Identificação de Possíveis Candidaturas de Fachada Utilizando Técnicas de Detecção de Anomalias

Thiago Guareschi[1], **Edimar Manica**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul - *Campus* Ibirubá
Rua Nelsi Ribas Fritsch, 1111 – CEP: 98200-000 – Ibirubá – RS – Brasil

Abstract. *Outlier Detection is a data mining technique, in which your objective is to identify data with attributes that stand out concerning others in an irregular way. This work pursuit to identify possible fake applications through anomaly detection tools applied to data about the candidates provided by the Tribunal Superior Eleitoral. Unsupervised algorithms were used to identify potential fake applications automatically. Then, a supervised algorithm was applied to find useful patterns that describe fake applications. With this, a total of 572 anomalies were identified in the 2018 and 2022 elections, which could represent fake applications.*

Resumo. *Detecção de Anomalias é uma das técnicas presentes na mineração de dados, que tem por objetivo identificar dados com atributos que se destacam em relação aos demais de maneira irregular. Este trabalho procura identificar possíveis candidaturas de fachada através de técnicas de detecção de anomalias, aplicadas aos dados sobre os candidatos disponibilizados pelo Tribunal Superior Eleitoral. Foram utilizados algoritmos não supervisionados para identificar potenciais candidaturas de fachada de forma automática. Após, foi aplicado um algoritmo supervisionado para encontrar padrões úteis que descrevem candidaturas de fachada. Com isso, foram identificadas um total de 572 anomalias nas eleições de 2018 e 2022, que podem representar candidaturas de fachada.*

1. Introdução

O processo democrático das eleições brasileiras conta com a utilização de dinheiro público para a realização de campanhas políticas. Esse financiamento, denominado Fundo Especial de Financiamento de Campanha (FEFC), é previsto na Lei nº 13.487/2017 (TSE, 2017), e é popularmente conhecido como fundo eleitoral. O valor do fundo para as eleições do ano de 2022 foi de 4,9 bilhões de reais (MOLITERNO; RODRIGUES, 2022), e seus recursos foram distribuídos aos partidos para a realização de suas campanhas políticas de acordo com as regras estabelecidas pela lei.

Para ter acesso ao fundo eleitoral, os partidos devem atender a um conjunto de requisitos. Um desses requisitos é a Lei nº 9.504/1997 artigo 10º, parágrafo 3º, que estabelece que cada partido ou coligação deve preencher o mínimo de 30% e o máximo de 70% para candidaturas de cada sexo (TSE, 1997). Além disso, de acordo com a Emenda Constitucional nº 117/2022, no mínimo 30% dos recursos do fundo eleitoral devem ser destinados às candidaturas femininas (BRASIL, 2022).

Para desviar dinheiro do fundo eleitoral ou preencher cotas estabelecidas por lei, alguns partidos praticam candidaturas de fachada. Uma candidatura de fachada, popularmente conhecida como candidatura "laranja", é aquela que entra nas eleições sem a intenção real de concorrer a uma das vagas (OLIVEIRA, 2022).

De acordo com um estudo publicado pelo jornal Gaúcha Zero Hora, apenas nas últimas eleições municipais ocorridas no ano de 2020, verificou-se indícios de mais de 5000 candidaturas de fachada do sexo feminino (TURTELLI; GOMES, 2020). Um outro levantamento realizado pelo Jornal Nacional nas eleições de 2018, identificou cerca de 51 candidatos(as) que possivelmente serviram como candidaturas de fachada levando em consideração a relação custo por voto (G1, 2019).

A utilização de candidaturas de fachada para obtenção ilícita do dinheiro do fundo eleitoral para benefício próprio ou do partido, são inegáveis de acordo com Wylie, Santos e Marcelino (2019). O preenchimento ilegítimo das vagas destinadas ao sexo feminino faz com que o avanço no número de mulheres a participar do quadro político não avance como esperado com a aprovação da Lei nº 9.504/1997 (TSE, 1997).

Após o período eleitoral, os candidatos e os partidos políticos devem realizar a prestação de contas para detalhar os seus gastos de campanha. Essas prestações de contas ficam disponíveis publicamente no Portal de Dados do TSE (TSE, 2022). Através desses dados, é possível verificar se os gastos de campanha pertinentes a algum candidato condizem com a verba recebida e se o número de votos recebidos ficou com um custo médio parecido com de outros candidatos.

Atualmente, a identificação dessas candidaturas fictícias de maneira manual por meio dos dados disponibilizados pelo Tribunal Superior Eleitoral (TSE) é extremamente demorada e ineficiente, visto a quantidade de dados a serem analisados e a falta de uma metodologia para realizar essa identificação. Com a técnica de detecção de anomalias aplicada a esses dados, seria possível identificar os padrões de uma candidatura e verificar o que difere das demais candidaturas, possibilitando assim recomendar as principais candidaturas que devem ser alvo de análise pelos órgãos competentes.

Nesse contexto, o objetivo deste trabalho é identificar possíveis candidaturas de fachada empregando técnicas de detecção de anomalias. Para isso, foi criado um *dataset* não rotulado e combinados dois algoritmos não supervisionados para classificar automaticamente as candidaturas como fachada ou não, gerando assim um *dataset* rotulado. Após, foi aplicado um algoritmo supervisionado no *dataset* rotulado e gerada uma árvore de decisão a fim de verificar os atributos que mais contribuem para a identificação de possíveis candidaturas de fachada.

O trabalho proposto realizou a aplicação de algoritmos não supervisionados utilizando técnicas de detecção de anomalias em dados eleitorais. Esses dados eleitorais foram coletados no formato CSV do Portal de Dados Abertos do TSE, e com eles, criado um *dataset* contendo as informações mais relevantes para o estudo de caso. Foi realizado todo o processo de seleção, limpeza e transformação dos dados para a aplicação dos algoritmos. Após a aplicação dos algoritmos, foi gerada uma árvore de decisão para verificar e descobrir os padrões de uma possível candidatura de fachada.

Com a aplicação dos algoritmos não supervisionados, foi possível identificar 572 candidaturas das quais ambos os algoritmos identificaram como anomalia, sendo classi-

ficadas automaticamente como possíveis candidaturas de fachada. Com a identificação dessas candidaturas, foi criado um *dataset* rotulado. Com o *dataset* rotulado, foram geradas as árvores de decisão, das quais mostraram que o atributo custo por voto é um dos mais relevantes para a identificação alvo do estudo.

O restante deste artigo está organizado como segue. A Seção 2 define os conceitos e algoritmos necessários para compreensão deste trabalho. A Seção 3 compara os trabalhos relacionados. A Seção 4 descreve a metodologia utilizada. A Seção 5 realiza uma análise dos resultados obtidos. Por fim, a Seção 6 apresenta a conclusão e sugere trabalhos futuros.

2. Fundamentação Teórica

Nesta seção, é apresentada a fundamentação teórica contendo as tecnologias e técnicas utilizadas na execução do trabalho. A Seção 2.1 trás a técnica de identificação de anomalias e os algoritmos utilizados na aplicação do trabalho, que são o *Isolation Forest* e o *Local Outlier Factor*. A Seção 2.2, explica o algoritmo Decision Tree, utilizado para obter a árvore de decisão do modelo.

2.1. Identificação de Anomalias

A mineração de dados é um processo computadorizado da inteligência de negócios que conduz buscas em grandes quantidades de dados e informações para tentar descobrir relações previamente desconhecidas, mas valiosas (TURBAN; VOLONINO, 2013). A identificação de anomalias é uma das técnicas presentes na mineração de dados.

Segundo Hawkins (1980), a definição de anomalia é: “Uma observação que se desvia tanto de outras observações a ponto de levantar suspeitas de que foi gerado por um mecanismo diferente”, ou seja, é algum dado que se destaca em relação a outros a ponto de parecer que não pertence ao conjunto. Na Figura 1, os pontos circundados em vermelho representam anomalias.

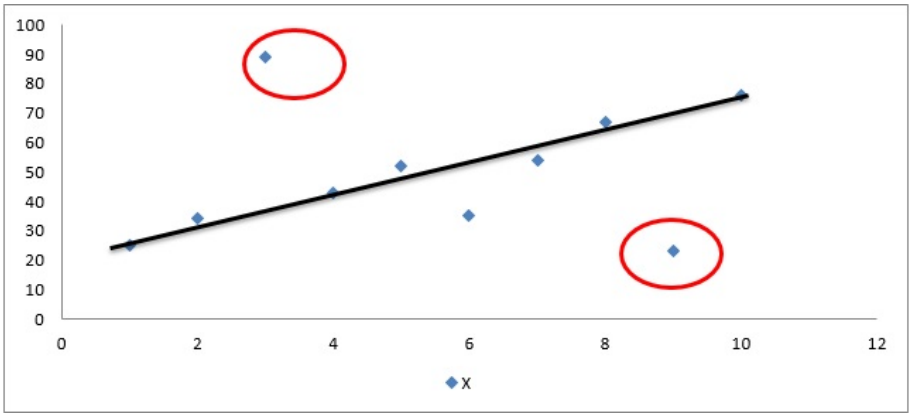

Figura 1. Exemplo de Anomalia.
Fonte: (CHETTY, 2017)

Dentro da técnica de detecção de anomalias, existem várias aplicações úteis. Por exemplo, a detecção de intrusão em um sistema, o monitoramento de saúde em monitores de frequência cardíaca, o diagnóstico de falhas em máquinas e equipamentos, a detecção de fraudes em benefícios fornecidos pelo estado, entre outras citadas por Hodge e Austin (2004).

Para este estudo de caso, deve-se levar em conta que foram utilizados algoritmos não supervisionados, onde estes não precisam de bases rotuladas. Além disso, nessa técnica não há uma distinção entre o *dataset* de treinamento e de teste. A ideia é que um algoritmo de detecção de anomalias não supervisionado pontue os dados unicamente com base nas propriedades dos atributos no conjunto de dados. Normalmente, as distâncias ou densidades são utilizadas para dar uma estimativa do que é normal e do que é uma anomalia (GOLDSTEIN; UCHIDA, 2016).

A Figura 2 mostra a diferença entre algoritmos supervisionados (a) e algoritmos não supervisionados (b). Interpreta-se que, os algoritmos supervisionados possuem seus resultados mais definidos, pois é realizado um treinamento com uma base de dados rotulada. Nos algoritmos não supervisionados, o resultado nem sempre é trivial, pois como não há os dados rotulados, e não há um modelo treinado para identificar um determinado conjunto de características, essa identificação de anomalia é dada unicamente através da interpretação dos dados presentes no *dataset* de acordo com o algoritmo selecionado.

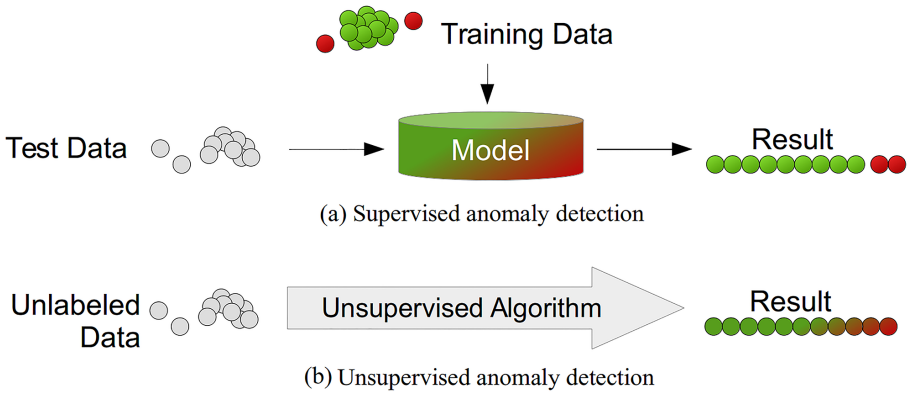

Figura 2. Diferença entre tipos de algoritmos.
Fonte: (GOLDSTEIN; UCHIDA, 2016)

O restante desta Subseção está dividida entre a Subseção 2.1.1 onde é explicado o funcionamento do algoritmo *Isolation Forest* e a Subseção 2.1.2 que é explicado o funcionamento do algoritmo *Local Outlier Factor*.

2.1.1. *Isolation Forest*

O algoritmo *Isolation Forest* é um método baseado em algoritmos de árvore não supervisionado, esse utilizado para a detecção de anomalias. Ele está disponível na biblioteca scikit-learn na linguagem de programação Python (SCIKIT-LEARN, 2022a).

Esse algoritmo trabalha isolando observações e selecionando aleatoriamente um

atributo para testar. Após, é selecionando aleatoriamente um valor dividido entre os valores máximo e mínimo do atributo selecionado. Essa partição recursiva é representada por uma estrutura em árvore, onde o número de partições necessárias para isolar uma amostra é equivalente ao comprimento do caminho desde o nó raiz até o nó terminal (SCIKIT-LEARN, 2022b).

Nesse algoritmo, a medida de normalidade é de acordo com a distância entre os nós, produzindo assim uma pontuação. Portanto, uma amostra que é isolada rapidamente em poucas divisões, recebe uma pontuação negativa e é considerada uma anomalia, já uma amostra que precisa de várias divisões para ser isolada, recebe uma pontuação positiva e então pode ser considerada uma observação válida.

2.1.2. *Local Outlier Factor*

O algoritmo *Local Outlier Factor* é um método não supervisionado baseado no cálculo de densidade em pontos de dados. Esse algoritmo está disponível na biblioteca scikit-learn na linguagem de programação Python (SCIKIT-LEARN, 2022a).

Esse algoritmo define a distância entre vizinhos e calcula o desvio da densidade local de um determinado ponto de dados em relação a seus vizinhos. São consideradas como anomalias as observações que têm uma densidade substancialmente menor do que seus vizinhos (SCIKIT-LEARN, 2022c).

Portanto, esse algoritmo faz o cálculo da densidade local dos valores presentes e compara com os seus vizinhos. Se a densidade desses valores for expressivamente menor do que de seus vizinhos, então é considerada uma possível anomalia.

2.2. Árvores de Decisão

O aprendizado em árvore de decisão é utilizado para a aproximação de funções-alvo discretas, em que a função aprendida é representada por uma árvore de decisão (MITCHELL, 1997). O objetivo desse algoritmo é mostrar as decisões tomadas no aprendizado e gerar a árvore de decisão para melhor analisar os resultados.

As árvores de decisão são métodos supervisionados de aprendizado, que utilizam modelos de classificação e regressão. O modelo de classificação tem o objetivo de prever uma classe associada com uma variável de entrada contendo determinados atributos (FONTANA, 2020). Um exemplo de algoritmo de classificação é o *Decision Tree*, presente na biblioteca scikit-learn (SCIKIT-LEARN, 2022d) e que é utilizado neste trabalho.

Um dos critérios utilizado para divisão dos nós da árvore de decisão é o *GINI*, que calcula o grau ou probabilidade de uma determinada variável ser classificada erroneamente quando é escolhida aleatoriamente. Esse grau varia entre 0 e 1, em que quanto mais próximo de 0, maior o nível de pureza da classificação, isto significa que todos os elementos contidos no nó são de uma classe única e esse nó não será dividido novamente. O GINI é calculado pela Equação 1, onde o *n* representa o número total de classes e o *Pi* representa a probabilidade de selecionar um dado que pertença a classe *i*.

$$GINI = 1 - \sum_{i=1}^{n} (P_i)^2 \tag{1}$$

O algoritmo *Decision Tree* é dividido em duas etapas, treinamento e predição. Basicamente o algoritmo seleciona um nó inicial e divide entre outros dois nós de acordo com um teste realizado no atributo selecionado. A cada novo nó, um novo teste de valores no atributo selecionado é realizado, onde a etapa anterior se repete até chegar ao resultado final, o chamado nó folha. Por fim, pode ser gerada uma imagem da árvore de decisão mostrando o percurso realizado até chegar ao resultado final.

3. Trabalhos Relacionados

Nesta Seção, são apresentados quatro artigos relacionados ao trabalho proposto que utilizam técnicas de detecção de anomalias.

O primeiro trabalho de Filho (2020) utilizou o algoritmo *Isolation Forest* aplicado aos dados eleitorais do ano de 2018 retirados do Portal de Dados Abertos do TSE. Os resultados obtidos permitiram a identificação de discrepâncias baseadas na relação custo por voto, ou seja, geralmente quanto maior o custo de uma campanha, maior será a quantidade de votos, e a discrepância seria um gasto desproporcional se comparado com a quantidade de votos obtidos. Por fim, foi gerada uma árvore de decisão com os perfis dos candidatos identificados como anomalias.

O segundo trabalho de Souza et al. (2021) também utilizou o algoritmo *Isolation Forest*, esse aplicado a uma base de dados disponibilizada pela empresa Accenture com informações de ambientes de produção real. O objetivo do trabalho era validar a criação de uma técnica de detecção de anomalias para ambiente de monitoramento de um sistema. Esse trabalho identifica potenciais problemas que podem surgir durante a execução de determinadas cadeias de processos, como, por exemplo, o tempo de execução. O autor concluiu que obteve bons resultados na capacidade de generalização para a detecção de anomalias nos processos do sistema.

O terceiro trabalho de Silva et al. (2021) utilizou os algoritmos *Isolation Forest*, *DBSCAN* e *k-means* aplicados a uma base de dados cedida pela empresa SENFIO. O objetivo foi encontrar anomalias em variações de temperatura em refrigeradores onde são armazenados produtos hospitalares, tais como vacinas e bolsas de sangue. Após a aplicação dos algoritmos, o que obteve melhor desempenho foi o *DBSCAN*, esse teve uma acurácia calculada em 76,7%, e identificou 946 anomalias nos 38 mil registros presentes na base de dados utilizada.

O último trabalho de Bueno et al. (2021) utilizou os algoritmos *Isolation Forest* e *k-means* aplicados aos dados eleitorais do ano de 2018 retirados do Pòrtal de Dados Abertos do TSE. Seu objetivo é identificar anomalias de maneira geral nesses dados. Para isso, foi utilizada uma combinação experimental entre dois algoritmos, o *Isolation Forest* e o *K-means*. Após a aplicação dos algoritmos observou-se que a aplicação isolada do *Isolation Forest* obteve um desempenho melhor se comparado com a combinação entre ele e o *K-means*.

A Tabela 1 trás uma comparação entre os trabalhos relacionados e o trabalho proposto. Todos os trabalhos selecionados utilizaram o algoritmo *Isolation Forest* para a tarefa de detecção de anomalias. O trabalho de Filho (2020) e o trabalho de Bueno et al. (2021) coletaram os dados utilizados no mesmo local do trabalho proposto, no caso o Portal de Dados Abertos do TSE. Os demais trabalhos também trazem experimentos com algoritmos de detecção de anomalias, mas esses aplicados a outros contextos.

Tabela 1. Comparação dos Trabalhos Relacionados.

Artigo	[de Albuquerque Filho 2020]	[de Souza et al. 2021]	[da Silva et al. 2021]	[Bueno et al. 2021]	Trabalho Proposto
Objetivo	Identificar candidaturas de fachada	Estimar a probabilidade de ocorrência de uma anomalia na execução de um sistema e classificar as com maior impacto	Identificar anomalias em variações de temperaturas de um refrigerador	Estudar uma combinação de dois algoritmos aplicados na detecção de anomalias em conjuntos de dados eleitorais	Identificar possíveis candidaturas de fachada
Algoritmo(s)	*Isolation Forest*	*Isolation Forest*	*Isolation Forest, DBSCAN, k-means*	*Isolation Forest, k-means*	*Isolation Forest, Local Outlier Factor, Decision Tree*
Resultados	Identificadas discrepâncias em 20 candidatos levando em conta a relação custo por voto	Dos cinco modelos criados, o modelo denominado rt obteve o melhor resultado, com acurácia variando de ~82% à ~99%	O algoritmo *DBSCAN* obteve o melhor resultado com 76,7% de acurácia, este detectou 946 anomalias nos 38 mil registros da base de dados	A combinação dos algoritmos teve um desempenho pior se comparado com a aplicação apenas do *Isolation Forest*	Identificadas 265 anomalias nas eleições de 2018, e identificadas 307 anomalias nas eleições de 2022
Base de Dados	Base de dados retirada do Portal de Dados Abertos do TSE, esta no formato CSV	Base de dados fornecida pela empresa Accenture, esses contidos em uma planilha	Base de dados fornecida pela empresa SENFIO, esses contidos em uma planilha	Base de dados retirada do Portal de Dados Abertos do TSE, esta no formato CSV	Base de dados retirada do Portal de Dados Abertos do TSE, esta no formato CSV

O trabalho de Filho (2020) possuí o mesmo objetivo do trabalho proposto, no caso, identificar possíveis candidaturas de fachada. O trabalho proposto identificou outros atributos que possam colaborar na identificação de possíveis candidaturas de fachada, e aplicou uma técnica de rotulação automática dos dados utilizando dois algoritmos para classificar candidaturas como de fachada ou não, criando assim um *dataset* rotulado. Com essa técnica foi possível utilizar um algoritmo supervisionado para gerar árvores de decisão e identificar padrões úteis e os atributos mais relevantes para esse estudo de caso.

4. Metodologia

Esta Seção apresenta a metodologia empregada neste trabalho com o objetivo de encontrar padrões que possam auxiliar na identificação de possíveis candidaturas de fachada.

A Figura 3 ilustra o fluxo utilizado na realização deste trabalho. Primeiramente, foi criada uma base de dados contendo as informações dos candidatos coletadas do Portal de Dados Abertos do TSE. Nas etapas seguintes, aplicou-se os processos de pré processamento e transformação dos dados, onde foi realizada a limpeza dos dados e a transformação dos atributos para valores numéricos. Em seguida, realizou-se a aplicação dos algoritmos não supervisionados para a rotulação automática das candidaturas como de fachada ou válidas. Após, efetuou-se o balanceamento dos dados, uma vez que o número de instâncias de candidaturas de fachada é 98,5 % menor que as candidaturas válidas. Na próxima etapa, foi realizada a mineração de dados utilizando um algoritmo supervisionado para o treinamento de um modelo. Posteriormente, o modelo foi analisado, onde verificou se os resultados condiziam com o objetivo. Em casos de resultados irrelevantes para o estudo de caso, retornou-se para a etapa anterior para alterar iterações ou atributos a fim de obter melhores resultados. Por fim, gerou-se uma árvore de decisão para verificar-se os atributos mais pertinentes para o estudo de caso. Cada etapa é descrita nas subseções seguintes.

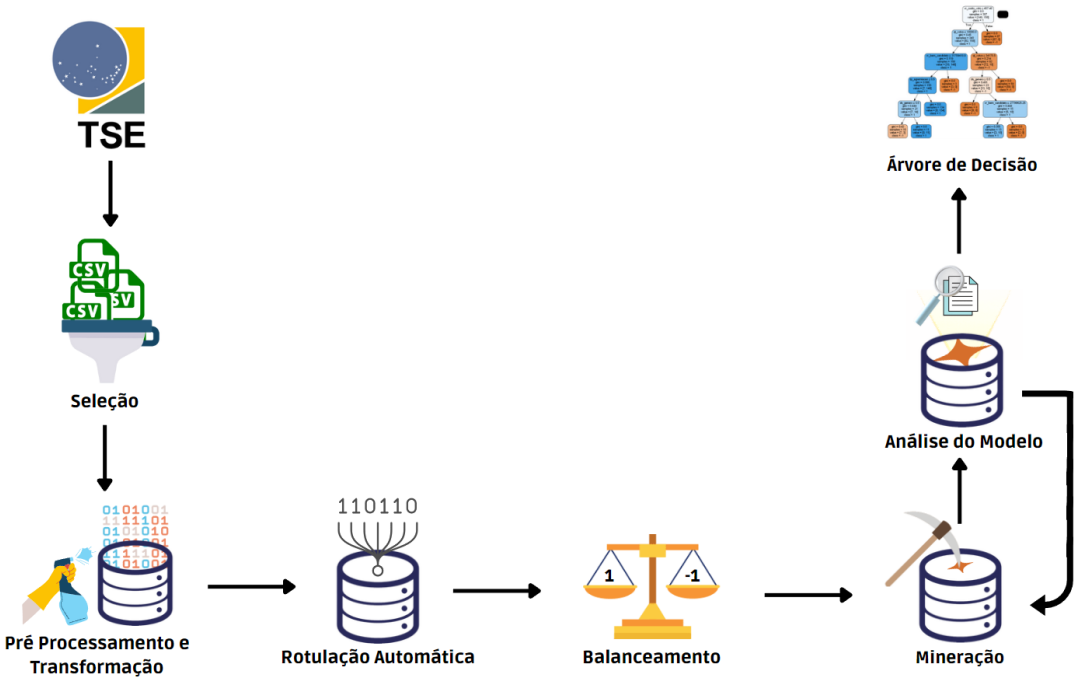

Figura 3. Fluxograma da metodologia adotada
Fonte: Autor

4.1. Seleção

O objetivo desta etapa foi buscar os dados necessários para a realização do trabalho e criar uma base de dados. A construção da base de dados se deu através da coleta manual de arquivos no formato CSV disponibilizados no Portal de Dados Abertos do TSE. Os arquivos selecionados para o estudo de caso foram os arquivos que continham informações dos candidatos, bens dos candidatos, prestações de contas eleitorais (despesas contratadas) e resultados das eleições.

Na etapa de seleção dos dados, foi definido que seriam utilizados apenas dados de eleições federais e estaduais sobre candidatos a deputados federais, estaduais e distritais, visto que é onde a maioria dos casos de candidaturas de fachada ocorrem nessas eleições. Também foram excluídos candidatos que tiveram suas candidaturas indeferidas pelo TSE, pois candidaturas que não participaram do processo eleitoral pelos mais variados motivos, não são relevantes para este estudo de caso, visto que o objetivo é identificar candidaturas de fachada, das quais a sua identificação só é possível com os resultados e dados posteriores à eleição. Para a aplicação dos algoritmos, foram selecionados os dados referentes as eleições de 2018 e 2022.

4.2. Pré Processamento e Transformação

Após a coleta dos dados, foi realizada uma limpeza de caracteres inválidos ao formato UTF-8. Em seguida, foram removidos dados incompletos, repetidos e irrelevantes para o estudo de caso, como por exemplo, dados de identificação pessoal do candidato (CPF, Título de Eleitor, entre outros).

Com a limpeza de dados concluída, foi criado também um atributo para calcular

o custo por voto de cada candidato. Alguns processos de transformação também foram realizados para favorecer os algoritmos na identificação de anomalias.

Uma das transformações realizadas foi a conversão da idade dos candidatos para faixa etária com um intervalo de 5 anos cada. Outra transformação foi a modificação do atributo custo por voto em candidatos que tiveram despesas maiores que zero e que não obtiveram nenhum voto. Para estes casos, foi definindo que o valor do custo por voto seria o total das despesas contratadas multiplicado por -1, para assim, dar ênfase nessas candidaturas.

A versão final da base de dados conta com duas tabelas, uma contendo candidaturas a deputado de 2018 e outra contendo candidaturas a deputado de 2022. A tabela de candidatos de 2018 conta com 24196 registros e a tabela de candidatos de 2022 conta com 24077 registros, totalizando assim um total de 48273 candidaturas analisadas pelos algoritmos.

4.3. Rotulação Automática

A rotulação automática dos dados definindo qual candidatura é possivelmente de fachada ou não, foi adotada com o objetivo de conseguir um *dataset* rotulado e poder aplicar um algoritmo de classificação para encontrar padrões úteis que descrevem candidaturas de fachada. A rotulação automática foi necessária uma vez que não existem dados públicos rotulados sobre o tema e não é uma tarefa trivial que possa ser feita manualmente por um usuário não especializado.

Para essa rotulação, foram utilizados dois algoritmos não supervisionados, no caso o *Isolation Forest* e o *Local Outlier Factor*. Ambos os algoritmos foram utilizados com os parâmetros já pré estabelecidos pela biblioteca (SCIKIT-LEARN, 2022a).

A regra da rotulação automática para este estudo de caso foi definida como: se uma candidatura for avaliada como de fachada por ambos os algoritmos utilizados, então essa é considerada de fachada. Caso uma candidatura não seja detectada como de fachada por nenhum dos algoritmos, então essa é considerada uma candidatura válida. E por fim, caso houver divergências entre os algoritmos para alguma candidatura, ou seja, caso um identifique como anomalia e o outro não, essas foram marcadas como talvez. O Algoritmo 1 exemplifica essa etapa da rotulação automática.

Algoritmo 1: Rotulação Automática de Dados

Entrada: $resIF$ - resultado Isolation Forest
$resLOF$ - resultado Local Outlier Factor
Saída: $resultado$ - resultado da rotulação

if *(resIF = fachada AND resLOF = fachada)* **then**
  resultado = fachada
else if *(resIF = não fachada AND resLOF = não fachada)* **then**
  resultado = válida
else
  resultado = talvez
end
return resultado

Após a rotulação, na base eleitoral de 2018, 19217 candidaturas foram rotuladas como candidaturas válidas, 265 foram rotuladas como possíveis candidaturas de fachada e 4714 foram rotuladas como talvez. Na base eleitoral de 2022, 19071 candidaturas foram rotuladas como candidaturas válidas, 307 foram rotuladas como possíveis candidaturas de fachada e 4699 foram rotuladas como talvez. As candidaturas das duas eleições que foram rotuladas como talvez foram eliminadas da base de dados. Com essa tática, foi possível obter um *dataset* rotulado.

4.4. Balanceamento

O objetivo dessa etapa foi realizar o balanceamento para equilibrar os dados pertinentes as duas classes alvo (candidaturas válidas e de fachada). O balanceamento de dados é necessário quando o conjunto de dados está desequilibrado, ou seja, algumas classes têm muito mais casos do que outras. O desequilíbrio desses dados tem um sério impacto sobre o desempenho de classificadores, onde algoritmos de aprendizagem que não consideram o equilíbrio de classes, tendem a ser sobrecarregados pela classe majoritária e acabam ignorando a classe minoritária (LIU; WU; ZHOU, 2008).

O balanceamento dos dados se deu necessário após a obtenção dos resultados, onde o número de instâncias da classe fachada foi cerca de 1,5% do total de candidaturas classificadas pelos algoritmos. Para resolver esse problema, também conhecido como *undersampling*, foi utilizado um algoritmo chamado *Random Undersampling* disponível na biblioteca imbalanced-learn (IMBALANCED-LEARN, 2022). Esse algoritmo opera excluindo observações de maneira aleatória da classe majoritária, nesse caso a classe de candidaturas classificadas como válidas.

4.5. Mineração

Com o *dataset* rotulado e o processo de balanceamento realizado, foi aplicado o algoritmo *Decision Tree* para criar um modelo que identificasse padrões úteis que descrevessem candidaturas de fachada. A criação desse modelo envolveu o treinamento e a realização de testes. Foi utilizado o parâmetro max_deph presente na biblioteca para definir a profundidade limite da árvore para 5, pois por padrão não possuía um limite, e acabava gerando árvores muito grandes para serem analisadas manualmente.

Vários atributos tiveram de ser descartados durante os experimentos, pois não geravam árvores com propósitos fim de identificação de candidaturas de fachada. Um desses atributos descartados foi o número do partido, onde observando na árvore de decisão, este gerava validações equivocadas, como por exemplo: caso o número do partido for menor que “X”, então esta observação é considerada uma candidatura de fachada. Com isso notou-se que atributos de identificação não são pertinentes no escopo do trabalho.

Para o treinamento do modelo foram definidas as variáveis com as características/dados referentes às candidaturas, e foi definido a variável alvo para predição do modelo. O *dataset* foi dividido em dois, sendo 50% utilizado para treinamento e os outros 50% utilizado para teste. Após o treinamento foi gerada as métricas de avaliação e a árvore de decisão para análise do modelo gerado.

A geração das árvores de decisão se deu através da aplicação do algoritmo *Decision Tree*. A geração das versões finais das árvores de decisão se deu após várias

aplicações dos algoritmos de detecção, com o intuito de se chegar a um resultado mais próximo ao estudo de caso selecionado.

4.6. Análise do Modelo

O objetivo dessa etapa foi analisar a árvore de decisão criada, verificar se novos ajustes deveriam ser realizados e extrair padrões úteis para identificar possíveis candidaturas de fachada.

Durante a análise dos primeiros modelos criados, notou-se que a árvore gerada possuía testes em atributos que levavam a resultados contestáveis, e que após a análise desses atributos, notou-se que os algoritmos não supervisionados não são eficazes quando testados em atributos de identificação, como, por exemplo, o número do partido. Isso se dá, pois os algoritmos não supervisionados realizam os seus testes levando em conta o valor máximo e mínimo de atributos e sua densidade, e atributos usados apenas para identificação de determinada categoria acabam interferindo negativamente nos resultados. Portanto, esses atributos tiveram de ser descartados e a etapa de mineração ser refeita.

Para avaliar o modelo criado pelo *Decision Tree*, foram utilizadas métricas tradicionais de classificação: precisão, revocação, acurácia e F1-score. Essas métricas são obtidas a partir da matriz de confusão, que é uma forma simples de apresentar os resultados de algoritmos de classificação. Ela indica a quantidade de ocorrências em quatro categorias (MARIANO et al., 2021), são elas:

- Verdadeiros Positivos (VP): Candidaturas identificadas como de fachada e que de fato eram de fachada;
- Falsos Positivos (FP): Candidaturas identificadas como de fachada e que na verdade eram candidaturas válidas;
- Verdadeiros Negativos (VN): Candidaturas identificadas como válidas e que de fato eram candidaturas válidas;
- Falsos Negativos (FN): Candidaturas identificadas como válidas e que na verdade eram de fachada.

A partir dessas variáveis, é possível calcular as métricas de eficácia para verificar se o modelo está valido ou não. As métricas utilizadas foram:

- Precisão: Porcentagem dada pela razão entre a quantidade de candidaturas classificadas corretamente como de fachada e o total de candidaturas classificadas como de fachada, essa, representada pela Equação 2;

$$Precisão = \frac{VP}{VP + FP} \tag{2}$$

- Acurácia: Porcentagem dada pela soma da quantidade de candidaturas classificadas corretamente, dividido pelo total de candidaturas, essa, representada pela Equação 3;

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{3}$$

- Revocação ou *Recall*: Porcentagem dada pela razão entre a quantidade de candidaturas classificadas corretamente como de fachada e a soma das candidaturas

classificadas corretamente como de fachada e candidaturas classificadas erroneamente como válidas, essa, representada pela Equação 4;

$$Revocação = \frac{VP}{VP + FN} \tag{4}$$

- F1-score: Porcentagem definida pela média harmônica entre as métricas de precisão e revocação, esta, representada pela Equação 5.

$$F1 = \frac{2 * Precisão * Revocação}{Precisão + Revocação} \tag{5}$$

5. Resultados e Discussão

Nesta Seção, são abordados os resultados obtidos com a realização dos experimentos. A Subseção 5.1 trás a análise da eficácia do modelo criado. A Subseção 5.2 apresenta uma análise das árvores de decisão geradas. Por fim, a Subseção 5.3 analisa os atributos mais relevantes para o estudo de caso.

5.1. Análise da Eficácia do Modelo

A análise da eficácia do modelo se deu em duas etapas. Na primeira etapa, foram executados os algoritmos não supervisionados para identificar as possíveis candidaturas de fachada e rotular o *dataset*, essa primeira análise foi feita de maneira superficial levando em conta se estava identificando candidaturas com maior custo por voto. Na segunda etapa, foi executado o algoritmo de balanceamento e, em seguida, o algoritmo de árvore para possibilitar a análise dos resultados.

Após gerar o modelo da árvore de decisão, foi possível obter as métricas de avaliação. A Tabela 2 apresenta as métricas obtidas na geração da árvore de decisão da base eleitoral de 2018, enquanto a Tabela 3 mostra as métricas obtidas na geração da árvore de decisão da base eleitoral de 2022.

Tabela 2. Métricas para a Base Eleitoral de 2018.

Métrica	Resultado
Acurácia	96,98%
Precisão	96,27%
Revocação	97,73%
F1 Score	96,99%

Tabela 3. Métricas para a Base Eleitoral de 2022.

Métrica	Resultado
Acurácia	94,14%
Precisão	94,56%
Revocação	93,29%
F1 Score	93,92%

O número total de anomalias identificadas pelos algoritmos na base eleitoral de 2018 foi de 265 possíveis candidaturas de fachada. Dessas candidaturas identificadas, destacam-se que foram identificadas 44 candidaturas com valor do custo por voto maior que R$ 500,00 reais, sendo que dessas 44 candidaturas, cerca de 93% são do sexo feminino.

Na base eleitoral de 2022, foram identificadas 307 possíveis candidaturas de fachada. Dessas candidaturas identificadas, 132 delas possuem o valor do custo por voto maior que R$ 500,00 reais, sendo que dessas 132 candidaturas, cerca de 74% são do sexo feminino. Com esses resultados, aponta-se indícios de uma maior utilização de candidaturas de fachada para o preenchimento de cotas para o gênero feminino de maneira irregular. A Figura 4 apresenta esses resultados em forma de gráfico para melhor compreensão dessa relação.

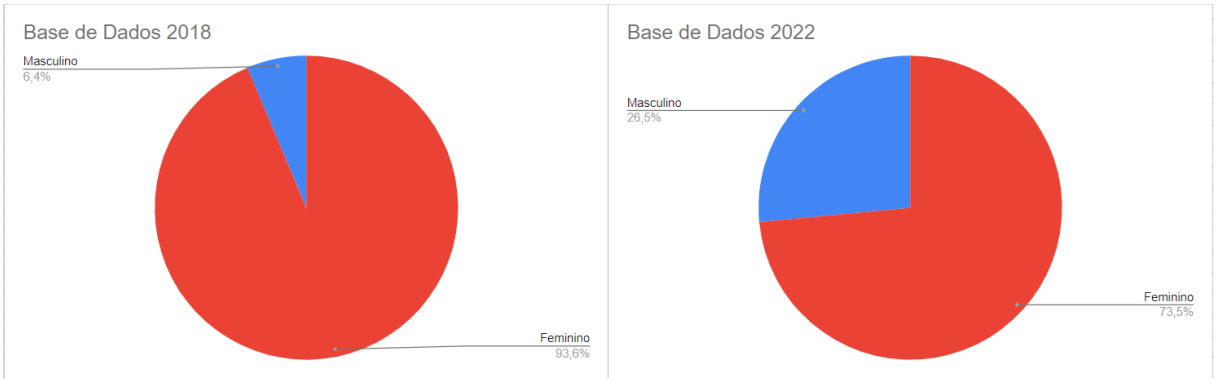

Figura 4. Quantidade de candidaturas por sexo com custo por voto maior que R$ 500,00 reais.
Fonte: Autor

5.2. Análise da Árvore de Decisão

As árvores de decisão foram geradas com o intuito de visualizar os padrões encontrados e os atributos mais relevantes para o estudo de caso. A Figura 5 representa a árvore de decisão gerada com a base eleitoral do ano de 2018, enquanto a Figura 6 ilustra a árvore de decisão gerada para a base eleitoral do ano de 2022.

Para análise das árvores, deve-se levar em conta a interpretação das seguintes variáveis de saída de cada nodo: A variável *gini* representa o grau de pureza da classificação; A variável *samples* representa a quantidade de amostras analisadas naquele nó; A variável *value* trás em ordem as observações que foram classificadas naquele nó e que se enquadram em cada uma das classes. E a variável *class* representa a classe em que está classificada aquele nó, nesse caso, a *class* = -1 representa uma possível candidatura de fachada, e a *class* = 1 representa uma candidatura válida. Para análise do gênero, lê-se que valor da variável = 1 representa candidaturas do sexo masculino e valor = 0 representa candidaturas do sexo feminino. Para análise do tipo de agremiação lê-se que valor da variável = 1 representa que a candidatura está em um partido isolado e valor = 0 representa que a candidaturas pertence a uma coligação.

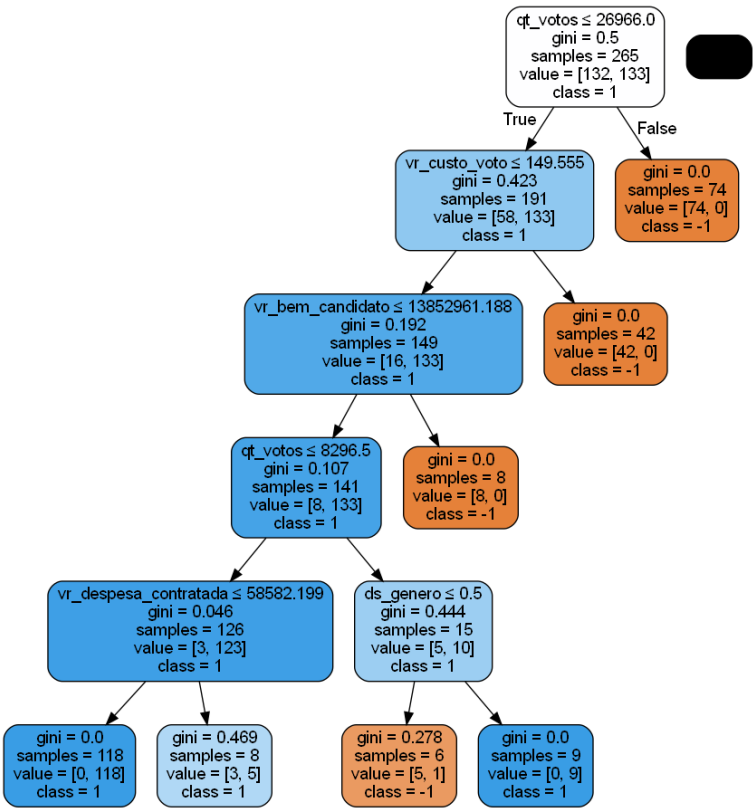

Figura 5. Árvore de Decisão para a Base Eleitoral 2018.
Fonte: Autor

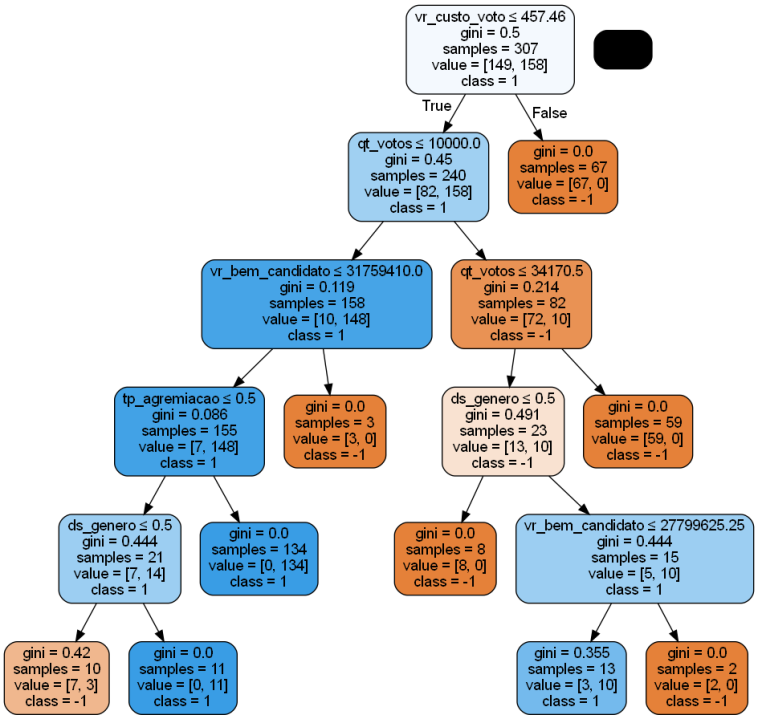

Figura 6. Árvore de Decisão para a Base Eleitoral 2022.
Fonte: Autor

Com as árvores de decisão ficou visivelmente claro os atributos que mais influenciaram nas tomadas de decisão. Os principais atributos de interesse identificados após a geração das árvores foram: quantidade de votos, valor total de despesas contratadas, valor total de bens dos candidatos, valor do custo por voto, tipo de agremiação e o gênero dos candidatos.

Ao analisar as árvores, é possível verificar que candidaturas com um valor de custo por voto maior que a média geral dos dados são identificadas como anomalias, mostrando assim a relevância do atributo para esse estudo de caso, como já identificado anteriormente por outras pesquisas (G1, 2019). A quantidade de votos e o valor total de bens dos candidatos também se demonstraram muito relevantes para o estudo de caso. Uma observação que pode ser tirada também está na relação com o gênero, onde após alguns testes realizados pela árvore, esse atributo influenciou na escolha do nó final, onde em ambas as árvores, observou-se que caso o gênero fosse feminino, era considerado uma possível candidatura de fachada.

Nota-se também que algumas validações realizadas pela árvore não condizem necessariamente com uma candidatura de fachada. O que acontece é que os algoritmos não supervisionados não são triviais, ou seja, eles acabam identificando outras anomalias que podem não ser relevantes para esse estudo de caso, como, por exemplo, se o atributo valor total de bens do candidato ou o atributo quantidade de votos for muito superior a média presente em outros candidatos, esse poderá ser considerado uma anomalia.

5.3. Análise dos Atributos Mais Relevantes

O objetivo desta Subseção é comparar os atributos identificados como mais relevantes no presente trabalho, com os atributos identificados no trabalho relacionado de Filho (2020). Ambos os trabalhos identificaram como atributos relevantes para o estudo de caso o valor do custo por voto, o valor total de bens dos candidatos, o valor total das despesas contratadas e o gênero das candidaturas.

Em ambos os trabalho foram geradas árvores de decisão para verificar os atributos mais relevantes, porém, alguns atributos distintos foram identificados entre os trabalhos. O trabalho proposto identificou outros atributos que podem ser relevantes, tais como a quantidade de votos e o tipo de agremiação. Já no trabalho de Filho (2020) foi identificado que o atributo grau de instrução também pode ser relevante para o estudo de caso. Acredita-se que a diferença entre os atributos identificados deve-se a técnica utilizada para rotulação das candidaturas e aos modelos treinados para geração das árvores de decisão.

6. Conclusão

O trabalho teve a proposta de identificar possíveis candidaturas de fachada utilizando técnicas de detecção de anomalias aplicadas a dados eleitorais disponibilizados pelo TSE. Para isso, foi criada uma base de dados contendo as informações dos candidatos coletadas do Portal de Dados Abertos do TSE. Após, foram aplicados algoritmos não supervisionados para a detecção e rotulação das candidaturas identificadas automaticamente. Foi aplicado também um algoritmo para realizar o balanceamento dos dados e aplicado um algoritmo supervisionado para gerar as árvores de decisão e verificar os padrões encontrados.

A técnica de rotulação automática para esse cenário onde não há um *dataset* rotulado, mostrou-se eficaz para a identificação das anomalias. Com esse *dataset* rotulado de maneira automática, possibilita-se a utilização de outros algoritmos supervisionados para auxiliar no processo de validação dos resultados e verificar-se assim se os atributos e parâmetros utilizados na detecção não supervisionada estão acertados.

No total, foram identificadas 572 possíveis candidaturas de fachada e foram verificados alguns padrões que compuseram essa identificação. Foram geradas duas árvores de decisão, uma para a base eleitoral de 2018 e outra para a base eleitoral de 2022. Em ambas as árvores, ficou claro que o atributo custo por voto é um dos que mais influenciou na detecção. Atributos como quantidade de votos e valor total dos bens dos candidatos também tiveram relevância para esse estudo de caso.

Observa-se também que em ambas as árvores geradas, o atributo gênero foi utilizado para chegar a algumas decisões dos nós folha, onde candidaturas do sexo feminino eram detectadas como de fachada nesse caso em combinação com outros atributos. Com isso, há indícios de uma maior utilização de candidaturas de fachada para o preenchimento ilegal das cotas de gênero. Os resultados obtidos pela rotulação dos dados e pela geração das árvores de decisão mostram que candidaturas que obtiveram maior custo por voto, tendem a serem detectadas como possíveis candidaturas de fachada.

Como trabalhos futuros, sugere-se a utilização de mais algoritmos de detecção de anomalias para poder realizar uma rotulação automática com maior diversidade de técnicas e assim melhorar a precisão dessa etapa. Outra possibilidade, é a utilização de outros dados disponíveis no portal de dados abertos do TSE, como, por exemplo, os dados processuais referentes aos candidatos, para possivelmente detectar algum candidato que já teve processos judiciais e verificar se essa informação pode ou não afetar na identificação de possíveis candidaturas de fachada. Outra possibilidade, é filtrar e separar o valor das despesas contratadas por categoria, como, por exemplo, valor gasto em propaganda de rádio, valor gasto com panfletos, valor gasto com repasses para outros candidatos, entre outros, para verificar se há um padrão de gastos nas candidaturas de fachada.

Referências

BRASIL. *Emenda Constitucional nº 117, de 5 de Abril de 2022*. 2022. Disponível em: http://www.planalto.gov.br/ccivil_03/Constituicao/Emendas/Emc/emc117.htm. Último Acesso: 09 de Junho de 2022.

BUENO, L. F. et al. Uma combinação dos algoritmos Isolation Forest e K-Means aplicada às Eleições Brasileiras. *Anais do Simpósio Brasileiro de Pesquisa Operacional;*, LIII Simpósio Brasileiro de Pesquisa Operacional (SBPO 2021), v. 53, n. 4, p. 1–13, 2021.

CHETTY, P. *Representation of dataset X for outlier identification*. 2017. Disponível em: https://www.projectguru.in/detect-outliers-dataset/. Último Acesso: 17 de Junho de 2022.

FILHO, J. E. de A. Detecção de anomalia nas Eleições de 2018 com Isolation Forest. *Revista de Engenharia e Pesquisa Aplicada*, Revista de Engenharia e Pesquisa Aplicada, v. 5, n. 2, p. 104–109, 2020.

FONTANA Éliton. *Introdução aos Algoritmos de Aprendizagem Supervisionada*. [S.l.]: Universidade Federal do Paraná - UFPR Departamento de Engenharia Química, 2020.

G1. *Levantamento identifica pelo menos 51 candidatos 'laranjas' na eleição do ano passado*. 2019. Disponível em: https://g1.globo.com/politica/noticia/2019/02/15/levantamento-identifica-pelo-menos-51-candidatos-laranjas-na-eleicao-do-ano-passado.ghtml. Último Acesso: 02 de Junho de 2022.

GOLDSTEIN, M.; UCHIDA, S. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS One*, PLOS One, v. 11, n. 4, p. e0152173, 2016.

HAWKINS, D. M. *Identification of Outliers*. 1ª. ed. Chapman and Hall, London, UK: Springer Dordrecht, 1980.

HODGE, V. J.; AUSTIN, J. A survey of outlier detection methodologies. *Artificial Intelligence Review*, Kluwer Academic Publishers, v. 22, n. 1, p. 85–126, 2004.

IMBALANCED-LEARN. *RandomUnderSampler*. 2022. Disponível em: https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html. Último Acesso: 29 de Dezembro de 2022.

LIU, X.-Y.; WU, J.; ZHOU, Z.-H. Exploratory Undersampling for Class-Imbalance Learning. *IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS*, IEEE, v. 39, n. 4, p. 539 – 550, 2008.

MARIANO, D. C. B. et al. *Data mining*. 1ª. ed. [S.l.]: SAGAH, 2021.

MITCHELL, T. M. *Machine Learning*. 1ª. ed. [S.l.]: McGraw-Hill Science/Engineering/Math, 1997.

MOLITERNO, D.; RODRIGUES, L. *Entenda o fundo eleitoral, que vai distribuir aos partidos R$ 4,9 bilhões para campanhas*. 2022. CNN Brasil. Disponível em: https://www.cnnbrasil.com.br/politica/entenda-o-fundo-eleitoral-que-vai-distribuir-aos-partidos-r-49-bilhoes-para-campanhas/. Último Acesso: 08 de Junho de 2022.

OLIVEIRA, M. L. P. *Candidaturas laranjas: o que são e como funcionam*. 2022. Disponível em: https://www.politize.com.br/candidaturas-laranjas/. Último Acesso: 21 de Setembro de 2022.

SCIKIT-LEARN. *scikit-learn Machine Learning in Python*. 2022. Disponível em: https://scikit-learn.org/stable/. Último Acesso: 13 de Agosto de 2022.

SCIKIT-LEARN. *sklearn.ensemble.IsolationForest*. 2022. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html. Último Acesso: 28 de Dezembro de 2022.

SCIKIT-LEARN. *sklearn.neighbors.LocalOutlierFactor*. 2022. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html. Último Acesso: 28 de Dezembro de 2022.

SCIKIT-LEARN. *sklearn.tree.DecisionTreeClassifier*. 2022. Disponível em: https://scikit-https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. Último Acesso: 28 de Dezembro de 2022.

SILVA, D. M. da et al. Criação de Modelo de Detecção de Anomalias para Termômetro IoT Usado em Refrigeradores Hospitalares. *Revista de Engenharia e Pesquisa Aplicada*, Revista de Engenharia e Pesquisa Aplicada, v. 6, n. 5, p. 120–128, 2021.

SOUZA, A. A. de et al. Detecção de Anomalias em Aplicações de Monitoramento de Sistemas utilizando Isolation Forest. *Revista de Engenharia e Pesquisa Aplicada*, Revista de Engenharia e Pesquisa Aplicada, v. 6, n. 5, p. 100–109, 2021.

TSE. *Lei das Eleições – Lei nº 9.504, de 30 de setembro de 1997*. 1997. DOU de 1º.10.1997. Disponível em: https://www.tse.jus.br/legislacao/codigo-eleitoral/lei-das-eleicoes/lei-das-eleicoes-lei-nb0-9.504-de-30-de-setembro-de-1997. Último Acesso: 06 de Junho de 2022.

TSE. *Lei nº 13.487, de 6 de outubro de 2017*. 2017. DOU de 6.10.2017. Disponível em: https://www.tse.jus.br/legislacao/codigo-eleitoral/leis-ordinarias/lei-no-13-487-de-6-de-outubro-de-2017. Último Acesso: 06 de Junho de 2022.

TSE. *Portal de Dados Abertos do TSE*. 2022. Disponível em: https://dadosabertos.tse.jus.br/. Último Acesso: 27 de Dezembro de 2022.

TURBAN, E.; VOLONINO, L. *Tecnologia da informação para gestão : em busca do melhor desempenho estratégico e operacional*. 8ª. ed. [S.l.]: Bookman Editora, 2013.

TURTELLI, C.; GOMES, B. *Estudo indica ao menos 5 mil candidatas laranjas*. 2020. Disponível em: https://gauchazh.clicrbs.com.br/politica/noticia/2020/11/estudo-indica-ao-menos-5-mil-candidatas-laranjas-ckhghpcz3001301hxcsfe7s9b.html. Último Acesso: 02 de Junho de 2022.

WYLIE, K.; SANTOS, P. dos; MARCELINO, D. Extreme non-viable candidates and quota maneuvering in brazilian legislative elections. *Revista do CESOP*, OPINIÃO PÚBLICA, v. 25, n. 1, p. 1–28, 2019.