamentais relacionados à captura e descoberta de anomalias em diversas áreas do conhecimento. Diante do fato de nosso objetivo principal ser construir uma ferramenta que indique possível risco das contas eleitorais e auxilie na tomada de decisão de órgãos fiscalizadores, faz-se necessário uma revisão de situações encontradas na natureza e quais as abordagens são aplicadas na literatura. A ideia principal deste capítulo é rever importantes publicações que tragam conteúdo e riqueza de detalhes sobre os melhores métodos e algoritmos que possam ser aplicados para detecção de anomalias, inclusive podendo ser diferentes do nosso problema, a fim de entender quais as melhores ferramentas disponíveis para uso. A seguir serão apresentadas algumas definições mais precisas sobre o evento de anomalia em si e posteriormente serão expostos os tratamentos das aplicações.

Como dito anteriormente, detecção de anomalias refere-se ao problema de encontrar padrões em dados que não estão em conformidade ao comportamento esperado. Esses padrões não conformes são frequentemente chamados de anomalias, *outliers*, observações discordantes, exceções, etc. em diferentes domínios de aplicação. 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 (M.A., 1887). Ao longo do tempo, uma variedade de técnicas de detecção de anomalias foram desenvolvidas 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.

A Figura 2.1 ilustra anomalias em um conjunto de dados bidimensional simples. Os dados tem duas regiões normais, N_1 e N_2, já que a maioria das observações se encontra nessas duas regiões. Pontos que estão suficientemente longe destas regiões, por exemplo, os

pontos o_1 e o_2, e pontos na região O_3, são anomalias. Anomalias podem ser induzidas nos dados por vários motivos, por exemplo, fraude bancária, intrusão cibernética, atividade terrorista ou discriminação de um sistema, mas todas as razões têm a característica comum de que são interessantes para o pesquisador.

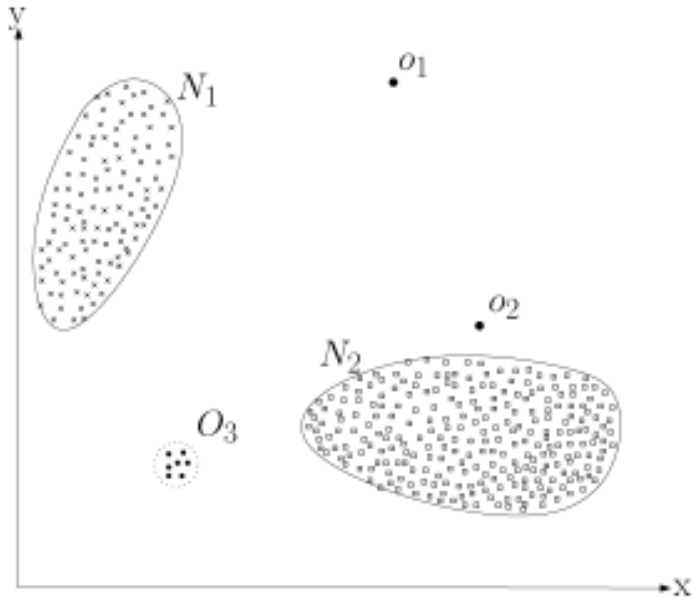

FIGURA 2.1 – Anomalia Pontual. Fonte: (CHANDOLA *et al.*, 2009)

Outro fato importante é que a detecção de anomalias está relacionada à identificação e remoção de ruídos (TENG *et al.*, 1990b) e ajuste de ruído (ROUSSEEUW; LEROY, 1987). Ruído pode ser definido como um fenômeno em dados que não é de interesse para o problema, mas atua como um obstáculo à análise de dados. A remoção de um ruído é impulsionada pela necessidade de eliminar os objetos indesejados antes que qualquer análise de dados seja executada.

2.1 Desafios na Detecção de Anomalias

Como dito anteriormente, uma anomalia não obedece ao comportamento normal esperado. Uma abordagem simples de detecção de anomalias, portanto, é definir uma região representando o comportamento normal e declarar qualquer observação nos dados que não pertencem a esta região normal como uma anomalia. Entretanto, vários fatores fazem essa abordagem aparentemente simples ser muito desafiadora:

1) Definir uma região normal que englobe todos os comportamentos normais possíveis é muito difícil. Além disso, o limite entre o comportamento normal e anômalo é frequentemente impreciso. Assim, uma observação anômala que fica perto do limite ou na fronteira pode na verdade ser normal e vice-versa.

2) Quando as anomalias são o resultado de ações maliciosas, os indivíduos malfeitores frequentemente adaptam-se para fazer com que as observações anômalas pareçam normais, dificultando a tarefa de definir o comportamento normal. Por exemplo, em nosso estudo, caso as contas eleitorais não estejam completas ou haja má fé dos candidatos em ocultar parte dos custos de campanha, isso pode nos levar a resultados enganosos.

3) Em muitos domínios, o comportamento normal continua evoluindo e uma noção atual de comportamento pode não ser suficientemente representativa no futuro. Um exemplo é o aumento das contas pagas no cartão de crédito. Se o cliente passa a consumir cinco vezes mais limite de um mês para o outro, significa que houve uma forte mudança no seu comportamento normal (habitual). Por outro lado, se o mesmo cliente aumentar seu consumo gradativamente durante um intervalo de tempo, poderia indicar uma evolução de suas despesas.

4) A noção exata de uma anomalia é diferente para diferentes domínios de aplicação. No caso de um estudo médico, um pequeno desvio do padrão normal (por exemplo, flutuações na temperatura corporal) pode ser uma anomalia, enquanto desvios similares no mercado financeiro (por exemplo, flutuações no valor de uma ação) podem ser considerados normais. Assim, aplicar uma técnica desenvolvida em um domínio pode não ser adequada em outro contexto.

5) A disponibilidade de dados rotulados para treinamento e validação de modelos usando técnicas de detecção de anomalias é geralmente uma questão relevante, pois normalmente não há conteúdo suficiente para determinar um bom ajuste. Por exemplo, neste projeto, há informações muito relevantes sobre o comportamento dos gastos de campanha, contudo não há uma classificação de risco para os candidatos. Ou seja, não é possível prever todos os comportamentos anômalos do conjunto de dados e qualquer resultado gerado precisará de avaliação humana para um veredicto.

6) Frequentemente, é comum que os dados apresentem ruídos que tendem ser semelhantes às anomalias propriamente ditas e por isso, é difícil distingui-los e removê-los. Por conta desse fator, é muito importante ter um profundo conhecimento do que se está modelando.

Dos fatores acima citados, os maiores desafios deste projeto concentram-se nos itens 1, 2, 5 e 6. Ou seja, dificuldade em definir a região normal dos dados, possível ocultação de informações por parte dos candidatos políticos, falta de dados rotulados da variável de interesse para treinamento e validação do modelo e, por fim, eventuais ruídos nos dados que possam existir.

Devido a esses desafios, o problema de detecção de anomalias, em sua forma mais geral, não é fácil de resolver. De fato, a maioria das técnicas de detecção de anomalias existentes resolve uma formulação específica do problema. Pesquisadores costumam ado-

tar conceitos de diversas disciplinas como estatística, aprendizado de máquina, mineração de dados, teoria da informação, etc. para resolução destes problemas. Voltando ao nosso projeto, dadas as situações encontradas, serão adotados algoritmos não supervisionados para resolução do problema, com enfoque nos métodos de agrupamento e isolamento de observações.

2.2 Aspectos do problema de detecção de anomalias

Esta seção identifica e discute os diferentes aspectos da detecção de anomalias. Como mencionado anteriormente, uma formulação específica do problema é determinada por vários fatores, como a natureza dos dados de entrada, a disponibilidade ou indisponibilidade de rótulos, bem como as restrições e requisitos induzidos pelo domínio da aplicação. Além disso, a seção também traz alguns domínios conhecidos onde se justifica a aplicação de técnicas de detecção de anomalias.

2.2.1 Natureza dos dados de entrada

Um aspecto fundamental de qualquer técnica de detecção de anomalias é a natureza dos dados de entrada. Entrada é geralmente uma seleção de um conjunto de dados (também referida como objeto, registro, ponto, vetor, padrão, evento, caso, amostra, observação ou entidade) (TAN *et al.*, 2005). Cada instância de dados pode ser descrita usando um conjunto de atributos (também chamado de variável, característica, recurso, campo ou dimensão). Os atributos podem ser de diferentes tipos, como binário, categórico ou contínuo. Cada instância de dados pode consistir em apenas um atributo (univariada) ou múltiplos atributos (multivariada). No caso de um conjunto de dados multivariado, todos os atributos podem ser do mesmo tipo ou podem ser uma mistura de tipos de dados.

Segundo (CHANDOLA *et al.*, 2009), a natureza dos atributos determina a aplicabilidade das técnicas de detecção de anomalias. Por exemplo, para técnicas estatísticas, diferentes modelos estatísticos devem ser usados para dados contínuos e categóricos. Da mesma forma, para técnicas baseadas em vizinhos mais próximos (ou KNN, *K-Nearest Neighbors*), a natureza dos atributos determina a medida de distância a ser usada. Muitas vezes, ao invés dos dados reais, a distância entre pares de variáveis podem ser usadas na forma de uma matriz de distâncias ou similaridades. Nesses casos, técnicas que exigem conjunto de dados originais não são aplicáveis, como exemplo alguns modelos estatísticos e técnicas baseadas em classificação.

Os dados de entrada também podem ser categorizados com base no relacionamento presente entre as instâncias de dados (TAN *et al.*, 2005). A maioria das técnicas de detecção

de anomalias existentes lidam com dados de registro (ou dados de ponto), nos quais nenhum relacionamento é assumido entre as instâncias de dados.

Em geral, os conjuntos de dados podem estar relacionados entre si. Alguns exemplos são dados sequenciais, dados espaciais e dados gráficos. Em dados sequenciais, o conjunto de dados são ordenados linearmente, por exemplo, dados de séries temporais, sequências do genoma e sequências de proteínas. Em relação a dados espaciais, cada conjunto de dados é relacionado às instâncias vizinhas, por exemplo, dados de tráfego de veículos e dados ecológicos. Quando os dados espaciais têm um componente temporal (sequencial) são denominados como dados espaço-temporais, por exemplo, dados de clima.

2.2.2 Tipos de Anomalia

Um aspecto importante de uma técnica de detecção de anomalias é a natureza da anomalia. As anomalias podem ser classificadas nas seguintes três categorias:

2.2.2.1 Anomalias pontuais

Se um conjunto de dados individual puder ser considerado anômalo em relação ao restante dos dados, esse conjunto é denominado uma anomalia pontual. Este é o tipo mais simples de anomalia, sendo o foco da maioria das pesquisas sobre detecção de anomalias.

Por exemplo, na Figura 2.1 apresentada no início da seção 2, os pontos o_1 e o_2, assim como os pontos na região O_3, ficam do lado de fora limite das regiões normais e, portanto, são anomalias pontuais, uma vez que são diferentes dos pontos de dados normais. Como um exemplo da vida real, considere a detecção de fraudes com cartões de crédito. Supondo que o conjunto de dados corresponde às transações de cartão de crédito de um indivíduo e, por uma questão de simplicidade, vamos supor também que os dados sejam definidos usando apenas um atributo: valores gastos no cartão. Uma transação para a qual o montante gasto é muito elevado, em comparação com a gama normal de despesas para esse indivíduo, será considerada uma anomalia pontual.

2.2.2.2 Anomalias Contextuais

Se um conjunto de dados for anômalo e depender de um contexto específico, então será denominada uma anomalia contextual (também referida como anomalia condicional (SONG *et al.*, 2007)). A noção de um contexto é induzida pela estrutura no conjunto de dados e tem que ser especificado como parte da formulação do problema. Cada conjunto de dados é definido seguindo dois conjuntos de atributos:

1. Atributos contextuais. Os atributos contextuais são usados para determinar o

contexto (ou vizinhança) para um conjunto de dados. Por exemplo, em conjuntos de dados espaciais, a longitude e latitude de um local são os atributos contextuais.

2. Atributos comportamentais. Os atributos comportamentais definem as características não-contextuais de um conjunto de dados. Por exemplo, em um conjunto de dados espaciais que descreve a média de chuvas de todos os países do mundo, a quantidade de chuvas em qualquer local (ou país) é considerado um atributo comportamental.

O comportamento anômalo é determinado usando os valores dos atributos comportamentais dentro de um contexto específico. Um conjunto de dados pode ser uma anomalia contextual (em um determinado contexto), mas um conjunto de dados idêntico (em termos de atributos comportamentais) poderia ser considerado normal em um contexto diferente. Esta propriedade é a chave para identificar o contexto e os atributos comportamentais para uma técnica de detecção de anomalias contextuais. Anomalias contextuais têm sido mais comumente exploradas em dados de séries temporais (FOX, 1972; WEIGEND *et al.*, 1995; SALVADOR *et al.*, 2004) e dados espaciais (SHEKHAR *et al.*, 2001; KOU *et al.*, 2006). A Figura 2.2 mostra um desses exemplos para uma série temporal de temperatura mensal de uma determinada área nos últimos anos.

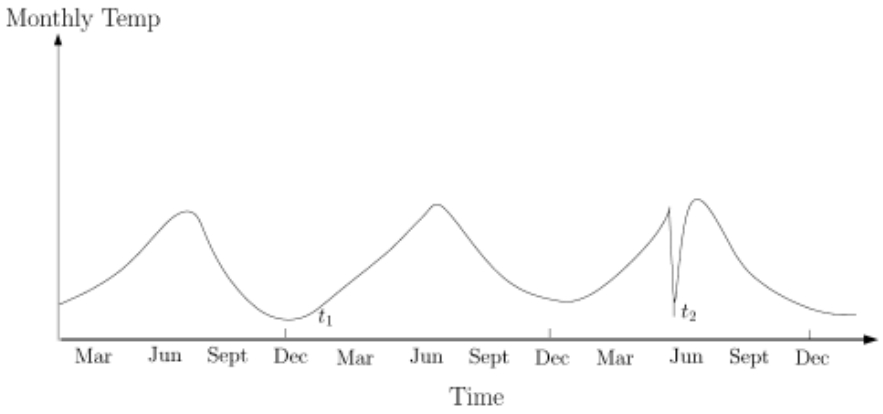

FIGURA 2.2 – Exemplo de Anomalia Contextual. Fonte: (CHANDOLA *et al.*, 2009).

A Figura 2.2 apresenta uma anomalia contextual t_2 em uma série temporal de temperatura. Nota-se que a temperatura no tempo t_1 é a mesma como no tempo t_2, mas ocorre em um contexto diferente e, portanto, não é considerada como uma anomalia.

Tomando como exemplo a temperatura do clima mediterrâneo, 2 a 5°C podem ser normais durante o inverno (no tempo t_1) naquele local, mas o mesmo valor durante o verão (no tempo t_2) seria uma anomalia. Um exemplo semelhante pode ser encontrado no domínio de detecção de fraude de cartão de crédito. Um atributo contextual do cartão de crédito pode ser o momento da compra. Suponha que um indivíduo geralmente tem uma conta de compras semanais de $R\$250,00$, exceto durante a semana do Natal, quando atinge $R\$5.000,00$. Uma nova compra de $R\$5.000,00$ em uma semana de julho será considerada uma anomalia contextual, uma vez que não é similar ao comportamento

normal do indivíduo no contexto do tempo, embora o mesmo montante gasto durante a semana de Natal será considerado normal.

A escolha de aplicar uma técnica de detecção de anomalias contextuais é determinada pela significância das anomalias contextuais no domínio de aplicação de destino. Outro fator chave é a disponibilidade de atributos contextuais. Em vários casos, definindo um contexto, a aplicação de uma técnica de detecção de anomalias contextuais é simples de ser realizada. Em outras situações, quando definir um contexto não é fácil, isso torna a aplicação de técnicas de detecção mais complexa.

2.2.2.3 Anomalias Coletivas

Se uma seleção parcial de dados relacionados for anômala em relação a todo o conjunto de dados, ela é denominada anomalia coletiva. Os dados individuais em uma anomalia coletiva podem não ser anomalias entre si, mas sua ocorrência conjunta como uma coleção é anômala. A Figura 2.3 é um exemplo que mostra um eletrocardiograma humano (GOLDBERGER *et al.*, 2000). A região destacada é considerada uma anomalia, porque o mesmo valor ocorre por um longo tempo anormalmente (contração atrial prematura). Anomalias coletivas foram exploradas para dados de sequência (WARRENDER *et al.*, 1999), dados gráficos (NOBLE; COOK, 2003) e dados espaciais (SHEKHAR *et al.*, 2001).

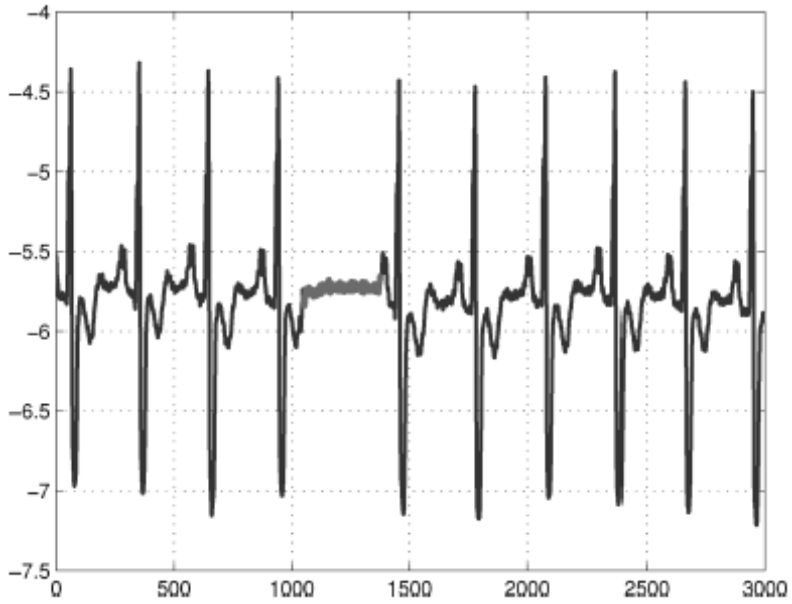

FIGURA 2.3 – Exemplo de Anomalia Coletiva. Contração atrial prematura. Fonte: (CHANDOLA *et al.*, 2009).

Deve-se notar que, embora anomalias pontuais possam ocorrer em qualquer conjunto de dados, anomalias coletivas podem ocorrer apenas em conjuntos de dados nos quais os mesmos estão relacionados. Em contrapartida, a ocorrência de anomalias contextuais

dependem da disponibilidade de atributos de contexto nos dados. Uma anomalia pontual ou uma anomalia coletiva também pode ser uma anomalia contextual se analisados em relação a um contexto. Assim, um problema de detecção de anomalias pontuais ou coletivas pode ser transformado em um problema de detecção de anomalia contextual incorporando a informação de contexto. Ressalta-se que as técnicas usadas para detectar anomalias coletivas são muito diferentes das técnicas de detecção de anomalias pontuais e contextuais.

2.2.3 Variável de Interesse

A correta classificação dos dados indica se esse conjunto de informações é normal ou se apresenta uma anomalia. Salienta-se que a obtenção de todos os tipos de comportamento dos dados é muitas vezes caro, uma vez que esta marcação geralmente é feita manualmente por um especialista humano e, portanto, um esforço substancial é necessário para obter a classificação da variável de interesse de todo o conjunto de dados. Normalmente, obter um conjunto rotulado de dados anômalos que cubram todos os tipos possíveis de comportamento anômalo é mais difícil do que classificar o comportamento normal. Além disso, o comportamento anômalo é frequentemente dinâmico na natureza. Por exemplo, novos tipos de anomalias podem surgir, para as quais não há dados de treinamento. Em certos casos, como a segurança do tráfego aéreo, casos anômalos podem significar eventos catastróficos e, portanto, são muito raros.

Com base na forma em que os rótulos da variável de interesse estejam disponíveis, as técnicas de detecção de anomalias podem operar em um dos três modos a seguir:

2.2.3.1 Detecção de anomalias via métodos supervisionados

As técnicas de modelagem supervisionada assumem que há disponibilidade de um conjunto de dados de treinamento com classes de dados normais e de anomalia. Uma abordagem típica em tais casos é construir um modelo preditivo para classes normais versus anomalias. Qualquer conjunto de dados não classificados é comparado com os modelos para determinar a qual classe ele pertence. Um grande problema que surge em detecção de anomalia supervisionada ocorre porque os conjuntos anômalos são muito menos frequentes quando comparados aos conjuntos normais nos dados de treinamento. Estes temas de distribuições desequilibradas de classes foram abordados na literatura através de mineração de dados e aprendizado de máquina (WEISS; HIRSH, 1998; JOSHI *et al.*, 2001; VILALTA; MA, 2002; CHAWLA *et al.*, 2004; PHUA *et al.*, 2004). Obter classificações precisas e representativas, especialmente para a classe de anomalia é geralmente um desafio. Técnicas têm sido propostas onde são adicionadas anomalias artificiais em um conjunto de dados

normal para se obter um melhor conjunto de dados de treinamento (THEILER; CAI, 2003; STEINWART *et al.*, 2005; ABE *et al.*, 2006).

2.2.3.2 Detecção de anomalias via métodos semi supervisionados

Técnicas que operam de forma semi supervisionada supõem que os dados de treinamento apresentam apenas o conjunto de dados de classe normal. Como não exigem uma classificação para a classe de anomalia, tais técnicas são mais amplamente aplicáveis do que técnicas supervisionadas. Por exemplo, na detecção de falhas de naves espaciais (FUJIMAKI *et al.*, 2005), um cenário de anomalia significaria um acidente, que traz grande complexidade e dificuldade de se modelar. A abordagem típica usada em tais técnicas é construir um modelo para a classe correspondente ao comportamento normal e em seguida usar o modelo para identificar anomalias nos dados de teste. Existe um conjunto limitado de técnicas de detecção de anomalias que pressupõe a disponibilidade do conjunto de anomalias para treinamento (WARRENDER *et al.*, 1999; DASGUPTA; NINO, 2000; DASGUPTA; MAJUMDAR, 2002). Tais técnicas não são comumente usadas, principalmente, porque é difícil obter um conjunto de dados de treinamento que cubra todos os comportamentos anômalos possíveis que podem ocorrer nos dados.

2.2.3.3 Detecção de anomalias via métodos não supervisionados

Os métodos não supervisionados têm como principal característica o fato de não necessitarem de fornecimento de conjuntos de dados rotulados para treinamento. Por este motivo apresentam a vantagem de serem amplamente aplicáveis. A obtenção dos rótulos é particularmente cara quando a classificação deve ser realizada por seres humanos. Além disso, lidar com uma distribuição de classes fortemente desequilibradas, que é inerente à detecção de anomalias, também pode afetar a eficiência dos algoritmos supervisionados (JAPKOWICZ; STEPHEN, 2002), o que favorece estratégias não supervisionadas.

As técnicas não supervisionadas fazem a suposição implícita de que os conjuntos normais são muito mais frequentes do que anomalias nos dados de teste. Se essa suposição não for verdadeira, então essas técnicas sofrem com uma alta taxa de falso positivo, ou seja, podem gerar um grande erro de classificação do modelo, identificando ocorrências normais como sendo anômalas. Muitas técnicas semi supervisionadas podem ser adaptadas para operar em um modo não supervisionado usando uma amostra do conjunto de dados não identificados como dados de treinamento. Tal adaptação assume que os dados de teste contêm pouquíssimas anomalias e o modelo construído durante o treinamento é robusto para essas poucas discrepâncias (CHANDOLA *et al.*, 2009).

Por conta dos pontos descritos anteriormente, como a inexistência de dados rotulados,

e, dada a flexibilidade e aplicabilidade dos algoritmos não supervisionados existentes, este projeto focará na aplicação deste tipo de técnicas na resolução do problema de anomalias nas contas eleitorais.

2.2.4 Resultado da detecção de anomalia

Um aspecto importante para qualquer técnica de detecção de anomalias é a maneira como estes eventos serão representados. Normalmente, os resultados e saídas produzidos pelos algoritmos de detecção de anomalias são apresentados de duas maneiras, descritas a seguir:

Pontuações: como o próprio nome sugere, estas técnicas de pontuação atribuem uma pontuação ou score de anomalia a cada indivíduo no conjunto de dados de teste. Dependendo do grau de classificação, esse indivíduo é considerado como uma anomalia. Portanto, a saída dessas técnicas é uma lista ordenada de anomalias como um *ranking*. Um especialista pode optar em analisar algumas poucas anomalias identificadas ou usar um ponto de corte para selecionar as anomalias mais graves ou relevantes.

Rótulos: os resultados das técnicas dessa categoria atribuem um rótulo (normal ou anômalo) para cada indivíduo do conjunto de dados de teste, ou seja, o pesquisador terá em mãos como resultado duas únicas marcações de evento (*0* e *1*, *Sim* ou *Não*, etc.).

Em suma, as técnicas de detecção de anomalias baseadas em pontuação permitem que o analista use um domínio específico para selecionar as anomalias mais relevantes. Já as técnicas que fornecem rótulos binários para os indivíduos de teste não permitem que os especialistas definam pontos de corte, embora isso possa ser controlado indiretamente por meio de opções de parâmetros dentro da técnica aplicada.

Neste projeto, diante do tipo de problema que será abordado no caso de anomalias em contas eleitorais, os resultados serão apresentados em forma de pontuação. Acreditamos que isso seja mais adequado à ferramenta de tomada de decisão proposta uma vez que expressa uma riqueza maior de informações para o pesquisador, além do fato de que o mesmo pode definir um ponto de corte de atuação, conforme tenha capacidade de análise.

2.3 Aplicações de detecção de anomalia

Nesta seção são apresentadas, de forma mais ampla, algumas aplicações da detecção de anomalias, como detecção de invasão de redes, fraudes, saúde pública e na indústria. A proposta é apresentar o contexto da anomalia, a natureza dos dados, desafios associados à sua detecção e as principais técnicas aplicadas. Com isso, esta revisão visa gerar insumos para decidirmos situações e técnicas favoráveis à resolução do nosso problema. Ressalta-se

que na pesquisa da literatura não foram encontrados registros de trabalhos relacionados à análise de anomalias em bases eleitorais. A seguir serão apresentadas algumas das situações de interesse relatadas em (CHANDOLA *et al.*, 2009).

2.3.1 Detecção de invasão

Detecção de invasão (ou intrusão) refere-se à detecção de atividades maliciosas e outras formas de interferência em um sistema relacionado com computador (PHOHA, 2002). Essas atividades maliciosas ou invasões são interessantes do ponto de vista de segurança de sistemas computacionais. Uma invasão é diferente do comportamento normal do sistema e, portanto, técnicas de detecção de anomalia são aplicáveis no domínio de detecção de invasão. O principal desafio para a detecção de anomalias nesse domínio é o enorme volume de dados. As técnicas de detecção de anomalias precisam ser computacionalmente eficientes para lidar com grande volume de entradas de dados. Além disso, os dados geralmente são transmitidos em fluxo contínuo, requerendo análise online, ou seja, juntamente com a coleta de dados. Outra questão que surge devido ao grande volume de entradas é a taxa de alarmes falsos. Como são milhões de objetos de dados, alguns alarmes falsos podem tornar a análise desafiadora para um analista. Dados classificados com comportamento normal sempre estão disponíveis, enquanto os classificados como invasão geralmente não estão. Portanto, técnicas de detecção de anomalias semi supervisionadas e não supervisionadas são preferidas nesse domínio. Os principais sistemas de detecção de invasão são os baseados em *host* e de rede. Em (GOGOI *et al.*, 2010), ou autores exploram as abordagens supervisionada e não supervisionada para análise de deteção de anomalias de dados de intrusão. (MUNIYANDI *et al.*, 2012) abordam o problema através de métodos de classificação enquanto (KAREV *et al.*, 2017) aplicam o algoritmo *Isolation Forest* não supervisionado em um novo contexto de intrusão cibernética.

2.3.2 Detecção de fraude

Detecção de fraude refere-se à detecção de atividades criminosas que ocorrem em organizações comerciais como bancos, empresas de cartão de crédito, agências de seguros, empresas de telefonia celular, mercado de ações e assim por diante. Os usuários malintencionados podem ser os clientes reais da organização ou podem ser outros indivíduos se passando por clientes (falsidade ideológica). A fraude ocorre quando esses usuários consomem os recursos fornecidos pela organização de maneira ilegal. As organizações estão interessadas na detecção imediata de tais fraudes para evitar perdas econômicas. (FAWCETT; PROVOST, 1999) introduzem o termo monitoramento de atividades como uma abordagem geral para detecção de fraudes nesses domínios. A abordagem típica de técni-

cas de detecção de anomalias é manter um perfil de uso para cada cliente e monitorá-los para observar quaisquer desvios. (NGAI *et al.*, 2011) realizam uma primeira revisão sistemática, identificável e abrangente da literatura acadêmica das técnicas de mineração de dados aplicadas à detecção de fraudes financeiras. Algumas das aplicações específicas de detecção de fraudes serão discutidas a seguir.

2.3.2.1 Detecção de fraude de cartão de crédito

As técnicas de detecção de anomalias são aplicadas para detectar aplicativos fraudulentos de cartão de crédito ou mesmo o uso fraudulento de cartão de crédito (associado a roubos ou clonagens de cartão). A detecção de aplicativos de cartão de crédito fraudulentos é semelhante à detecção de fraudes de seguros (GHOSH; REILLY, 1994). Os dados são tipicamente compostos por registros definidos em várias dimensões, como identificação do usuário, valor gasto, tempo entre o uso consecutivo do cartão e assim por diante. As fraudes aparecem normalmente em registros transacionais e correspondem a pagamentos, compra de itens nunca comprados pelo usuário, alto valor de compra e assim por diante. Como as empresas de crédito têm dados completos disponíveis para acompanhar o perfil de uso do cliente no cartão de crédito, as técnicas baseadas em cluster ou de identificação de perfil são normalmente usadas nesse caso. O desafio associado à detecção do uso não autorizado de cartões de crédito é que requer uma abordagem *online* de captura de fraude (ou seja, assim que a transação fraudulenta ocorrer, evitando assim perda financeira). Para isso, basicamente cada cliente tem um perfil de uso construído com o histórico de transações do cartão de crédito. Qualquer nova transação realizada é comparada ao perfil do usuário e, caso sinalize algo estranho, é classificada como uma anomalia a ser tratada. (WHITROW *et al.*, 2009) apresentam uma estratégia para agregação mais eficaz de transações de cartões, além de aplicar alguns métodos de classificação como máquinas de vetores de suporte (SVM, de *Support Vector Machine*), *random forest* e regressão logística para efeito de comparação de resultados. Já (RYMAN-TUBB *et al.*, 2018) mostram como a inteligência artificial e o aprendizado de máquina impactam na detecção de fraudes com cartões de pagamento e (CHAUDHARY *et al.*, 2012; SAGADEVAN *et al.*, 2018) mostram como as técnicas de mineração de dados podem ser combinadas com êxito para obter uma alta cobertura de fraude reduzindo a taxa de erro.

2.3.2.2 Detecção de fraude de reivindicação de seguro

Um problema importante na indústria de seguros é avaliar o sinistro, por exemplo, se houve fraude no seguro de automóvel. Segurados e fornecedores de má fé tentam manipular informações com o processamento de reclamações não autorizadas e ilegais. A detecção de tal fraude tem sido muito importante para as seguradoras para evitar per-

das financeiras. Os dados disponíveis são geralmente os documentos apresentados pelos segurados. As técnicas extraem diferentes características, com variáveis tanto categóricas quanto contínuas. Normalmente, as seguradoras e seus investigadores avaliam essas fraudes manualmente e depois são usadas como exemplos de classificação para uso de técnicas semi supervisionadas. Em (NEUMANN *et al.*, 2019) é apresentada a aplicabilidade do aprendizado de máquina no mercado de seguros de automóveis.

2.3.2.3 Detecção de informações privilegiadas

Outra aplicação recente de técnicas de detecção de anomalias ocorreu na detecção de *insider trading*. O uso de informações privilegiadas é um fenômeno encontrado nos mercados de capitais, onde as pessoas obtêm lucros ilegais agindo (ou vazando) informação privilegiada antes que a mesma se torne pública. O *insider trading* pode ser detectado identificando atividades comerciais anômalas no mercado.

Os dados disponíveis são de várias fontes heterogêneas, como dados de negociação de opções, dados de negociação de ações e notícias de mídia. Os dados possuem associações temporais, uma vez que são coletados continuamente. As técnicas de detecção de anomalias nessa categoria são necessárias para mitigar as fraudes e impedir que pessoas e organizações lucrem de forma ilícita. Em (KULKARNI *et al.*, 2017) são coletadas informações privilegiadas disponibilizadas pelas Comissões de Câmbio e Valores Mobiliários dos EUA (SEC) com o objetivo de estudar uma abordagem de redes de relacionamento para resolver o problema de identificação de *insider trading*.

2.3.3 Detecção de anomalias médicas e de saúde pública

A detecção de anomalias no meio médico e de saúde pública funciona tipicamente com o registro dos pacientes. Os dados podem ter anomalias devido a várias razões, como anomalias na condição do paciente, erros de instrumentação, entre outros. Várias técnicas também são aplicadas na detecção de surtos de doenças em uma região específica. Portanto, a detecção de anomalias é um problema muito crítico na área da saúde e por isso requer um alto grau de precisão.

Neste tipo de aplicação os dados geralmente consistem em registros que podem ter vários tipos diferentes de informação, como idade do paciente, grupo sanguíneo e peso. Os dados também podem ter um aspecto espacial. A maioria das técnicas atuais de detecção de anomalias na saúde visa detectar registros anômalos (anomalias pontuais). Normalmente os dados pertencem aos pacientes saudáveis, portanto a maioria das técnicas adota uma técnica semi supervisionada para essa abordagem. O aspecto mais desafiador do problema de detecção de anomalias na área da saúde é o impacto de se classificar

corretamente uma anomalia, visto que pode ser muito alto (risco de vida de um paciente).

Outra abordagem interessante aplicada no meio da saúde realizada por (BUSCEMA *et al.*, 2016; MENNINI *et al.*, 2017) está relacionada aos gastos eficientes com as verbas destinadas à área pelo governo. Os autores mostram através de RNAs (Redes Neurais Artificiais) potenciais ganhos e estratégias para confrontar o uso ineficiente dos recursos quando aplicadas a dados orçamentários de saúde da Itália.

2.3.4 Detecção de danos industriais

Indústrias sofrem danos devido ao uso contínuo e desgaste normal de suas peças e máquinas. Tais danos precisam ser detectados precocemente para evitar perdas e interrupção da produção. Os dados nesse domínio são geralmente referidos como dados de sensores, porque são gravados e coletados usando diferentes equipamentos para análise. A detecção de anomalias de danos industriais pode ser ainda classificada em dois domínios: um que lida com defeitos na mecânica de componentes como motores, turbinas, etc. e outro que lida com defeitos em estruturas, como rachaduras em vigas ou deformações em fuselagens, etc.

Como os dados nessa área geralmente têm um aspecto temporal, por exemplo devido ao desgaste de peças, algumas análises de séries temporais também podem ser usadas para captura de anomalias. Em (BASU; MECKESHEIMER, 2007) os autores estudam o problema de detectar valores incomuns em séries temporais onde sugerem uma aplicação para identificação de anomalias de sensores em um avião. Uma abordagem de detecção de anomalias através de *Machine Learning* foi aplicada a um conjunto de dados industrial relacionado à fabricação de semicondutores em (SUSTO *et al.*, 2017).

3 Métodos Utilizados para Detecção de Anomalias

Neste capítulo serão abordadas várias das principais técnicas de detecção de anomalias registradas na literatura. A maioria dessas abordagens são baseadas em modelos existentes, construindo um perfil de conjuntos de dados normais e, em seguida, identificando aqueles que não estão em conformidade com o perfil normal, ou seja, anomalias. Exemplos notáveis são os métodos estatísticos, métodos baseados em classificação e métodos baseados em agrupamento. Além disso, também será explorado um método alternativo proposto por (LIU *et al.*, 2008) que tem por premissa o isolamento de informações anômalas, chamado de *Isolation Forest*, que propõe um isolamento explícito das anomalias ao invés de definir perfis normais no conjunto de dados.

De forma geral, os métodos de detecção de anomalias têm sido o tópico de vários levantamentos e artigos de revisão bem como livros. (HODGE; AUSTIN, 2004) fornecem uma extensa pesquisa de técnicas de detecção de anomalias desenvolvidas em aprendizado de máquina e domínios estatísticos. (AGYEMANG *et al.*, 2006) apresentam o uso das técnicas de detecção de anomalias para dados numéricos e simbólicos. Uma extensa revisão de técnicas de detecção de anomalias utilizando redes neurais e abordagens estatísticas foi apresentada em (MARKOU; SINGH, 2003a) e (MARKOU; SINGH, 2003b), respectivamente. Uma grande quantidade de pesquisas estatísticas sobre detecção de *outliers* foram produzidas, bem como uma revisão em vários livros e artigos (ROUSSEEUW; LEROY, 1987; PINCUS, 1995; HAWKINS, 1980; BECKMAN; COOK, 1983; BAKAR *et al.*, 2006). Em (CHANDOLA *et al.*, 2009), os autores fornecem uma visão geral estruturada e extensiva da pesquisa de técnicas de detecção de anomalias abrangendo múltiplas áreas de pesquisa e aplicação em diferentes domínios. Uma detalhada pesquisa sobre o desafio da dimensionalidade em algoritmos de detecção de *outliers* foi estudada em (ZIMEK *et al.*, 2012). (Emmott *et al.*, 2015) apresentam uma metanálise completa do problema de detecção de anomalias, identificando e comparando algoritmos da literatura e produzindo um grande arcabouço de *benchmarks* de detecção de anomalias. Por fim, (DOMINGUES *et al.*, 2017) aplicam algoritmos de aprendizado de máquina não supervisionados no contexto da detecção de *outliers*, comparando-os em conjuntos de dados reais e da literatura, submetendo-os a

extensos testes de escalabilidade, consumo de memória e robustez, a fim de construir uma visão completa das características dos algoritmos.

A seguir serão apresentados de maneira geral os principais métodos de análise de detecção de anomalias utilizados na literatura.

3.1 Métodos de Classificação

A técnica de classificação (TAN *et al.*, 2005) é usada para aprender um determinado modelo (classificador) a partir de um conjunto de dados já identificados (treinamento) e, em seguida, classificar um conjunto de dados de teste em uma das classes (normal ou anomalia) usando o modelo aprendido (teste). A detecção de anomalia baseada em classificação opera em duas fases. A fase de treinamento aprende através de um classificador usando os dados de treinamento disponíveis. A fase de teste classifica um novo conjunto de dados de teste como normal ou anômala, usando o classificador.

Por conta da necessidade das técnicas de classificação possuírem dados de treinamento identificados previamente, o uso delas neste trabalho seria totalmente limitado, uma vez que não há essa informação a priori em nosso conjunto de dados. Por esta razão, as mesmas não serão desenvolvidas em nosso contexto. Todavia, com intuito de completude, será apresentado um breve resumo das principais técnicas de detecção de anomalias baseadas nestes métodos.

1) Baseado em Redes Neurais Artificiais (RNAs): têm sido aplicadas à detecção de anomalias em muitas situações. Uma técnica básica de detecção de anomalias de múltiplas classes usando redes neurais opera em duas etapas. Primeiro, uma rede neural é construída nos dados normais de treinamento para aprender diferentes classes normais. Isto é feito resolvendo um problema de ajuste de dados conhecidos a um modelo construído por composição de funções, associadas às camadas da rede neural. A segunda etapa consiste em utilizar cada conjunto de teste como uma entrada para a rede neural. Basicamente, se a rede neural aceitar a entrada de teste, o dado é considerado normal, caso contrário, se a rede neural rejeitar uma entrada de teste é classificada como uma anomalia (STEFANO *et al.*, 2000). Há uma série de variantes da técnica de redes neurais, entre elas: perceptrons multi-camadas, redes de árvores, redes oscilatórias, redes de *Hopfield* etc. Maiores detalhes sobre a construção, treinamento e validação das RNAs podem ser encontrados em (BUSCEMA *et al.*, 2018).

2) Baseado em Redes Bayesianas: uma técnica básica para um conjunto de dados categóricos univariados usando estimativas de redes bayesianas naive consiste em calcular a probabilidade a posteriori de se observar uma classificação de um indivíduo do conjunto de dados como normal e a classificação do mesmo como uma anomalia, dado um conjunto

de dados de teste. A classificação com maior probabilidade a posteriori é escolhida como a classe estimada para o conjunto de teste proposto. A verossimilhança e a priori são estimadas pelo treinamento dos dados. As probabilidades zero, especialmente para a classe de anomalias, são suavizadas usando a suavização de Laplace. A técnica básica pode ser generalizada para conjuntos de dados categóricos multivariados, agregando as probabilidades posteriores por atributo para cada conjunto de dados de teste e usando o valor agregado para atribuir uma classificação aos mesmos (CHANDOLA *et al.*, 2009). Diversas variantes dessa técnica foram propostas para detecção de intrusão de rede (BARBARá *et al.*, 2001; SEBYALA *et al.*, 2002; VALDES; SKINNER, 2000; YE *et al.*, 2000; Bronstein *et al.*, 2001), para detecção em vídeos de vigilância (DIEHL, 2002) e para detecção de anomalias em dados de texto (BAKER *et al.*, 1999).

3) Baseado em Máquinas de Vetores de Suporte (SVM ou *Support Vector Machines*): baseiam-se em técnicas de aprendizagem de separação de classes a partir de um conjunto de dados de treinamento (RÄTSCH *et al.*, 2002), na qual é determinada uma região limite que contém as instâncias de dados de treino. As versões mais simples destes métodos fazem separações por hiperplanos das regiões do espaço. Neste tipo de técnica podem ser usados *kernels*, como o da função de base radial (RBF), para aprender regiões complexas. Para cada conjunto de dados de teste, a técnica determina se os dados estão dentro ou fora da região de treino. Se o conjunto de pontos estiver dentro da mesma, ela será declarada como normal, caso contrário, será declarada como anômala. Variantes da técnica básica têm sido propostas para detecção de anomalias em dados de sinal de áudio (DAVY; GODSILL, 2002), e detecção de intrusão de chamadas (ESKIN *et al.*, 2002; HELLER *et al.*, 2003; LAZAREVIC *et al.*, 2003). (MOURAO-MIRANDA *et al.*, 2011) utilizou SVM para reconhecimento de discrepância em padrões faciais e mais recentemente (GARCIA-FONT *et al.*, 2018) aplicaram o método SVM para captura de ataques a sensores de rede.

4) Baseado em Regras: normalmente esta abordagem é aplicada quando se tem pouco conhecimento sobre o conjunto de dados. Este tipo de técnica de detecção de anomalias consiste na criação de regras que capturam o comportamento normal de um sistema. Uma instância de teste que não é coberta por essa regra é considerada uma anomalia. Uma técnica baseada em regras de várias classes consiste em duas etapas. A primeira etapa é aprender regras a partir dos dados de treinamento usando um algoritmo de aprendizado de regras, como RIPPER, Árvores de Decisão e assim por diante. Cada regra possui um valor de confiança associado que é proporcional ao número de instâncias de treinamento classificadas corretamente e o número total de instâncias de treinamento cobertas pela regra. A segunda etapa é encontrar, para cada conjunto de dados de teste, a regra que melhor classifique a instância de dados. Como resultado, quanto menor for esta relação, maior a chance da instância de dados ser uma anomalia. Diversas variantes da técnica baseada em regras foram propostas (TENG *et al.*, 1990a; FAN *et al.*, 2001; LEE *et*

al., 1997; HELMER *et al.*, 1998; SALVADOR *et al.*, 2004).

3.2 Métodos de Agrupamento

A técnica de agrupamento é usada para agrupar pontos semelhantes em *clusters*, em outras palavras, grupos de dados homogêneos entre si. O agrupamento é basicamente uma técnica não supervisionada, cujo propósito é agrupar automaticamente dados segundo alguma métrica de associação (por exemplo, através de centroides, que são os pontos médios dentro de um *cluster*). Segundo (BANO; KHAN, 2018), as técnicas de agrupamento de dados são normalmente categorizadas em dois grandes tipos: procedimentos de particionamento e procedimentos hierárquicos. O procedimento hierárquico visa criar uma hierarquia de *clusters*, semelhante à árvore de decisão e seus resultados são mostrados na forma de dendrogramas. Já o agrupamento por métodos de particionamento visa criar várias partições no conjunto de dados e as avalia segundo algum padrão.

Embora métodos de agrupamento e detecção de anomalia pareçam ser fundamentalmente diferentes, várias técnicas de detecção de anomalias baseadas em clusterização foram desenvolvidas (CHANDOLA *et al.*, 2009). Isso é possível, pois tais métodos permitem separar conjunto de dados normais de pontos discrepantes que eventualmente serão classificados como anomalias. Seguem algumas vantagens desse tipo de técnica para investigação de anomalias:

1) As técnicas baseadas em agrupamento podem operar em um modo não supervisionado;

2) Tais técnicas podem ser adaptadas frequentemente a outros tipos de dados complexos, simplesmente conectando um algoritmo de clusterização e ajustando ao novo conjunto de dados que se queira testar;

3) A fase de testes para técnicas baseadas em agrupamento é rápida, pois o número de *clusters* com os quais cada conjunto de dados de teste precisa ser comparada é pequena;

Por ser um método de fácil aplicação, muitas técnicas de detecção de anomalias usando clusterização foram desenvolvidas para diferentes domínios de aplicação. E cada uma delas segue um conjunto diferente de regras para definir a similaridade entre os conjuntos de dados. Alguns dos principais são: métodos de conectividade, centroides, distribuição e densidade. Abaixo está um breve resumo de cada um deles, retirado de (XU; TIAN, 2015):

1) Modelos de conectividade: a ideia básica desse tipo de algoritmo de agrupamento é construir um relacionamento hierárquico entre os dados. Suponha que cada ponto de dados represente um *cluster* individual no início e, em seguida, os dois *clusters* mais próximos sejam agrupados em um novo *cluster* até que ao término do processo exista

apenas um único resultante. Outra forma é classificando todos os dados como um único *cluster* e, em seguida, os particiona à medida que a sua distância aumenta. Exemplos conhecidos de *cluster* hierárquico são os métodos BIRCH e CURE.

2) Modelos centroides: são algoritmos iterativos de clusterização nos quais a noção de similaridade é derivada pela proximidade de um ponto de dados ao centroide dos *clusters*. Os algoritmos de clusterização *K-Means* e *K-Medoids* são métodos populares que se enquadram nessa categoria. Nestes modelos, o número de *clusters* é previamente parametrizado no algoritmo, o que torna importante ter conhecimento anterior do conjunto de dados. Além disso, esses modelos são executados iterativamente para encontrar o ótimo local. Na seção 3.2.1 o algoritmo *K-Means* será apresentado em detalhes.

3) Modelos de distribuição: esses modelos de armazenamento em *cluster* são baseados na noção de que é provável que todos os pontos de dados no *cluster* pertençam à uma mesma distribuição conhecida (por exemplo: Normal-Gaussiana, Exponencial, etc.). Esses modelos geralmente sofrem *overfitting*. Exemplo desses modelos são os algoritmos DBCLASD e GMM.

4) Modelos de densidade: a ideia básica desse tipo de algoritmo de agrupamento é que os dados que estão em uma região com alta densidade do espaço de dados são considerados pertencentes a um mesmo *cluster*. Dessa forma, isolam várias regiões de densidade diferentes e atribuem os pontos de dados dentro dessas regiões no mesmo *cluster*. Exemplos populares de modelos de densidade são o DBSCAN, OPTICS e *Mean-shift*.

3.2.1 Algoritmo *K-Means*

O termo termo *K-Means* foi usado pela primeira vez por (MACQUEEN, 1967) e o algoritmo padrão foi proposto de maneira independente por (LLOYD, 1957) e (FORGY, 1965) e, por essa razão, é normalmente descrito como método de Lloyd-Forgy. Seu principal objetivo é encontrar similaridades entre os dados e agrupá-los conforme o número de *clusters* desejado. Por ser extremamente conhecido e apresentar uma abordagem prática de desenvolvimento, conclui-se ser um bom método para aplicação neste projeto.

Conceitos Gerais: em primeiro lugar serão apresentados alguns conceitos básicos sobre o *cluster* e a sua notação. Um *cluster* $C = \{C_1, C_2, ..., C_k\}$ consiste em uma partição de um conjunto de dados em subconjuntos C_i (chamados *clusters*), no qual os elementos de cada *cluster* apresentam alguma semelhança entre si e diferem dos elementos que se encontram nos outros *clusters*.

Considere um conjunto de dados $D = \{x_j\}_{i=i}^{n}$ com n pontos num espaço de dimensão d. Seja $C = \{C_1, C_2, ..., C_k\}$ um *cluster* do conjunto de dados. Para cada cluster C_i existe um ponto z_i que o representa e que é designado por centroide, que será definido como a

média de todos os elementos dentro do *cluster*, isto é:

$$z_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j, \tag{3.1}$$

onde n_i é o número de elementos de cada *cluster*.

Na aplicação dos algoritmos de *cluster* é necessário avaliar iterativamente a qualidade do mesmo e, para tal, recorre-se à soma dos quadrados dos resíduos (*Sum of Squared Errors*) SSE e que mede a semelhança de cada elemento através da sua distância aos centroides:

$$SSE(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} D(x_j, z_i) \tag{3.2}$$

Esta função também pode ser vista como uma avaliação de dispersão dos elementos dentro de um *cluster*.

Descrição do Algoritmo: o *K-Means* aplica uma abordagem iterativa para encontrar o *cluster* que minimiza o SSE convergindo para uma solução local que pode não ser um *cluster* ótimo global. Esse algoritmo começa por distribuir aleatoriamente os pontos do conjunto de dados D em k *clusters* e calcula os centroides através da média dos pontos do *cluster* C_i. Em seguida, são aplicadas duas fases iterativamente: a atribuição de *clusters* e a atualização dos centroides.

Na atribuição de *clusters*, cada ponto $x_j \in D$ é associado ao *cluster* que tem o centroide z_i mais próximo do ponto, isto é, x_j é atribuído ao *cluster* C_{j*} quando:

$$j^* = argmin_{i=1,\ldots,k}\{\|x_j - z_i\|_2^2\}. \tag{3.3}$$

Em seguida, atualizam-se os centroides através da média de todos os pontos que se encontram no cluster C_i. Essas duas fases são realizadas iterativamente até alcançarem um mínimo local, isto é, o *K-Means* converge se os centroides não mudarem após uma iteração. Aqui é possível impor uma condição de parada ao algoritmo, como por exemplo, $\sum_{i=1}^{k} \left\|z_i^t - z_i^{t-1}\right\|_2^2 \leq \epsilon$, onde $\epsilon > 0$ é o limite de convergência e z_i^t é o centroide do *cluster* C_i na iteração t.

O pseudo-código deste método encontra-se no apêndice A. Este é um exemplo de atribuição única, o que significa que cada ponto só pertence a um *cluster*.

Convergência: em (BOTTOU; BENGIO, 1994), os autores mostram as propriedades

de convergência do algoritmo de agrupamento *K-Means* onde apresentam os resultados via métodos do gradiente, algoritmo EM e método de Newton. Uma revisão recente sobre os aspectos de convergência pode ser encontrada na tese de mestrado de (NUNES, 2016).

Seja $D = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^d$ um conjunto finito de elementos, que pretendem ser agrupados em k subconjuntos, de acordo com a semelhança entre eles. Uma formulação para este problema é a seguinte:

$$\begin{aligned} &\text{Min} \quad f(\boldsymbol{W}, \boldsymbol{Z}) = \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} D(x_j, z_i) \\ &\text{s.a.:} \quad \sum_{i=1}^{k} w_{ij} = 1, j = 1, 2, ..., n \\ &\qquad w_{ij} = 1 \quad ou \quad w_{ij} = 0, \qquad i = 1, 2, ..., k \qquad j = 1, 2, ..., n \end{aligned} \tag{3.4}$$

tal que $\boldsymbol{Z} = [z_1, z_2, ..., z_k] \in R^{d \times k}$, z_i é o centroide do *cluster* i e $\boldsymbol{W} = [w_{ij}]$ sendo uma matriz real de dimensão $k \times n$. Se w_{ij} tomar o valor 1 significa que o ponto x_j pertence ao *cluster* i e verifica-se o contrário quando o seu valor é 0. A função D mede a semelhança entre o ponto x_j e o centroide z_i que, no algoritmo *K-Means*, será dado por:

$$D(x_j, z_i) = \|x_j - z_i\|_2^2 = \sum_{l=1}^{d} (x_{lj} - z_{li})^2 \tag{3.5}$$

Em (NUNES, 2016), o autor demonstra que a função de semelhança $D(x_j, z_i) = \|x - z_i\|_2^2$ usada pelo *K-Means* é convexa e, após algumas etapas, chega na convergência da mesma.

Ordem de Complexidade: (NUNES, 2016) faz uma revisão sobre o assunto: seja D o conjunto de dados que contém, no total, n pontos e cada ponto apresenta uma dimensão d. Ao ser aplicado um método de *cluster* a D, esse fica dividido em k *clusters*, $C = \{C_1, C_2, ..., C_k\}$, sendo cada *cluster* representado pelo respectivo centroide. Denota-se por t o número de iterações que o algoritmo leva até convergir, cada uma delas composta por duas fases: a fase de atribuição dos pontos aos *clusters* e a fase da atualização do centroide de cada *cluster*. É importante considerar o esforço computacional envolvido em cada uma delas. A fase de atribuição dos n pontos aos k *clusters* necessita calcular as n distâncias dos pontos de dimensão d aos k centroides, o que leva um tempo de $O(nkd)$. Denotando por n_i a dimensão do *cluster* C_i, e tendo em conta que $\sum_{i=1}^{k} n_i = n$, é possível avaliar a ordem de complexidade da fase de atualização dos *clusters*. O centroide do cluster C_i é atualizado com o cálculo da média dos seus n_i pontos (de dimensão d) originando um tempo de $O(n_i d)$. Deste modo, o tempo total desta fase é $O(nd)$. Tendo em conta que cada iteração t necessita executar cada uma dessas fases, conclui-se que esse algoritmo tem uma complexidade computacional de $O(tnkd)$.

A proposta de uso desse método consiste em avaliar nos *clusters* resultantes do algoritmo, quais os pontos mais distantes de seu centroide. Desta maneira, entende-se que quanto maior for a distância, maiores serão as chances deste dado ser uma anomalia.

3.3 Método *Isolation Forest*

Como discutido anteriormente, a maioria das abordagens de detecção de anomalias são baseadas em modelos que definem um perfil para o conjunto de dados normais e, em seguida, identifica as instâncias de dados que não estão em conformidade com o perfil normal. O método *Isolation Forest* propõe um isolamento explícito das anomalias ao invés de definir perfis normais no conjunto de dados.

O uso deste tipo de isolamento permite que o método proposto por (LIU *et al.*, 2008), *iForest*, explore sub-amostragens de uma forma que não é viável nos métodos anteriores, criando um algoritmo com uma baixa complexidade linear de tempo e de baixa memória computacional. Esse método se mostra muito versátil, pois também funciona em problemas de alta dimensão que possuem um grande número de atributos irrelevantes e em situações em que o conjunto de treinamento não contém anomalias classificadas. Uma pesquisa nesse contexto foi proposta por (PUGGINI; MCLOONE, 2018) para avaliação do uso de dados de espectroscopia de emissão ótica (OES), onde a alta dimensão e a natureza correlacionada dos dados do OES podem limitar o desempenho de busca dos sistemas de detecção de anomalias. Estudos interessantes aplicando o algoritmo na captura de ameaças de intrusão cibernéticas foram propostos por (KAREV *et al.*, 2017; AHMED *et al.*, 2019; SIDDIQUI *et al.*, 2019). (Fan *et al.*, 2017) aplicaram este método de detecção para previsão de falências em bancos poloneses. (Weng; Liu, 2019) propõe uma adaptação do algoritmo para abordagem em anomalias coletivas para fluxos multidimencionais em segurança de equipamento celulares. (FILIPPOV *et al.*, 2018) propõem o uso do algoritmo para autenticação de usuários em dispositivos celulares via reconhecimento de padrões de desbloqueio. Aplicações dessa técnica em outras áreas podem ser vistas em (LIU *et al.*, 2012), também em (WU *et al.*, 2018) com detecção de movimentações atípicas em sensores de dados, (ZAREAPOOR *et al.*, 2012; SURYANARAYANA *et al.*, 2018; SHARMILA *et al.*, 2019) com detecção de fraudes em cartões de crédito.

Basicamente, o método tira proveito das propriedades quantitativas observadas em anomalias, a saber:

(i) Elas são a minoria, o que consiste em menos instâncias;

(ii) Elas possuem valores de atributos que são muito diferentes daqueles de instâncias normais.

Em outras palavras, as anomalias são "poucas e diferentes", o que as tornam mais suscetíveis ao isolamento do que os pontos normais. Os autores mostram que uma estrutura de árvores pode ser construída de forma eficaz para isolar cada instância de dados. E, por conta de sua suscetibilidade ao isolamento, as anomalias são isoladas mais próximas da raiz da árvore; enquanto pontos normais são isolados na extremidade mais profunda da árvore. Esta característica de isolamento da árvore constitui a base do método *Isolation Forest* para detectar anomalias, e que os autores também denominam de isolamento de árvores ou simplesmente *iTree*.

O método proposto chamado *Isolation Forest* ou *iForest* ou ainda isolamento de floresta, constrói um conjunto de *iTrees* para um determinado conjunto de dados, e em seguida, as anomalias são classificadas como as instâncias que 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. 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.*, 2008). Neste trabalho, os valores dos parâmetros foram definidos conforme as recomendações dos autores (ou seja, números de árvores igual a 100 e tamanho mínimo de $2^8 = 256$ subamostras dos dados).

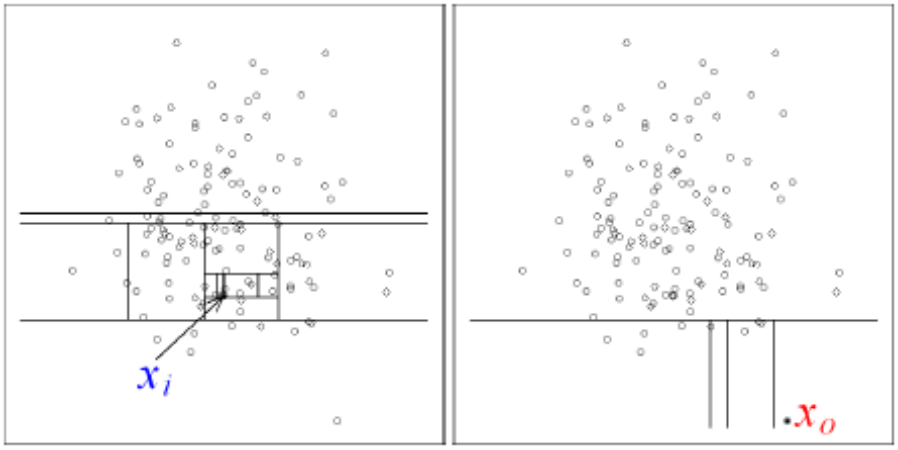

FIGURA 3.1 – Partições necessárias para separarem um ponto normal x_i (à esquerda) e um ponto anômalo x_0 (à direita). Fonte:(LIU *et al.*, 2008).

A Figura 3.1 mostra que anomalias são mais suscetíveis ao isolamento e, portanto, têm comprimentos curtos. Dada uma distribuição normal bivariada (com 135 observações), um ponto normal x_i requer ao todo doze partições aleatórias para que seja isolado. Enquanto uma anomalia x_0 requer apenas quatro partições para que seja isolada.

Esse particionamento aleatório produz caminhos menores perceptíveis para anomalias, pois as poucas ocorrências de anomalias resultam em um número menor de partições, ou seja, caminhos mais curtos em uma estrutura de árvore, enquanto que instâncias com

valores de atributos normais têm maior probabilidade de serem particionadas em um número elevado de iterações.

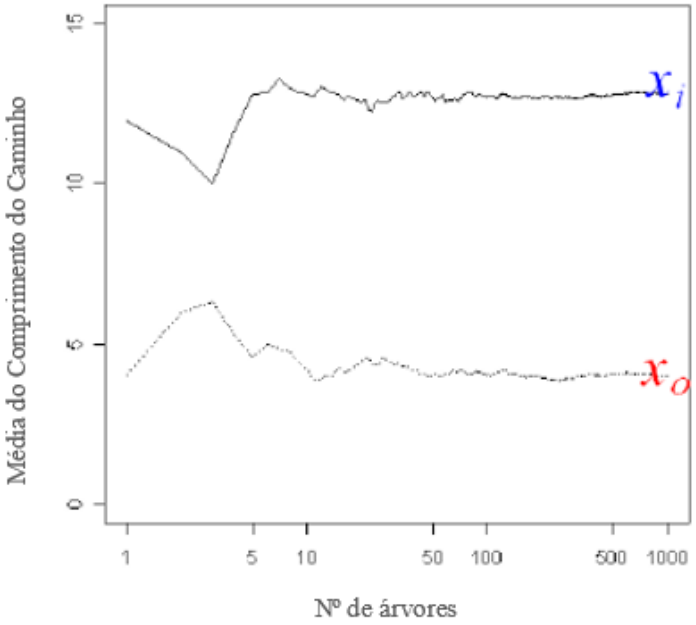

FIGURA 3.2 – Comprimentos médios de convergência para x_i e x_0 quando o número de árvores aumenta. Fonte: (LIU *et al.*, 2008).

Considerando o exemplo anterior, foi repetido o isolamento para 1.000 simulações. Como cada partição foi gerada aleatoriamente, as árvores individuais foram geradas com diferentes conjuntos de partições. A Figura 3.2 apresenta os comprimentos médios de convergência para os pontos x_i e x_0, normal e anômalo respectivamente, a medida que o número de árvores aumenta (simulações considerando de 1 a 1.000 árvores). Medindo o comprimento do caminho (eixo y) é possível encontrar o comprimento esperado de x_0 e x_i (ambos convergem quando o número de árvores aumenta). Nota-se que a partir da 100ª árvore simulada, os comprimentos médios ficam estabilizados e, considerando 1.000 árvores simuladas, os comprimentos médios do caminho de x_0 e x_i convergem para 4.02 e 12.82, respectivamente. Isso mostra que as anomalias (x_0) apresentam comprimentos de caminho menores que as instâncias normais (x_i).

Definição da Árvore de Isolamento (*iTree*): é uma árvore binária aleatória com dois tipos de nós: nós externos sem filhos e nós internos com exatamente dois filhos, conforme mostrado na Figura 3.3. 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 $q < p$.

Seja $X = \{x_1, ..., x_n\}$ o conjunto de dados de uma distribuição d-variável de n instâncias. Uma amostra de ψ instâncias $X' \subset X$ é usada para construir uma árvore de isolamento (*iTree*). Para isso, divide-se recursivamente X' selecionando aleatoriamente

um atributo q e um valor de divisão p, até que:

(i) $|X'| = 1$, ou seja, o nó possui apenas uma instância de dados; ou

(ii) todos os dados em X' têm os mesmos valores.

Em suma, o desenvolvimento de construção da *iTree* é definido como um processo de divisão do espaço de entrada inicial recursivamente em subespaços menores até que todas as instâncias de dados estejam classificadas.

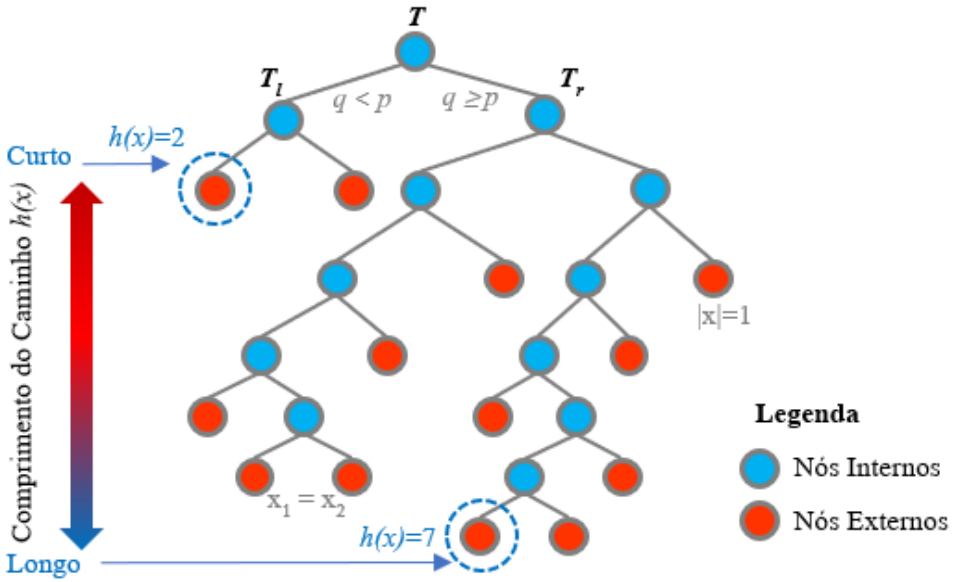

FIGURA 3.3 – Representação de uma *iTree*.

Como dito anteriormente, verifica-se pela Figura 3.3 que uma *iTree* é uma árvore binária, na qual cada nó possui exatamente zero ou dois nós filhos. Assumindo que todas as instâncias são distintas, cada uma delas é isolada para um nó externo quando uma *iTree* está completa. Além disso, nota-se que, dadas ψ instâncias, o número de nós externos também será ψ, enquanto o número de nós internos será ψ - 1 e, por fim, o número total de nós de uma *iTree* será 2ψ - 1. Dessa forma, o requisito de memória para cada *iTree* aumenta linearmente com o tamanho ψ do conjunto de dados de entrada.

Com intuito de mostrar um caso didático do funcionamento da *iTree*, a seguir será apresentado um exemplo simples de caráter ilustrativo. Seja uma instância de dados Y, composta pelos seguintes elementos $Y = \{8, 8, 17, 20, 20, 31, 40, 58, 60, 60, 65, 72, 95\}$, ou seja, de tamanho $\psi = 13$. A Figura 3.4 mostra os passos de construção de uma *iTree* aleatória. Os registros em azul representam os elementos do vetor Y. A medida que a árvore vai crescendo, as instâncias de dados são separadas continuamente (estão representadas pela cor vermelha), conforme o teste aleatório $q < p$. O algoritmo finaliza quando todos os pontos do conjunto de dados estejam isolados.

A Figura 3.5 representa instâncias do *Isolation Forest*. O *iForest* é um conjunto de

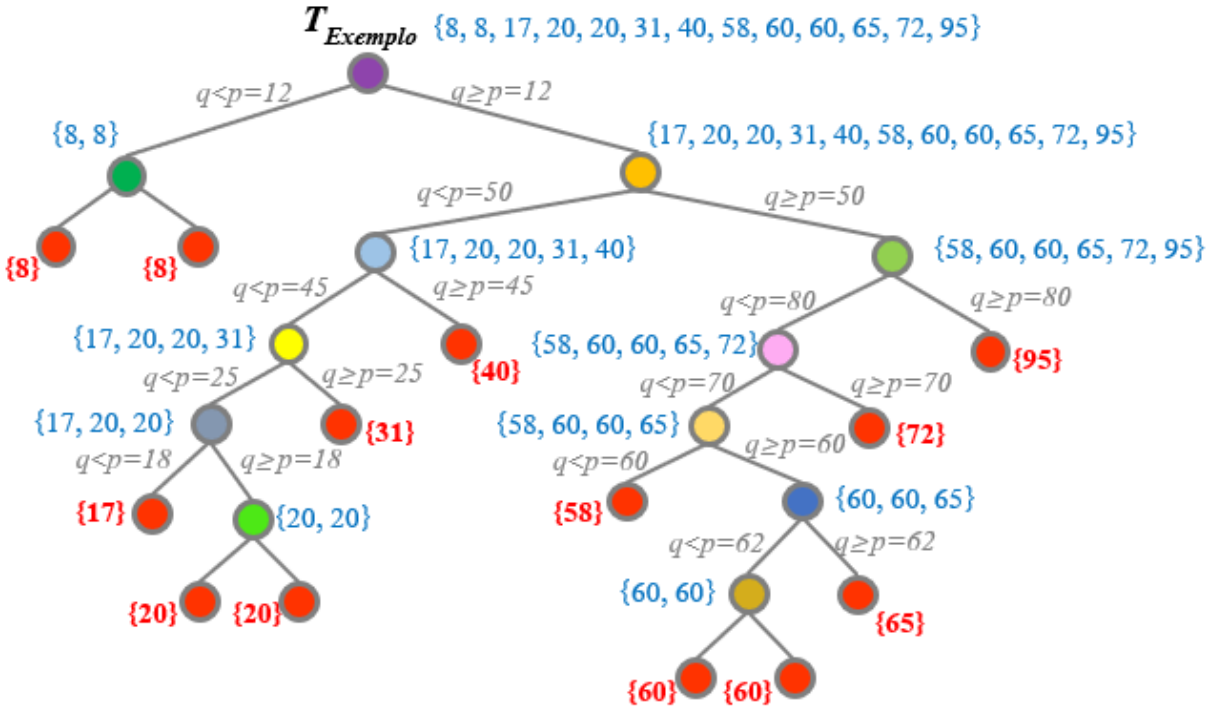

FIGURA 3.4 – Exemplo ilustrativo de construção de uma *iTree*.

iTrees aleatórias. Se as árvores de isolamento aleatórias produzem comprimentos de caminho mais curtos para algumas instâncias de dados específicas, é muito provável que sejam anomalias. Caso contrário, sendo o caminho mais longo, é provável que sejam uma observações normais. Ainda observando a Figura 3.5, uma mesma instância classificada pela cor amarela representa diferentes caminhos tomados em cada *iTree* gerada aleatoriamente.

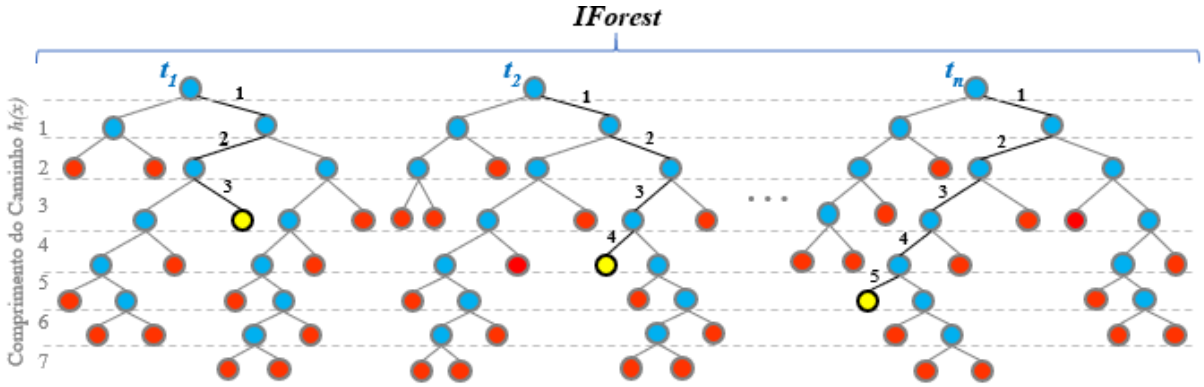

FIGURA 3.5 – Representação de um *iForest*.

A tarefa de detecção de anomalias é fornecer uma classificação que reflita o grau de anomalia. Assim, uma maneira de detectar anomalias é classificar instâncias de dados, de acordo com seus comprimentos de caminho. (LIU *et al.*, 2008) definiram o comprimento do caminho e a pontuação da anomalia que estão descritas a seguir.

Definição do Comprimento do Caminho *h(x)*: de um ponto x é medido pelo

número de arestas x que atravessa uma *iTree* do nó raiz até que o percurso seja terminado em um nó externo. Exemplos de comprimentos de caminho estão ilustrados na Figura 3.3.

Como mencionado anteriormente, os pontos de dados anômalos são mais suscetíveis ao isolamento e, portanto, têm comprimentos de caminho curtos. Por outro lado, os pontos de dados normais geralmente exigem que mais partições sejam isoladas e, portanto, têm comprimentos de caminho longos. Assim, as pontuações de anomalias podem ser definidas como funções dos comprimentos dos caminhos.

A técnica *Isolation Forest* é definida pela aplicação de um numeroso conjunto de árvores de isolamento aleatórias (*iTrees*), vide a Figura 3.5. Extrapolando, seja $F = \{i_1, ..., i_n\}$, para cada árvore i é possível calcular o número de iterações $h_i(x)$ necessário para isolar uma amostra x. O número médio de etapas necessárias para isolar uma amostra x em uma floresta é dada por:

$$h(x) = \frac{1}{N} \sum_{i \in F} (h_i(x)) \tag{3.6}$$

Voltando à Figura 3.5, é possível calcular o comprimento do caminho $h(x)$ para a instância de dados da cor amarela através da Equação 3.6. Considerando a informação presente na Figura 3.5 e, tomando apenas os três caminhos disponíveis, temos $h(x)$=1/3 *(3+4+5)=4, ou seja, um $h(x)$=4.

Definição de pontuação de anomalia: é simplesmente um score de classificação de anomalias. Um ponto importante é que como as *iTrees* são construídas por subconjuntos de dados divididos por atributos aleatórios, consequentemente os scores de anomalia construídos também são aleatórios. O número de etapas necessárias para isolar uma observação x é influenciado pelo número de amostras ψ no conjunto de dados. Uma pontuação de anomalia normalizada $s(x, \psi)$ é definida como:

$$s(x, \psi) = 2^{-\frac{E(h(x))}{c(\psi)}} \tag{3.7}$$

Onde $E(h(x))$ é a média de $h(x)$ de um conjunto de árvores de isolamento. Como o *iTree* tem uma estrutura equivalente à *Binary Search Tree* (BST ou árvore de busca binária), a estimativa da média $h(x)$ para terminações de nós externos usada será a mesma. Então, $c(\psi)$ é dado por:

$$c(\psi) = \begin{cases} 2H(\psi - 1) - 2(\psi - 1)/\psi, & \text{se } \psi > 2, \\ 1, & \text{se } \psi = 2, \\ 0, & \text{caso contrário,} \end{cases} \tag{3.8}$$

em que $H(i)$ é o número harmônico e pode ser estimado por $ln(i)$ + 0,5772156649 (cons-

tante de Euler). Pode ser provado que $c(\psi)$ é o número médio de passos necessários para isolar uma amostra das outras ψ amostras (LIU *et al.*, 2008). Nesse sentido, pode ser usada como um fator de normalização de $h(x)$ o que torna o valor de s independente do número de amostras (ψ).

Na Equação 3.7, as seguintes condições fornecem três valores especiais de pontuação de anomalias:

- Quando $E(h(x)) \rightarrow c(\psi), \quad s \rightarrow 0,5$;
- Quando $E(h(x)) \rightarrow 0, \quad s \rightarrow 1$; e
- Quando $E(h(x)) \rightarrow \psi - 1, \quad s \rightarrow 0$.

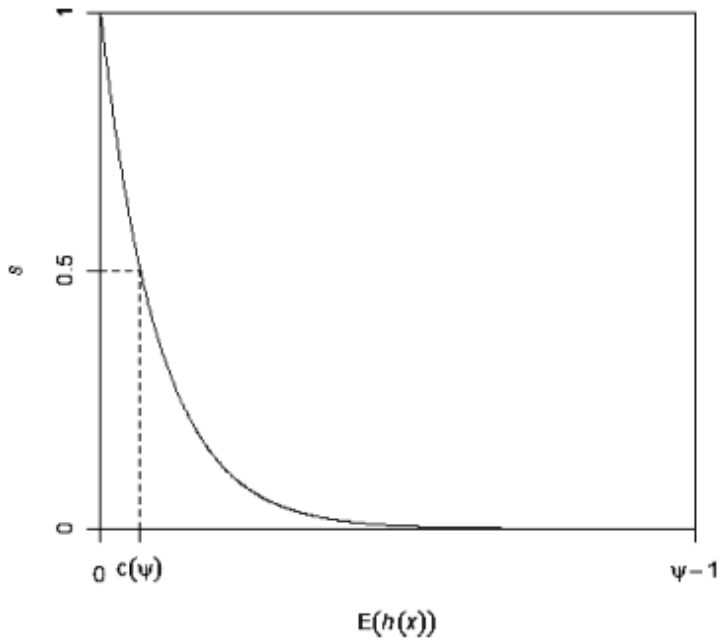

FIGURA 3.6 – Relação do comprimento esperado do caminho $E(h(x))$ e pontuação s da anomalia. Fonte: (LIU *et al.*, 2008).

Além disso, s é monótono para $h(x)$. A Figura 3.6 ilustra o relacionamento entre $E(h(x))$ e s e as seguintes condições aplicadas onde $0 < s \leq 1$ para $0 < h(x) \leq \psi - 1$. Usando as pontuações de anomalia s, é possível fazer a seguinte avaliação:

(i) Se as instâncias retornarem s muito próximas de 1, elas serão definitivamente anomalias;

(ii) Se as instâncias forem muito menores que 0,5, elas terão grandes chances de serem consideradas instâncias normais;

(iii) Se todas as instâncias retornarem $s \approx 0,5$, a amostra de dados provavelmente não terá nenhuma anomalia distinta.

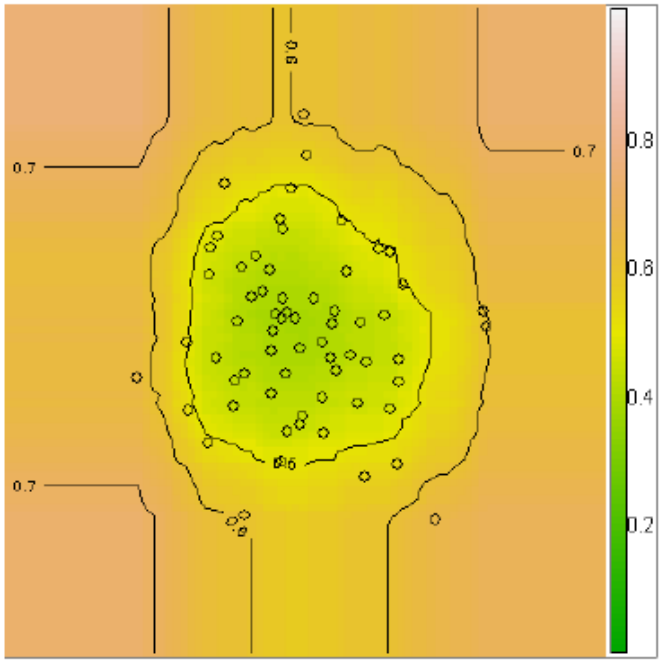

FIGURA 3.7 – Contagem de anomalias do *iForest* para uma distribuição normal de 64 observações. Fronterias de contorno para $s = 0,5; 0,6; 0,7$ são ilustradas. Fonte: (LIU *et al.*, 2008).

Uma fronteira de contorno pode ser produzida para o score de anomalias, basta aplicar a amostra de dados a um conjunto de árvores de isolamento. Isso facilita uma análise detalhada do resultado da detecção. A Figura 3.7 mostra um exemplo dessa fronteira de contorno (usando uma distribuição normal com 64 observações), permitindo que um usuário visualize e identifique anomalias no espaço da instância de dados. Usando a fronteira de contorno, é possível identificar claramente três pontos, onde $s \geq 0,6$, como potenciais anomalias.

Detecção de Anomalias usando o *Isolation Forest*: a detecção de anomalias usando *iForest* é um processo que consiste em dois estágios. No primeiro ocorre o desenvolvimento de uma *iTree*, que é construída particionando recursivamente o conjunto de dados de treinamento até que as instâncias sejam isoladas, ou seja, atingida uma altura de árvore específica da qual resulta um modelo parcial. Detalhes do estágio de treinamento podem ser encontrados nos Pseudo Algoritmos 1 e 2, apresentados nas Figuras 3.8 e 3.9. Existem dois parâmetros de entrada para o algoritmo *iForest*. Eles são o tamanho da subamostragem ψ (que controla o tamanho dos dados de treinamento) e o número de árvores t (que controla o tamanho do conjunto de árvores).

O segundo estágio é o de avaliação, onde uma pontuação de anomalia s é derivada do comprimento do caminho esperado $E(h(x))$ para cada instância de teste. $E(h(x))$ são gerados através de cada *iTree* em um *iForest*. Um único comprimento de caminho $h(x)$ é derivado da contagem do número de arestas e do nó raiz para um nó de terminação à

Algorithm 1 : *iForest*(X, t, ψ)

Inputs: X - input data, t - number of trees, ψ - subsampling size
Output: a set of t *iTrees*

```
Initialize Forest
for i = 1 to t do
    X' ← sample(X, ψ)
    Forest ← Forest ∪ iTree(X')
end for
return Forest
```

FIGURA 3.8 – Pseudo algoritmo *iForest*. Fonte: (LIU *et al.*, 2008)

Algorithm 2 : *iTree*(X')

Inputs: X' - input data
Output: an *i*Tree

```
if X' cannot be divided then
    return exNode{Size ← |X'|}
else
    let Q be a list of attributes in X'
    randomly select an attribute q ∈ Q
    randomly select a split point p between the max and min values of attribute
    q in X'
    X_l ← filter(X', q < p)
    X_r ← filter(X', q ≥ p)
    return inNode{Left ← iTree(X_l),
                  Right ← iTree(X_r),
                  SplitAtt ← q,
                  SplitValue ← p}
end if
```

FIGURA 3.9 – Pseudo algoritmo *iTree*. Fonte: (LIU *et al.*, 2008)

medida que a instância x atravessa uma *iTree*. Quando o percurso é definido para um limite de altura predefinido (*hlim*), o valor de retorno é e mais um ajuste $c(Size)$. Esse ajuste considera a estimativa de um comprimento médio de caminho de uma subárvore aleatória que poderia ser construída usando dados de tamanho (*Size*) além do limite de altura da árvore. Quando $h(x)$ é obtido para cada árvore do conjunto, uma pontuação de anomalia é produzida calculando $s(x, \psi)$ na Equação 3.7. Esse processo pode ser acompanhado pelo Pseudo Algoritmo 3, apresentado na Figura 3.10.

Ordem de Complexidade: de acordo com (LIU *et al.*, 2008), o *iForest* tem uma ordem de complexidade de tempo $O(t\psi log(\psi))$ no estágio de treinamento e $O(ntlog(\psi))$ no estágio de avaliação, onde n é o tamanho dos dados de teste. A rápida execução do *iForest* com baixa exigência de memória é um resultado direto da construção de modelos parciais e requer apenas um tamanho de amostra significativamente pequeno em comparação com o conjunto de treinamento fornecido. Essa capacidade apresenta grande vantagem no domínio da detecção de anomalias.

Algorithm 3 : *PathLength*$(x, T, hlim, e)$

Inputs : x - an instance, T - an *i*Tree, $hlim$ - height limit, e - current path length; to be initialized to zero when first called

Output: path length of x

1: **if** T is an external node or $e \geq hlim$ **then**
2: return $e + c(T.size)$ {$c(.)$ is defined in Equation 1}
3: **end if**
4: $a \leftarrow T.splitAtt$
5: **if** $x_a < T.splitValue$ **then**
6: return $PathLength(x, T.left, hlim, e + 1)$
7: **else** $\{x_a \geq T.splitValue\}$
8: return $PathLength(x, T.right, hlim, e + 1)$
9: **end if**

FIGURA 3.10 – Pseudo algoritmo do comprimento do caminho. Fonte: (LIU *et al.*, 2008)

3.4 Método Combinado *K-Means+Isolation Forest*

Como mencionado na Seção 1.2.3, com o intuito de melhorar a assertividade da captura de casos anômalos em conjunto de dados, foi estabelecido um método combinando o algoritmo de agrupamento *K-Means* e a técnica de isolamento de observações *Isolation Forest*. O *KM+IF* consiste em duas etapas: na primeira são criados *clusters* utilizando como parâmetro a distância euclidiana e na segunda é aplicado o *Isolation Forest* em cada partição para avaliar e determinar a existência de anomalias. Para exemplificar como esta abordagem pode ser útil, a Figura 3.11 ilustra três situações simuladas que expressam diferentes níveis de complexidade para captura de anomalias.

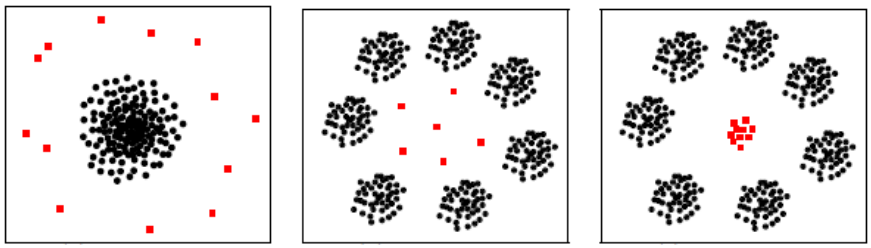

FIGURA 3.11 – Desafios e complexidade na detecção de anomalias. Fonte: (CHANDOLA *et al.*, 2009).

Nota-se pela Figura 3.11 as seguintes situações: à esquerda consta uma simulação de um conjunto de dados normal onde foram acrescentados pontos anômalos (em vermelho) distantes e ao redor da distribuição; no quadro ao meio há um conjunto de dados gerados com sete variações formando grupos homogêneos e ao centro pontos anômalos esparsos; por fim, à direita os mesmos conjuntos de dados formando sete variações homogêneas e uma oitava considerando uma nuvem coesa de pontos anômalos. Na primeira situação, uma série de métodos podem capturar com certa facilidade as anomalias presentes no conjunto de dados, entretanto, nas situações seguintes é muito difícil definir uma região

normal e que separe as anomalias.

Para os exemplos apresentados na Figura 3.11, uma vantagem observada é que fazendo um agrupamento inicial do conjunto de dados, permite que a dimensionalidade do problema seja reduzida e assim é possível gerar subgrupos homogêneos de informação (através do *K-Means*) e em seguida investigar se há anomalias locais presentes nos mesmos. A ideia dessa técnica híbrida é detectar anomalias de forma mais sensível na região específica de um *cluster* e tornar as anomalias locais em anomalias globais após aplicação do *Isolation Forest*.

Os trabalhos de (Kurnianingsih *et al.*, 2018) e (GAO *et al.*, 2019) mostram que a aplicação conjunta das técnicas trazem benefícios para assertividade geral do ajuste. Os autores 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. Com base nessas informações e resultados obtidos, e por acreditar que essa abordagem possa ser mais assertiva em determinadas situações, serão aplicados os modelos combinados *K-Means* e *Isolation Forest* para os conjuntos de dados presentes no Capítulo 4.

3.4.1 Métricas de Performance

Para avaliação dos resultados de ajuste dos modelos, uma série de métricas de desempenho são aplicadas na literatura. Em (ZHU *et al.*, 2017) os autores sugerem o uso das seguintes medidas de performance para mensurar detecção de fraudes: Acurácia, Precisão, *Recall* ou Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e F1-score. Essas medidas são derivadas de uma matriz de confusão, apresentada na Tabela 3.1. A seguir estão alguns detalhes de cada medida de performance.

TABELA 3.1 – Matriz de Confusão.

	Condição Verdadeira	
Condição da Predição	**Positivo (Anomalia)**	**Negativo (Não Anomalia)**
Positivo (Anomalia)	Verdadeiro Positivo (TP)	Falso Positivo (FP)
Negativo (Não Anomalia)	Falso Negativo (FN)	Verdadeiro Negativo (TN)

1. **Acurácia:** é a razão entre os dados classificados corretamente e o número total de observações, ou seja, consiste na taxa de acerto do modelo onde casos de anomalias verdadeiras são corretamente classificadas como anomalias e casos de não anomalias são classificados como normais. É calculada usando a Equação 3.9. Quanto maior seu valor indica que melhor será a performance do modelo, contudo deve ser avaliada

com cautela quando aplicada em dados desbalanceados pois pode levar à uma decisão errada.

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.9}$$

2. **Precisão:** é a razão entre as anomalias classificadas corretamente e o total de anomalias detectadas pelo modelo. Dessa forma, quanto maior for seu valor, melhor será a performance do modelo. É calculada usando a Equação 3.10.

$$Precisão = \frac{TP}{TP + FP} \tag{3.10}$$

3. ***Recall* ou Taxa de Verdadeiros Positivos (TPR):** é a proporção entre as anomalias classificadas corretamente e o número total de anomalias. Em outras palavras, consiste na volumetria de eventos de risco capturados. É calculada usando a Equação 3.11. Quanto maior a taxa, melhor será a qualidade de ajuste do modelo.

$$TPR = \frac{TP}{TP + FN} \tag{3.11}$$

4. **Taxa de Falsos Positivos (FPR):** é a razão entre as não anomalias classificadas incorretamente como anomalias e o número total de não anomalias, ou seja, é a volumetria de casos de não risco classificados como anomalias. É calculada usando a Equação 3.12. Quanto menor seu valor, melhor será o resultado do ajuste.

$$FPR = \frac{FP}{FP + TN} \tag{3.12}$$

5. **F-measure ou F1-score:** combina a Taxa de Verdadeiros Positivos e a Precisão através de uma média harmônica, sendo que quanto maior seu valor, melhor é a qualidade de ajuste do modelo. É calculada usando a Equação 3.13.

$$F1score = \frac{2 * Precisão * TPR}{Precisão + TPR} \tag{3.13}$$

Um fato importante relacionado a essas medidas é que podem apresentar diferenças relevantes de um problema para outro. Na literatura não encontramos um *benchmark* com valores de assertividade mínima para se considerar uma boa qualidade de ajuste do modelo na abordagem não supervisionada. Além disso, é indicado o uso com parcimônia dessas medidas, principalmente a acurácia, visto que não é suficiente para julgar o desempenho dos modelos quando a varável de interesse está desbalanceada. Normalmente, é necessário a combinação das métricas de avaliação para se obter uma decisão.

4 Estudo de Caso

Neste capítulo serão apresentadas as aplicações dos métodos descritos anteriormente em conjuntos de dados reais. Na seção 4.1 será mostrado um problema muito conhecido na literatura acadêmica, relacionado ao conjunto de dados Iris (FISHER, 1936). Já na seção 4.2 será detalhado o problema principal deste trabalho. Através de dados divulgados das últimas campanhas eleitorais de 2018, obtidos por meio de fontes públicas, serão aplicados os algoritmos *Isolation Forest* e *KM+IF*, com intuito de criar uma ferramenta para apoio na tomada de decisão.

4.1 Experimento Inicial

A primeira análise realizada para aplicar os métodos *Isolation Forest* e *KM+IF* neste projeto foi no conjunto de dados Iris (FISHER, 1936). Esse conjunto é constituído de dados multivariados introduzido pelo estatístico e biólogo britânico Ronald Fisher, em seu artigo de 1936, para uso de múltiplas medições em problemas taxonômicos, como um exemplo de análise discriminante linear. Esses dados foram coletados por (ANDERSON, 1935) para quantificar a variação morfológica das flores de íris de três espécies relacionadas. Duas das três espécies foram coletadas na Península de Gaspé "todas do mesmo pasto e colhidas no mesmo dia e medidas ao mesmo tempo pela mesma pessoa com o mesmo aparelho".

O conjunto de dados consiste em 50 amostras de cada uma das três espécies de flores Iris (setosa, virgínica e versicolor), que podem ser observadas na Figura 4.1. Quatro características foram medidas a partir de cada amostra: o comprimento e a largura das sépalas e pétalas, em centímetros. Com base na combinação dessas quatro características, (FISHER, 1936) desenvolveu um modelo discriminante linear para distinguir as espécies umas das outras.

Para efeito comparativo, o primeiro teste realizado para avaliação do conjunto de dados foi uma análise descritiva através de *box-plot* simples para identificar qual a distribuição das pétalas e sépalas de forma univariada. Os testes realizados partem do pressuposto que não há a priori a variável de espécie de iris para treinamento, ou seja, dentro do conjunto de dados, irá se buscar a classificação das espécies e ao mesmo tempo os eventuais pontos

FIGURA 4.1 – Imagem das três espécies de Iris (Setosa, Virgínica e Versicolor).

anômalos.

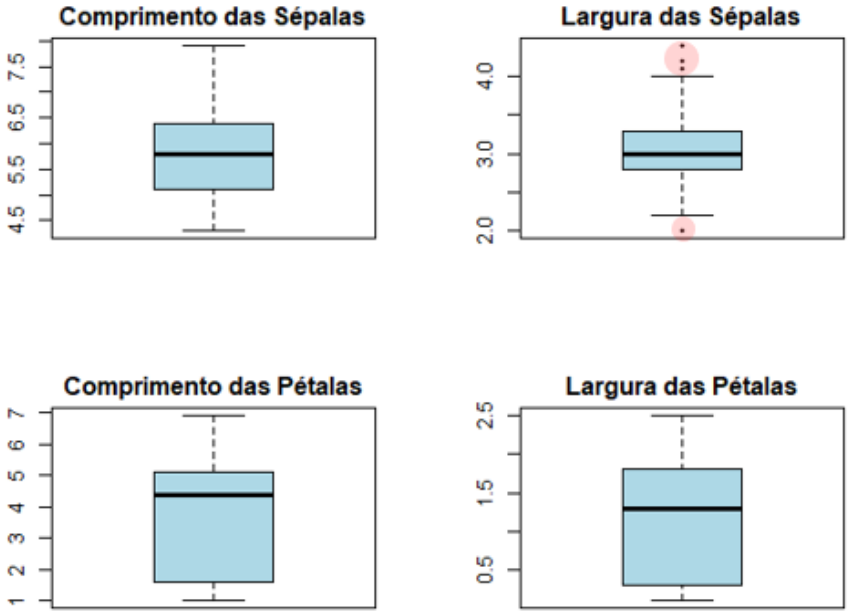

FIGURA 4.2 – Box-plot do conjunto de dados Iris. Comprimento e Largura das Sépalas e Pétalas. Há presença de 4 *ouliers* na largura das sépalas.

Observa-se pela Figura 4.2, de forma univariada, a distribuição das medidas das sépalas apresentam valores discrepantes, segundo definição do Box-Plot, que considera limites inferiores e superiores como $LimiteInferior : max\{min(dados); Q_1 - 1,5(Q_3 - Q1)\}$ e $LimiteSuperior : min\{max(dados); Q_1 + 1,5(Q_3 - Q1)\}$.

Através da Figura 4.3, nota-se que após a classificação dos grupos em *clusters* pelo método *K-Means*, a distribuição das observações ficam bem ajustadas quando comparadas pelo comprimento e largura das pétalas. Para as variáveis de comprimento e largura das sépalas o ajuste não é perfeito, uma vez que há erros de classificação para as espécies virgínica e versicolor.

Como o interesse é descobrir se os métodos *K-Means* e *Isolation Forest* conseguem ser assertivos na captura de anomalias dentro dos grupos de espécies, será incluído de forma aleatória novas observações com características anômalas no conjunto de dados. A

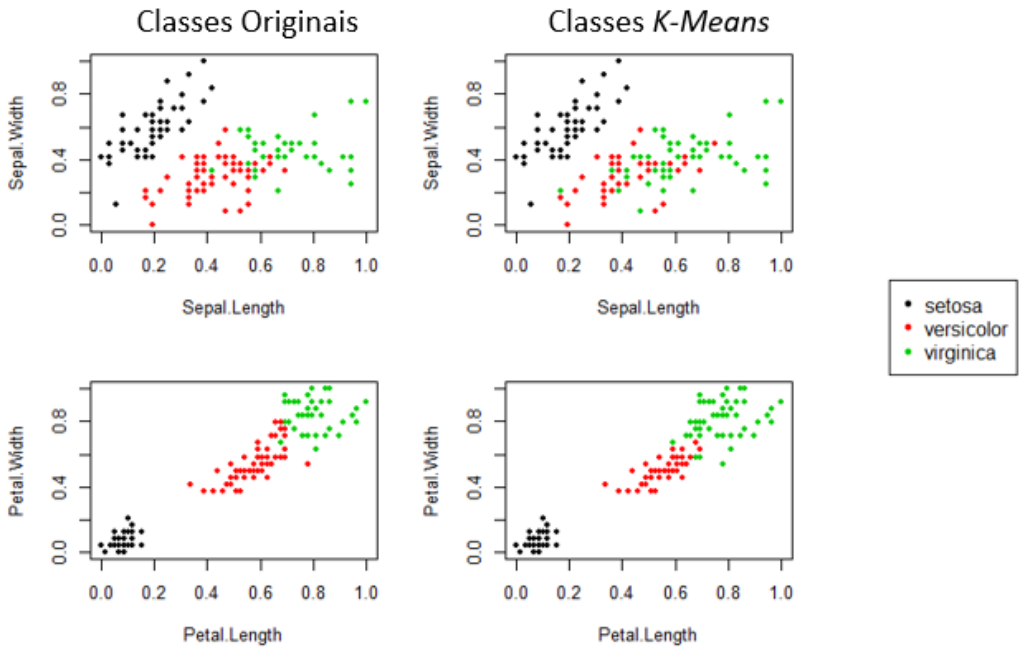

FIGURA 4.3 – Comparação entre classes originais e após aplicação do método *K-Means* com k=3.

Figura 4.4 representa a distribuição dos *clusters* das espécies após aplicado o algoritmo *K-Means*. As marcações em "x" azuis em destaque representam as anomalias adicionadas artificialmente no conjunto de dados. A Figura 4.5 contém as observações com as maiores distâncias do centroide (pontos azuis) do *cluster*. Para defini-las utilizou-se o critério de maior distância euclidiana com um ponto de corte igual ao percentil 95% da distribuição de observações.

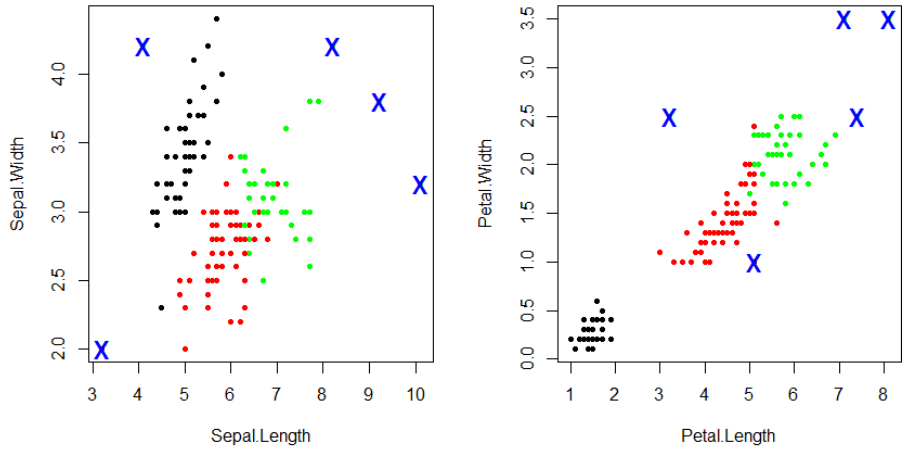

FIGURA 4.4 – *Outliers* simulados para aplicação do modelo *K-Means*.

Verifica-se pela Figura 4.5, com o ponto de corte igual ao percentil 95% da distribuição das maiores distâncias, que todas as anomalias simuladas foram capturadas pelo método *K-Means*. Além das cinco anomalias geradas artificialmente, outras três observações do conjunto de dados também foram classificadas como anomalias.

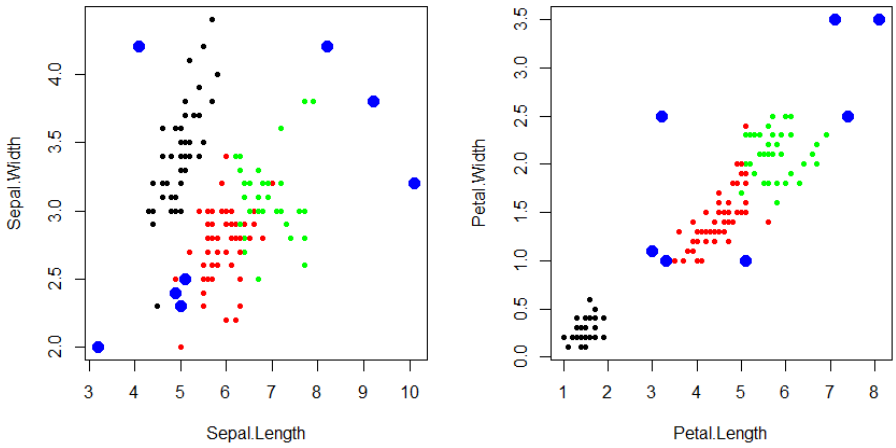

FIGURA 4.5 – Classificação de Anomalias após ajuste do modelo *K-Means*.

O próximo teste será realizar o ajuste do algoritmo *Isolation Forest* para este conjunto de dados a fim de avaliar se existem possíveis anomalias por este método. Para tal, foi considerada como anomalia os pontos onde o score do modelo foi maior que o percentil 95% da distribuição. Os resultados podem ser observados na Figura 4.6. Assim como no método *K-Means*, o algoritmo *Isolation Forest* capturou as cinco anomalias simuladas e mais três observações do conjunto inicial dos dados.

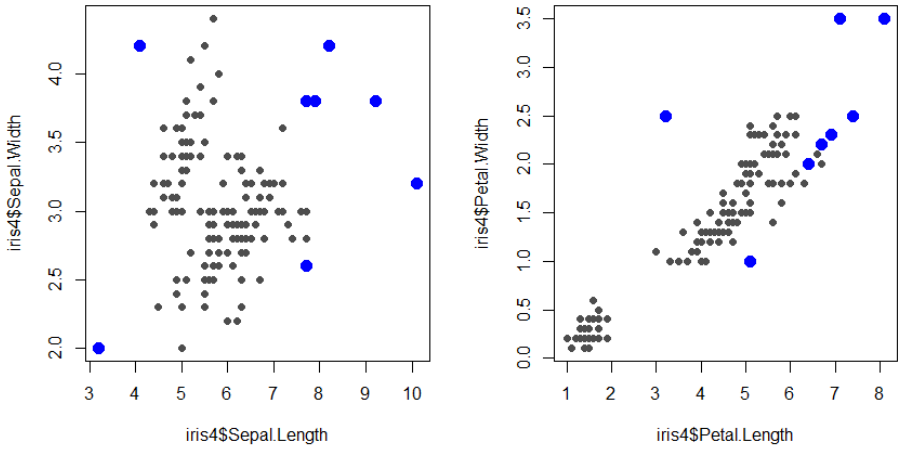

FIGURA 4.6 – Classificação de Anomalias após ajuste do modelo *Isolation Forest*.

O último teste será aplicar os métodos combinados *K-Means + Isolation Forest* no conjunto de dados. Dessa forma, utilizando os *clusters* gerados no primeiro resultado, serão agora submetidos individualmente à execução do algoritmo *Isolation Forest* para comparação da solução. A Figura 4.7 contém o ajuste resultante do modelo combinado *K-Means + Isolation Forest*, onde cada *cluster* está separado com suas respectivas anomalias classificadas. Observa-se que todas as cinco anomalias geradas artificialmente foram capturadas pelo método combinado dos algoritmos e além delas mais cinco anomalias foram identificadas no conjunto dos *clusters*.

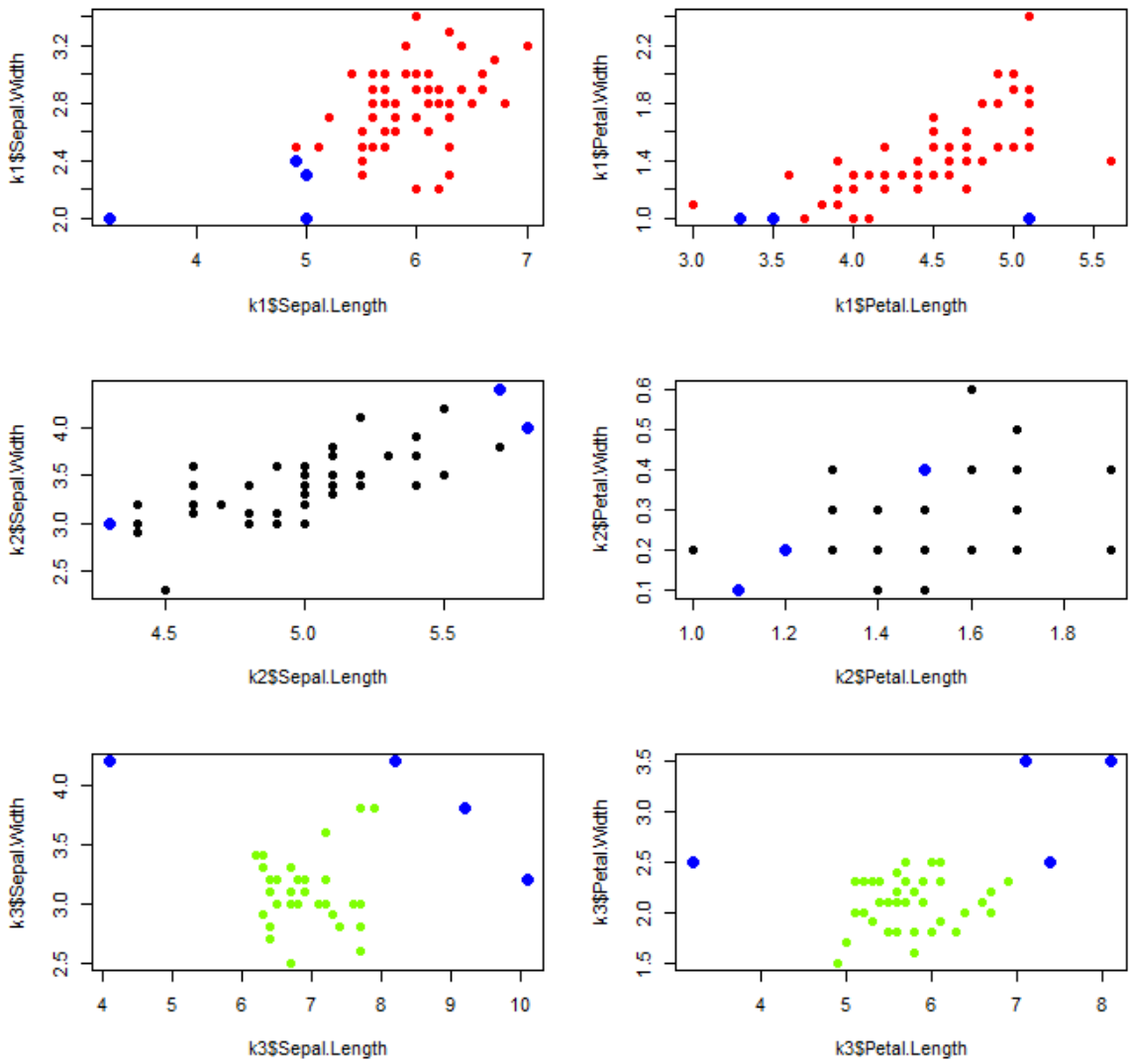

FIGURA 4.7 – Classificação de Anomalias após ajuste do modelo combinado *KM+IF*.

Como resultados gerais da aplicação dos métodos *K-Means*, *Isolation Forest* e *KM+IF* nesse problema, nota-se que todos tiveram boa performance uma vez que classificaram corretamente as anomalias artificiais acrescentadas ao conjunto de dados inicial.

4.2 Estudo de Caso Principal

O principal estudo que será discutido neste trabalho está relacionado com as informações das campanhas eleitorais ocorridas no Brasil, no ano de 2018. A maior motivação para isso é o momento de grande instabilidade ocorrida no Brasil nos últimos anos. A crise política combinada com estratégias econômicas questionadas por um número significativo de opositores ao longo dos anos fizeram com que a credibilidade do cidadão comum perante aqueles que os representa tivesse forte queda. Foram tantos escândalos gerados no cenário político nas últimas décadas que fica difícil mensurar o impacto negativo no

desenvolvimento do país.

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.

Além disso, notícias recentes disponíveis na mídia envolvendo irregularidades de candidatos e parlamentares eleitos geraram alertas para as entidades governamentais de fiscalização, sendo importante apurar os fatos e ter clareza desses acontecimentos.

Por essas razões, acredita-se que um diagnóstico mais aprofundado das candidaturas eleitorais possa fornecer resultados interessantes sobre divergências nas campanhas políticas. Para tal, um método eficiente de análise será através das técnicas de detecções de anomalias.

4.2.1 Descrição do Problema

O primeiro ponto relevante do problema de análise das campanhas eleitorais é entender as leis às quais estão inseridas. 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, por exemplo, quebra de decoro parlamentar, deixar de comparecer à terça parte das sessões legislativa ordinárias, etc.). 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.

Uma vez que estão dispostas leis, regras, normas, etc. é plausível imaginar que haverá alguém ou grupo de indivíduos que tentarão burlar o sistema através de mecanismos ardilosos com objetivo claro de cometer fraudes, desvios da verba pública, entre outros crimes.

Para tratar desse tipo de fenômeno, uma forma prática e eficiente de investigação é correlacionar a maior série de dados disponíveis e avaliar incompatibilidades e discrepâncias entre o que é esperado vis-à-vis possíveis atipicidades.

A investigação científica que será realizada neste trabalho tem grande interesse em mapear e levantar indivíduos que se enquadram em situações atípicas e anômalas. Como exemplo, houve recentemente uma série de divulgações por meio do COAF (Conselho de Controle de Atividades Financeiras - antes ligado aos Ministérios da Fazenda e Justiça e a partir da medida provisória n^o 893/2019 está ligado ao Banco Central do Brasil, como Unidade de Inteligência Financeira (UIF)) de uma série de candidatos que receberam ou transferiram montantes financeiros incompatíveis com suas campanhas eleitorais.

Vale ressaltar que esse trabalho não visa ser um mecanismo de acusação, apenas um instrumento para que os órgãos responsáveis tenham mais facilidade para identificar prioridades nas investigações. Além disso, salienta-se novamente que não foram encontradas pesquisas anteriores que se propõem analisar atipicidades em contas eleitorais e essa particularidade reforça que esse estudo pode contribuir com o avanço e controle na gestão de recursos federais, além de motivar outras pesquisas semelhantes.

4.2.2 Fontes de Informação

Todos os dados coletados foram retirados do site do TSE, por meio de uma *API* (Interface de Programação de Aplicação) desenvolvida na linguagem *Python*. A descrição destes dados no TSE é a seguinte: *"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"*. Os arquivos fornecidos estão em formato .TXT e podem ser importados para programas estatísticos ou de manipulação de base de dados.

A seguir estão listadas as principais bases disponíveis, bem como um resumo dos seus metadados:

1. **Base de Candidatos:** contém informações relacionadas ao perfil dos candidatos nas eleições, declarações de bens e dados sobre os partidos, as coligações e as vagas por cargo e por unidade eleitoral;
2. **Base de Eleitorado:** contém informações relacionadas ao perfil do eleitorado de cada pleito até o grão de zona eleitoral;
3. **Base de Partidos:** contém informações relacionadas aos órgãos partidários e seus membros;

4. **Base de Pesquisas Eleitorais:** contém informações sobre pesquisas eleitorais, notas fiscais dos serviços, questionários e localidades onde foram realizadas;
5. **Base de Prestação de Contas:** contém arquivos de prestação de contas (receitas e despesas de campanha) de candidatos, de partidos e de comitês.
6. **Bases Processuais:** contém informações de processos eleitorais, decisões e recursos referente ao pleito de candidatos;
7. **Bases de Resultados:** contém informações dos resultados das eleições (resultados totalizados e boletins de urna).

Além dessas bases, será utilizada também uma base contendo processos de crimes eleitorais capturada via TSE. Essa informação será muito importante para uso na avaliação da qualidade do ajuste do modelo proposto.

4.2.3 Estrutura do Conjunto de Dados

O conjunto de dados inicial do projeto contém registros de mais de 29.000 candidaturas, desde deputados estaduais até o nível de presidente do Brasil. Neste trabalho, a proposta será trabalhar apenas com os deputados federais (que somam ao todo 8.588 candidatos) por ter abrangência nacional e os tetos de gastos serem os mesmos e iguais a $R\$2.5$ milhões. Por exemplo, o teto de gastos de uma campanha presidencial é de $R\$70$ milhões apenas para o 1º turno das eleições. Isso implicaria que ao analisar todos esses cargos (de deputados a presidente), seria gerado de partida grandes discrepâncias e, portanto, o controle do perfil do público de entrada possibilita resultados mais assertivos.

Em relação ao conjunto de informações do banco de dados, há um vasto número de variáveis disponíveis para construção dos modelos. Além de variáveis cadastrais (tais como, idade, cidade, status da candidatura, partido, gênero e instrução), há outras que informam bens declarados (empresas, imóveis e investimentos), despesas de campanha (alimentação, deslocamento, eventos, estrutura, serviço pessoal, publicidade, financeiras e transferências), receitas (através de doações eletrônicas, recursos políticos diversos e recursos próprios) e votos (totais, por municípios e por zonas eleitorais). Adicionais à essas, também foram construídas variáveis interagidas com as informações iniciais. Como exemplo, custo por voto: razão entre todas as despesas e a quantidade de votos obtida pelo candidato; despesa total: soma de todas as despesas declaradas pelo candidato; bens total: soma de todos os bens declarados pelo candidato; bens de alto risco: bens considerados de risco elevado para lavagem de dinheiro, tais como, dinheiro em espécie e obras de arte; inadimplência: pagamentos atrasados pelo candidato.

4.2.4 Análise Descritiva

Com intuito de apresentar algumas informações em maiores detalhes, a Tabela 4.1 contém informações descritivas de variáveis relevantes do conjunto de dados. Por exemplo, observando a variável "Custo por Voto", enquanto a mediana é igual a $R\$3,46$, houve uma parcela de 1% de candidatos que tiveram um valor igual ou maior do que $R\$19.500,00$ para cada voto. Outra variável interessante é a de "Despesa Total", observa-se que a média de gastos é de $R\$148.000,00$ (e mediana de $R\$5.000,00$), enquanto houve um candidato com uma despesa de $R\$2.591.858,00$, ou seja, ultrapassou o teto de gastos da campanha de Deputados Federais em quase $R\$92.000,00$.

TABELA 4.1 – Distribuição de variáveis de Despesa ($R\$$), Receita ($R\$$) e Votos dos Candidatos.

Variáveis	Qtde	Média	DP	Mín	Mediana	90%	95%	99%	Máx
Despesas Alimentação	8.588	937,07	6.204,61	0,00	0,00	525,00	3.464,92	22.065,00	194.771,20
Despesas Deslocamento	8.588	13.194,30	49.827,63	0,00	0,00	26.244,83	76.714,01	242.287,42	763.519,50
Despesas Diversas	8.588	1.991,86	17.042,64	0,00	0,00	648,60	4.243,52	50.634,35	648.894,95
Despesas Estruturas	8.588	2.305,72	10.266,96	0,00	0,00	4.461,40	12.512,88	43.430,38	234.181,53
Despesas Eventos	8.588	17.863,67	92.538,47	0,00	0,00	12.539,80	72.356,50	453.244,10	1.878.045,00
Despesas Outros	8.588	172,02	3.425,34	0,00	0,00	29,40	269,66	2.625,03	200.783,18
Despesas Serviços	8.588	45.135,28	158.170,72	0,00	630,10	87.341,15	247.828,64	901.018,94	2.299.650,00
Despesas Publicidade	8.588	6.309,17	32.951,72	0,00	0,00	7.200,00	27.000,00	150.000,00	861.000,00
Despesas Financeiras	8.588	224,99	909,67	0,00	0,00	408,66	1.197,07	4.340,99	21.663,90
Despesas Transferências	8.588	11.699,63	72.863,45	0,00	0,00	0,00	20.000,00	363.436,28	1.596.044,60
Despesas Vinc. Campanha	8.588	48.172,70	139.111,22	0,00	985,34	131.461,34	325.350,46	693.368,06	1.939.475,00
Receitas Eletrônicas	8.588	133.580,44	370.299,82	0,00	3.500,00	369.033,51	949.991,65	1.998.434,00	2.513.936,58
Receitas de Risco	8.588	1.448,33	13.630,63	0,00	0,00	1.941,30	5.000,00	23.806,35	750.000,00
Receitas Diversas	8.588	21.896,85	131.337,83	0,00	1.650,61	25.517,27	61.699,94	318.401,88	2.299.660,00
Votos Totais	8.588	10.667	36.745	0	1.058	26.586	63.998	132.409	1.843.735
Despesa Total	8.588	148.006,41	399.869,60	0,00	5.000,00	425.864,38	1.080.817,36	2.097.396,54	2.591.858,00
Custo por Voto	8.588	2.040,00	35.143,26	0,00	3,46	50,38	120,52	19.598,76	47.598,15

4.2.5 Construção do Modelo

Da base inicial de candidatos a deputados federais contendo 8.588 indivíduos, nota-se a existência de candidaturas "vazias", ou seja, sem votos e sem nenhum valor movimentado na campanha (receitas e despesas). Para evitar distorções por conta dessas candidaturas e como esse perfil pouco agrega em informações para o problema principal, foram aplicados simultaneamente dois filtros de materialidade no conjunto inicial de dados:

(i) Quantidade total de votos menor do que 1.000;

(ii) Despesas + Receitas inferior 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 do modelo. Estudos preliminares incluindo estas candidaturas indicavam ruídos nos resultados e ao separar esses casos permitiu uma

análise mais precisa sobre a movimentação das contas eleitorais. Como consequência, o conjunto de dados final para treinamento ficou com uma volumetria igual 3.066 candidatos a deputado federal.

Foram construídos os algoritmos de *Isolation Forest* e o modelo combinado *K-Means + Isolation Forest* no conjunto de dados. Para ajuste do modelo combinado, primeiramente foi aplicado algoritmo *K-Means*, onde foram utilizadas algumas variáveis que melhor representam as divulgações das contas eleitorais, entre as quais: gastos/despesas, receitas e votos. A partir disso foi gerado o gráfico do "cotovelo" (Figura 4.8), também conhecido por *elbow method*, que é um método heurístico para encontrar o número apropriado de *clusters* em um conjunto de dados.

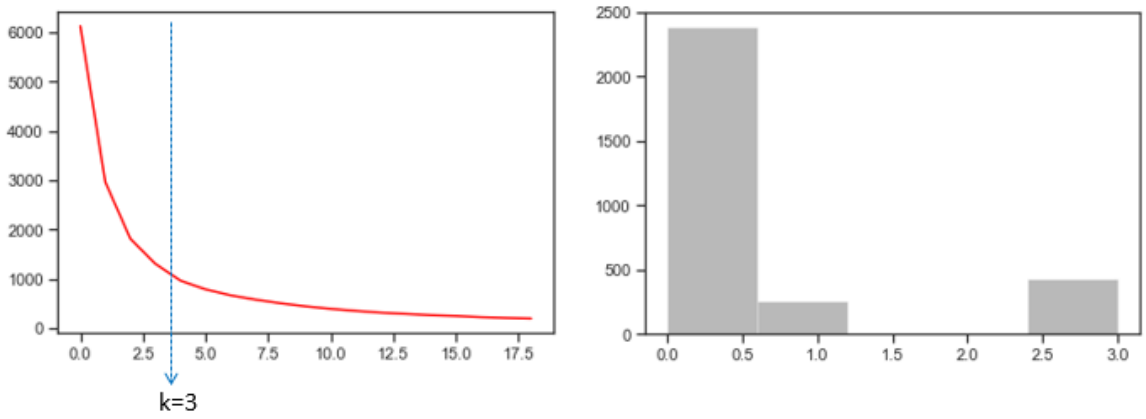

FIGURA 4.8 – Gráfico do "cotovelo" para definir o número de *clusters* (à esquerda) e Volumetria de cada *cluster* gerado (à direita).

Nota-se pela Figura 4.8 que o número ideal de *clusters* é k=3. Além disso, após o agrupamento realizado pelo algoritmo, obtêm-se a seguinte volumetria de observações para cada *cluster*: k_1=2.383, k_2=254 e k_3=429.

Outro 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 *K-Means + Isolation Forest* valores acima do percentil 90% da distribuição dos candidatos, ou seja, esses indivíduos apresentam maiores chances de serem anomalias por apresentarem maiores pontuações.

A Figura 4.9 representa a distribuição dos scores *Isolation Forest* (à esquerda) e *KM+IF* (à direita) dos candidatos. Assumindo o ponto de corte igual ao percentil 90% da distribuição para ambos casos, resulta respectivamente, valores de score acima de 0,52 e 0,51.

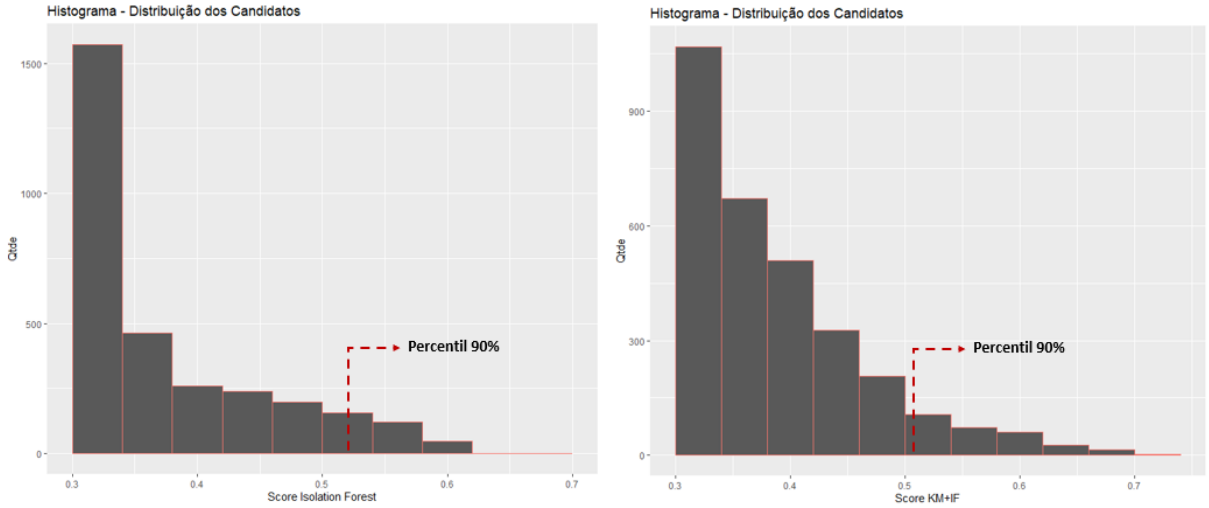

FIGURA 4.9 – Distribuição do Score dos Candidatos e marcação do percentil 90%.

As Figuras 4.10 e 4.11 representam respectivamente, as distribuições dos scores dos algoritmos *Isolation Forest* e *K-Means + Isolation Forest*, classificadas para os 3.066 candidatos a deputado federal. Além disso, com intuito de ser mais informativo, os gráficos dos scores estão associados a algumas variáveis relevantes: bens de alto risco, custo por voto, votos totais, bens totais, inadimplência e despesa total. Para efeito de visualização, dividiu-se a distribuição através da mediana e estão marcados respectivamente, em vermelho e azul, as observações à esquerda e à direita da mesma.

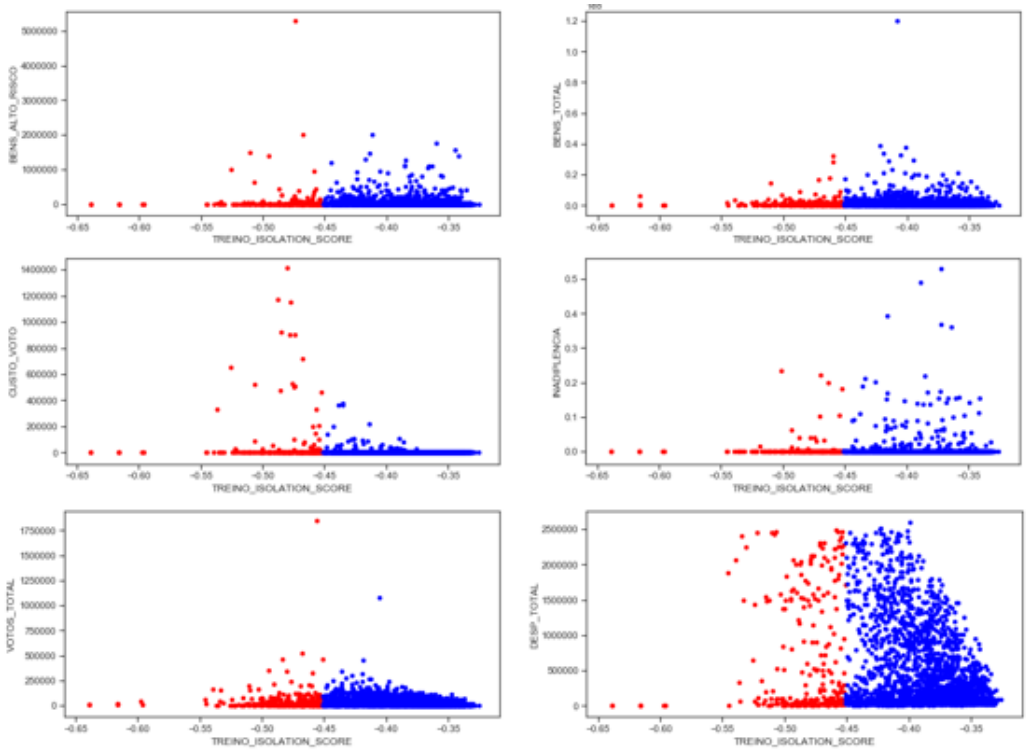

FIGURA 4.10 – Distribuição do score *Isolation Forest* e variáveis: bens de alto risco, custo por voto e votos total (à esquerda); bens total, inadimplência e despesa total (à direita)

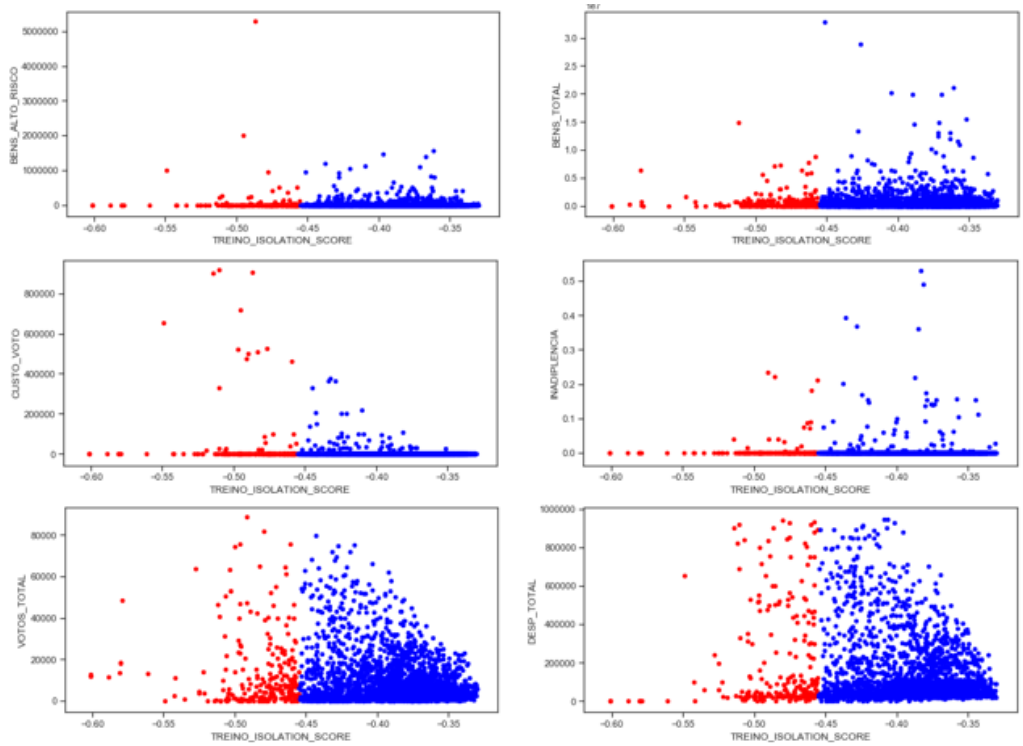

FIGURA 4.11 – Distribuição do score do modelo combinado *K-Means + Isolation Forest* e variáveis: bens de alto risco, custo por voto e votos total (à esquerda); bens total, inadimplência e despesa total (à direita)

4.2.6 Construção de uma Variável de Interesse

Para avaliar o resultado do modelo, 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 4.2 mostra a distribuição de candidatos com algum processo na justiça (inclui criminal, tributário, trabalhista cível/administrativo e eleitoral) e considera duas situações: processos adquiridos historicamente e processos ativos na referência de outubro de 2019. Para este projeto serão considerados apenas os crimes eleitorais e por essa razão as volumetrias são mostradas separadamente na Tabela 4.2.

Ainda analisando a Tabela 4.2, verifica-se que historicamente os 1.212 candidatos marcados com "Sim, Eleitorais" possuem ao todo mais de 10.000 processos de crimes eleitorais e quase 35.000 processos de outras origens. Observa-se também que há 97 candidatos com 404 processos de crimes eleitorais ativos na referência de outubro/19, 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.

TABELA 4.2 – Distribuição de Candidatos e Processos Judiciais.

	Processos Históricos			**Processos Ativos (Out/19)**		
Possui Processos	**Candidatos**	**Proc. Totais**	**Proc. Eleitorais**	**Candidatos**	**Proc. Totais**	**Proc. Eleitorais**
Sim, Outros	774	10.726	0	772	3.318	0
Sim, Eleitorais	1.212	34.783	10.095	97	404	142
Nenhum	1.080	-	-	-	-	-
Total	**3.066**	**45.509**	**10.095**	**869**	**3.722**	**142**

A Tabela 4.3 contém um resumo do percentual de candidatos com processos na justiça. Observa-se que 35% dos candidatos nunca tiveram nenhum processo, 25% apresentam algum processo diferente de crime eleitoral (por exemplo cível/trabalhista) e 40% possuem algum processo de crime eleitoral, dos quais 3% estão com processos ativos quando observados no mês de outubro/19.

TABELA 4.3 – Percentuais dos Candidatos com Processos de Crimes Eleitorais.

Possui Processos	**Qtde**	**%**	**Ativos (out/19)**
Sim, Outros	774	25%	772
Sim, Eleitoral	1.212	40%	**97**(3%)
Nenhum	1.080	35%	-
Total	**3.066**	**100%**	**869**

A Tabela 4.4 contém uma visão histórica da quantidade de processos de crimes eleitorais dos candidatos. Nota-se que 18% da base (220 candidatos) possuem 10 ou mais processos nessa categoria.

TABELA 4.4 – Quantidade de Processos de Crimes Eleitorais Históricos dos Candidatos.

Qtde de Processos	**Qtde de Candidatos**	**% de Candidatos**
1	218	18%
2	232	19%
3	139	11%
4	106	9%
5	98	8%
6	62	5%
7	61	5%
8	40	3%
9	36	3%
≥ 10	**220**	18%
Total	**1.212**	**100%**

Dadas as informações apresentadas nas Tabelas 4.2, 4.3 e 4.4, foi construída a variável de interesse para avaliação dos resultados dos modelos. Ela é baseada na combinação

dos 97 candidatos que, possuem processos de crimes eleitorais ativos (Tabela 4.3), com 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 e, 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 pode ser visto na Tabela 4.5 e é igual a 270 observações (valores em negrito - 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 4.5 – Variável de interesse: Processos Eleitorais Totais (270 candidatos).

	Processos Ativos (out/19)		
Processos Históricos (≥ 10)	**Não**	**Sim**	**Total**
Não	2.796	**50**	2.846
Sim	**173**	**47**	220
Total	2.969	97	3.066

4.2.7 Resultados

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* e *(K-Means + Isolation Forest)*. A Tabela 4.6 contém as métricas de performance descritas anteriormente na seção 3.4.1.

TABELA 4.6 – Performance dos Algoritmos aplicados às Contas Eleitorais.

Métricas de Performance (%)					
Algoritmos	**TPR**	**FPR**	**Precisão**	**Acurácia**	**F1-score**
Isolation Forest	30,0%	8,0%	26,5%	86,5%	28,1%
KM+IF	15,9%	9,4%	14,1%	84,0%	14,9%

Nota-se pela Tabela 4.6 que os resultados indicam que o método *Isolation Forest* sozinho apresenta indicadores superiores quando comparado ao método combinado *(K-Means + Isolation Forest)*. Analisando as métricas uma a uma, observa-se os seguintes resultados:

(i) a Taxa de Verdadeiros Positivos (TPR) é praticamente o dobro quando comparados *IF x KM+IF* (30,0% contra 15,9%), ou seja, a volumetria de casos anômalos capturados é quase 2 vezes maior no *Isolation Forest*;

(ii) a Taxa de Falsos Positivos (FPR) é parecida, 8,0% contra 9,4%, com o *Isolation Forest* se saindo levemente superior na classificação de casos de não risco como anomalias;

(iii) o indicador de Precisão mostra resultados bastante superiores do *Isolation Forest* frente ao *KM+IF*, 26,5% contra 14,1%, ou seja, praticamente o dobro da captura das anomalias verdadeiras sobre o total de anomalias classificadas pelos modelos;

(iv) a Acurácia é muito semelhante, 86,5% contra 84,0%, com o *Isolation Forest* se saindo levemente superior no resultado. Este indicador 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*, pois está classificando melhor as observações não anômalas;

(v) o resultado do F1-score é praticamente o dobro no algoritmo *Isolation Forest*. Como ele combina a TPR e a Precisão através de uma média harmônica e esses resultados se mostram superiores, já era esperado que o F1-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* como razoável. Isso porque 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 (baseado na variável "Processos Totais"). Resumidamente, o algoritmo é capaz de capturar quase 3,5 vezes mais anomalias nesse público quando comparado à uma escolha aleatória.

Entretanto, ressalta-se que o resultado do algoritmo combinado *K-Means + Isolation Forest* mostrou desempenho insatisfatório e abaixo do esperado para esse conjunto de dados. Analisando em detalhes esse efeito, há uma hipótese de que essa performance esteja diretamente associada ao nível de acerto do algoritmo *K-Means* aplicado no primeiro passo. 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 4.7.

Verifica-se pela Tabela 4.7 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 nos *clusters*.

TABELA 4.7 – Taxa de Risco em cada *Cluster* e no Conjunto de Dados Total.

Cluster	**Processos Totais**	**Qtde**	**%**
k_1	Não	2.292	96%
	Sim	91	**4%**
	Subtotal	2.383	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	2.796	91%
	Sim	270	**9%**
	Total	3.066	100%

Com intuito de avaliar o resultado do método combinado *KM+IF* dentro de cada *cluster*, nas Tabelas 4.8, 4.9 e 4.10 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 entres os resultados do algoritmo e um potencial benefício de seu uso como parte da solução.

TABELA 4.8 – Performance dos Algoritmos aplicados para o *Cluster* k_1.

Métricas de Performance (%)					
Algoritmos	**TPR**	**FPR**	**Precisão**	**Acurácia**	**F1-score**
Isolation Forest	2,2%	0,6%	13,3%	95,7%	3,8%
KM+IF	30,8%	9,2%	11,8%	88,5%	17,0%

TABELA 4.9 – Performance dos Algoritmos aplicados para o *Cluster* k_2.

Métricas de Performance (%)					
Algoritmos	**TPR**	**FPR**	**Precisão**	**Acurácia**	**F1-score**
Isolation Forest	57,1%	61,6%	26,3%	43,6%	36,0%
KM+IF	10,1%	10,0%	27,9%	67,8%	14,8%

Através das Tabelas 4.8, 4.9 e 4.10 é possível notar que o método combinado *(K-Means + Isolation Forest)* apresenta melhores resultados para o *cluster* k_1. Em contrapartida, para os *clusters* k_2 e k_3 prevaleceu o melhor ajuste do algortimo *Isolation Forest*. Ressalta-se que as métricas de performance mudaram consideravelmente quando comparadas individualmente para cada *cluster*. Por exemplo, no k_1 a Taxa de Verdadeiros Positivos (TPR)

TABELA 4.10 – Performance dos Algoritmos aplicados para o *Cluster* k_3.

Métricas de Performance (%)					
Algoritmos	**TPR**	**FPR**	**Precisão**	**Acurácia**	**F1-score**
Isolation Forest	18,3%	10,8%	34,4%	72,4%	23,9%
KM+IF	5,0%	11,3%	12,0%	68,9%	7,1%

é 14 vezes maior no método *KM+IF* quando comparado ao *Isolation Forest* sozinho (30,8% contra 2,2%) e, além disso, o F1-score é 4,5 vezes maior para o método *KM+IF* (17,0% contra 3,8%). Na prática, o método *Isolation Forest* 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 o método combinado *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 F1-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".

A seguir, será apresentada uma série de Figuras 4.12 à 4.31, com resultados gráficos do ajuste dos modelos *Isolation Forest* e *KM+IF*. As Figuras 4.12 à 4.29 apresentam uma visão bivariada entre o score dos modelos em conjunto com as variáveis: despesas total, custos total e custos por voto. É possível verificar através das Figuras 4.12 à 4.17 uma avaliação dos *clusters* com a distribuição dos modelos *IF e KM+IF* e a variável de interesse Processos Totais. Já as Figuras 4.18 à 4.23 apresentam a classificação final dos modelos *IF e KM+IF*, ou seja, marcação de anomalia ou não, pelo modelo. Além disso, as Figuras 4.24 à 4.29 contêm uma comparação dos resultados das classes dos modelos *IF e KM+IF* versus a variável de interesse Processos Totais. Por fim, as Figuras 4.30 e 4.31 compreendem uma comparação entre os scores dos modelos *IF* e *KM+IF* e as suas classificações.

Como dito anteriormente, nota-se pelas figuras que o agrupamento executado pelo *K-Means* conseguiu identificar relativamente bem os grupos com presença de anomalias, ou seja, percebe-se que candidatos que possuem processos de crimes eleitorais têm correlação com o perfil de anomalias identificadas nas partições dos *clusters*. Dessa forma, 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, a quantidade de anomalias no conjunto se destacam mais dentro do grupo e portanto são mais facilmente identificadas. Isso faz com que o método combinado se saia melhor neste grupo do que o *Isolation Forest* aplicado sozinho. Já para os outros dois *clusters* a taxa de anomalias é muito alta, 28% e 24%. Isso faz com que o conjunto de dados anômalos (do conjunto completo) seja uma quan-

tidade considerável do subconjunto, não sendo portanto anômalos dentro daquele grupo. Um exemplo extremo seria o caso que houvesse um *cluster* formado sobretudo por dados anômalos e apenas um dado normal. Obviamente, com uma visão local do *cluster*, seria o dado normal que deveria ser classificado como discrepante. Isso faz com que o *Isolation Forest* quando aplicado ao *cluster* não capture bem a informação, o que implica no baixo desempenho do método nos *clusters* k_2 e k_3. Como existe quase o dobro de anomalias nos *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 o *Isolation Forest* sozinho.

É evidente que os resultados poderiam ser melhores caso as anomalias identificadas por esses algoritmos fossem classificadas por algum órgão regulador, visto que toda a avaliação de performance está baseada numa variável de interesse baseada em processos judiciais e não na análise minuciosa das contas eleitorais. De toda forma, esse experimento expôs uma situação interessante, na qual o uso das técnicas combinadas *K-Means + Isolation Forest* podem apresentar resultados inferiores dependendo de sua aplicação e que não foram observadas na pesquisas anteriores de (Kurnianingsih *et al.*, 2018) e (GAO *et al.*, 2019).

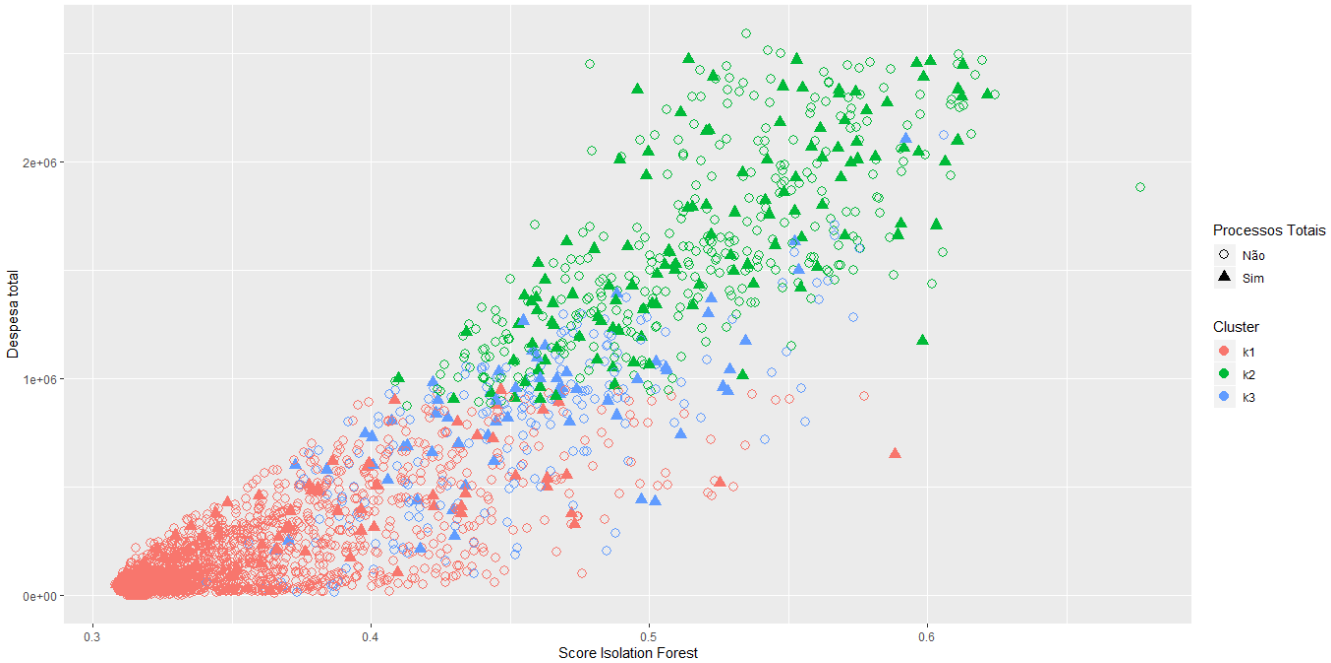

FIGURA 4.12 – Score do modelo *Isolation Forest* x Despesas Total

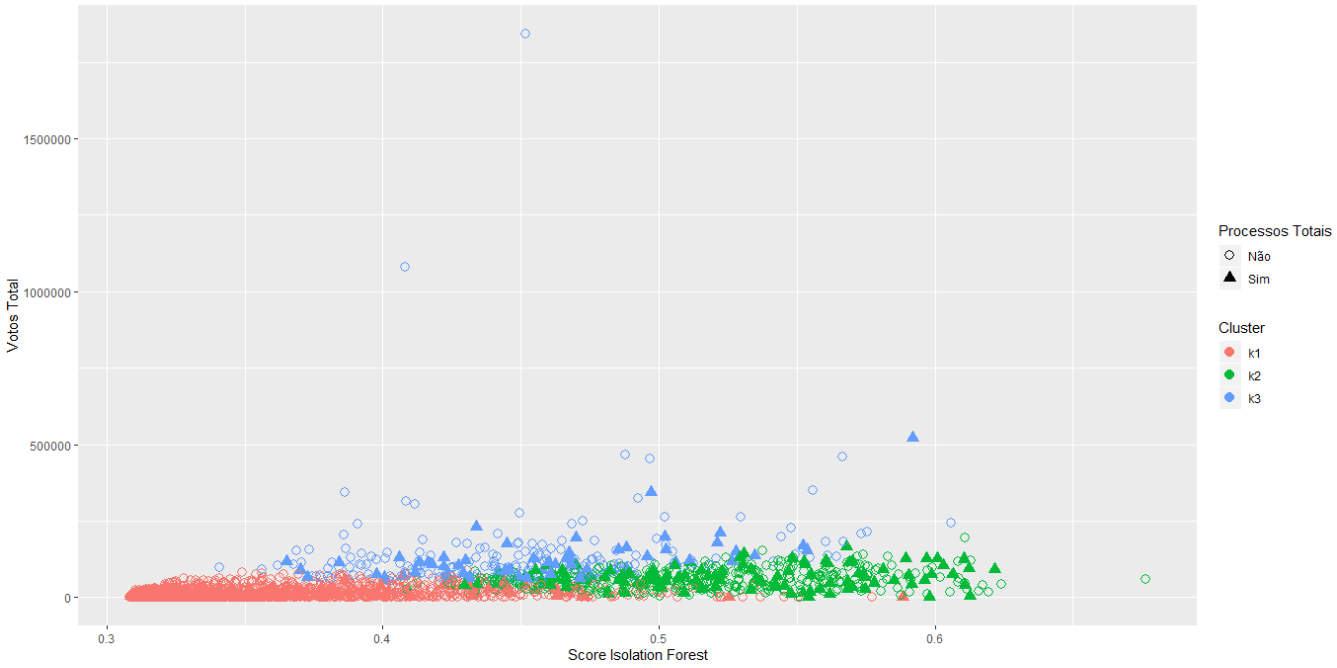

FIGURA 4.13 – Score do modelo *Isolation Forest* x Votos Total

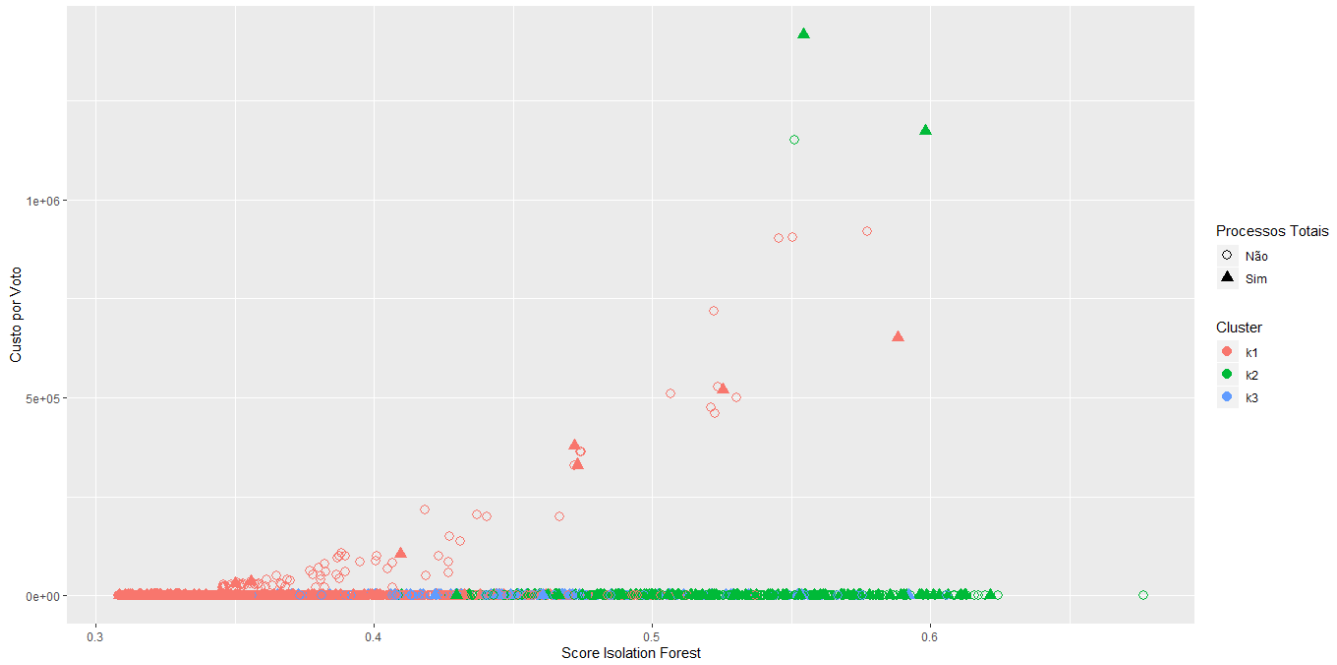

FIGURA 4.14 – Score do modelo *Isolation Forest* x Custo por Voto

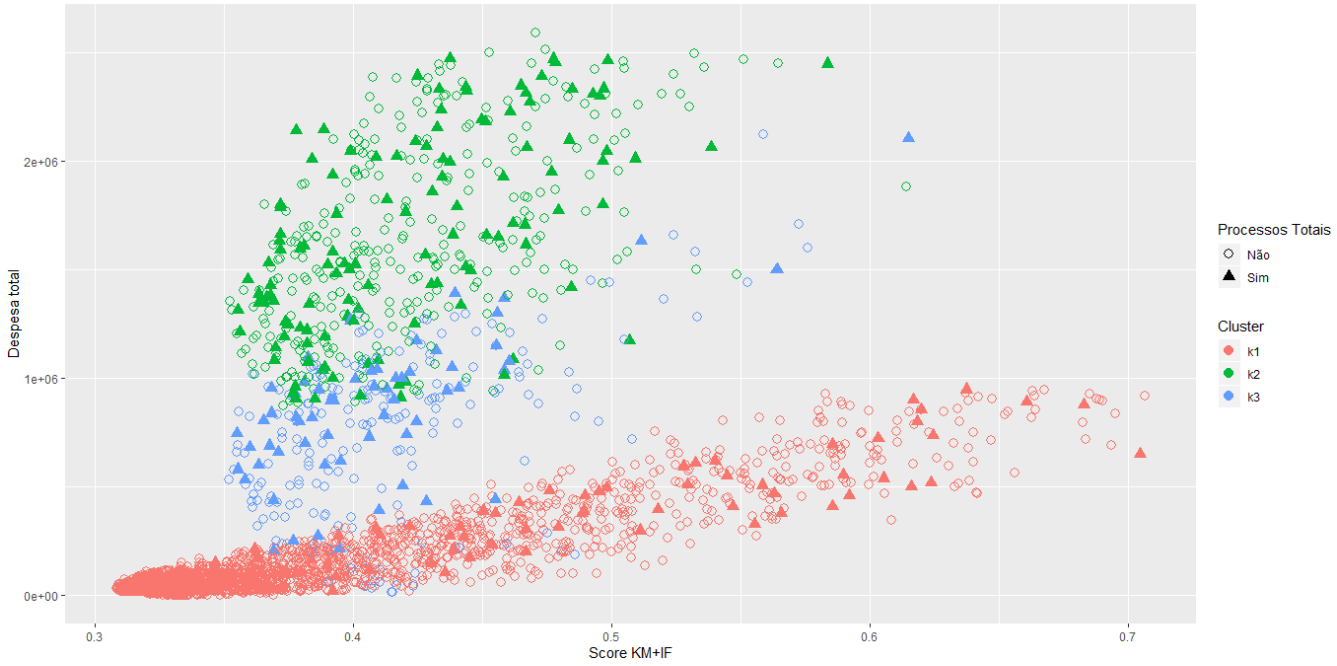

FIGURA 4.15 – Score do modelo combinado *KM+IF* x Despesas Total

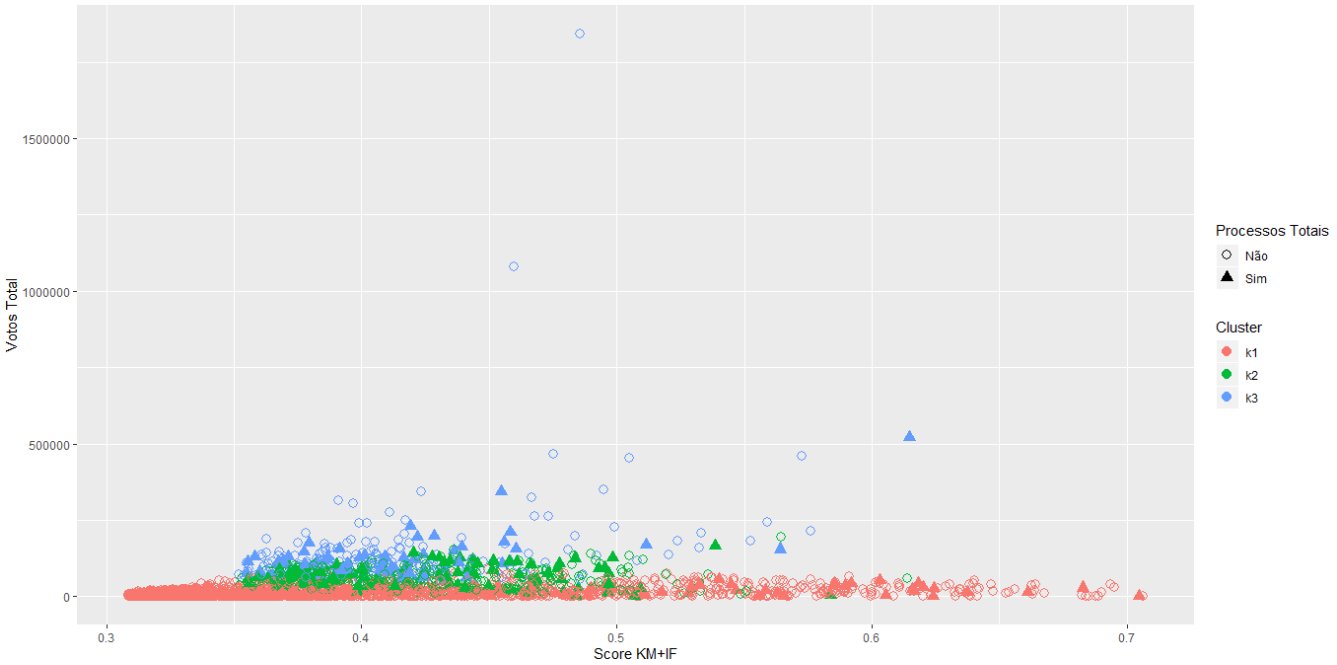

FIGURA 4.16 – Score do modelo combinado *KM+IF* x Votos Total

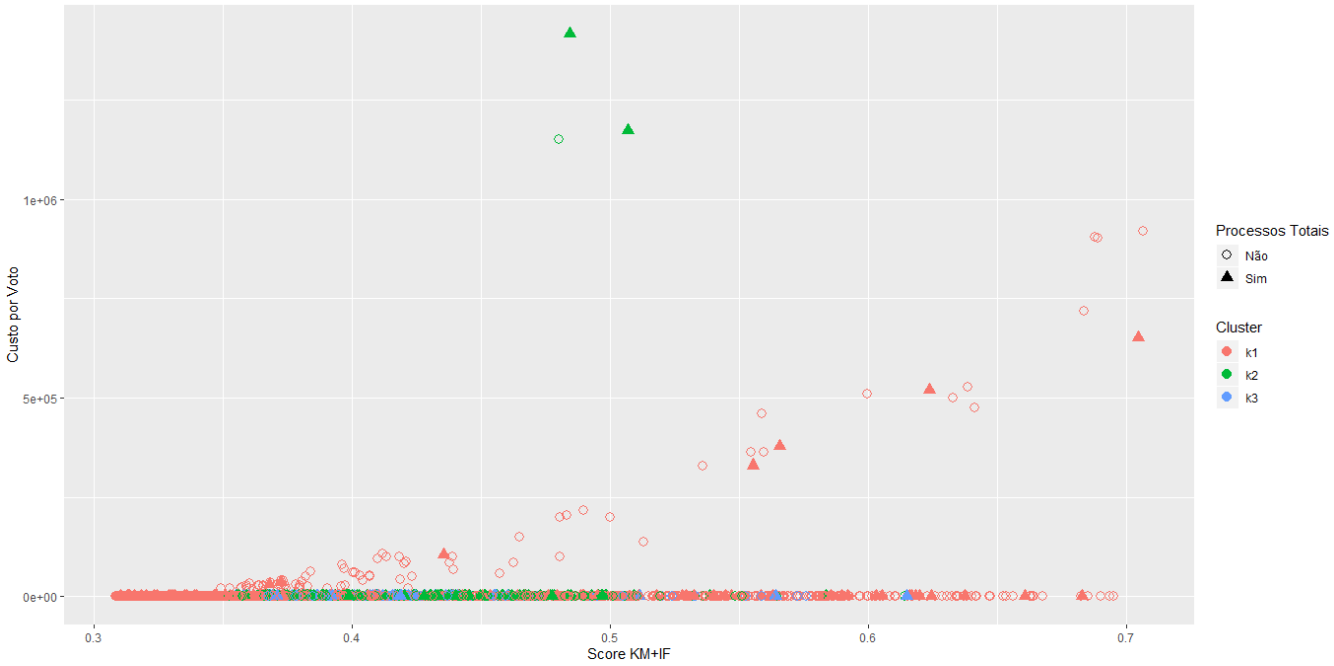

FIGURA 4.17 – Score do modelo combinado *KM+IF* x Custo por Voto

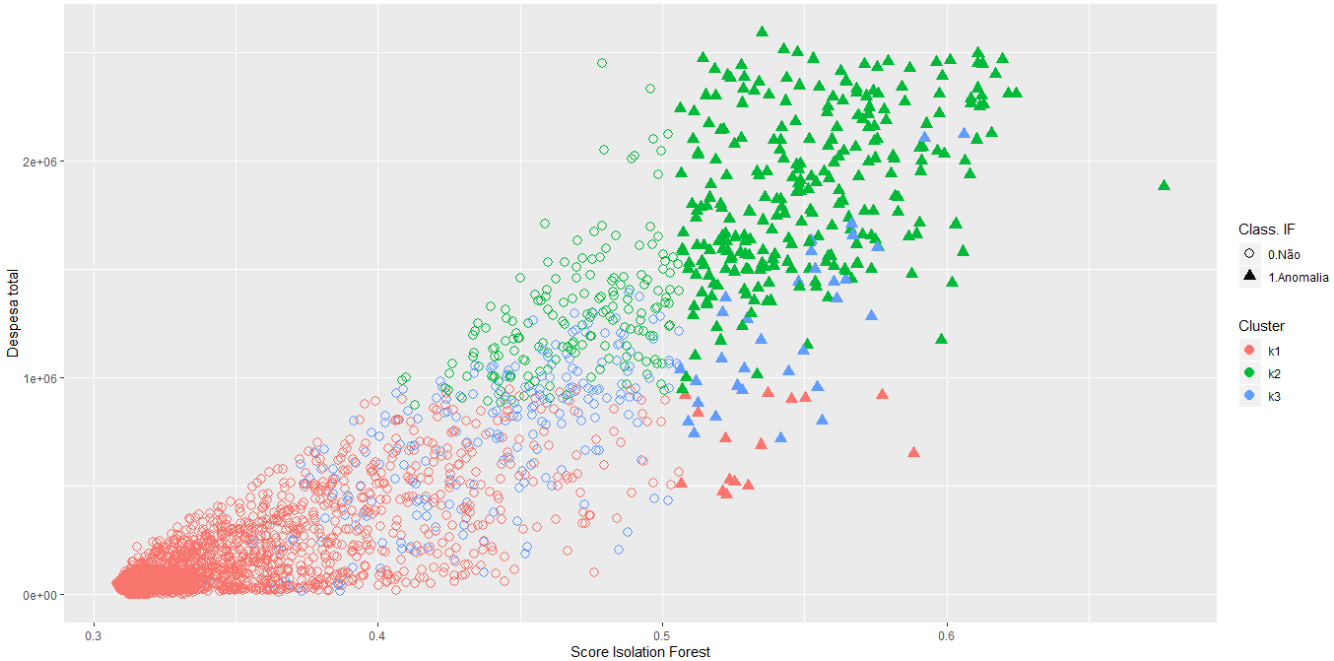

FIGURA 4.18 – Score e Classificação do modelo *Isolation Forest* x Despesas Total

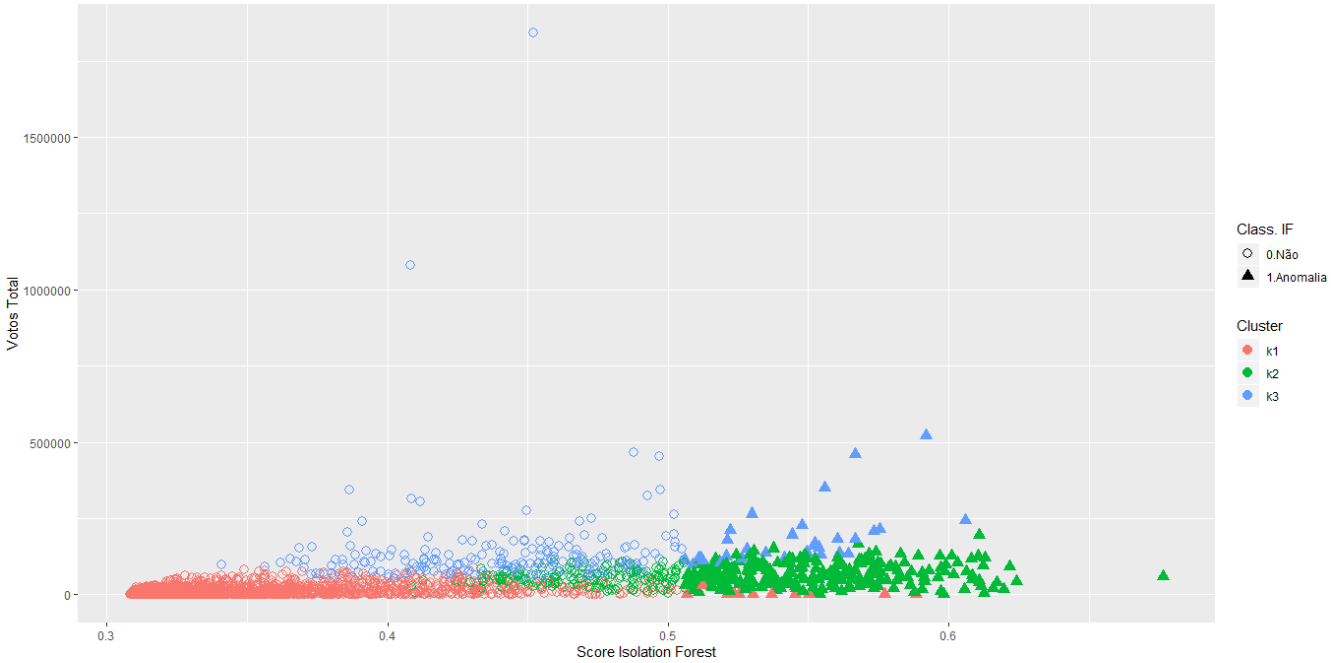

FIGURA 4.19 – Score e Classificação do modelo *Isolation Forest* x Votos Total

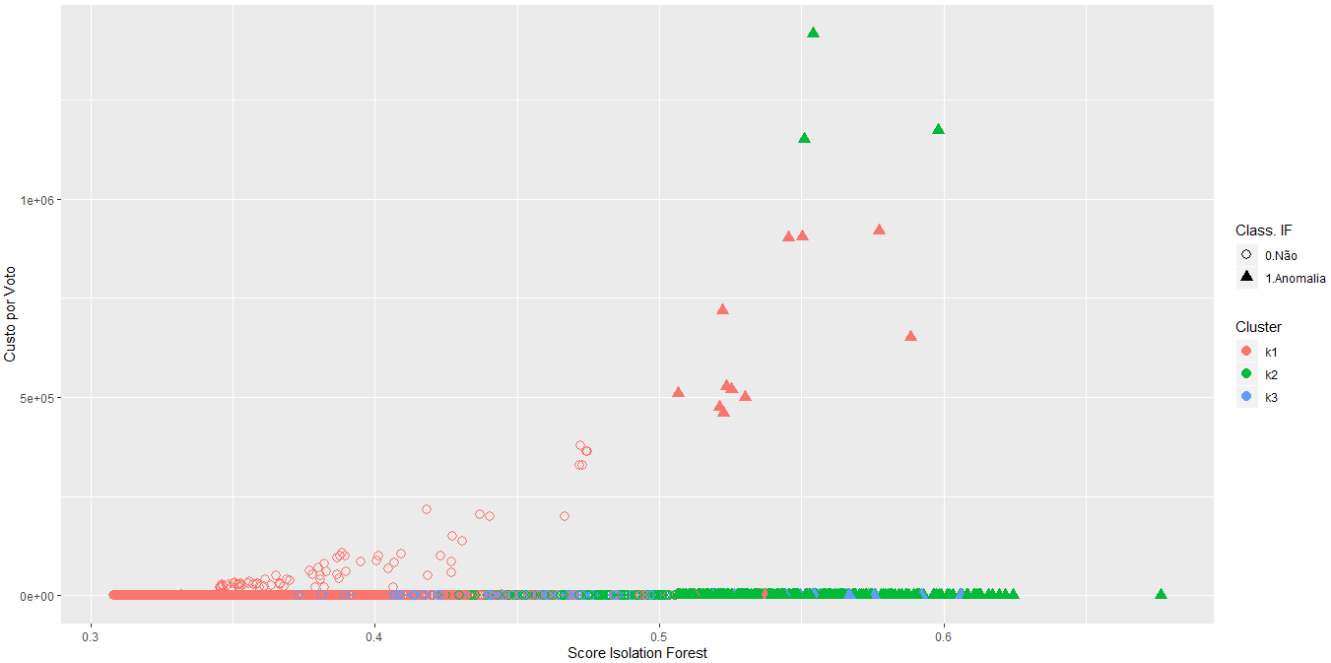

FIGURA 4.20 – Score e Classificação do modelo *Isolation Forest* x Custo por Voto

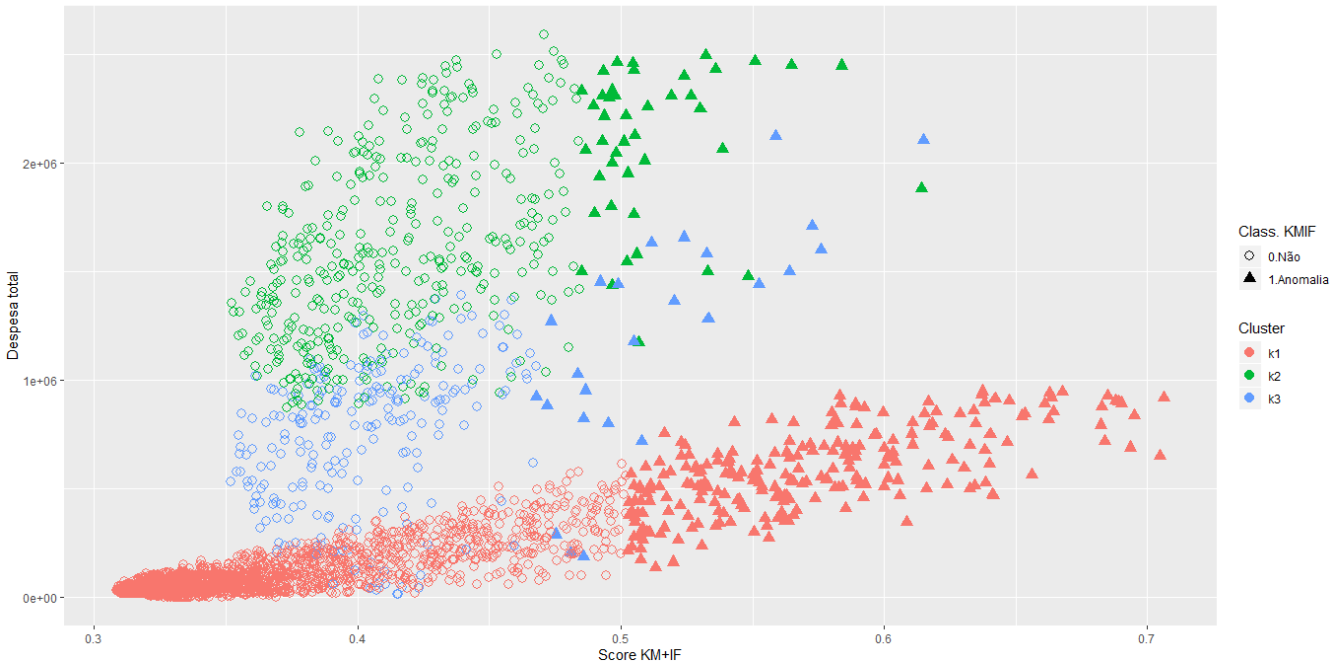

FIGURA 4.21 – Score e Classificação do modelo *KM+IF* x Despesas Total

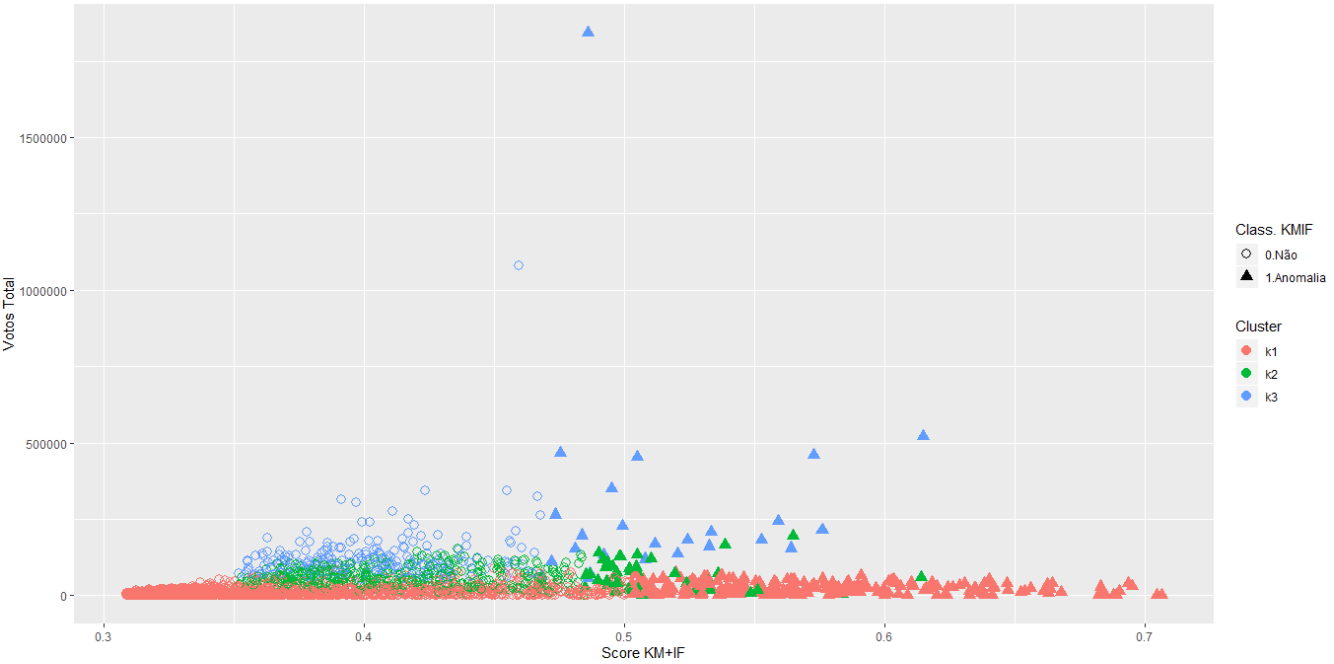

FIGURA 4.22 – Score e Classificação do modelo *KM+IF* x Votos Total

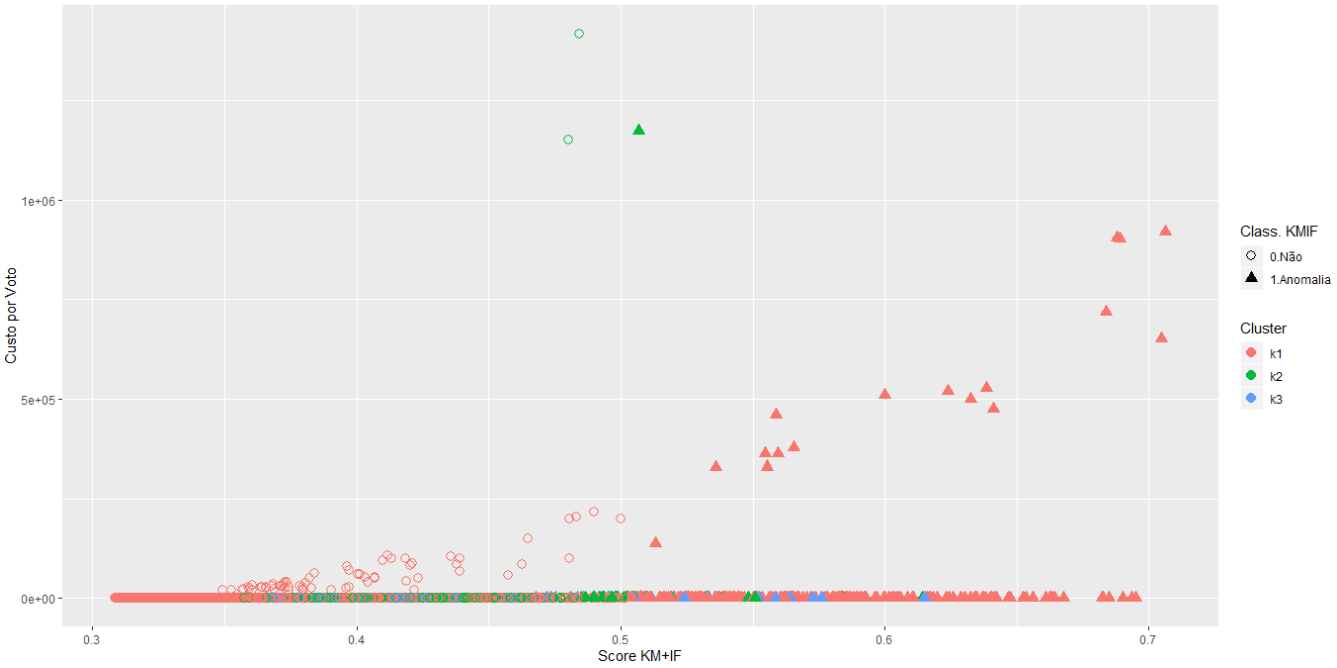

FIGURA 4.23 – Score e Classificação do modelo *KM+IF* x Custo por Voto

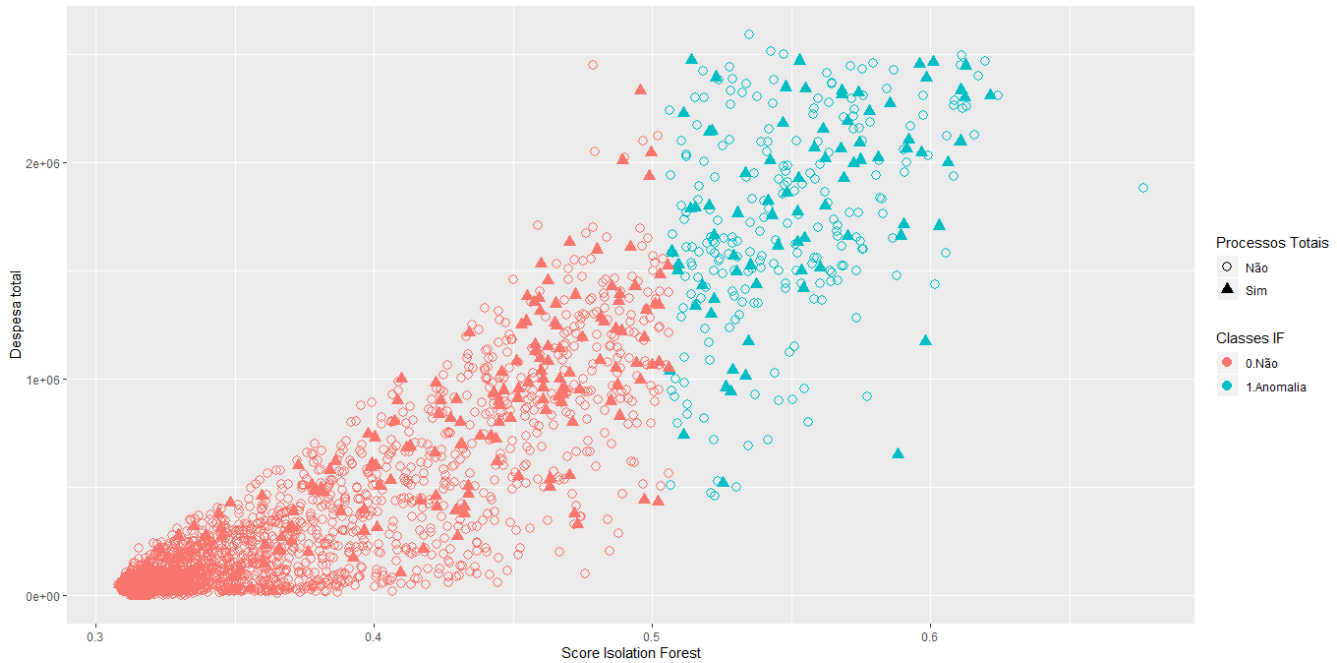

FIGURA 4.24 – Comparação de Processos Totais e Classificação do modelo *IF* x Despesas Total

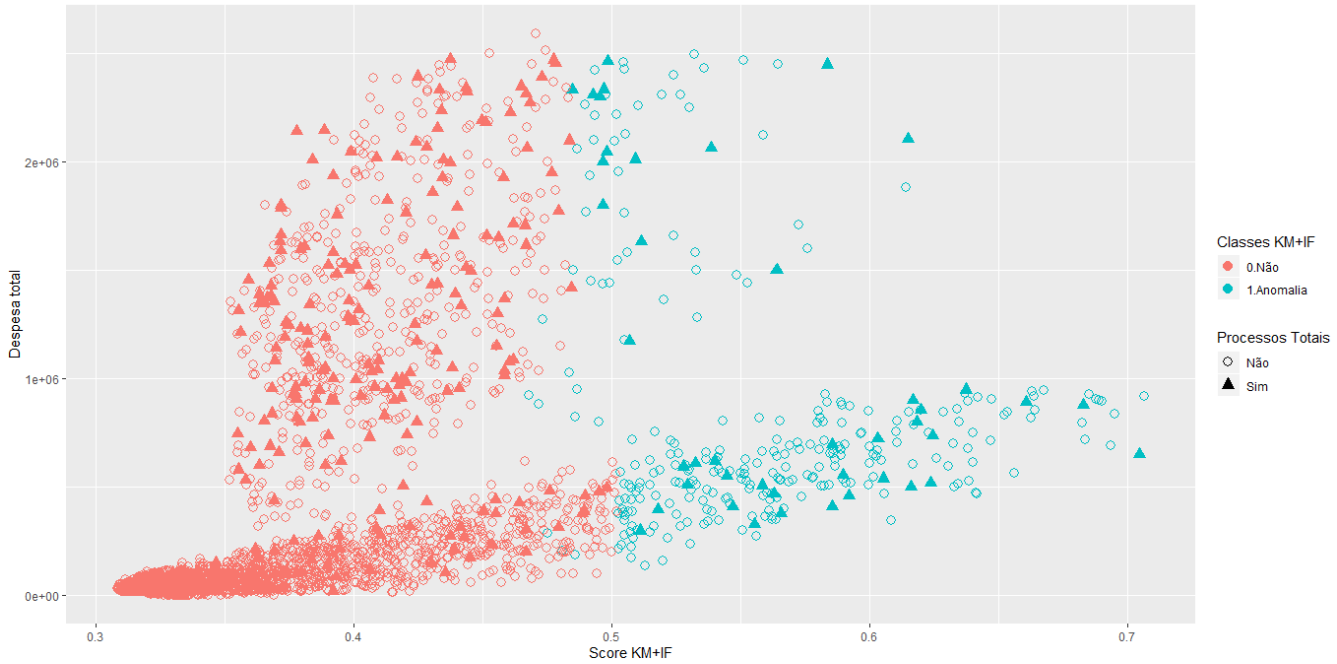

FIGURA 4.25 – Comparação de Processos Totais e Classificação do modelo *KM+IF* x Despesas Total

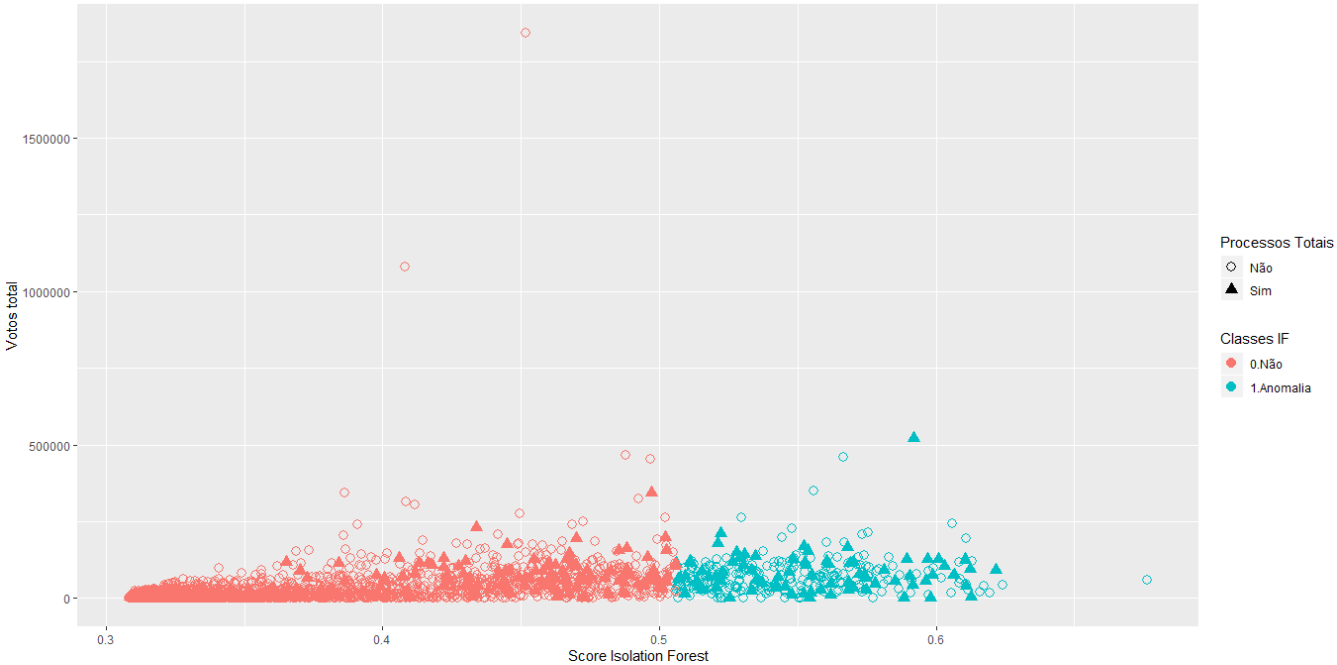

FIGURA 4.26 – Comparação de Processos Totais e Classificação do modelo *IF* x Votos Total

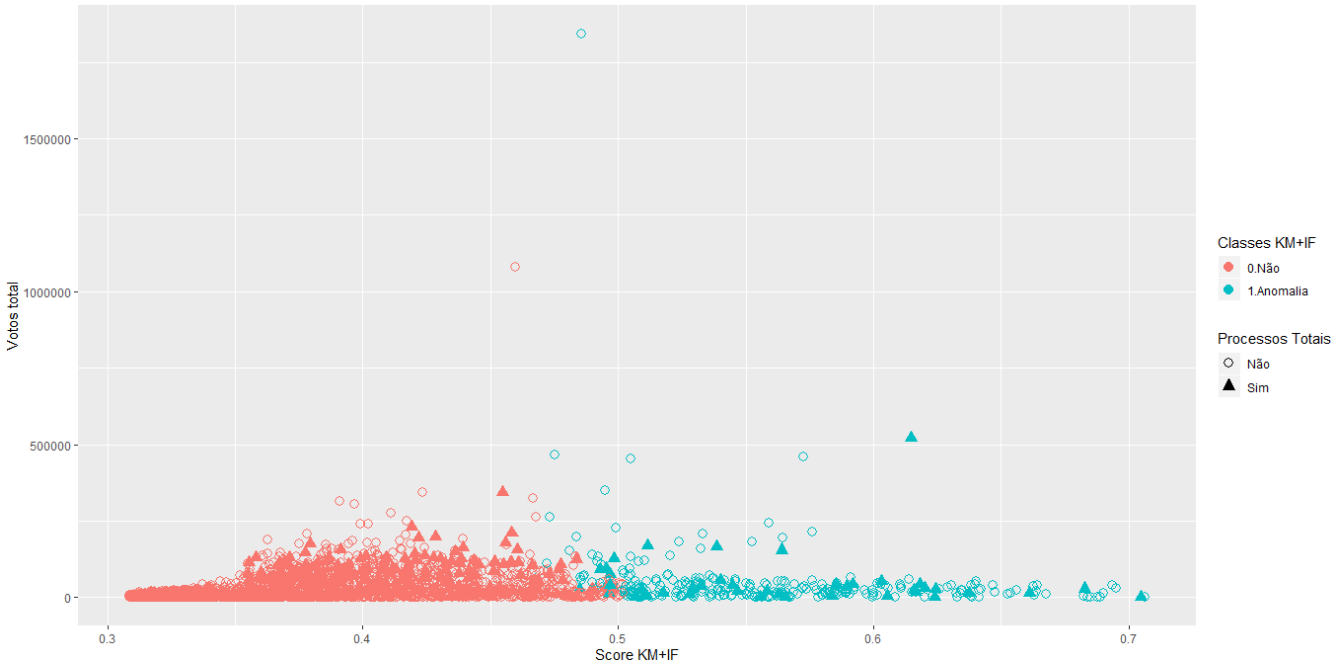

FIGURA 4.27 – Comparação de Processos Totais e Classificação do modelo *KM+IF* x Votos Total

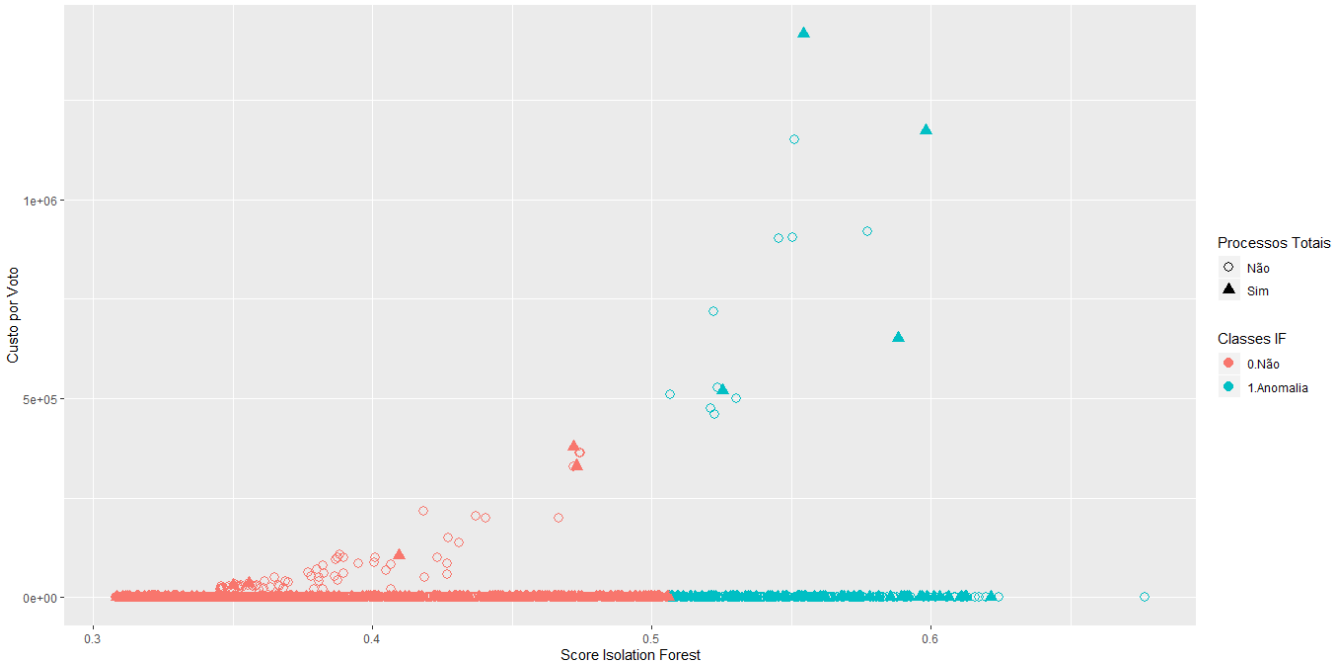

FIGURA 4.28 – Comparação de Processos Totais e Classificação do modelo *IF* x Custo por Voto

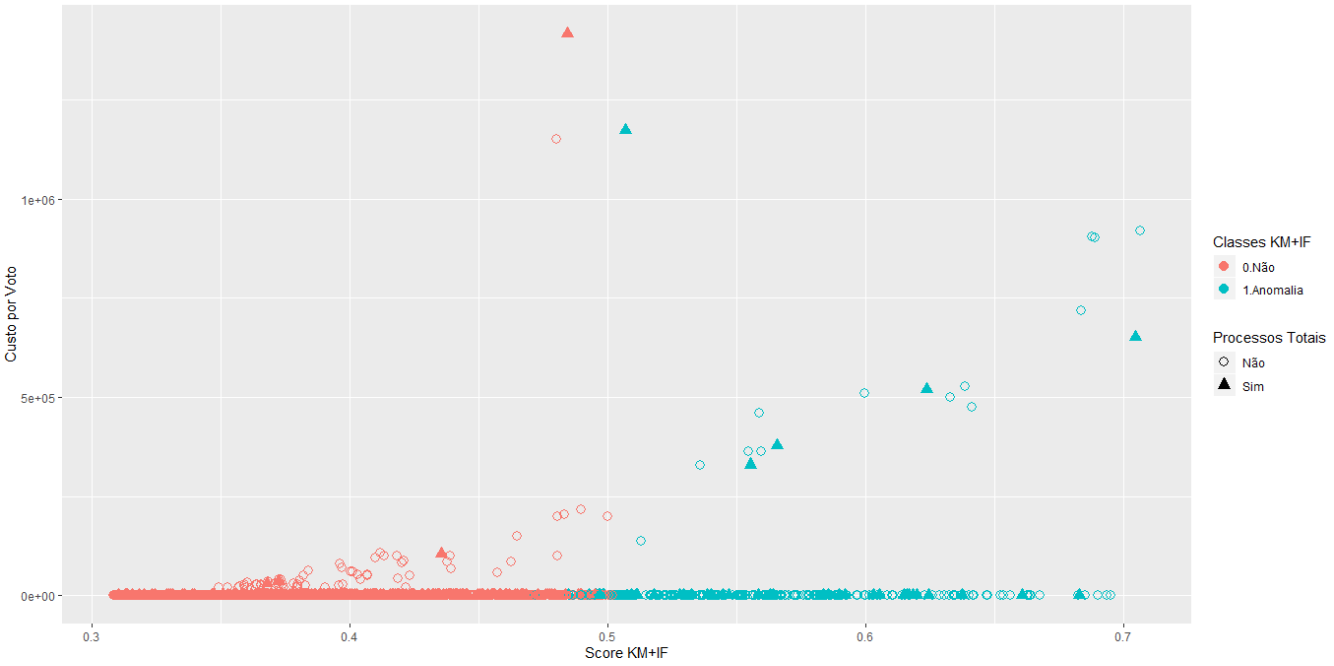

FIGURA 4.29 – Comparação de Processos Totais e Classificação do modelo *KM+IF* x Custo por Voto

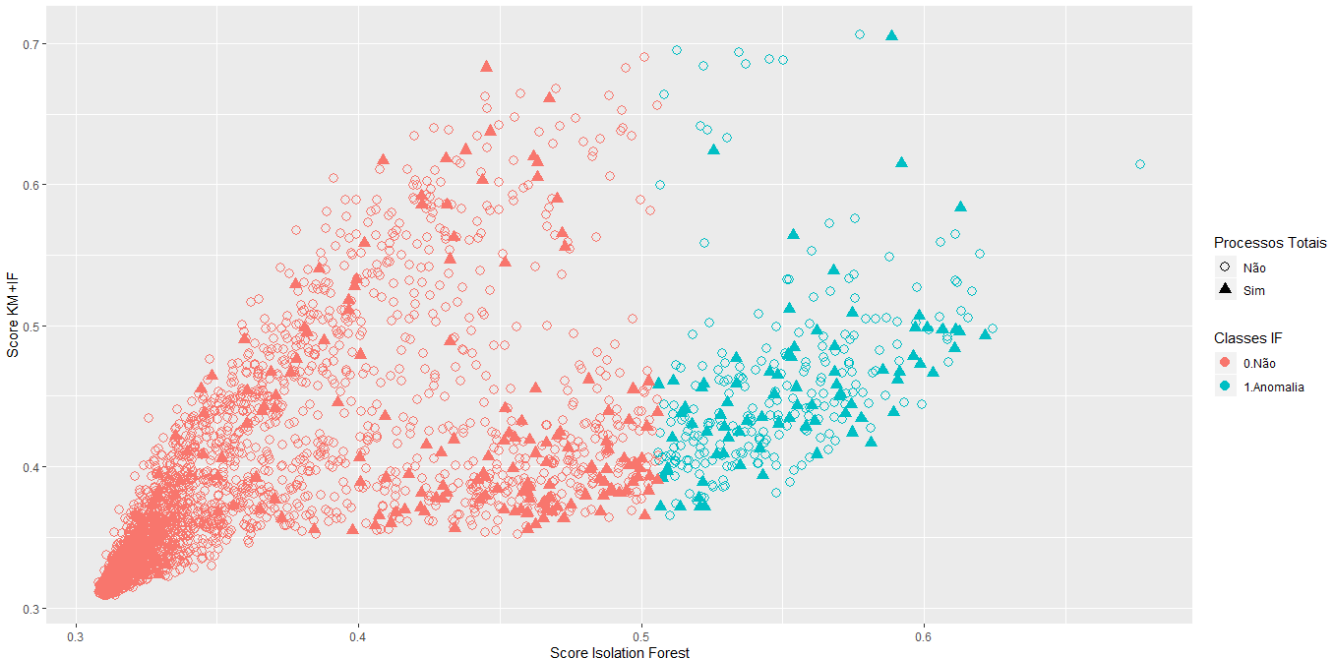

FIGURA 4.30 – Comparação entre os modelos *IF* e *KM+IF* x Classificação IF

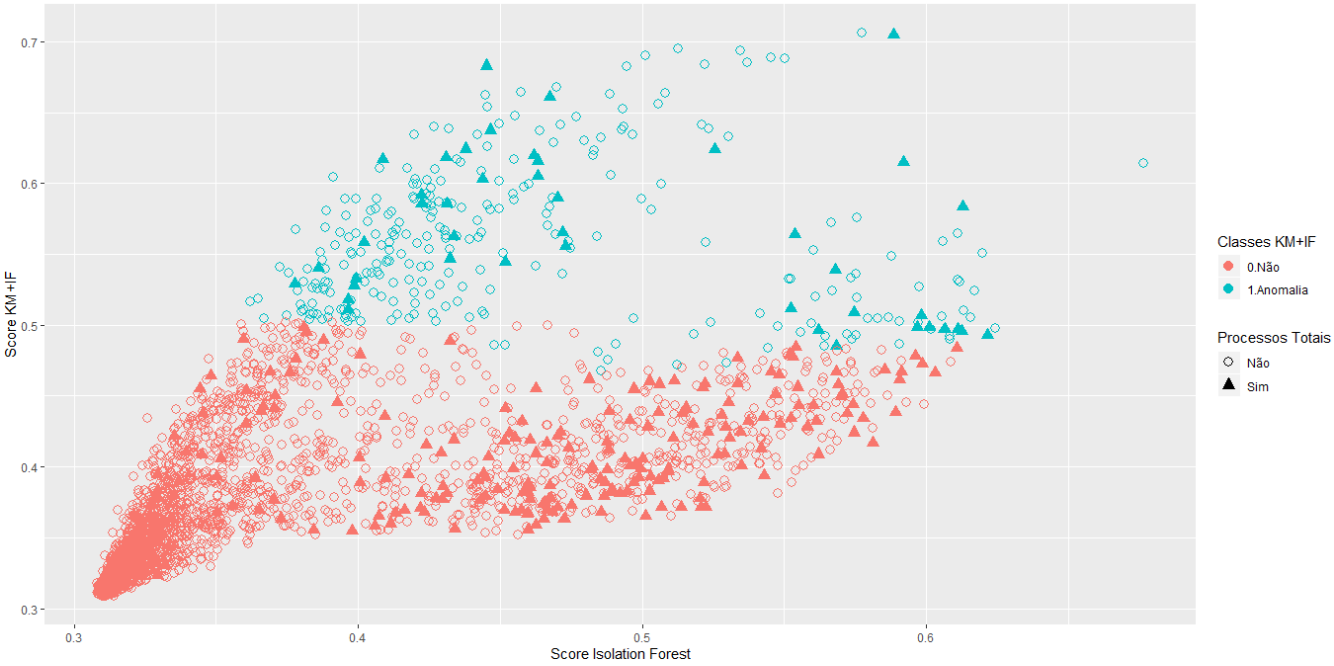

FIGURA 4.31 – Comparação entre os modelos *IF* e *KM+IF* x Classificação KMIF

5 Conclusão e Perspectivas Futuras

Nesse trabalho foi realizado um estudo sobre a detecção de anomalias, bem como um levantamento na literatura das metodologias utilizadas para sua captura. Além disso, foram discutidas as características gerais do fenômeno anomalia, sua tipologia, importância na sua localização, desafios na detecção e aplicações em diversas áreas da natureza. Por fim, foi dada ênfase nos métodos estudados no projeto, mais especificamente, nos algoritmos não supervisionados *K-Means* e *Isolation Forest*.

Utilizando os dois métodos *K-Means* e *Isolation Forest*, foi sugerida uma proposta de combiná-los para se obter resultados mais assertivos. Essa abordagem híbrida foi discutida recentemente por (Kurnianingsih *et al.*, 2018) e (GAO *et al.*, 2019) e, segundo os autores, têm algumas vantagens, como eliminar o custo computacional do cálculo de distâncias, além da possibilidade de ser mais eficiente quando aplicado em grandes conjuntos de dados. Outra vantagem é que a combinação de métodos permite detectar mais facilmente as anomalias locais, visto que o algoritmo *K-means* particiona o conjunto de dados em grupos menores *(clusters)*, dessa forma as anomalias locais antes do agrupamento são transformadas em anomalias globais.

Para avaliar a proposta citada, foram analisados dois estudos de caso. No primeiro, relativo ao conjunto de dados de espécies de plantas Iris (FISHER, 1936), foi utilizado para teste de conceito da aplicação. Foram acrescentados alguns pontos atípicos no conjunto de dados inicial e, a partir disso, foi avaliada a captura dessas observações pelos algoritmos, e que apresentaram resultados satisfatórios. Já a segunda aplicação está relacionada ao tema principal deste projeto - a detecção de anomalias em conjunto de dados de candidaturas eleitorais. Para avaliar os resultados e, sabendo inicialmente que não haveria uma variável resposta rotulada, foi proposta a construção de uma variável de interesse baseada em processos de crimes eleitorais que os candidatos possuem (processos históricos e ativos). É possível observar pelos resultados que a combinação dos métodos *KM+IF* trouxe vantagens em situações onde o *cluster* apresenta menores taxas de anomalias, obtendo melhores performances de captura. Entretanto, para os *clusters* com altas concentrações de anomalias, o modelo combinado *KM+IF*, apresentou resultados inferiores, visto que o próprio *K-Means* capturou e separou de forma satisfatória um grande volume de anomalias nessas partições de *clusters*. É importante ressaltar que à medida em que o conjunto

de dados têm *clusters* bem definidos, o método de agrupamento tende a funcionar bem; já para os casos onde isso não ocorre, provavelmente os resultados serão inferiores.

Uma contribuição deste trabalho relaciona-se à escolha da abordagem não supervisionada para aplicação em um conjunto de dados de candidaturas eleitorais. Essa proposta tende a ser mais complexa por apresentar desafios na avaliação dos resultados. Outro fato relevante observado nesse estudo é que o uso das técnicas combinadas *K-Means + Isolation Forest* podem apresentar resultados inferiores às técnicas aplicadas separadamente *K-Means* e *Isolation Forest*. Isso dependerá dos dados e do contexto da aplicação e essa conclusão não foi identificada nas pesquisas anteriores de (Kurnianingsih *et al.*, 2018) e (GAO *et al.*, 2019).

Sugere-se, como trabalhos futuros, a aplicação da metodologia *KM+IF* em outros conjuntos de dados sem variável de interesse ou informações rotuladas. As conclusões que (Kurnianingsih *et al.*, 2018) e (GAO *et al.*, 2019) citam são baseadas em situações onde possuem a variável de interesse classificada para treinamento dos algoritmos. Como foi descrito, nessa pesquisa não havia uma variável de treinamento para os algoritmos, e, desta maneira, ressalta-se que uma possível justificativa dos resultados não apresentarem altos índices de acerto seja simplesmente porque talvez a variável de interesse construída (processos criminais dos candidatos) não possa ser explicada pelos dados de campanha. Isso por sinal pode indicar que este trabalho tenha capturado uma irregularidade ainda não avaliada pela justiça. Ademais, outra abordagem interessante é a aplicação do mesmo método para os demais cargos de candidatos das eleições de 2018, como deputados estaduais e governadores. Através das classificações resultantes, ficaria a sugestão de uma análise de casos por parte de órgãos reguladores, com intuito de se avaliar se as candidaturas de fato apresentaram alguma anomalia como indicaram os modelos.

Referências

ABE, N.; ZADROZNY, B.; LANGFORD, J. Outlier detection by active learning. In: . [S.l.: s.n.], 2006. v. 2006, p. 504–509.

AGYEMANG, M.; BARKER, K.; ALHAJJ, R. A comprehensive survey of numeric and symbolic outlier mining techniques. **Intell. Data Anal.**, IOS Press, Amsterdam, The Netherlands, The Netherlands, v. 10, n. 6, p. 521–538, dez. 2006. ISSN 1088-467X. Disponível em: <http://dl.acm.org/citation.cfm?id=1609942.1609946>.

AHMED, S.; LEE, Y.; HYUN, S.; KOO, I. Unsupervised machine learningâbased detection of covert data integrity assault in smart grid networks utilizing isolation forest. **IEEE Transactions on Information Forensics and Security**, PP, p. 1–1, 03 2019.

ALESKEROV, E.; FREISLEBEN, B.; RAO, B. Cardwatch: A neural network based database mining system for credit card fraud detection. In: . [S.l.: s.n.], 1997. p. 220 – 226. ISBN 0-7803-4133-3.

ANDERSON, E. The Irises of the Gaspe Peninsula. **Bulletin of the American Iris Society**, v. 59, p. 2–5, 1935.

BAKAR, Z. abu; MOHEMAD, R.; AHMAD, A.; DERIS, M. M. A comparative study for outlier detection techniques in data mining. In: . [S.l.: s.n.], 2006. p. 1 – 6.

BAKER, L. D.; HOFMANN, T.; MCCALLUM, A. K.; YANG, Y. **A Hierarchical Probabilistic Model for Novelty Detection in Text**. 1999.

BANO, S.; KHAN, N. A survey of data clustering methods. **International Journal of Advanced Science and Technology**, v. 113, 04 2018.

BARBARá, D.; COUTO, J.; JAJODIA, S.; WU, N. Adam: A testbed for exploring the use of data mining in intrusion detection. **SIGMOD Rec.**, ACM, New York, NY, USA, v. 30, n. 4, p. 15–24, dez. 2001. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/604264.604268>.

BASU, S.; MECKESHEIMER, M. Automatic outlier detection for time series: An application to sensor data. **Knowl. Inf. Syst.**, v. 11, p. 137–154, 02 2007.

BECKMAN, R. J.; COOK, R. D. Outlier . s. **Technometrics**, Taylor Francis, v. 25, n. 2, p. 119–149, 1983. Disponível em: <https://doi.org/10.1080/00401706.1983.10487840>.

BOTTOU, L.; BENGIO, Y. Convergence properties of the k-means algorithms. In: . [S.l.: s.n.], 1994. p. 585–592.

Bronstein, A.; Das, J.; Duro, M.; Friedrich, R.; Kleyner, G.; Mueller, M.; Singhal, S.; Cohen, I. Self-aware services: using bayesian networks for detecting anomalies in internet-based services. In: **2001 IEEE/IFIP International Symposium on Integrated Network Management Proceedings. Integrated Network Management VII. Integrated Management Strategies for the New Millennium (Cat. No.01EX470)**. [S.l.: s.n.], 2001. p. 623–638.

BUSCEMA, M.; MAURELLI, G.; MENNINI, F.; GITTO, L.; RUSSO, S.; RUGGERI, M.; CORETTI, S.; CICCHETTI, A. Artificial neural networks and their potentialities in analyzing budget health data: an application for italy of what-if theory. **Quality Quantity**, 03 2016.

BUSCEMA, P. M.; MASSINI, G.; BREDA, M.; LODWICK, W. A.; NEWMAN, F.; ASADI-ZEYDABADI, M. An introduction. In: _____. **Artificial Adaptive Systems Using Auto Contractive Maps: Theory, Applications and Extensions**. Cham: Springer International Publishing, 2018. p. 1–9. ISBN 978-3-319-75049-1. Disponível em: <https://doi.org/10.1007/978-3-319-75049-1_1>.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 41, n. 3, p. 15:1–15:58, jul. 2009. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/1541880.1541882>.

CHAUDHARY, K.; YADAV, J.; MALLICK, B. A review of fraud detection techniques: Credit card. **International Journal of Computer Applications**, v. 45, 01 2012.

CHAWLA, N.; JAPKOWICZ, N.; KOTCZ, A. Editorial: Special issue on learning from imbalanced data sets. **SIGKDD Explorations**, v. 6, p. 1–6, 06 2004.

DASGUPTA, D.; MAJUMDAR, N. Anomaly detection in multidimensional data using negative selection algorithm. In: . [S.l.: s.n.], 2002. v. 2, p. 1039 – 1044. ISBN 0-7803-7282-4.

DASGUPTA, D.; NINO, L. F. A comparison of negative and positive selection algorithms in novel pattern detection. In: . [S.l.: s.n.], 2000. v. 1, p. 125 – 130 vol.1. ISBN 0-7803-6583-6.

DAVY, M.; GODSILL, S. Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation. In: **2002 IEEE International Conference on Acoustics, Speech, and Signal Processing**. [S.l.: s.n.], 2002. v. 2, p. II–1313–II–1316. ISSN 1520-6149.

DIEHL, C. P. Real-time object classification and novelty detection for collaborative video surveillance. In: . [S.l.: s.n.], 2002.

DOMINGUES, R.; FILIPPONE, M.; MICHIARDI, P.; ZOUAOUI, J. A comparative evaluation of outlier detection algorithms: Experiments and analyses. **Pattern Recognition**, v. 74, 09 2017.

Emmott, A.; Das, S.; Dietterich, T.; Fern, A.; Wong, W.-K. A Meta-Analysis of the Anomaly Detection Problem. **arXiv e-prints**, p. arXiv:1503.01158, Mar 2015.

ESKIN, E.; ARNOLD, A.; PRERAU, M.; PORTNOY, L.; STOLFO, S. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In: **Applications of Data Mining in Computer Security**. [S.l.]: Kluwer, 2002.

Fan, S.; Liu, G.; Chen, Z. Anomaly detection methods for bankruptcy prediction. In: **2017 4th International Conference on Systems and Informatics (ICSAI)**. [S.l.: s.n.], 2017. p. 1456–1460.

FAN, W.; MILLER, M.; STOLFO, S. J.; LEE, W.; CHAN, P. K. Using artificial anomalies to detect unknown and known network intrusions. In: **Proceedings of the 2001 IEEE International Conference on Data Mining**. Washington, DC, USA: IEEE Computer Society, 2001. (ICDM '01), p. 123–130. ISBN 0-7695-1119-8. Disponível em: <http://dl.acm.org/citation.cfm?id=645496.658057>.

FAWCETT, T.; PROVOST, F. Activity monitoring: Noticing interesting changes in behavior. **Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 12 1999.

FILIPPOV, A.; IUZBASHEV, A.; KURNEV, A. User authentication via touch pattern recognition based on isolation forest. In: . [S.l.: s.n.], 2018. p. 1485–1489.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v. 7, n. 7, p. 179–188, 1936.

FORGY, E. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. **Biometrics**, v. 21, n. 3, p. 768–769, 1965.

FOX, A. J. Outliers in time series. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 34, n. 3, p. 350–363, 1972. ISSN 00359246. Disponível em: <http://www.jstor.org/stable/2985071>.

FUJIMAKI, R.; YAIRI, T.; MACHIDA, K. An approach to spacecraft anomaly detection problem using kernel feature space. In: **Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining**. New York, NY, USA: ACM, 2005. (KDD '05), p. 401–410. ISBN 1-59593-135-X. Disponível em: <http://doi.acm.org/10.1145/1081870.1081917>.

GAO, R.; ZHANG, T.; SUN, S.; LIU, Z. Research and improvement of isolation forest in detection of local anomaly points. **Journal of Physics: Conference Series**, IOP Publishing, v. 1237, p. 052023, jun 2019. Disponível em: <https://doi.org/10.1088%2F1742-6596%2F1237%2F5%2F052023>.

GARCIA-FONT, V.; GARRIGUES, C.; POUS, H. Difficulties and challenges of anomaly detection in smart cities: A laboratory analysis. **Sensors**, v. 18, p. 3198, 09 2018.

GHOSH, S.; REILLY, D. L. Credit card fraud detection with a neural-network. In: **HICSS**. [S.l.: s.n.], 1994.

GOGOI, P.; BORAH, B.; BHATTACHARYYA, D. K. Anomaly detection analysis of intrusion data using supervised unsupervised approach. **JCIT**, v. 5, p. 95–110, 02 2010.

GOLDBERGER, A. L.; AMARAL, L. A. N.; GLASS, L.; HAUSDORFF, J. M.; IVANOV, P. C.; MARK, R. G.; MIETUS, J. E.; MOODY, G. B.; PENG, C.-K.; STANLEY, H. E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. **Circulation**, v. 101, n. 23, p. e215–e220, 2000. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

HAWKINS, D. **Identification of outliers**. London [u.a.]: Chapman and Hall, 1980. (Monographs on applied probability and statistics). ISBN 041221900X. Disponível em: <http://gso.gbv.de/DB=2.1-/CMD?ACT=SRCHASRT=YOPIKT=1016TRM=ppn+02435757Xsourceid=fbw_bibsonomy>.

HELLER, K. A.; SVORE, K.; KEROMYTIS, A. D.; STOLFO, S. One class support vector machines for detecting anomalous windows registry accesses. 12 2003.

HELMER, G.; WONG, J.; HONAVAR, V.; MILLER, L. Intelligent agents for intrusion detection. In: . [S.l.: s.n.], 1998. p. 121 – 124. ISBN 0-7803-9914-5.

HODGE, V.; AUSTIN, J. A survey of outlier detection methodologies. **Artif. Intell. Rev.**, Kluwer Academic Publishers, Norwell, MA, USA, v. 22, n. 2, p. 85–126, out. 2004. ISSN 0269-2821. Disponível em: <https://doi.org/10.1023/B:AIRE.0000045502.10941.a9>.

JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: A systematic study. **Intell. Data Anal.**, v. 6, p. 429–449, 11 2002.

JOSHI, M. V.; AGARWAL, R.; KUMAR, V. Mining needle in a haystack: Classifying rare classes via two-phase rule induction. In: . [S.l.: s.n.], 2001. v. 30, p. 91–102.

KAREV, D.; MCCUBBIN, C.; VAULIN, R. Cyber threat hunting through the use of an isolation forest. In: **Proceedings of the 18th International Conference on Computer Systems and Technologies**. New York, NY, USA: ACM, 2017. (CompSysTech'17), p. 163–170. ISBN 978-1-4503-5234-5. Disponível em: <http://doi.acm.org/10.1145/3134302.3134319>.

KOU, Y.; LU, C.-T.; CHEN, D. Spatial weighted outlier detection. In: . [S.l.: s.n.], 2006. v. 2006.

KULKARNI, A.; MANI, P.; DOMENICONI, C. Network-based anomaly detection for insider trading. **CoRR**, abs/1702.05809, 2017. Disponível em: <http://arxiv.org/abs/1702.05809>.

Kurnianingsih; Nugroho, L. E.; Widyawan; Lazuardi, L.; Prabuwono, A. S. Detection of anomalous vital sign of elderly using hybrid k-means clustering and isolation forest. In: **TENCON 2018 - 2018 IEEE Region 10 Conference**. [S.l.: s.n.], 2018. p. 0913–0918.

LAZAREVIC, A.; ERTOZ, L.; KUMAR, V.; OZGUR, A.; SRIVASTAVA, J. A comparative study of anomaly detection schemes in network intrusion detection. In: _____. **Proceedings of the 2003 SIAM International Conference on Data**

Mining. [s.n.], 2003. p. 25–36. Disponível em: <https://epubs.siam.org/doi/abs/10.1137/1.9781611972733.3>.

LAZZARI, E. A. Por que os brasileiros não confiam em partidos políticos? 2016.

LEE, W.; STOLFO, S.; CHAN, P. Learning patterns from unix process execution traces for intrusion detection. 05 1997.

LIU, F. T.; TING, K.; ZHOU, Z.-H. Isolation-based anomaly detection. **ACM Transactions on Knowledge Discovery From Data - TKDD**, v. 6, p. 1–39, 03 2012.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. In: **Proceedings of the 2008 Eighth IEEE International Conference on Data Mining**. Washington, DC, USA: IEEE Computer Society, 2008. (ICDM '08), p. 413–422. ISBN 978-0-7695-3502-9. Disponível em: <http://dx.doi.org/10.1109/ICDM.2008.17>.

LLOYD, S. P. **Least squares quantization in PCM**. [S.l.], 1957.

M.A., F. E. Xli. on discordant observations. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, Taylor Francis, v. 23, n. 143, p. 364–375, 1887. Disponível em: <https://doi.org/10.1080/14786448708628471>.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **In 5-th Berkeley Symposium on Mathematical Statistics and Probability**. [S.l.: s.n.], 1967. p. 281–297.

MARKOU, M.; SINGH, S. Novelty detection: A review - part 1: Statistical approaches. **Signal Processing**, v. 83, p. 2003, 2003.

MARKOU, M.; SINGH, S. Novelty detection: A review - part 2: Neural network based approaches. **Signal Processing**, v. 83, p. 2499–2521, 2003.

MENNINI, F. S.; GITTO, L.; RUSSO, S.; CICCHETTI, A.; RUGGERI, M.; CORETTI, S.; MAURELLI, G.; BUSCEMA, P. M. Does regional belonging explain the similarities in the expenditure determinants of italian healthcare deliveries?: An approach based on artificial neural networks. **Economic Analysis and Policy**, v. 55, p. 47 – 56, 2017. ISSN 0313-5926. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0313592616300704>.

MOURAO-MIRANDA, J.; HARDOON, D.; MARQUAND, A.; WILLIAMS, S.; SHAWE-TAYLOR, J.; BRAMMER, M. Patient classification as an outlier detection problem: An application of the one-class support vector machine. **NeuroImage**, v. 58, p. 793–804, 06 2011.

MUNIYANDI, A.; RAMACHANDRAN, R.; RAJARAM, R. Network anomaly detection by cascading k-means clustering and c4.5 decision tree algorithm. **Procedia Engineering**, v. 30, p. 174–182, 12 2012.

NEUMANN Åukasz; NOWAK, R. M.; OKUNIEWSKI, R.; WAWRZYÅSKI, P. Machine learning-based predictions of customersâ decisions in car insurance. **Applied Artificial Intelligence**, Taylor Francis, v. 33, n. 9, p. 817–828, 2019. Disponível em: <https://doi.org/10.1080/08839514.2019.1630151>.

NGAI, E.; HU, Y.; WONG, Y.; CHEN, Y.; SUN, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. **Decision Support Systems**, v. 50, p. 559–569, 02 2011.

NOBLE, C. C.; COOK, D. J. Graph-based anomaly detection. In: **Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2003. (KDD '03), p. 631–636. ISBN 1-58113-737-0. Disponível em: <http://doi.acm.org/10.1145/956750.956831>.

NUNES, D. H. F. **Um breve estudo sobre o algoritmo K-means**. 2016. 60 f. Dissertação (Mestrado em Matematica) — Universidade de Coimbra, Cidade de Coimbra, 2016.

PARMAR, J. D.; PATEL, J. T. Anomaly detection in data mining: A review. In: . [S.l.: s.n.], 2017.

PHOHA, V. **Internet Security Dictionary**. [S.l.: s.n.], 2002. ISBN 978-0-387-95261-1.

PHUA, C.; ALAHAKOON, D.; LEE, V. Minority report in fraud detection: Classification of skewed data. **SIGKDD Explorations**, v. 6, p. 50–59, 01 2004.

PINCUS, R. Barnett, v., and lewis t.: Outliers in statistical data. 3rd edition. j. wiley & sons 1994, xvii. 582 pp., £49.95. **Biometrical Journal**, v. 37, n. 2, p. 256–256, 1995. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.4710370219>.

PUGGINI, L.; MCLOONE, S. An enhanced variable selection and isolation forest based methodology for anomaly detection with oes data. **Engineering Applications of Artificial Intelligence**, v. 67, p. 126–135, 01 2018.

RÄTSCH, G.; MIKA, S.; SCHÖLKOPF, B.; MÜLLER, K.-R. Constructing boosting algorithms from svms: an application to one-class classification. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 9, p. 1184–1199, set. 2002.

RIBEIRO, H. V.; ALVES, L. G. A.; MARTINS, A. F.; LENZI, E. K.; PERC, M. The dynamical structure of political corruption networks. **Journal of Complex Networks**, v. 6, n. 6, p. 989–1003, 01 2018. ISSN 2051-1329. Disponível em: <https://dx.doi.org/10.1093/comnet/cny002>.

ROUSSEEUW, P. J.; LEROY, A. M. **Robust Regression and Outlier Detection**. New York, NY, USA: John Wiley & Sons, Inc., 1987. ISBN 0-471-85233-3.

RYMAN-TUBB, N.; KRAUSE, P.; GARN, W. How artificial intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. **Engineering Applications of Artificial Intelligence**, v. 76, p. 130–157, 11 2018.

SAGADEVAN, S.; MALIM, N.; YEE, O. Credit card fraud detection using machine learning as data mining technique. In: . [S.l.: s.n.], 2018.

SALVADOR, S.; CHAN, P.; BRODIE, J. Learning states and rules for time series anomaly detection. In: **FLAIRS Conference**. [S.l.: s.n.], 2004.

SEBYALA, A. A.; OLUKEMI, T.; SACKS, D. L. Active platform security through intrusion detection using nai?ve bayesian network for anomaly detection. In: . [S.l.: s.n.], 2002.

SHARMILA, V.; R, K. K.; R, S.; D, S.; R, H. Credit card fraud detection using anomaly techniques. In: . [S.l.: s.n.], 2019. p. 1–6.

SHEKHAR, S.; LU, C.-T.; ZHANG, P. Detecting graph-based spatial outliers: algorithms and applications (a summary of results. In: . [S.l.: s.n.], 2001. p. 371–376.

SIDDIQUI, M.; STOKES, J.; SEIFERT, C.; ARGYLE, E.; MCCANN, R.; NEIL, J.; CARROLL, J. Detecting cyber attacks using anomaly detection with explanations and expert feedback. In: . [S.l.: s.n.], 2019. p. 2872–2876.

SONG, X.; WU, M.; JERMAINE, C.; RANKA, S. Conditional anomaly detection, knowledge and data engineering. **IEEE Transactions on**, v. 19, p. 631–645, 01 2007.

SPENCE, C.; PARRA, L.; SAJDA, P. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In: **Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA'01)**. Washington, DC, USA: IEEE Computer Society, 2001. (MMBIA '01), p. 3–. ISBN 0-7695-1336-0. Disponível em: <http://dl.acm.org/citation.cfm?id=882464.882797>.

STEFANO, C. D.; SANSONE, C.; VENTO, M. To reject or not to reject: That is the question-an answer in case of neural classifiers. **Trans. Sys. Man Cyber Part C**, IEEE Press, Piscataway, NJ, USA, v. 30, n. 1, p. 84–94, fev. 2000. ISSN 1094-6977. Disponível em: <http://dx.doi.org/10.1109/5326.827457>.

STEINWART, I.; HUSH, D.; SCOVEL, C. A classification framework for anomaly detection. **Journal of Machine Learning Research**, v. 6, p. 211–232, 02 2005.

SURYANARAYANA, S.; GN, B.; RAO, G. Machine learning approaches for credit card fraud detection. **International Journal of Engineering and Technology(UAE)**, v. 7, 05 2018.

SUSTO, G. A.; TERZI, M.; BEGHI, A. Anomaly detection approaches for semiconductor manufacturing. **Procedia Manufacturing**, v. 11, p. 2018–2024, 12 2017.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining, (First Edition)**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367.

TENG, H.; CHEN, K.; LU, S. Adaptive real-time anomaly detection using inductively generated sequential patterns. In: **Proceedings. 1990 IEEE Computer Society Symposium on Research in Security and Privacy**. [s.n.], 1990. p. 278–284. Exported from https://app.dimensions.ai on 2019/03/17. Disponível em: <https://app.dimensions.ai/details/publication/pub.1086373304>.

TENG, H. S.; CHEN, K.; LU, S. C. Adaptive real-time anomaly detection using inductively generated sequential patterns. In: **Proceedings of the 1990 IEEE**

Symposium on Research in Computer Security and Privacy. [S.l.: s.n.], 1990. p. 278–284.

THEILER, J.; CAI, M. Resampling approach for anomaly detection in multispectral images. **Proc SPIE**, v. 5093, 04 2003.

TI. **Global corruption report 2004: political corruption. Transparency Internacional, 2004**. 2004. Disponível em: <https://www.transparency.org/research/gcr/gcr_political_corruption/0>. Acesso em: 20 jul. 2019.

TI. **Global corruption report 2006: corruption and health. Transparency Internacional, 2006**. 2006. Disponível em: <https://www.transparency.org/research/gcr/gcr_health/0>. Acesso em: 20 jul. 2019.

TI. **Global corruption report 2011: climate change. Transparency Internacional, 2011**. 2011. Disponível em: <https://www.transparency.org/research/gcr/gcr_climate_change/0>. Acesso em: 20 jul. 2019.

TI. **Corruption Perceptions Index 2018. Transparency Internacional, 2018**. 2018. Disponível em: <https://www.transparency.org/cpi2018/results>. Acesso em: 20 jul. 2019.

VALDES, A.; SKINNER, K. Adaptive, model-based monitoring for cyber attack detection. In: **Proceedings of the Third International Workshop on Recent Advances in Intrusion Detection**. London, UK, UK: Springer-Verlag, 2000. (RAID '00), p. 80–92. ISBN 3-540-41085-6. Disponível em: <http://dl.acm.org/citation.cfm?id=645838.670720>.

VILALTA, R.; MA, S. Predicting rare events in temporal domains. In: **Proceedings of the 2002 IEEE International Conference on Data Mining**. Washington, DC, USA: IEEE Computer Society, 2002. (ICDM '02), p. 474–. ISBN 0-7695-1754-4. Disponível em: <http://dl.acm.org/citation.cfm?id=844380.844743>.

WARRENDER, C.; FORREST, S.; PEARLMUTTER, B. Detecting intrusions using system calls: Alternative data models. In: **Proceedings of the 1999 IEEE Symposium on Security and Privacy**. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 1999. v. 1999-January, p. 133–145.

WEIGEND, A. S.; MANGEAS, M.; SRIVASTAVA, A. N. Nonlinear gated experts for time series: discovering regimes and avoiding overfitting. **International journal of neural systems**, v. 6 4, p. 373–99, 1995.

WEISS, G.; HIRSH, H. Learning to predict rare events in event sequences. 01 1998.

Weng, Y.; Liu, L. A collective anomaly detection approach for multidimensional streams in mobile service security. **IEEE Access**, v. 7, p. 49157–49168, 2019.

WHITROW, C.; HAND, D.; JUSZCZAK, P.; WESTON, D. J.; ADAMS, N. Transaction aggregation as a strategy for credit card fraud detection. **Data Mining and Knowledge Discovery**, v. 18, p. 30–55, 02 2009.

WU, T.; ZHANG, Y.-J.; TANG, X. Isolation forest based method for low-quality synchrophasor measurements and early events detection. In: . [S.l.: s.n.], 2018. p. 1–7.

XU, D.; TIAN, Y. A comprehensive survey of clustering algorithms. **Annals of Data Science**, v. 2, n. 2, p. 165–193, Jun 2015. ISSN 2198-5812. Disponível em: <https://doi.org/10.1007/s40745-015-0040-1>.

YE, N.; XU, M.; EMRAN, S. M. Probabilistic networks with undirected links for anomaly detection. 07 2000.

ZAREAPOOR, M.; K.R., S.; ALAM, A. Analysis on credit card fraud detection techniques: Based on certain design criteria. **International Journal of Computer Applications**, v. 52, p. 35–42, 08 2012.

ZHU, X.; TAO, H.; WU, Z.; CAO, J.; KALISH, K.; KAYNE, J. **Fraud Prevention in Online Digital Advertising**. [S.l.: s.n.], 2017.

ZIMEK, A.; SCHUBERT, E.; KRIEGEL, H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. **Statistical Analysis and Data Mining**, v. 5, p. 363–387, 10 2012.

Apêndice A - Pseudocódigo dos Algoritmos

Algorithm 1 K-means

procedure K-MEANS($\boldsymbol{D}, k, \epsilon$) ▷ $\boldsymbol{D}$ conjunto dos pontos, k número de clusters, ϵ erro

    $iteracao \leftarrow 0$

    Inicializar os $\boldsymbol{\mu}_i$, $i = \{1, \ldots, k\}$ (Com pontos aleatórios de $\boldsymbol{D}$)

    repeat

        $C_i \leftarrow \emptyset, \quad \forall j = 1 \ldots k$

        for $\boldsymbol{x}_j \in \boldsymbol{D}$ **do**

            $j^* \leftarrow \underset{i}{\operatorname{argmin}} \left\{ \|\boldsymbol{x}_j - \boldsymbol{\mu}_i^{iteracao}\|^2 \right\}$

            $C_{j^*} = C_{j^*} \cup \{\boldsymbol{x}_j\}$

        end for

        for $i = 1$ até k **do**

            $\boldsymbol{\mu}_i^{iteracao} = \frac{1}{|C_i|} \sum_{\boldsymbol{x}_j \in C_i} \boldsymbol{x}_j$

        end for

    until $\sum_{i=1}^{k} \|\boldsymbol{\mu}_i^{iteracao} - \boldsymbol{\mu}_i^{iteracao-1}\|^2 < \epsilon$

end procedure

FIGURA A.1 – Pseudo algoritmo do método *K-Means*. Fonte: [Nunes 2016]

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO Nº	4. Nº DE PÁGINAS
DM	06 de janeiro de 2020	DCTA/ITA/DM-101/2019	95

5. TÍTULO E SUBTÍTULO:

Aplicação de algoritmos não supervisionados em dados eleitorais

6. AUTOR(ES):

Mateus Vendramini Polizeli

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica – ITA / Universidade Federal de São Paulo – UNIFESP

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Detecção de anomalias; *Outliers*; Dados eleitorais; *Isolation Forest*; *K-Means*

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Processamento de dados; Anomalias; Eleição; Detecção; Computação; Pesquisa operacional

10. APRESENTAÇÃO: (**X**) **Nacional** ( ) **Internacional**

ITA/UNIFESP, São José dos Campos. Curso de Mestrado. Programa de Pós-Graduação em Engenharia de Produção. Área de Metodos de Otimização e Ciência de Dados. Orientador: Prof. Dr. Luís Felipe Cesar da Rocha Bueno. Defesa em 10/12/2019. Publicada em 2019.

11. RESUMO:

Diante da busca incessante da sociedade por clareza nos gastos públicos, eficiência na gestão e transparência com uso da máquina pública, torna-se relevante a estruturação de trabalhos que possibilitem uma apuração aprofundada para acompanhamento eficiente dessas ações. A partir de um estudo inicial na literatura, verificou-se a existência de uma série de controles e divulgação de prestação de contas de setores e órgãos públicos. Contudo, apesar de iniciativas como essas, ainda há poucos trabalhos considerando uma investigação mais aprofundada para capturar possíveis irregularidades do meio político. Dessa forma, o objetivo deste projeto é estudar alguns mecanismos de detecção de anomalias associados ao conjunto de dados das candidaturas eleitorais de 2018. As metodologias propostas são baseadas nos algoritmos não supervisionados *K-Means* e *Isolation Forest* como tentativa de criar uma ferramenta de apoio à tomada de decisão para os reguladores, visando direcionar os recursos humanos para investigação. É sugerida também uma combinação desses algoritmos, denominado aqui como *KM+IF*, com intuito de melhorar a acurácia e diminuir as taxas de erro associadas aos modelos. Os resultados observados neste projeto indicam que a proposta *KM+IF* mostra bom desempenho para situações onde estão disponíveis as variáveis de interesse. Entretanto, pode apresentar resultados insatisfatórios quando tais não estão disponíveis. No estudo de caso realizado para o conjunto de candidaturas eleitorais, o resultado geral do algoritmo *KM+IF* foi inferior ao resultado individual das técnicas *K-Means* e *Isolation Forest*.

12. GRAU DE SIGILO:

(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **SECRETO**