UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DOIS VIZINHOS
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

CESAR RICARDO VELASQUE TRINDADE

DETECÇÃO DE ANOMALIAS EM SISTEMAS DE ADMINISTRAÇÃO DE FROTAS PÚBLICAS MUNICIPAIS

TRABALHO DE CONCLUSÃO DE CURSO DE ESPECIALIZAÇÃO

DOIS VIZINHOS
2021

CESAR RICARDO VELASQUE TRINDADE

DETECÇÃO DE ANOMALIAS EM SISTEMAS DE ADMINISTRAÇÃO DE FROTAS PÚBLICAS MUNICIPAIS

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Marcelo Teixeira

DOIS VIZINHOS
2021

CESAR RICARDO VELASQUE TRINDADE

DETECÇÃO DE ANOMALIAS EM SISTEMAS DE ADMINISTRAÇÃO DE FROTAS PÚBLICAS MUNICIPAIS

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Data de aprovação: 13/dezembro/2021

Marcelo Teixeira
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco

Francisco Pereira Júnior
Mestrado
Universidade Tecnológica Federal do Paraná - Câmpus Cornélio Procópio

Rafael Gomes Mantovani
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Apucarana

DOIS VIZINHOS
2021

Dedico esta monografia à minha amada esposa Susane e meu querido filho Daniel Ricardo pelo incentivo e apoio na realização deste projeto.

AGRADECIMENTOS

A Deus pelo dom da vida.

À Universidade Tecnológica Federal do Paraná pela primazia na oferta do Curso de Especialização em Ciência de Dados.

Ao meu orientador, Professor Marcelo Teixeira, pelo suporte, dedicação e conhecimento dispensados na realização deste projeto.

Aos meus pais pela maneira como me educaram, ensinando a mim o caminho correto a seguir na estrada da vida.

Quando você tem uma meta, o que era um obstáculo passa a ser uma etapa de um dos planos. (BOEHME, Gerhard Erich).

RESUMO

A corrupção endêmica é de difícil detecção e tratamento, sendo elo sólido de impunidade. Nos sistemas informatizados, que integram a complexa rede de gestão pública, é possível que dados legítimos sejam manipulados com intenção não legítima, sombreando a realidade com o propósito de burlar o fisco. Detectar tais padrões de ilegitimidade não é uma tarefa trivial e, em geral, está associada a aspectos interpretativos sobre dados legitimados perante o sistema. Como essa interpretação depende de rastrear e cruzar um amplo e complexo emaranhado de informações, de múltiplas naturezas, como web, bases externas, interfaces públicas estadual e federal, etc., tal tarefa é muitas vezes inviável, se tornando a espinha dorsal da corrupção endêmica. Neste artigo, técnicas de ciência de dados são aplicadas para subsidiar a extração de padrões de irregularidades e tendências em bases reais de dados públicos. Em particular, é aplicado o algoritmo Local Outlier Factor (LOF) , baseado no conceito de densidade local através dos k vizinhos mais próximos, onde a distância é usada para avaliar a densidade, sobre bases de dados públicos relativos à manutenção de frotas, para identificar transações potencialmente fraudulentas. Duas análises foram conduzidas para ilustrar o método, uma relativa a abastecimentos de veículos a gasolina e outra referente a abastecimentos de veículos com óleo diesel. Resultados dão conta de que o desempenho do algoritmo evidencia a existência de registros de abastecimentos de veículos acima da capacidade do tanque de combustível, sendo 27 para abastecimentos com gasolina e 90 para abastecimentos com óleo diesel.

Palavras-chave: corrupção endêmica. fraudulentas. irregularidades.

ABSTRACT

Endemic corruption is difficult to detect and treat, being a solid link of impunity. In computerized systems, which are part of the complex public management network, it is possible that legitimate data are manipulated with an unlawful intention, shadowing reality with the purpose of evading the tax authorities. Detecting such patterns of illegitimacy is not a trivial task and, in general, it is associated with interpretive aspects of legitimate data before the system. As this interpretation depends on tracking and crossing a wide and complex tangle of information, of multiple natures, such as the web, external bases, state and federal public interfaces, etc., this task is often unfeasible, becoming the backbone of endemic corruption . In this article, data science techniques are applied to support the extraction of irregularity patterns and trends in real public databases. In particular, the LOF algorithm is applied, based on the concept of local density through k nearest neighbors, where distance is used to assess density, on public databases relating to fleet maintenance, to identify potentially transactions fraudulent. Two analyzes were carried out to illustrate the method, one concerning the refueling of vehicles with gasoline and the other concerning the refueling of vehicles with diesel oil. Results show that the performance of the algorithm evidences the existence of refueling records for vehicles above the fuel tank capacity, with 27 for refueling with gasoline and 90 for refueling with diesel oil.

Keywords: endemic corruption. fraudulent. irregularities.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ALICE	Análise de Licitações e Editais
CRISP-DM	Cross-Industry Standard Process for Data Mining
INFOSAS	Sistema de Mineração de Dados para controle da produção do Sistema Único de Saúde
IPC	Índice de Percepção da Corrupção
KDD	Knowledge Discovery in Databases
LOF	Local Outlier Factor
MONICA	Monitoramento Integrado para o Controle de Aquisições
SEFAZ	Secretaria de Estado de Fazenda
SOFIA	Sistema de Orientação sobre Fatos e Indícios para o Auditor

SUMÁRIO

1 INTRODUÇÃO

Em sociedades subdesenvolvidas, como é o caso do Brasil, práticas de corrupção endêmica são recorrentes. Diferente do que normalmente se toma como corrupção, ela não incide apenas sobre a esfera pública-política. O senso comum passa a ideia de que o cidadão não se nega a obter vantagem quando lhe é conveniente, indicando que o *jeitinho* e a *malandragem* são heranças culturais. Isso, em conjunção com a ineficácia das leis que disciplinam e penalizam atos de corrupção, institucionalizam o sistema corruptivo brasileiro.

De acordo com (INTERNATIONAL, 2020) o IPC 2020 assinala um panorama bastante desanimador no que diz respeito ao combate à corrupção mundial. Grande parte dos países progrediu pouco ou nada em quase uma década, sendo que mais de dois terços dos países obtiveram uma pontuação abaixo de 50, em uma escala que vai de 0 a 100, em que 0 significa altamente corrupto e 100 significa muito íntegro. Cabe ressaltar que o Brasil figurou em 2020 na 94ª posição num ranking de 180 países, com 38 pontos. Em relação ao ano anterior, houve uma pequena melhora, já que em 2019, a pontuação foi de 35 pontos, demonstrando que a percepção da corrupção no Brasil permanece estagnada em patamar muito ruim.

Na prática, a corrupção produz cinco efeitos devastadores (WARDE, 2018): transforma o Estado e as suas funções em coisas no mercado; desnatura as demais instituições para as submeter aos fins próprios da corrupção; usurpa, ao se apropriar do Estado, a energia vital dos trabalhadores; falseia a concorrência entre os agentes econômicos, para incrementar o poder de mercado de uns em detrimento de outros, até o seu expurgo dos mercados e por fim é um obstáculo ao desenvolvimento das nações, promovendo a pobreza e afrontando a dignidade das pessoas.

O combate à corrupção deve ser, portanto, visto como um ato de intolerante reação ao sistema, passível de ampla e severa punição, forçando assim aqueles que ocupam posições oficiais de trabalho ao exercício digno e incorruptível de suas atribuições. Entretanto, a tarefa de reagir ao sistema e punir a corrupção esbarra, sobretudo, em detectar atos ilícitos. Quando praticados de modo explícito, a detecção costuma ser simples e acompanhada de provas cabais e, em geral, materiais. Porém, para uma vasta classe de atos modernos de ilicitude, eventuais indícios são de ordem não material, muitas vezes mascarados em fragmentos de dados cuja derivação do valor semântico depende de complexo processamento computacional. Dessa forma, os avanços nas áreas de hardware e software, muito embora agreguem vantagens sem precedentes à sociedade (ELEUTÉRIO; MACHADO, 2011), também subsidiam indiretamente a engenharia de novas práticas ilegais, criminosas e fraudulentas.

De fato, avanços nas áreas de hardware e de comunicação geram uma superabundância de dados (CASTRO; FERRARI, 2016), de modo tal que a capacidade de registrar dados supera a habilidade de análise e extração do conhecimento destes dados. Nesse sentido, a possibilidade de aplicar técnicas e ferramentas que convertam, de forma automática e prática, dados em

informações úteis, tem se tornado estratégica. É desse contexto que emergem recursos avançados de computação que culminam na estruturação de toda uma área de pesquisa: a ciência de dados, essencialmente pautada em descobrir e expor conhecimento em grandes volumes de dados, usando para isso uma vasta lista de métodos, algoritmos, ferramentas e perfis profissionais.

Este artigo aplica o algoritmo LOF , baseado no conceito de densidade local através dos *k* vizinhos mais próximos, onde a distância é usada para avaliar a densidade sobre bases de dados públicos relativos à manutenção de frotas e mostra como certos padrões de anomalia podem ser detectados e associados a padrões ilegais em base de dados. Para isso, será usado a base de dados de manutenção de frotas de uma Prefeitura Municipal, o método KDD e a ferramenta RapidMiner. [1]

1.1 Problema de Pesquisa

O uso de computadores em benefício do crime ocorre há bastante tempo, no entanto, a legislação brasileira não apresenta algo concreto em termos de tipificação de crimes cibernéticos. Para (ELEUTÉRIO; MACHADO, 2011) os equipamentos computacionais são utilizados como ferramenta de apoio aos crimes convencionais ou como meio para realização do crime. Na primeira modalidade, o computador é apenas um simples coadjuvante no cometimento do crime enquanto que na segunda é a peça principal para a existência do delito.

Deste modo, é imprescindível identificar se uma transação é passível de ser ou não fraudulenta. Em parte, detectar este tipo de situação torna-se improvável tendo em vista que transações consideradas fraudulentas em muito se assemelham a transações normais e pelo fato de que a descoberta de uma fraude deve ocorrer de maneira tempestiva.

Isto posto, uma questão teórica surge como indagação: técnicas de mineração de dados podem ser úteis na descoberta de fraudes em sistemas de administração de frotas em bases de dados públicos municipais?

1.2 Objetivos

Os principais objetivos do trabalho são descritos na sequência.

1.2.1 Objetivo Geral

Apresentar técnicas de data mining (mineração de dados) para auxiliar na detecção de outliers em sistemas de administração de frotas geridos por uma base de dados públicos municipal.

1.2.2 Objetivos Específicos

- Entender o método de descoberta de conhecimento em base de dados (KDD);

[1] https://rapidminer.com/

- Empregar uma ferramenta para uso de algoritmos de data mining;
- Assimilar e empregar as técnicas essenciais de data mining; e
- Construir um modelo para analisar uma base de dados na investigação de tendências e padrões.

1.3 Justificativa

O aumento do ataque de fraudadores é notório. Tirando vantagens de vulnerabilidades em sistemas, realizam os mais diversos tipos de golpes, com o propósito de enganar ou iludir quem quer que seja. Observa-se, na literatura, um crescimento expressivo de ferramentas direcionadas para o combate a fraudes, no entanto, processos voltados para análise de dados municipais ainda são de números reduzidos, possivelmente pelo motivo de que tais dados são guardados localmente, muitas vezes de difícil acesso ao cidadão comum, permanecendo alcançável apenas pequenas amostras que na maioria das vezes não refletem a realidade. Desta forma, se justifica o presente trabalho a fim de analisar dados municipais e incluir na literatura uma referência a mais.

1.4 Materiais e Métodos

No presente trabalho efetuou-se uma análise qualitativa e exploratória, através da revisão bibliográfica e experimentos, utilizando como base de consulta teórica livros sobre mineração de dados e artigos acadêmicos disponíveis on-line, associando e confrontando os assuntos obtidos nas fontes de consulta com os resultados consolidados nos registros armazenados em um sistema de controle e administração de frotas de uma Prefeitura Municipal. Os dados foram tabulados em uma plataforma de software de ciência de dados denominado RapidMiner. O framework fornece um ambiente para preparar os dados e organizar a mineração de dados a fim de encontrar dados anormais no conjunto de dados e as possíveis causas do desvio.

1.5 Organização do Trabalho

O trabalho é organizado da seguinte forma. O Capítulo 2, Revisão da Literatura, onde são apresentados os conceitos referentes ao processo de encontrar as respostas para os objetivos que se propõe o trabalho. Desta maneira, o capítulo discorre sobre temas e trabalhos que possuem relação com a pesquisa, bem como a teoria sobre o algoritmo LOF e o seu emprego na detecção de anomalias. O Capítulo 3 descreve os materiais e métodos propostos, onde é descrito o pré-processamento de dados, técnicas e ferramentas utilizadas. O Capítulo 4 apresenta os resultados obtidos com os experimentos e questionamentos sobre as anomalias detectadas. O Capítulo 5 descreve a conclusão do trabalho, com suas limitações, proposta de

trabalhos futuros, dificuldades que foram encontradas durante o processo de pesquisa e as considerações finais, que destacam as conclusões do pesquisador.

2 REVISÃO DE LITERATURA

De acordo com (REPÚBLICA, 2016), que Institui a Política de Dados Abertos do Poder Executivo federal, dados abertos são aqueles acessíveis ao público, em meio digital, processáveis por máquina, referenciados na internet e disponibilizados sob licença que permita sua livre utilização, consumo ou cruzamento, limitando-se a creditar a autoria ou a fonte. Análises sobre dados abertos não costumam ser prioritárias pela gestão pública, mas tornam-se exequíveis a qualquer indivíduo, viabilizando, por exemplo, a construção de ferramentas de reconhecimento de fraudes.

Iniciativas nessa direção incluem a "Operação Serenata de Amor" (BRASIL, 2021a), um projeto de inteligência artificial que usa a ciência de dados para fiscalizar gastos públicos e compartilhar estas informações com os cidadãos. Nessa vertente, a Serenata de Amor originou a "Rosie" (BRASIL, 2021b), uma máquina artificialmente inteligente que analisa gastos que são reembolsados por deputados federais e senadores durante o exercício de sua função, pela Cota para Exercício da Atividade Parlamentar, identificando transações suspeitas e estimulando a população a contestá-las.

De maneira díspar aos projetos Serenata de Amor e Rosie, que focam na análise dos gastos públicos de parlamentares, existem outras abordagens que colaboram na identificação e combate às irregularidades já no início do processo. Por exemplo, os Robôs "Mônica" (Monitoramento Integrado para o Controle de Aquisições), "Sofia" (Sistema de Orientação sobre Fatos e Indícios para o Auditor) e "Alice" (Análise de Licitações e Editais), aplicam inteligência artificial em dados do Tribunal de Contas da União para identificar e combater irregularidades em licitações.

Vale considerar que, embora essas soluções sejam adotadas como forma de evitar o desvio de recursos e dano ao erário, a operação Serenata de amor tem um caráter mais voltado para a moralidade e honestidade, enquanto que as tecnologias empregadas pelo Tribunal de Contas da União são mais direcionadas ao controle de operações.

Ainda outros trabalhos aplicam mineração de dados para identificar anomalias semânticas, o que se alinha aos propósitos deste artigo, embora não sejam específicos para o domínio de aplicação aqui considerado.

A partir de estudos de aprendizagem de máquina com algoritmos supervisionados, (LUBAMBO, 2008) apresenta um modelo capaz de classificar empresas suspeitas de operarem exportações fictícias, fundamentada na metodologia Cross-Industry Standard Process for Data Mining (CRISP-DM), realizando experimentos através de um estudo de caso, baseados nos algoritmos de indução de regras APRIORI e TERTIUS, com dados reais da Secretaria da Fazenda de Pernambuco (SEFAZ-PE).

Já (KINTOPP, 2017) identifica possíveis anomalias nos dados disponíveis no portal da transparência. Para isso, este trabalho aplica a metodologia de Descoberta de Conhecimentos

em Bases de Dados (Knowledge Discovery in Databases – KDD) por meio do algoritmo LOF com o objetivo de atribuir uma pontuação de anormalidade para cada instância da base, utilizando técnicas baseadas nos k-vizinhos mais próximos.

Em direção similar, mas se utilizando de outras técnicas e outras bases de dados públicos, (LOPES, 2019) visa classificar gastos públicos para a detecção de possíveis fraudes. Para tal, o trabalho compara a performance das técnicas de regressão logística, árvore de decisão, random forest, gradient boosting e lightGBM.

Já o trabalho de (ASSUNÇÃO et al., 2016) descreve o INFOSAS, um sistema interativo de detecção de anomalias no sistema de pagamento aos prestadores de serviços ao Sistema Único de Saúde (SUS) para posterior auditoria. O sistema procura detectar dois tipos de discrepâncias: um valor médio excessivo cobrado por procedimentos dentro de um alvo e um número excessivo na produção de um procedimento por parte de um estabelecimento. Um dos resultados mais importantes do INFOSAS é a sua capacidade de análise do volume total de produção de serviços de saúde em busca de anomalias, considerando cada prestador de serviço.

Constata-se, na literatura, um movimento significativo na ampliação das ferramentas voltadas ao combate a fraudes; contudo, aplicações em dados municipais são ainda incipientes, principalmente pelo fato de que esse domínio não possui o mesmo grau de transparência em comparação a dados federais ou mesmo estaduais. Fato é que dados municipais são, em geral, armazenados localmente, de difícil acesso ao cidadão, e apenas amostragens genéricas (em geral obrigatórias) e de difícil compreensão comum, são expostas. É dessa lacuna que emerge o objetivo deste trabalho, o qual envolve os conceitos descritos na seção seguinte.

De acordo com (TAN; STEINBACH; KUMAR, 2009) a extração de informação útil tem provado ser extremamente desafiadora, pois muitas vezes, ferramentas e técnicas tradicionais de análise de dados não podem ser utilizadas devido ao volume do conjunto dos dados ser muito grande. Outro óbice diz respeito ao árduo trabalho no tratamento inicial dos dados, ou pré-processamento, que incluem a junção de dados de várias fontes, limpeza dos dados com o objetivo de remover ruídos e seleção de registros que sejam importantes para a tarefa de mineração de dados.

A detecção de anomalias, ilustrada na Figura 1, é um processo que consiste em distinguir padrões em objetos que diferem do comportamento esperado, isto é, que se diferenciam de outros objetos integrantes de uma mesma amostra. Logo os objetos $o1$, $o2$ e $o3$ podem ser descritos com valores destoantes, os quais se encontram distantes dos valores considerados normais $c1$ e $c2$.

Segundo (CASTRO; FERRARI, 2016), em geral há diversas abordagens para detecção de anomalias, dependendo da área em que será aplicada e de certos aspectos que devem ser ponderados, porém, o fator que mais influencia uma abordagem de detecção de anomalias está relacionado à disponibilidade e ao uso dos rótulos de dados.

Desta forma, devemos compreender que, sendo a detecção de anomalias parte integrante do aprendizado de máquina, existem duas grandes divisões de formas pela qual o

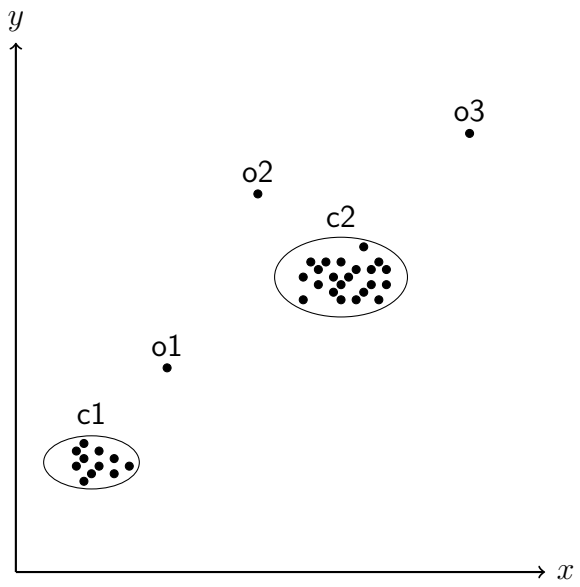

Figura 1 – Exemplo de anomalias

aprendizado pode ocorrer: supervisionado e não supervisionado.

O aprendizado supervisionado assume que objetos normais e anomalias são modelados, ou seja, há o emprego de um modelo de dados. Tal técnica necessita de um conjunto de treinamento com os objetos anômalos e normais. Para (STEVEN, 2017) o aprendizado supervisionado é o paradigma básico para problemas de classificação e regressão enquanto o não supervisionado procura encontrar estrutura nos dados, fornecendo rótulos ou valores para dar sentido a um conjunto de dados.

Diferentemente, no aprendizado não supervisionado, usado nesse trabalho, não existe nenhuma proposição sobre o rótulo de dados e as anomalias são identificadas sem o conhecimento prévio das classes ditas normais e anômalas. Desta forma, os dados mais distantes serão reconhecidos como possíveis anomalias. O aprendizado não supervisionado adquire conhecimento através de dados de teste que não foram rotulados, classificados ou categorizados antecipadamente. Tal técnica reconhece similaridades nos dados e reage com base na presença ou ausência das semelhanças em cada novo dado.

Um exemplo ocorre ao estabelecer um mercado para um produto totalmente novo, ainda não comercializado. Há algumas informações, mas não todas. Desta forma, um algoritmo apresenta alguns insights sobre o possível desempenho do novo produto através do aprendizado não supervisionado.

2.0.0.1 Algoritmo Local Outlier Factor - LOF

Os algoritmos baseados em distância local usam as distâncias existentes entre objetos do conjunto de dados que está sob análise, à procura de desvios entre determinado objeto e seus vizinhos.

Composto por diversas fases, o algoritmo LOF necessita de um conjunto de dados de

entrada, ou seja, dados para efetuar a transação. Os dados de entrada passam por uma etapa de pré-processamento, a fim de que se verifique a sua adequação para o propósito de detecção de anomalia, o que em geral é associado à precisão dos resultados após processamento.

O primeiro estágio do pré-processamento, denominado de limpeza de dados, serve para eliminar dados ditos sujos ou desnecessários; o segundo estágio implementa a transformação dos dados, ou seja, modificações que permitam processar cálculos numéricos; por fim, o terceiro estágio envolve a padronização dos dados para escalas normalizadas, por exemplo, quando um atributo possui valor que destoa de outros e deturpa o cálculo. O *Max-Min* é um dos métodos mais comuns de normalização e é utilizado neste trabalho. Sua aplicação consiste em transformar atributos em um mesmo intervalo de valores, por exemplo, [0, 1]. A equação do *Max-Min* é mostrada na Eq. (1).

$$a' = \frac{a - min_a}{max_a - min_a}(new\_max_a - new\_min_a) + new\_min_a \tag{1}$$

A partir de sua entrada, o algoritmo LOF depende da pesquisa de vizinhos mais próximos. O método realiza uma marcação em cada ponto de dados, efetuando um cálculo da proporção das densidades médias dos vizinhos do ponto com a densidade do próprio ponto. A densidade estimada de um ponto p é o número de vizinhos de p dividido pela soma das distâncias até os seus vizinhos, como mostra a Eq. (2).

$$f(p) = \frac{k}{\sum_{x \in N(p)^{d(p,x)}}} \tag{2}$$

em que:

$N(p)$ o conjunto de vizinhos do ponto p;

k é o número de pontos desse conjunto;

$d(p,x)$ é a distância entre os pontos p e x.

O LOF de um objeto é calculado da seguinte forma: havendo uma quantidade k de vizinhos mais próximos, utilizar a Eq. (3) para encontrar o LOF de cada objeto da base. Serão consideradas anomalias todos os objetos $x_i = 1, \ldots N$, cujo $\mathrm{LOF}_k(x_i) >> 1$.

$$\mathrm{LOF}_k(x_i) = \frac{1}{lrd(x_i)} \cdot \frac{\sum_{x_j \in N_{k^{(x_i)}}{}^{lrd(x_j)}}}{k} \tag{3}$$

A Figura 2 facilita entender como o algoritmo LOF se comporta. O ponto A tem um escore LOF elevado devido a sua densidade ser inferior em relação às densidades dos vizinhos. Os círculos pontilhados mostram a distância até o terceiro vizinho mais próximo de cada ponto. (BREUNIG et al., 2000) explica que o valor de LOF é aproximadamente 1 para objetos que se encontram dentro de um grupamento (cluster). Para os outros objetos são apresentados limites superiores e inferiores, os quais são interpretados através de uma heurística de classificação de objetos por seu maior valor de LOF no intervalo selecionado.

Através do algoritmo LOF é possível reconhecer anomalias em um conjunto de dados que não se enquadrariam como anomalias em outra região do conjunto de dados, devido à

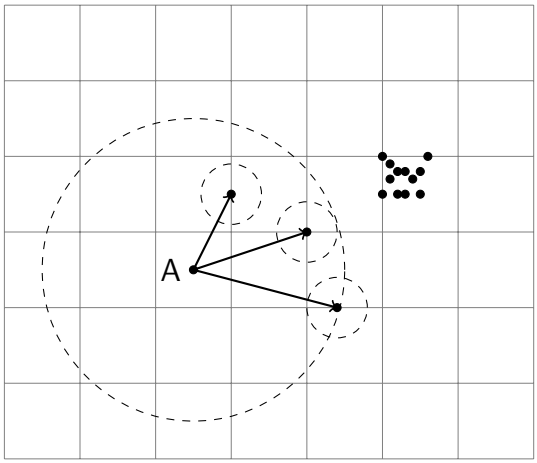

Figura 2 – Ilustração do algoritmo LOF

abordagem local. Um objeto a uma curta distância de um cluster com grande densidade é considerado anomalia, ao passo que outro objeto no interior de um cluster disperso pode apresentar distâncias parecidas com seus vizinhos e por consequência não ser caracterizado como anomalia.

3 MATERIAIS E MÉTODOS

Este trabalho analisou uma parcela do domínio em que a detecção de anomalias se enquadra. O conjunto de dados extraídos contém 22447 lançamentos de registros reais relacionados à frota de uma Prefeitura Municipal, cuja identificação é preservada por questões de exposição e eventual conflito com os propósitos deste trabalho, que é revelar possíveis anomalias.

3.1 Materiais

Para a etapa da mineração de dados foi empregado o software RapidMiner. No experimento, calculamos os outliers locais com LOF > 1.5.

A entidade $Item$ representa as características dos veículos e possui quinze atributos. Tendo em vista que a capacidade do tanque de combustível não se faz presente na entidade, tal informação, de valor fundamental para análise dos dados após a aplicação do algoritmo LOF , foi incluída na base de dados tendo como parâmetro os valores existentes nas fichas técnicas de cada veículo. Essas informações foram colhidas nos sites do fabricante.

Na Tabela 1, podemos verificar a quantidade de combustível consumida pelos veículos e o valor pago pela Prefeitura no período de 22 de outubro de 2010 a 30 de julho de 2016.

Combustível	**Qtd veículos**	**Consumo (lts)**	**Valor**
Óleo diesel	35	855.370,02	R$ 1.824.943,61
Gasolina	56	73.039,64	R$ 198.947,76

Tabela 1 – Quantidade de combustível e valor pago.

3.2 Métodos

No banco de dados em questão, tomou-se como base a entidade $MovimentoVeiculo$, que armazena registros de consumo de combustíveis e lubrificantes, serviços a executar e troca de pneus. Sobre essa tabela foi executada uma filtragem a fim de retornar um subconjunto T de atributos sobre os quais se tem interesse em determinar possíveis anomalias. Dessa forma, definiu-se

$\mathrm{T} = \langle cdVei, cdTipoMov, dtEv, hrEv, vlVal, qtMov, dsHis \rangle$.

Nesse esquema resultante, $cdVei$ é o código do veículo, $cdTipoMov$ se refere ao tipo de combustível, $dtEv$ representa a hora em que ocorreu o abastecimento, $hrEv$ representa o horário em que ocorreu o abastecimento, $vlVal$ é o valor pago pelo combustível, $qtMov$ se refere à quantidade abastecida de combustível e $dsHis$ representa a descrição da quantidade e tipo de combustível.

Na etapa de preparação da base de dados criou-se uma rotina computacional para reduzir a dimensão e selecionar apenas os dados que contém instâncias de abastecimentos de gasolina (2640 instâncias) e óleo diesel (8307 instâncias). Objetos com dados ausentes se mostraram presentes no atributo $vlVal$ (referente ao valor pago pelo abastecimento) em 53 instâncias de abastecimento com gasolina e em 94 de abastecimento com óleo diesel.

Observou-se que os dados ausentes se concentravam no período de 18 de novembro de 2010 a 1º de dezembro de 2010. Adotou-se a imputação de valores do tipo **hot-deck**, onde aos valores ausentes foram imputados o valor do mesmo atributo de um objeto similar selecionado, neste caso, o valor do preço do combustível à época.

A frota cadastrada na base de dados, 56 veículos com o combustível gasolina e 35 com óleo diesel teve seu domínio discretizado em 18 grupos conforme a capacidade do tanque a fim de permitir uma melhor análise dos dados. Esse procedimento foi realizado durante a seleção dos atributos para a análise.

A limpeza dos dados, com o intuito de imputar valores ausentes, reduzir a dimensão da base de dados, bem como padronizar e deixar os objetos em um formato compatível com o processo de análise foi empregada em especial nos atributos $dtEv$ que se refere a data em que ocorreu o abastecimento e $dsHis$ referente ao volume de combustível abastecido. Aplicou-se a técnica de normalização *Max-Min* referenciada na Eq. (1) com um intervalo de valores [0,1] no atributo $qtMov$. A distância selecionada no algoritmo foi a euclidiana.

4 RESULTADOS

Os resultados dos experimentos mostraram que a aplicação do algoritmo LOF para identificar anomalias como possíveis indícios de fraudes apresentou objetos Falsos Positivos, em virtude da variação entre o menor e o maior número de volume de combustível. Observamos que o combustível óleo diesel ficou disposto entre 5 e 800 litros, conforme a Figura 3, ao passo que o combustível gasolina ficou disposto entre 5 e 65 litros (Figura 4). Para uma melhor visualização dos gráficos, o volume de combustível foi dividido em múltiplos de 5.

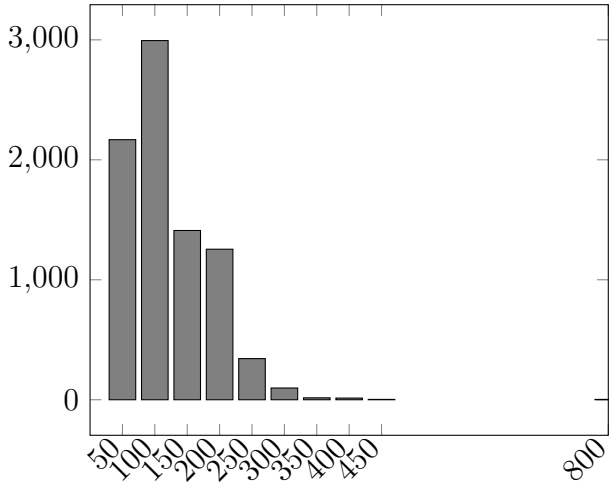

Figura 3 – Gráfico de abastecimentos com óleo diesel

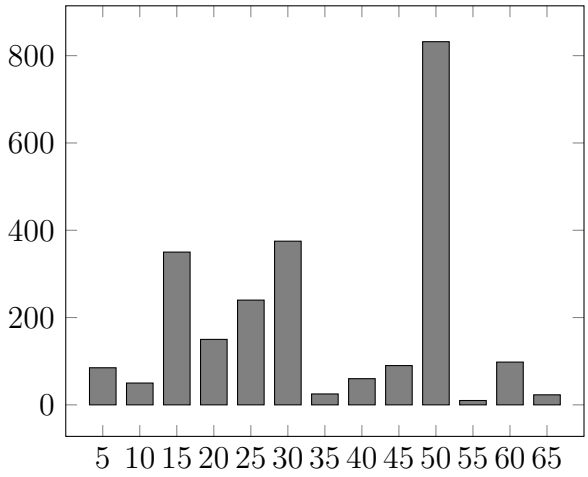

Figura 4 – Gráfico de abastecimentos com gasolina

Entretanto o algoritmo reconheceu instâncias Verdadeiros Positivos, demonstrando a existência de registros de abastecimentos acima do limite do tanque de combustível.

Ao separarmos as instâncias por veículos de maneira mais específica, isto é, que possuem um certo padrão na distribuição dos valores no abastecimento, o algoritmo LOF

distinguiu os valores anormais. Para o veículo identificado como Cod74, abastecido com óleo diesel, do total de 21 abastecimentos o algoritmo detectou 13 instâncias caracterizadas como outliers, sendo 10 Verdadeiros Positivos (VP) e 3 Falsos Positivos (FP), com uma acurácia de 85% e precisão de 77%.

Se verificarmos que, para este veículo, a capacidade máxima de combustível no tanque é 275 litros, os outliers identificados como *VP* caracterizam, em tese, uma provável fraude já que é impossível inserir um volume de combustível além da capacidade de um reservatório.

A Tabela 2 demonstra o total de outliers identificados, bem como as instâncias consideradas Falsos Positivos.

Abastecimentos	**Combustível (lt)**	**Outlier**	**Status**
2	80	2.559	normal
1	250	2.002	normal
6	400	6.831	outlier
1	600	2.344	outlier
3	800	3.516	outlier

Tabela 2 – Anomalias registradas para o veículo Cod74.

Para o veículo identificado como Cod04, abastecido com gasolina, em que o tanque permite no máximo 50 litros, os maiores outliers representaram 6 abastecimentos.

A Tabela 3 representa os maiores outliers identificados para o veículo Cod4.

Abastecimentos	**Combustível (lt)**	**Outlier**	**Status**
1	60	10.174	outlier
1	58	8.570	outlier
2	3	6.563	outlier
1	7	4.723	outlier
1	6	3.979	outlier

Tabela 3 – Anomalias registradas para o veículo Cod04.

Da mesma maneira, foram detectadas anomalias para pequenos abastecimentos, principalmente em veículos pesados, os quais possuem um reservatório de combustível de grande capacidade, tendo em vista as horas de trabalho do equipamento e o grande consumo de combustível. De alguma forma, foge aos padrões estes abastecimentos porque equipamentos deste tipo habitualmente necessitam de uma grande quantidade de combustível para operação. A Tabela 4 apresenta estes dados.

As Figuras 5 e 6 representam as anomalias detectadas pelo algoritmo LOF na base de dados para os veículos abastecidos com óleo diesel e gasolina. O eixo x caracteriza a quantidade de combustível e o eixo y as anomalias. Observamos que as maiores anomalias estão localizadas precisamente na área do gráfico que marca as maiores quantidades de combustível, algumas delas acima da capacidade do tanque.

Código	Abastecimentos	Combustível (lt)	Outlier	Tanque (lt)
8	1	10	15.703	160
10	1	10	6.980	160
11	1	10	4.919	340
17	2	10	3.648	150
39	1	10	4.585	105
46	1	10	3.604	103
47	1	10	4.424	103
63	1	5	4.593	380
73	1	10	4.909	160

Tabela 4 – Anomalias registradas para pequenos abastecimentos.

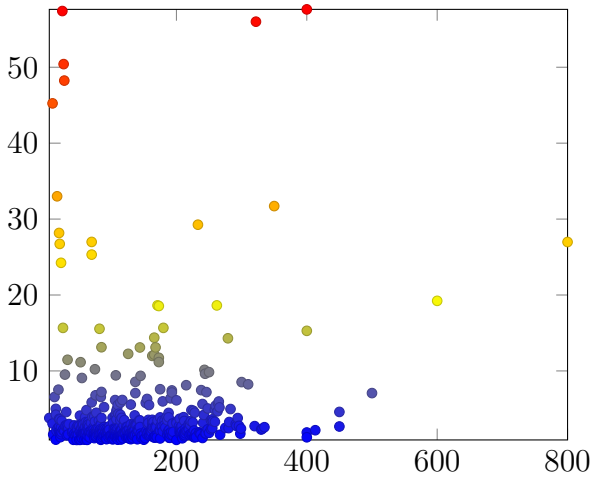

Figura 5 – Anomalias registradas para veículos abastecidos com diesel

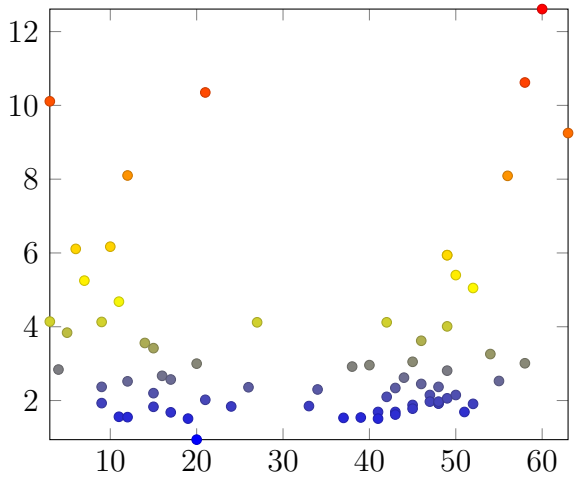

Figura 6 – Anomalias registradas para veículos abastecidos com gasolina

5 CONCLUSÃO

Concluímos que, de 10947 registros de abastecimentos, 117 foram identificados como Verdadeiros Positivos, sendo 1,07% do total, com uma média de 19 abastecimentos acima da capacidade do veículo a cada ano, no período de 22 de outubro de 2010 a 30 de julho de 2016.

5.1 Limitações

Reconhecer e restringir fraudes passou a ser uma questão essencial, principalmente na administração municipal, e sob este aspecto, no presente trabalho alguns pontos tornaram a pesquisa mais árdua. A base de dados utilizada, composta de 127 entidades, exigiu um grande esforço no pré-processamento dos dados. Embora o núcleo do experimento esteja direcionado sobretudo na entidade $MovimentoVeiculo$, representou um desafio ao processo, pela existência de dados faltantes.

Devido à ausência de um atributo para armazenar o marcador do odômetro e horímetro dos veículos na base de dados, não foi possível estabelecer um parâmetro para determinar o consumo médio de combustível.

5.2 Trabalhos Futuros

Devido à escolha de apenas um algoritmo para detecção de anomalias na base de dados, não se aplicou outros cenários com algoritmos semelhantes. Dessa maneira, como proposta de trabalhos futuros, sugere-se a perspectiva de análise com outros algoritmos, como por exemplo o kNN.

5.3 Considerações Finais

Não podemos afirmar que houve fraude nem qualquer tipo de irregularidade nos abastecimentos. É factível concluir que o algoritmo empregado indica que alguns veículos apresentaram quantidade de combustível abastecida além da capacidade do reservatório e consequentemente estão em desarmonia em relação aos demais.

Igualmente deve ser observado que, em se tratando de máquinas e equipamentos, é usual conduzir recipientes com combustível para situações de trabalho em áreas distantes de recursos de reabastecimento.

Referências

ASSUNÇÃO, R. M. et al. Detecção de anomalias nos pagamentos do sistema único de saúde. **J. health inform**, p. 459–468, 2016. Citado na página 17.

BRASIL, O. K. **Operação Serenata de Amor**. 2021. Disponível em: <https://serenata.ai/>. Citado na página 16.

BRASIL, O. K. **Rosie da Serenata**. 2021. Disponível em: <https://twitter.com/RosieDaSerenata/>. Citado na página 16.

BREUNIG, M. M. et al. Lof: identifying density-based local outliers. In: **Proceedings of the 2000 ACM SIGMOD international conference on Management of data**. [S.l.: s.n.], 2000. p. 93–104. Citado na página 19.

CASTRO, L. N. de; FERRARI, D. G. **Introcução à mineração de dados**. 1. ed. São Paulo: Editora Saraiva, 2016. Citado 2 vezes nas páginas 12 e 17.

ELEUTÉRIO, P. M. da S.; MACHADO, M. P. **Desvendando a computação forense**. 6. ed. São Paulo: Novatec, 2011. Citado 2 vezes nas páginas 12 e 13.

INTERNATIONAL, T. **Índice de percepção da corrupção 2020**. 2020. Disponível em: <https://cutt.ly/WQUSLbW>. Acesso em: 9 de agosto de 2021. Citado na página 12.

KINTOPP, P. M. **Aplicação de técnicas de aprendizado de máquina em dados públicos para detecção de anomalias**. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2017. Citado na página 16.

LOPES, M. A. **Aplicação de aprendizado de máquina na detecção de fraudes públicas**. Dissertação (B.S. thesis) — Universidade de São Paulo, 2019. Citado na página 17.

LUBAMBO, S. W. **Processo de Mineração de Dados como Apoio à Decisão no Controle de Gastos Públicos**. Dissertação (B.S. thesis) — Universidade Federal de Pernambuco, 2008. Citado na página 16.

REPÚBLICA, P. da. **Decreto nº 8777 de 11 de maio de 2016**. 2016. Disponível em: <https://cutt.ly/MvG9rOD>. Citado na página 16.

STEVEN, S. S. **Data Science Design Manual**. [S.l.]: Springer, 2017. Citado na página 18.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introcução ao Data Mining**. Rio de Janeiro: Ciência Moderna, 2009. Citado na página 17.

WARDE, W. **O espetáculo da corrupção: como um sistema corrupto e o modo de combatê-lo estão destruindo o país**. 1. ed. Rio de Janeiro: Leya, 2018. Citado na página 12.