UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

PABLO MEZZON KINTOPP

# APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA EM DADOS PÚBLICOS PARA DETECÇÃO DE ANOMALIAS

TRABALHO DE CONCLUSÃO DE CURSO

**MEDIANEIRA**

**2017**

PABLO MEZZON KINTOPP

# APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA EM DADOS PÚBLICOS PARA DETECÇÃO DE ANOMALIAS

Trabalho de Conclusão de Curso apresentado ao Departamento Acadêmico de Computação da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do título de "Bacharel em Ciência da Computação".

Orientador: Prof. Dr. Arnaldo Candido Junior

**MEDIANEIRA**

**2017**

Ministério da Educação
**Universidade Tecnológica Federal do Paraná**
Diretoria de Graduação e Educação Profissional
Coordenação do Curso de Ciência da Computação

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

---

**TERMO DE APROVAÇÃO**

# APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA EM DADOS PÚBLICOS PARA DETECÇÃO DE ANOMALIAS

Por

**PABLO MEZZON KINTOPP**

Este Trabalho de Conclusão de Curso foi apresentado às 15:50h do dia 13 de Junho de 2017 como requisito parcial para a obtenção do título de Bacharel no Curso de Ciência da Computação, da Universidade Tecnológica Federal do Paraná, Câmpus Medianeira. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

---

Prof. Dr. Arnaldo Candido Junior
UTFPR - Campus Medianeira

---

Prof. Dr. Paulo Lopes de Menezes
UTFPR - Campus Medianeira

---

Prof. Msc. Jorge Aikes Junior
UTFPR - Campus Medianeira

A folha de aprovação assinada encontra-se na Coordenação do Curso.

# RESUMO

KINTOPP, Pablo Mezzon. APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA EM DADOS PÚBLICOS PARA DETECÇÃO DE ANOMALIAS. 56 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2017.

O aprendizado de máquina (AM) tem sido aplicado cada vez mais no dia-a-dia da sociedade e das organizações. Com isso, muitos avanços em soluções e inovações têm surgido nos mais diversos domínios de aplicação. O objetivo deste trabalho é demonstrar a aplicação de uma técnica do AM com ênfase em detecção de anomalias em uma área de grande interesse para os cidadãos brasileiros, que é o uso do dinheiro público. Este trabalho visa encontrar anomalias nas informações de gasto divulgadas pelas prefeituras brasileiras, podendo assim indicar casos suspeitos de improbidade administrativa. Logo, para aplicar tais técnicas foram seguidas as etapas básicas do processo de extração de conhecimento, sendo estas: coleta dos dados, pré-processamento, aplicação e a análise dos resultados. Por fim, os resultados mostraram a identificação de várias anomalias. Sendo parte delas relativas a prefeituras que tiveram gastos que podem instigar o interesse dos cidadãos em averiguá-los. Assim, foi concluído que a utilização de AM pode ser útil no auxílio da fiscalização do uso do dinheiro público.

**Palavras-chave:** mineração de dados, inteligência artificial, aprendizado do computador

# ABSTRACT

KINTOPP, Pablo Mezzon. APPLICATION OF MACHINE LEARNING TECHNIQUES ON PUBLIC DATA FOR ANOMALY DETECTION. 56 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2017.

Machine Learning (ML) has been increasingly applied in the daily lives of society and organizations. Therefore, many advances in solutions and innovations have arisen among diverse application fields. The objective of this work is to demonstrate the application of a technique from ML focused on anomaly detection in an area of great interest to Brazilian citizens, which is the public funds usage. This study aims to find anomalies in the data of expenditure divulged by the Brazilian municipalities in order to indicate suspicious cases of administrative improbity. To apply such techniques, the following steps of the knowledge extraction process were used: data collection, pre-processing, application and result analysis. The results showed the identification of several anomalies. Some of them are related to cities that had the expenses which can instigate the interest of the citizens to investigate them. Hence, it was concluded that the use of ML could be useful in assisting the supervision of the public money usage.

**Keywords:** data mining, artificial intelligence, computational learning

## AGRADECIMENTOS

Agradeço a minha família, especialmente a minha mãe, minha irmã e meu pai pelo apoio, incentivo e confiança incondicional que sempre me forneceram, não apenas durante a graduação, mas durante toda a vida.

Agradeço ao professor Arnaldo, pela orientação, incentivo e sugestões que foram de suma importância para a construção e finalização deste trabalho.

Agradeço também aos professores e colegas de curso que de certa forma contribuíram em minha graduação e no término deste trabalho.

## LISTA DE FIGURAS

## LISTA DE SIGLAS

| | |
|---|---|
| AM | aprendizado de máquina |
| CBLOF | Cluster-Based Local Outlier Factor |
| CGU | Controladoria-Geral da União |
| COP | Correlation Outlier Probability |
| CSV | Comma-separated values |
| ELKI | Environment for Developing KDD-Aplications Supported by Index-Structures |
| EM | Expectation Maximization |
| ENEM | Exame Nacional do Ensino Médio |
| IDH | Índice de Desenvolvimento Humano |
| IFGF | Índice FIRJAN de Gestão Fiscal |
| KDD | Knowledge Discovery in Databases |
| LOF | Local outlier factor |
| MLE | Maximum Likelihood Estimates |
| OPTICSOF | Ordering Points To Identify the Clustering Structure with Outlier Factors |
| PCA | Principal Component Analysis |
| RBF | radial basis function |
| SICONFI | Sistema de Informações Contábeis e Fiscais |
| SOM | Self-Organizing Maps |
| SVM | Support Vector Machine |
| TCE | Tribunal de Contas do Estado |
| TIC | Tecnologias de Informação e Comunicação |

**SUMÁRIO**

# 1 INTRODUÇÃO

Aproximadamente do ano de 2014 até 2017, o Brasil vem atravessando uma crise política, onde a corrupção e a má gestão do dinheiro público são fatores que contribuem diretamente para o agravamento dessa situação.

De acordo com Transparency International (2016), para o ano de 2015, o Brasil ocupava a 76ª posição no quesito de menor nível de corrupção percebida no setor público entre os 168 países e territórios analisados. Apesar de o Brasil estar melhor classificado que pelo menos 80 países ele ainda está, segundo o estudo, entre os países em que mais houve aumento no nível de corrupção se comparado aos resultados de 2012, onde o Brasil ocupava o 69º lugar dentre 175 países que foram analisados naquele estudo (Transparency International, 2013). Este declínio pode ser relacionado aos escândalos de corrupção ocorridos na empresa de capital aberto Petrobrás. Além disso, gastos irregulares e atos de improbidade administrativa tem sido frequentemente relatados em portais WEB de órgãos governamentais tais como o portal da Controladoria-Geral da União (CGU) e dos Tribunais de Contas dos Estados (TCE), que são responsáveis por fiscalizar e denunciar o mau uso do dinheiro público.

Além disso, não somente a União e Estados tem tido problema, mas a situação das contas públicas municipais também não está bem. Segundo FIRJAN (2016), em 2015, 87% dos municípios brasileiros estavam em situação de gestão fiscal considerada crítica ou difícil. De fato, como mostra a Figura 1, a pontuação que determina a média da qualidade da gestão fiscal dos municípios brasileiros alcançou o pior resultado em 2015 quando comparado aos 10 anos anteriores.

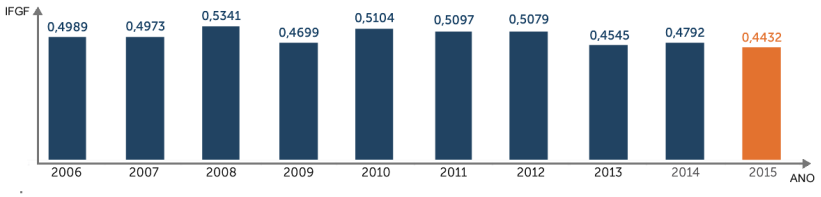



**Figura 1 – Evolução qualidade da gestão fiscal municipal em 10 anos**

**Fonte: Adaptado de (FIRJAN, 2016)**

Entretanto, a corrupção e o mau uso ou gestão do dinheiro público não são problemas exclusivos do Brasil, ou de regiões pobres. Trata-se de um problema mundialmente conhecido que somente tem sido tratado de melhor maneira ultimamente, devido ao maior acesso a informação proporcionado pelas Tecnologias de Informação e Comunicação (TIC), onde os cidadãos passaram a possuir um maior controle sobre meios para examinar os recursos públicos (SILVA; FLACH, 2013).

No Brasil, graças a Lei da transparência e a uma iniciativa da CGU, desde 2004 está acessível o Portal da Transparência[1], que reúne informações a respeito dos recursos públicos e suas destinações. O portal permite ao cidadão um acompanhamento integral de como o dinheiro público está sendo utilizado. A importância de tal ferramenta para os cidadãos torna-se cada vez mais evidente, e como afirma Carvalho (2001), quanto maior for a transparência das ações governamentais, que instiguem a informação na sociedade sobre tais ações, maior será o interesse desta população em participar e lutar por políticas mais justas.

No portal da transparência é possível fazer pesquisas para encontrar dados detalhados sobre a execução orçamentária e financeira do governo federal além de poder acompanhar os recursos públicos transferidos para o exterior, estados, municípios, instituições privadas, cidadãos e os gastos diretos do Poder Executivo Federal. Todas estas consultas e muitas outras informações podem ser acessadas e visualizadas no próprio portal. Entretanto, com o objetivo de facilitar as consultas e oferecer uma forma prática e rápida para o usuário armazenar os dados, o portal disponibiliza uma seção de Downloads onde é possível baixar alguns dados referentes aos anos anteriores a 2011 e dados completos relativos à 2011 até 2016 em um formato de arquivo apropriado.

De maneira similar ao portal da transparência, muitos outros órgãos governamentais, que também possuem portais na web, passaram a dar maior transparência ao uso do dinheiro público. Por exemplo, muitos *sites* relacionados aos tribunais de contas dos estados, disponibilizam dados do uso do dinheiro públicos em seus respectivos estados. Já no portal do Tesouro Nacional [2], é possível ter acesso aos dados dos exercícios fiscais, de cada ano, tanto das prefeituras, como dos estados e da própria união.

Com todos esses dados disponíveis, muitas opções de fiscalizar o uso do dinheiro público são dadas aos cidadãos. Entretanto, apesar de não parecer ser uma tarefa que requer muitas habilidades, analisar, compreender e fiscalizar esses dados não trata-se de algo tão trivial. O esforço e o tempo necessário para analisar uma quantidade razoável de informações, já são fatores comprometedores da motivação de qualquer cidadão. Portanto, essa tarefa não parece ser viável, sem o uso de alguma ferramenta ou método para auxiliar com tamanha quantidade

[1] http://www.portaltransparencia.gov.br/

[2] http://www.tesouro.fazenda.gov.br/

de dados.

Em computação existe uma subárea, pertencente ao ramo da inteligência artificial, que é conhecida por saber lidar muito bem com grande volume de dados, este campo de estudo é chamado de mineração de dados. Esse domínio é conhecido por utilizar técnicas de aprendizagem automática para descobrir padrões interessantes e extrair informações valiosas mediante uma grande quantidade de dados (RUSSELL et al., 2003). De fato, a utilização de técnicas de AM e mineração de dados têm auxiliado na solução de diversos tipos de problemas, assim como na melhoria e automatização de serviços pertencentes às mais diversas áreas. Alguns exemplos da utilização dessas técnicas são: previsão do tempo; tomada de decisões financeiras; processamento de imagens; geração de diagnósticos; marketing e vendas; processos sofisticados de fabricação e aplicações científicas nos mais diversos ramos (WITTEN et al., 2016).

Logo, utilizar essas técnicas para extrair informações interessantes dos dados referentes ao uso do dinheiro público pode ser de alguma utilidade. Com isso, este trabalho visa aplicar técnicas de aprendizagem automática com ênfase na detecção de anomalias sobre os dados disponibilizados em portais públicos tais como o portal da transparência ou o portal do tesouro nacional. Dessa maneira, informações possivelmente relevantes sobre a administração pública podem ser descobertas e expostas aos cidadãos.

Trabalhos similares já tem sido feitos, e resultados interessantes já tem aparecido. Por exemplo, a Operação Serenata de Amor[3], que vêm utilizando a inteligencia artificial para analisar os pedidos de reembolso dos deputados (BRIGADE, 2017). Isto é, por meio dos dados, disponibilizados no próprio portal da câmara dos deputados, referentes ao uso da cota parlamentar, a operação busca identificar casos com alta probabilidade de ilegalidade. Essa operação já trouxe a tona casos interessantes de reembolso, tal como: reembolso por bebida alcóolica em Las Vegas, reembolsos de 13 refeições feitas no mesmo dia e gastos mensais com gasolina de até R$ 6.000,00 reais, entre outros.

## 1.1 OBJETIVOS GERAL E ESPECÍFICOS

O objetivo geral deste trabalho é identificar, por meio da aplicação de técnicas da Inteligência Artificial nos dados disponíveis no portal da transparência, possíveis anomalias nos

[3] https://serenatadeamor.org/

gastos públicos. Este objetivo principal pode ser dividido nos seguintes objetivos específicos:

- Identificar os dados disponíveis nos portais relevantes para o contexto da aplicação;
- Criar um algoritmo para unificar os dados que são baixados separadamente, caso necessário;
- Fazer o pré-processamento dos dados visando adequar os atributos mais relevantes a análise;
- Selecionar um ou mais algoritmos de aprendizado de máquina que utilizam técnicas para detecção de anomalia;
- Aplicar os algoritmos selecionados nos dados por meio de uma ferramenta de extração de conhecimento;
- Exibir os resultados das técnicas aplicadas, destacando possíveis anomalias e mostrando informações relevantes.

## 1.2 JUSTIFICATIVA

Para averiguar que o dinheiro público não está sendo mal utilizado, uma boa solução é examinar todos os dados disponíveis em portais tais como o portal da transparência. Porém a quantidade de dados disponibilizada por esses portais é muito grande e em crescimento constante, o que acaba tornando inviável para uma ou mesmo várias pessoas analisarem tanta informação. Portanto, uma boa alternativa é utilizar a máquina para fazer isso. Os computadores podem processar grandes quantidades de dados e efetuarem cálculos complexos com velocidade muito mais rápida que os humanos. Além de que a máquina não requer pausas durante a análise dos dados, e tende a alcançar resultados muito mais precisos, com menos chances de falhar durante o processo da análise. Desta maneira, é possível executar uma verificação detalhada em dados disponíveis sobre os recursos públicos, o que possibilita destacar os possíveis casos considerados anormais ou fora de um padrão. E assim, este esforço visa ajudar a sociedade a utilizar os resultados desta mineração como guia para identificar possíveis casos de gastos irregulares, ou má gestão do dinheiro público, o que eventualmente poderá ajudar na diminuição e mitigação de tais práticas.

## 1.3 ORGANIZAÇÃO DO DOCUMENTO

Esse documento será organizado da seguinte forma: O Capítulo 2 apresenta inicialmente os principais conceitos do aprendizado de máquina. Em seguida, são apresentadas as principais técnicas para detecção de anomalias, com o intuito de apresentar as principais abordagens adotadas nesse campo de estudo. Os métodos e materiais utilizados se encontram no Capítulo 3, nele são descritas todas as etapas para o desenvolvimento do projeto. No Capítulo 4 é descrito um experimento preliminar com o objetivo de agregar informações sobre os métodos e materiais utilizados assim como aumentar o conhecimento sobre as ferramentas, técnicas e adversidades existentes nesse projeto. No capitulo 5 são exibidos e discutidos todos os resultados obtidos durante os experimentos executados. Por fim no capitulo 6 são descritas as considerações finais, por meio de uma conclusão e em seguida são sugeridas possíveis maneiras da continuação deste trabalho.

# 2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção será descrito o estado da arte do tema escolhido. Primeiramente será dada uma breve introdução sobre aprendizado de máquina e técnicas de detecção de anomalia. Em seguida serão apresentadas as abordagens mais utilizadas para detectar anomalias e, por fim, será descrito o funcionamento dos algoritmos escolhidos para serem aplicados nesse estudo.

## 2.1 APRENDIZADO DE MÁQUINA

Aprendizado de máquina (AM) é uma subárea de Inteligência Artificial que estuda e elabora algoritmos para aprender e fazer previsões de acordo com um conjunto de dados (KOHAVI; PROVOST, 1998). Onde, "um programa de computador é dito apreender com uma experiência *E* em relação a uma tarefa *T*, dado uma medida de desempenho *D*, somente se seu desempenho *D* em *T* é incrementado com a experiência *E*. " (MITCHELL, 1997). Isto é, a aprendizagem da máquina ocorre por meio de uma fase de treinamento para obter sua experiencia, e em seguida tem seu desempenho validado por uma fase de testes.

O AM visa responder muitas das mesmas questões levantadas na área de estatística ou no ramo da mineração de dados. Entretanto, há divergências entre esses campos de estudo, principalmente quanto aos objetivos de cada um. Enquanto métodos estatísticos visam alcançar um entendimento sobre o processo no qual foram gerados os dados, técnicas de AM concentram-se em gerar um sistema capaz de melhorar seu desempenho em cima de seus próprios resultados obtidos. Ou seja, em AM o programa deve ser capaz de adaptar sua estratégia de execução de acordo com as novas informações aprendidas (PATCHA; PARK, 2007). As técnicas de AM podem ser classificadas entre três categorias de acordo com a resposta que a máquina recebe sobre seu aprendizado (RUSSELL et al., 2003). Tais categorias são: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. Alguns autores consideram ainda a existência de uma quarta categoria, chamada aprendizado

semi-supervisionado.

Métodos que utilizam aprendizado supervisionado são treinados por meio de instâncias rotuladas. Isto é, o algoritmo de aprendizagem recebe um conjunto de dados de entrada junto com suas saídas esperadas. Com isso o algoritmo pode aprender comparando as saídas obtidas com as esperadas. Por fim o algoritmo modifica o modelo de acordo com o que foi aprendido. Já o aprendizado semi-supervisionado utiliza tanto dados rotulados quanto não rotulados. Esse tipo de método é útil quando o custo para rotular todos os dados é elevado. O que acaba por gerar uma fase de treinamento com somente parte das instâncias rotuladas. No aprendizado não supervisionado nenhum dado é rotulado. O sistema deve tentar aprender por conta própria. O objetivo nesse tipo de método geralmente é identificar padrões e estruturas nos dados. Por fim, aprendizado por reforço, é aquele em que o algoritmo aprende pelas tentativas e erros. Cada ação gera uma recompensa ou punição de acordo com o sucesso ou insucesso da ação executada. Assim a máquina tende a buscar sempre o sucesso para garantir uma recompensa melhor.

Apesar do AM, nos últimos tempos, ter se mostrado como uma ferramenta poderosa para obtenção automática de conhecimento, não existe um único algoritmo ou método que tenha demonstrado desempenho suficiente para resolver todos os problemas da área. Portanto, é importante a compreensão sobre as vantagens assim como as limitações dos mais diversos algoritmos de AM. E se possível, sempre utilizar metodologias que permitam avaliar os conceitos induzidos por eles. Com isso, é possível alcançar um melhor resultado nas aplicações do mundo real (MONARD; BARANAUSKAS, 2003).

## 2.2 ANOMALIAS

Um *outlier* pode ser definido como uma observação que difere tanto das outras que causa a impressão de ter sido gerada por um mecanismo diferente (HAWKINS, 1980). Isto é, anomalias, *outliers* ou pontos discrepantes são elementos que não estão em conformidade com o resto dos demais dados considerados normais. Geralmente são instâncias que não se comportam da maneira esperada. Por exemplo, como mostra a Figura 2, as instâncias $A_1$, $A_2$ e mesmo o pequeno subconjunto $A_3$ não pertencem a nenhum dos conjuntos identificados $N_1$ ou $N_2$, logo tais instâncias podem ser classificadas como anormais. O motivo da ocorrência de anomalias nos dados pode ser relacionado com diversos fatores, tais como: atividades

maliciosas, irregularidades, inconsistência e até mesmo erros nos dados. Entretanto todos *outliers* possuem certa relevância quando submetidos a uma análise que os tornam a chave para estudos de detecção de anomalia (AGGARWAL; YU, 2001).

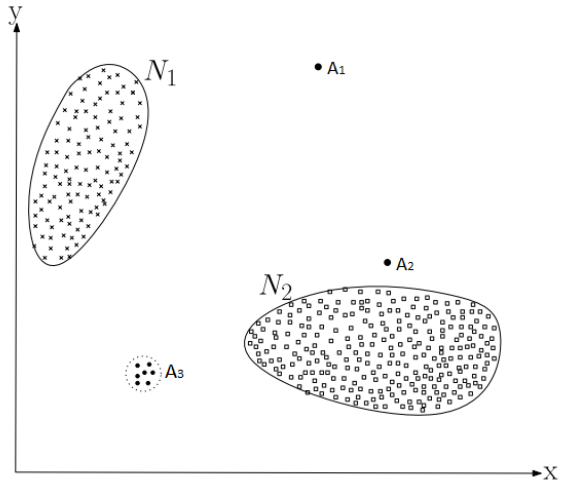



**Figura 2 – Anomalias em um conjunto de dados em 2D**
**Fonte: Adaptado de Chandola et al. (2009)**

## 2.3 TÉCNICAS PARA DETECÇÃO DE ANOMALIA

A detecção de anomalias está relacionada ao problema em encontrar padrões nos dados que não estejam se comportando conforme o esperado, ou seja, identificar os chamados *outliers*. A importância da detecção de anomalias está relacionada ao fato de que os *outliers* podem gerar informações úteis indiferentemente da área na qual provém os dados. Os mais conhecidos domínios de aplicação onde se é utilizada a detecção de anomalias são: fraudes financeiras com cartões de crédito, seguros e planos de saúde. Além de ser muito utilizada para detecção de falhas em sistemas críticos de segurança, identificação de intrusos na rede e vigilância militar (CHANDOLA et al., 2009).

Vários autores sobre detecção de anomalia procuram focar seus trabalhos em uma única área de pesquisa ou em domínio de aplicação específico. Entretanto, estudos comparativos (CHANDOLA et al., 2009; AGYEMANG et al., 2006; HODGE; AUSTIN, 2004) tendem a subcategorizar as técnicas de detecção de anomalia, sendo estas:

- Técnicas baseadas em classificação;
- Técnicas baseadas em agrupamento;
- Técnicas baseadas nos vizinhos mais próximos;
- Técnicas estatísticas;
- Técnicas da teoria da informação;
- Técnicas espectrais.

Cada uma dessas técnicas possuem características próprias e tendem a ser nomeadas de acordo com as abordagens adotadas como base para a aplicação e implementação das mesmas. Entretanto, mesmo nas subcategorias é comum ocorrer subdivisões conforme os princípios e suposições adotados pelos métodos.

### 2.3.1 Técnicas baseadas em classificação

Classificação é uma técnica de aprendizado de máquina na qual um modelo é induzido a partir de dados de treinamento e, posteriormente, esse modelo é utilizado para classificar dados novos, isto é, não vistos durante o seu treinamento (DUDA et al., 2012). A maneira com que a técnica de detecção de anomalia baseada em classificação opera é muito similar, onde primeiramente treina-se o classificador por meio dos dados rotulados disponíveis e em seguida é feita a classificação de uma instância teste como normal ou anormal. Técnicas para detecção de anomalia baseadas em classificação geralmente são subdivididas em técnicas multiclasses e com uma classe. Essa última, como o nome sugere, assume que todas as instâncias de treinamento pertencem a mesma classe. É utilizado então um algoritmo para delimitar e aprender a área em torno das instâncias consideradas normais. Portanto, como é possível ver na Figura 3a, qualquer objeto que não estiver contido dentro da região aprendida é considerado um ponto discrepante.

Já as técnicas multiclasses abordam os dados de uma maneira diferente. Elas assumem que as instâncias de treinamento podem ser rotuladas com várias classes. Isto é, os dados ali contidos podem pertencer a uma das $N$ classes normais existentes (STEFANO et al., 2000). Para essas técnicas multiclasses, será necessária a existência de um classificador capaz de identificar, para cada instância, se a mesma pertence a uma das classes. Portanto, uma nova instância teste poderá ser classificada em uma das classes existentes ou não. Caso o classificador não consiga atribuir essa nova instância a uma classe, como mostra a Figura 3b, ela é considerada uma anomalia.

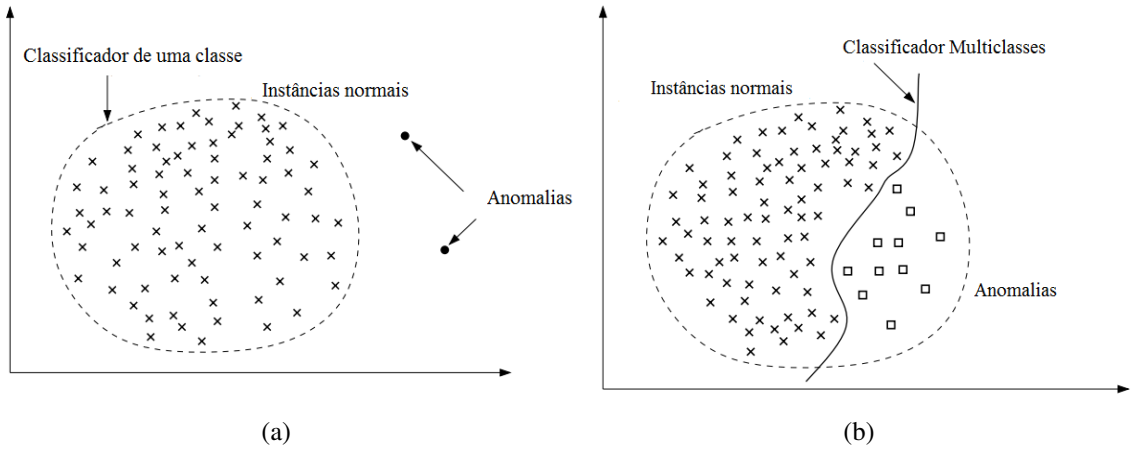



**Figura 3 – Técnicas para detecção de anomalias usando classificação.**
**Fonte: Adaptado de Chandola et al. (2009)**

De acordo com Chandola et al. (2009) as vantagens de técnicas para detecção de anomalias baseadas em classificação são estas:

- As técnicas, especialmente categorizadas como multiclasses, fazem um bom uso de poderosos algoritmos para identificar objetos pertencentes a diferentes classes;
- Uma vez que o modelo é gerado na fase de treinamento, classificar instâncias de teste é um processo considerado rápido.

E as desvantagens são:

- O sucesso de técnicas multiclasses dependem da precisão dos rótulos atribuídos para as várias classes normais, o que geralmente não ocorre;
- Atribuir rótulos para cada instância de teste pode atrapalhar na identificação de anomalias quando uma pontuação significante é necessária para caracterizar um objeto como ponto discrepante.

Muitas abordagens para detecção de anomalias utilizando classificação geralmente fazem uso das Máquinas de Vetores Suporte (Support Vector Machines -SVM). SVM é uma técnica de AM de classificação, geralmente para uma classe, que tem chamado atenção recentemente por mostrar resultados superiores que muitas outras técnicas de AM em diversas áreas, como no reconhecimento de padrões (LORENA; CARVALHO, 2007). Para detecção de anomalias as técnicas usando SVM aprendem a região em que as instâncias de treinamento estão contidas. Em casos de regiões mais complexas, é feita a utilização de *kernels*, tais como o *kernel* de função de base radial (*radial basis function* - RBF), dada pela Equação 1, onde $x_i$ e $x_j$ são dois pontos do espaço de entradas e $\sigma^2$ é a amplitude especificada. Para cada instância de teste é verificado se essa instâncias pertence a essa região. Esse dado será considerado normal

se pertencer a região ou anômalo caso não esteja contido nela (CHANDOLA et al., 2009). Essas delimitações criadas pela função *kernel* utilizada são chamadas fronteiras de decisão, pois são elas que separam as classes e mesmo os *outliers* presentes no conjunto de dados(LORENA; CARVALHO, 2007).

$$k(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2}||x_i - x_j||^2\right). \quad (1)$$

Segundo Amer et al. (2013), é possível atribuir um escore de anormalidade por meio do uso de SVM ao verificar a distância de cada instância em relação a fronteira de decisão. Logo, essa pontuação pode obtida ao aplicar a Equação 2, em que $g_{max}$ refere-se a distância máxima entre o conjunto de dados e a fronteira de decisão. Dessa maneira, as instâncias que estiverem próximo a essa fronteira receberão valores tendendo a 1,0, enquanto *outliers* receberão valores acima disso.

$$f(x) = \frac{g_{max} - g_x}{g_{max}} \quad (2)$$

### 2.3.2 Técnicas baseadas no vizinho mais próximo

A ideia básica utilizada em técnicas baseadas nos vizinhos mais próximos é que dados normais estão localizados em vizinhanças populosas, enquanto pontos discrepantes encontram-se mais isolados de seus vizinhos (CHANDOLA et al., 2009). Para a utilização destas técnicas é preciso garantir a existência de uma maneira de mensurar a semelhança entre instâncias ou objetos. Por exemplo, para dados com atributos contínuos geralmente é utilizada a distância euclideana como medida de dissimilaridade, entretanto como aponta Tan et al. (2013), para atributos categóricos geralmente é utilizado o coeficiente de casamento simples para medir similaridade. Com uma semelhança/diferença estimada para com seus vizinhos, é possível então atribuir uma pontuação para cada instância. Essa pontuação ou escore indica o grau de anormalidade de uma instância. A maneira em que se é calculada essa pontuação pode subcategorizar o método aplicado entre as seguintes categorias: técnicas baseadas no *k-ésimo* vizinho mais próximo e técnicas baseadas na densidade relativa.

As técnicas básicas baseadas nos $k$ vizinhos mais próximos, utilizam a distância de uma instância $p$ para com seu k-ésimo vizinho para atribuir uma pontuação de anomalia a

essa instância. Essa técnica para detectar anomalias tem sido muito utilizada nas mais diversas áreas de aplicação como pode ser verificado em estudos realizados por Byers e Raftery (1998), Guttormsson et al. (1999), Ramaswamy et al. (2000). As técnicas baseadas na densidade relativa utilizam uma abordagem um pouco diferente para atribuir uma pontuação de anormalidade à uma instância. Dada uma instância, a distância até seu *k-ésimo* vizinho é equivalente ao raio de uma hiperesfera, sendo o ponto central da hiperesfera a própria instância. Logo, é possível calcular a densidade do objeto relativa a sua vizinhança e identificar instâncias que são isoladas de seus vizinhos.

De acordo com Chandola et al. (2009) as vantagens de técnicas para detecção de anomalias baseadas nos vizinhos mais próximos são estas:

- São conduzidas fortemente pelos dados, ou seja elas não criam suposições a respeito do comportamento dos dados, nem requerem a supervisão humana para obter resultados;
- Essas técnicas podem ser utilizadas tanto de maneira não supervisionada como semi-supervisionada, e essa última geralmente possui melhor desempenho para encontrar anomalias;
- A adaptação dessas técnicas para tipos de dados diferentes do usual é geralmente fácil, considerando que basta apenas definir uma maneira de mensurar as distâncias entre os dados.

E as desvantagens são:

- Se os dados normais não possuem muitos vizinhos tais técnicas podem falhar ao classificá-los;
- A complexidade computacional de tais técnicas geralmente é $O(n^2)$, em que $n$ representa o número de instância;
- Para dados complexos pode ser difícil definir uma medida para calcular as distâncias.

### 2.3.3 Técnicas baseadas em agrupamento

Agrupamento ou *clustering* é a ação de organizar um conjunto heterogêneo de objetos em subgrupos mais homogêneos. Técnicas baseadas em agrupamento delimitam grupos de instâncias de maneira similar ao que ocorre em técnicas de classificação. Entretanto, o que diferencia agrupamentos de classificação é o fato de não existir a etapa de treinamento do modelo que ocorre antes da classificação, bem como a inexistência de classes pré-definidas.

Em agrupamento, os dados são imediatamente designados a um grupo de acordo com suas similaridades básicas com outros elementos daquele conjunto (BERRY; LINOFF, 1997). As técnicas para detecção de anomalias baseadas em agrupamentos geralmente seguem umas das seguintes premissas:

- Dados normais são encontrados em agrupamentos enquanto anomalias não pertencem a nenhum agrupamento;
- Dados normais estão localizados o mais próximo possível do centroide de um agrupamento, enquanto anomalias estão distantes do mesmo;
- Dados normais pertencem a densos e grandes agrupamentos, enquanto anomalias estão em agrupamentos pequenos ou dispersas.

Técnicas baseadas na primeira premissa geralmente utilizam algum algoritmo comum de agrupamento em um conjunto de dados. As instâncias não pertencentes a nenhum agrupamento criado são classificadas como *outliers*. Logo, como mostra a Figura 4, para este tipo de técnica, as instâncias *p1* e *p2* seriam consideradas anomalias, uma vez que não estão em nenhum dos agrupamentos criados (*C1* ou *C2*). Alguns dos algoritmos mais utilizados para isso são: DBSCAN (ESTER et al., 1996), ROCK (GUHA et al., 2000), SNN (LAZAREVIC et al., 2003) e o algoritmo *FindOut* (YU et al., 2002). Entretanto como aponta He et al. (2003) tais métodos têm como principal objetivo identificar agrupamentos e anomalias geralmente são consideradas como ruídos. Logo, técnicas baseadas na primeira suposição podem não ser tão eficazes para a identificação de instâncias anormais.

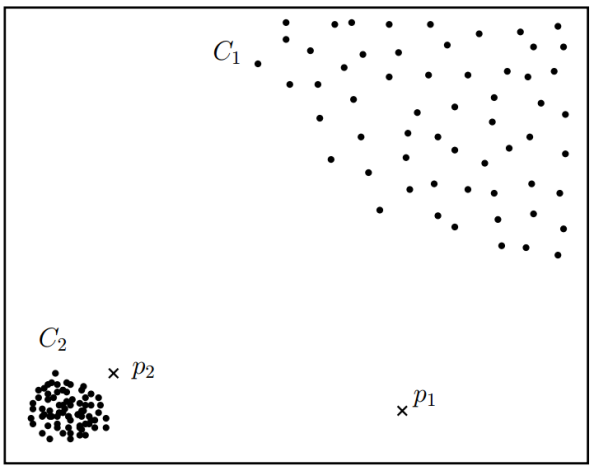



**Figura 4 – Técnicas de agrupamento para detecção de anomalias**

**Fonte: Chandola et al. (2009)**

Técnicas baseadas na segunda premissa consistem em duas etapas. Na primeira, um algoritmo de agrupamento é utilizado nos dados para gerar os grupos de instâncias. Na

segunda, é calculada a distância de cada objeto para com o centroide do agrupamento mais próximo. Vários algoritmos de agrupamento vêm sendo utilizados para aplicar tais técnicas. E como demonstra Smith et al. (2002) algoritmos como *Self-Organizing Maps* (SOM), *K-means Clustering* e *Expectation Maximization* (EM) podem ser usados para agrupar dados de treinamento e em seguida classificar instâncias de teste. Nesse tipo de técnica, os pontos *P1* e *P2* presentes na Figura 4, possivelmente seriam considerados anomalias, devido as suas distancias em relação aos centroides de *C1* e *C2*.

Técnicas baseadas na terceira premissa surgiram para cobrir as falhas das duas premissas anteriores. Isto é, quando anomalias formam um pequeno agrupamento e não são detectadas. Essas técnicas identificam como anomalias as instâncias pertencentes a agrupamentos cujo tamanho e densidade é inferior a um limiar. Muitos estudos com variações deste tipo de técnicas já foram propostos. Por exemplo, o método *FindCBLOF*, proposto por (HE et al., 2003) que atribui um grau de anormalidade para cada instância chamado *Cluster-Based Local Outlier Factor* (CBLOF). Esse método calcula o tamanho do agrupamento ao qual uma instância pertence e sua distância até o respectivo centroide.

De acordo com Chandola et al. (2009) as vantagens de técnicas para detecção de anomalias baseadas em agrupamento são:

- Essas técnicas podem ser operadas de modo não supervisionado;
- Elas geralmente podem ser adaptadas mesmo a tipos de dados complexos;
- A faze de treinamento dessas técnicas é considerada rápida.

E as desvantagens são:

- O desempenho dessas técnicas depende fortemente do algoritmo utilizado para agrupar as instâncias;
- Muitas técnicas dessa categoria não são otimizadas para detecção de anomalias.

Apesar do das técnicas de AM baseadas em agrupamento não serem por padrão otimizadas para a detecção de anomalias, alguns autores tem mostrado que ao combinar essas técnicas com outros tipos de abordagens, é possível otimizar a detecção de *outliers*. Um exemplo disso, é a abordagem chamada Ordering Points To Identify the Clustering Structure with Outlier Factors (OPTICSOF) proposta pelos autores Breunig et al. (1999). Nessa abordagem, o algoritmo de agrupamento OPTICS foi incrementado com a possibilidade de atribuir um fator de anormalidade para as instâncias. Isto é, o algoritmo OPTICS que encontra *clusters* baseado na densidade e ordena os pontos considerados próximos de maneira a mantê-los como vizinhos, passou a ter uma noção melhor de *outliers* ao considerar uma vizinhança local de cada objeto. Com isso, instâncias sem vizinhos próximos poderão receber uma pontuação elevada de anormalidade.

### 2.3.4 Técnicas estatísticas

Um conceito que define bem a visão que a estatística tem para com pontos discrepantes é de que "anomalia é uma observação suspeita de ser parcial ou completamente irrelevante devido a ela não ser gerada pelo modelo estocástico escolhido" (ANSCOMBE, 1960). De acordo com Chandola et al. (2009), técnicas estatísticas para detecção de anomalias são realizadas a partir de testes de inferência aplicados a um modelo estatístico ajustado de acordo com os dados. Este teste serve para indicar a probabilidade de um novo objeto pertencer ao modelo. Onde a interpretação dos resultados baseia-se na hipótese de que dados normais pertencem a regiões com valores de probabilidades elevada, enquanto pontos discrepantes ocorrem em áreas com baixo índice de probabilidade de acordo com um modelo estocástico. Geralmente, as técnicas estatísticas para detecção de pontos discrepantes são classificadas em dois grupos: paramétricas e não paramétricas.

As técnicas paramétricas fazem uso de parâmetros da distribuição para o cálculo estatístico. Elas assumem que os dados normais são gerados por uma distribuição paramétrica com parâmetros $\Theta$ e uma função densidade de probabilidade $f(x,\Theta)$. Os parâmetros $\Theta$ são estimados por meio dos dados de entrada. O objeto $x$ é uma instância teste ou observação. Para atribuir uma pontuação de anormalidade a uma instância $x$ basta utilizar o inverso da função $f(x,\Theta)$. Entretanto, como aponta Bamnett e Lewis (1994), testes de discordância podem ser utilizados para detectar pontos discrepantes. As anomalias são identificadas por meio da utilização de uma hipótese nula $H_0$. Basicamente $H_0$ assume que uma instância $x$ foi gerada usando uma distribuição estimada com parâmetros $\Theta$. Caso o teste estatístico rejeite $H_0$ conclui-se que $x$ é uma anomalia.

Em não paramétricas, geralmente o modelo não é pré-definido e somente é gerado a partir dos dados fornecidos. Isto é, métodos que utilizam tais técnicas não pressupõem a existência de uma distribuição estatística já conhecida, e sim que o mesmo precisa ser adaptado de acordo com os dados. Essa característica faz com que a quantidade de suposições sobre os dados sejam bem limitadas quando comparado com técnicas paramétricas.

Segundo Chandola et al. (2009) as vantagens de técnicas estatísticas para detecção de anomalias são estas:

- Se suposições a respeito da distribuição adjacente dos dados são verdadeiras, técnicas deste tipo geram soluções estatisticamente justificáveis para detectar anomalias;
- Uma pontuação de anormalidade proveniente de uma técnica estatística utiliza um intervalo de confiança que pode ser usado como informação adicional na tomada de

decisões para qualquer instância de teste;

- Se a etapa de estimar a distribuição for bem elaborada para lidar com anomalias nos dados, técnicas estatísticas podem operar de maneira não supervisionada.

As desvantagens são estas:

- Partir da suposição que os dados são gerados a partir de uma distribuição específica pode conduzir ao erro, principalmente em grandes conjuntos de dados;
- Projetar hipóteses para distribuições complexas não é um processo trivial.
- As técnicas mais fáceis de implementar geralmente não são capazes de identificar as interações entre diferentes atributos para dados multivariados.

Um exemplo de técnicas paramétricas para detecção de anomalias são as técnicas baseadas no Modelo Gaussiano. Essas técnicas assumem que os dados provém de uma distribuição Gaussiana. Ou seja, essa técnica assume que os dados estão dispostos de maneira em que a maioria desses dados estão agrupados no intervalo do meio dessa distribuição, enquanto o resto dos dados distribui-se simetricamente em direção aos extremos. Os parâmetros são estimados utilizando a máxima verossimilhança (Maximum Likelihood Estimates - (MLE)), que busca maximizar a probabilidade dos dados observados. Portanto, um escore de anormalidade pode ser atribuído a uma instância de acordo com a distância desse dado para com a média estimada para o modelo estatístico (CHANDOLA et al., 2009). Uma simples abordagem para identificar um limiar de anormalidade é atribuir como anomalias as instâncias que tenham distância maior que $3\sigma$ da média da distribuição, onde $\sigma$ é o desvio padrão dessa distribuição. O motivo disso é que estatisticamente, a região dessa distância contém 99,7% das instâncias do conjunto de dados (SHEWHART, 1931).

### 2.3.5 Técnicas da teoria da informação

A teoria da informação, originalmente conhecida como a teoria matemática da comunicação (SHANNON, 2001), é um ramo pertencente a matemática estatística. Essa teoria estuda a compreensão da comunicação por meio da medição da informação. A informação ou comunicação é avaliada de maneira quantitativa e qualitativa, utilizando-se para isto medidas próprias da teoria da informação.

Portanto, técnicas para detecção de anomalias baseadas na teoria da informação utilizam medidas criadas por estudos desta área para analisar a informação contida em um

conjunto de dados. De acordo com Chandola et al. (2009) as seguintes medidas podem ser usadas para detecção de anomalias: entropia, entropia condicional, entropia relativa e informação mutua.

Entropia é uma medida de incerteza ou impuridade atribuída a um conjunto de dados. Como pode ser visto na Equação 3, para um conjunto de dados $X$, onde cada instância $x$ pertence a uma classe $C$, a entropia $H$ é dada pela soma da probabilidade de ocorrência $P(x)$ de cada dado multiplicada pelo logaritmo natural do inverso dessa probabilidade. Logo, um valor alto para a entropia caracteriza a existência de dados mais diferentes enquanto um valor mais baixo indica dados mais similares (LEE; XIANG, 2001).

$$H(X) = \sum_{x \in C_x} P(x) log \frac{1}{P(x)} \tag{3}$$

Entropia condicional é o cálculo da entropia de um conjunto de dados $X$, condicionada pela existência de outro conjunto $Y$. Logo, como mostra a Equação 4, a entropia condicional de $X$, dado $Y$, é calculada pela soma probabilidade de ocorrência da combinação de dados $x$ e $y$ pertencentes a uma classe $C$. Portanto, para detecção de anomalias, a entropia condicional pode ser utilizada como medida de regularidade em elementos sequenciais. Por exemplo, em sequências de texto, seja “tinto” uma palavra e “vinho” outra palavra, deseja-se calcular a probabilidade de “tinto” ser observada dado que “vinho” foi observada.

$$H(X,Y) = \sum_{x,y \in C_x, C_y} P(x,y) log \frac{1}{P(x|y)} \tag{4}$$

A entropia relativa, também chamada de distância de *Kullback-Leibler*, é uma medida da distância entre duas distribuições de probabilidade. Essa distância pode ser calculada por meio da Equação 5, onde $p(x)$ e $q(x)$ referem-se as probabilidades de um dado ou evento $x$ ocorrer nas distribuições $p$ e $q$, sendo $x$ pertencente a classe $C$. Em detecção de anomalias, esse tipo de medida pode ser utilizada para verificar a similaridade entre o conjunto de dados do treinamento e o conjunto de dados de teste.

$$D(p|q) = \sum_{x \in C_x} p(x) log \frac{p(x)}{q(x)} \tag{5}$$

Informação mútua é uma medida que determina a qualidade dos atributos de uma instância. Sendo esses atributos utilizados pelo classificador para determinar essa instância a uma classe. Portanto, a informação mútua de um atributo $A$ pertencente a um conjunto de dados $X$ é dada pela Equação 6, onde $Valores(A)$ são os possíveis valores para $A$ e $X_v$ é um subconjunto de $X$ em que $A$ assume o valor $v$. Se todos atributos resultarem em baixa informação mútua, então o classificador terá um desempenho fraco devido ao elevado nível de entropia

existentes nas partições criadas durante essa fase. Ou seja, quanto maior for a informação mútua encontrada em atributos, melhor serão as chances de detectar anomalias (LEE; XIANG, 2001).

$$M(X,A) = H(X) \; - \sum_{v \in Valores(A)} \frac{|X_v|}{|X|} H(X_v) \tag{6}$$

Segundo Chandola et al. (2009) as vantagens desse tipo de técnicas para detecção de anomalias são:

- Essas técnicas podem operar de maneira não supervisionada;
- Elas não fazem o uso de suposições sobre a distribuição estatística do conjunto de dados.

As desvantagens são estas:

- O desempenho dessas técnicas pode cair muito dependendo da medida da teoria da informação escolhida;
- Essa técnicas dependem da informação relativa ao tamanho do conjunto de dados, o que pode não ser tão simples de obter;
- A atribuição de uma pontuação de anormalidade para uma instância usando essas técnicas não é um processo trivial .

### 2.3.6 Técnicas espectrais

O teorema espectral é utilizado na álgebra linear para permitir que uma matriz possa ser diagonalizada. Esse teorema também fornece uma decomposição canônica chamada decomposição espectral. Essa decomposição é responsável por fatorar uma matriz em uma forma canônica. Isto é, alterar a representação da matriz, por meio de uma maneira padrão, para a forma de uma expressão matemática (GOLUB; LOAN, 2012).

Técnicas espectrais para detectar anomalias baseiam-se na suposição que dados podem ser incorporados em subespaços mais baixos no qual instâncias normais e anomalias tendem a aparecer significantemente diferente. Logo, estabelecer tais subespaços possibilita a identificação de anomalias facilmente (AGOVIC et al., 2008). Várias das técnicas espectrais utilizam o *Principal Component Analysis* (PCA) para projetar dados em um espaço dimensional mais baixo. Com isso, é possível analisar a projeção para cada instância junto aos componentes principais com baixa variação. Instâncias normais que satisfaçam a correlação dos dados possuem um valor baixo para as projeções, enquanto anomalias que desviem da estrutura de

correlação possuem valores altos (DUTTA et al., 2007).

Técnicas espectrais que utilizam o PCA, podem também ser utilizadas para complementar técnicas com outras abordagens para detecção de anomalias. Por exemplo, o algoritmo chamadado Correlation Outlier Probility (COP) proposto por Kriegel et al. (2012). Esse algoritmo faz uso das técnicas baseadas nos $k$ vizinhos mais próximos em conjuntos com as espectrais, por meio do uso de um PCA para computar uma correlação local para cada instância. Dessa maneira, os objetos *outliers* receberão uma pontuação de desvio tendendo a 1, enquanto instâncias normais irão receber pontuações próximas ou iguais a 0.

Como proposto por Idé e Kashima (2004), também é possível detectar anomalias utilizando séries temporais em grafos. Onde cada grafo é representado como uma matriz adjacente para um determinado tempo. A cada intervalo de tempo, o componente principal da matriz é escolhido como o vetor de atividade para aquele grafo. As séries temporais dos vetores de atividade são consideradas como uma matriz e o vetor singular principal esquerdo é obtido para capturar as dependências normais nos dados ao longo do tempo. Para cada novo grafo de teste, é calculado o ângulo entre seu vetor de atividade e o vetor singular principal esquerdo obtido do grafo anterior. Esse ângulo é utilizado para determinar o grau de anormalidade de um grafo de teste.

## 2.4 LOCAL OUTLIER FACTOR

Vários estudos apontam *outlier* como sendo uma propriedade binária em meio a um conjunto de dados, ou seja, ele existe ou não. Entretanto, como aponta os autores do método *Local outlier factor* (LOF) Breunig et al. (2000), de acordo com o contexto, uma visão binária sobre anomalias nem sempre é a melhor solução. Portanto, LOF é um método para detecção de *outliers* que sugere a utilização de um fator local para determinar o quanto um dado pode ser considerado uma anomalia. Este grau de anomalia atribuído ao objeto está diretamente ligado ao seu isolamento em relação ao seus vizinhos.

Na abordagem do método LOF, a densidade local de um objeto e seu fator local são atribuídos de acordo com seus $k$ vizinhos mais próximos. Normalmente, essa densidade é estimada por meio de uma função de distância, sendo geralmente a distância euclidiana em que o ponto pode ser alcançado a partir de seus vizinhos. Ao comparar a densidade local de um dado com a densidade local de seus vizinhos é possível identificar tanto áreas com densidade similar assim como objetos com densidades mais baixas que seus vizinhos. Esses pontos com

baixa densidade são considerados anomalias (BREUNIG et al., 2000).

De acordo com Chandola et al. (2009), esse tipo de técnica, baseada nos vizinhos mais próximos, pode ser subcategorizada como uma técnica de detecção de anomalia baseada na densidade relativa. Com isso, para cada instância, é calculada a distância entre seu *k-ésimo* vizinho mais próximo. Essa distância é utilizada como um raio para considerar qualquer outras instâncias pertencentes a região da hiperesfera, com centro sendo a instância a ser calculada, como parte da vizinhança dessa instância. Assim, é possível calcular a densidade dessa instância e posteriormente, comparar com a densidade das instâncias vizinhas. Isto é, verificar para cada instância a densidade relativa local. No algoritmo LOF, essa comparação entre densidades para com os $k$ vizinhos determina um escore que pode ser utilizado para identificar o grau de anormalidade de cada instância.

Para determinar o fator local de um ponto $p$, é necessário empregar três distâncias: distância euclidiana ($d_e$); distância k ($d_k$); e distância de alcance ($d_a$). Primeiramente, a distância euclidiana é utilizada para cálculo da distância k. A distância $d_k(p)$ é definida como a distância euclidiana de $p$ em relação ao seu *k-ésimo* vizinho mais próximo. A seguir, é calculada a distância de alcance de $p$ em relação a um ponto de referência $o$ como mostrado na Equação 7, que será dada pelo máximo entre a distância $k$ de $p$ ou a distância euclidiana de $p$.

$$d_{a_k}(p,o) = max\{d_k(p), d_e(p,o)\} \tag{7}$$

Em seguida é calculada a densidade de alcance *den* local para o objeto $p$. Como a Equação 8 sugere, este valor é dado pelo inverso da distância de alcance média entre o ponto $p$ e seus $k$ vizinhos pertencentes ao conjunto $N$.

$$den(p) := 1/\frac{\sum_{o \in N_k(p)} d_{a_k}(p,o)}{|N_k(p)|} \tag{8}$$

Logo, como mostra a Equação 9, o fator local do ponto $p$ pode ser encontrado por meio do valor da densidade de alcance média de sua vizinhança $N$, dividido pela densidade local de $p$.

$$LOF_k(p) := \frac{\sum_{o \in N_k(p)} \frac{den(o)}{den(p)}}{|N_k(p)|} \tag{9}$$

A interpretação da pontuação resultante do cálculo do fator local para um objeto ocorre da seguinte maneira: Quanto mais próximo de 1 maior será o grau de similaridade daquele dado com seus vizinhos, logo não será classificado como *outlier*. Entretanto, quanto maior for o valor do fator local de um dado, menor sera a densidade local daquela região e consequentemente maior será o grau de anomalia. Por exemplo, a Figura 5 mostra que as instâncias isoladas receberam pontuação mais elevada, enquanto os pontos com muitos vizinhos

próximos receberam valores de fator local próximo de 1,0. Assim quanto menor for o valor deste fator local, maior será a densidade da região ao qual este objeto pertence (BREUNIG et al., 2000).

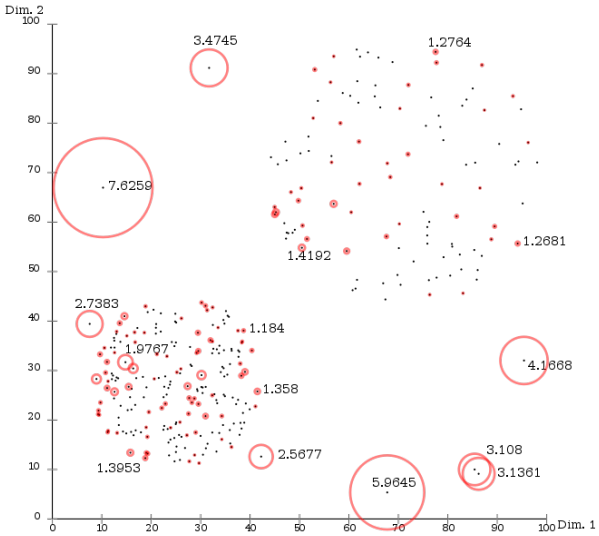



**Figura 5 – Fator local do algoritmo LOF em 2D**

**Fonte: Adaptado de ELKI (2017)**

Segundo Chandola et al. (2009), técnicas baseadas em densidade geralmente não funcionam adequadamente se o conjunto de dados tiver regiões com diferentes densidades. Isto é, caso os agrupamentos formados nos dados possuam vizinhanças com densidades diferentes, ou se as distâncias entre as instâncias presentes em cada grupo forem diferentes das outras. Entretanto, a abordagem de atribuir escores de anormalidade tomada pelo algoritmo LOF, de maneira relativa aos vizinhos locais, contorna este problema. De fato, este tipo de técnica, que faz o uso da densidade local, pode ser capaz de detectar anomalias que a abordagem que faz uso da densidade global não conseguiria.

# 3 MATERIAIS E MÉTODOS

Nesse capítulo são descritos os métodos, assim como os materiais utilizados para o desenvolvimento deste trabalho. A metodologia seguida neste trabalho foi de Descoberta de Conhecimentos em Bases de Dados (Knowledge Discovery in Databases - KDD). Portanto, quatro etapas principais foram seguidas, sendo estas:

1. Identificação, seleção e coleta dos dados;
2. Pré-processamento;
3. Mineração de anomalias;
4. Exibição, análise e comparação dos resultados.

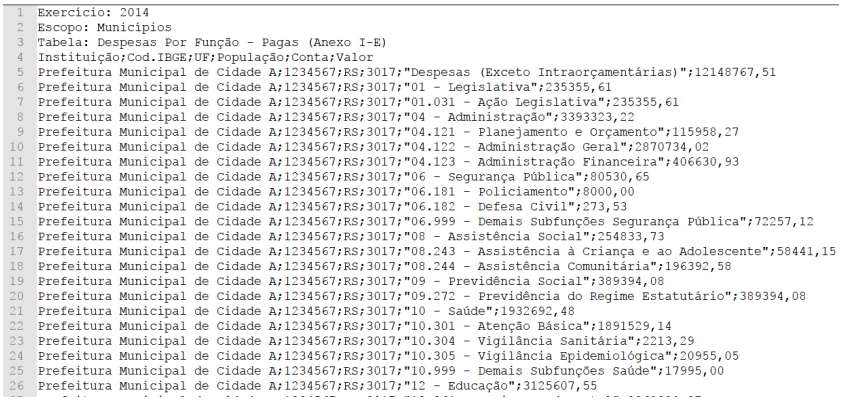
O processo será descrito na mesma sequência cronológica em que as etapas foram executadas. Para cada fase serão relatados os principais fundamentos e tecnologias a serem empregados.

## 3.1 IDENTIFICAÇÃO, SELEÇÃO E COLETA DOS DADOS

Nesta fase, inicialmente, foi feito um levantamento de portais brasileiros com dados públicos disponíveis. Em seguida foi efetuada uma busca detalhada nesses portais com o objetivo em identificar os portais com dados relevantes para a execução deste trabalho. Os dados selecionados para este trabalho foram os dados contábeis das prefeituras. Esses dados são disponibilizados pelo Sistema de Informações Contábeis e Fiscais[1] (SICONFI) do portal do Tesouro Nacional[2]. Os dados descarregados foram referentes as contas anuais de todas prefeituras brasileira que declararam para os anos de 2013, 2014 e 2015. Isto é, os dados escolhidos são a respeito das despesas empenhadas pelas prefeituras para cada um desses anos. Esses dados são disponibilizados pelo SICONFI no formato *comma-separated values* (CSV) e

[1] https://siconfi.tesouro.gov.br/
[2] http://www.tesouro.fazenda.gov.br/

são referentes as despesas das prefeituras pagas por função. Trata-se de um único arquivo para cada ano contendo milhares de linhas, cada linha, como mostra a Figura 6, refere-se a um valor gasto em uma categoria feito por uma prefeitura. Na mesma linha, também são informados o código da cidade dado pelo IBGE assim como a população e estado dessa cidade. Portanto, várias linhas são necessárias para identificar os gastos de uma mesma cidade devido às diversas categorias existentes, tais como administração, saúde, educação, entre várias outras.



```
Exercício: 2014
Escopo: Municípios
Tabela: Despesas Por Função - Pagas (Anexo I-E)
Instituição;Cod.IBGE;UF;População;Conta;Valor
Prefeitura Municipal de Cidade A;1234567;RS;3017;"Despesas (Exceto Intraorçamentárias)";12148767,51
Prefeitura Municipal de Cidade A;1234567;RS;3017;"01 - Legislativa";235355,61
Prefeitura Municipal de Cidade A;1234567;RS;3017;"01.031 - Ação Legislativa";235355,61
Prefeitura Municipal de Cidade A;1234567;RS;3017;"04 - Administração";3393323,22
Prefeitura Municipal de Cidade A;1234567;RS;3017;"04.121 - Planejamento e Orçamento";115958,27
Prefeitura Municipal de Cidade A;1234567;RS;3017;"04.122 - Administração Geral";2870734,02
Prefeitura Municipal de Cidade A;1234567;RS;3017;"04.123 - Administração Financeira";406630,93
Prefeitura Municipal de Cidade A;1234567;RS;3017;"06 - Segurança Pública";80530,65
Prefeitura Municipal de Cidade A;1234567;RS;3017;"06.181 - Policiamento";8000,00
Prefeitura Municipal de Cidade A;1234567;RS;3017;"06.182 - Defesa Civil";273,53
Prefeitura Municipal de Cidade A;1234567;RS;3017;"06.999 - Demais Subfunções Segurança Pública";72257,12
Prefeitura Municipal de Cidade A;1234567;RS;3017;"08 - Assistência Social";254833,73
Prefeitura Municipal de Cidade A;1234567;RS;3017;"08.243 - Assistência à Criança e ao Adolescente";58441,15
Prefeitura Municipal de Cidade A;1234567;RS;3017;"08.244 - Assistência Comunitária";196392,58
Prefeitura Municipal de Cidade A;1234567;RS;3017;"09 - Previdência Social";389394,08
Prefeitura Municipal de Cidade A;1234567;RS;3017;"09.272 - Previdência do Regime Estatutário";389394,08
Prefeitura Municipal de Cidade A;1234567;RS;3017;"10 - Saúde";1932692,48
Prefeitura Municipal de Cidade A;1234567;RS;3017;"10.301 - Atenção Básica";1891529,14
Prefeitura Municipal de Cidade A;1234567;RS;3017;"10.304 - Vigilância Sanitária";2213,29
Prefeitura Municipal de Cidade A;1234567;RS;3017;"10.305 - Vigilância Epidemiológica";20955,05
Prefeitura Municipal de Cidade A;1234567;RS;3017;"10.999 - Demais Subfunções Saúde";17995,00
Prefeitura Municipal de Cidade A;1234567;RS;3017;"12 - Educação";3125607,55
```

**Figura 6 – Arquivo com as despesas das prefeituras**

**Fonte: Autoria própria**

## 3.2 PRÉ-PROCESSAMENTO

Nesta etapa foi todo o tratamento dos dados coletados com o objetivo de adequar esses dados para serem utilizados posteriormente na aplicação de algoritmos de AM por uma ferramenta KDD. Esse tratamento foi quase que inteiramente feito utilizando a linguagem de programação Java. Foi criado um algoritmo contendo diversos métodos genéricos, tendo vários parâmetros de entrada, para efetuar as diversas transformações necessárias nos dados no formato CSV. Em geral, foram criados métodos para remover colunas ou linhas desnecessárias, e mesclar várias linhas referentes a uma instância em uma única linha. Dessa maneira, foi mantido os atributos dessa cidade como sendo as colunas da linha a qual ela pertence. Além disso, foram elaborados métodos necessário para as diversas abordagem de normalização experimentadas. A normalização geralmente consiste em transformar os valores dos atributos originais em um intervalo específico, por exemplo, o intervalo [0,1]. As principais técnicas utilizadas neste

trabalho foram a suavização, a normalização pela média zero e desvio padrão unitário e a transformação dos atributos referentes aos gastos absolutos em gastos relativos a população (por pessoa). A suavização consistiu em reduzir a escala dos dados por população, buscando utilizar valores próximos para cidades de tamanho similar. Para isso, foi utilizado o algoritmo na base decimal. A transformação em gastos relativos ocorreu ao dividir os valores gastos em cada categoria de uma instância pelo valor da população. Assim, esse método gerou valores de quanto foi gasto para cada função por habitante. Os métodos de normalização pela média e desvio padrão foram utilizados como complementares aos outros citados em praticamente todos os experimentos. Para automatizar o processo dessa etapa e das outras, um *script* foi elaborado para executar os métodos de tratamento de dados aqui descritos, conforme se fizeram necessários ao longo dos experimentos executados.

## 3.3 MINERAÇÃO DE ANOMALIAS

Nesta etapa foram executados os experimentos aplicando principalmente o algoritmo de AM LOF, em conjunto com a ferramenta para extração de conhecimento *Environment for Developing KDD-Aplications Supported by Index-Structures* (ELKI). O ELKI[3] é um arcabouço para extração de conhecimentos e mineração de dados. Ele foi desenvolvido na Universidade de Munique utilizando a linguagem de programação Java e possui código aberto. Vários algoritmos de AM foram implementados e integrados ao ELKI, o que torna mais simples o processo de utilização de um ou mais desses algoritmos, bastando apenas configurar o parâmetros gerais da ferramenta e os parâmetros específicos do algoritmo, caso existam. O foco dessa ferramenta está na aplicação de algoritmos de AM com ênfase em métodos não supervisionados para agrupamentos e detecção de *outliers* (ACHTERT et al., 2008). Como mostra a Figura 7, esta etapa pode, inicialmente, ser aplicada por meio do uso da interface gráfica do ambiente ELKI. Onde foi feita a seleção do algoritmo a ser executado e configuração dos parâmetros gerais, tais como o arquivo de entrada e o arquivo de saída, assim como parâmetros específicos do algoritmo, por exemplo, o valor de $k$ para os algoritmos baseados nos $k$ vizinhos mais próximos. Eventualmente, a chamada dessa ferramenta foi automatizada ao inclui-la no *script* elaborado durante a etapa anterior.

O principal algoritmo escolhido para ser aplicado durante essa etapa foi o LOF, cujo

[3] http://elki.dbs.ifi.lmu.de/

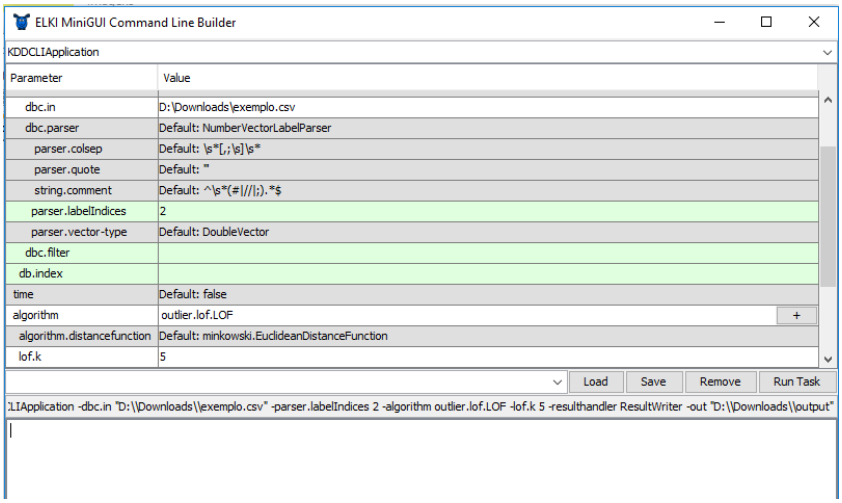



**Figura 7 – Interface Gráfica ELKI**

**Fonte: Autoria própria**

o propósito é atribuir uma pontuação de anormalidade a cada instância utilizando as técnicas baseadas nos k-vizinhos mais próximos. A função utilizada para calcular a distância de cada instância para com seus k vizinhos foi a distância euclidiana. O valor do parâmetro *k* escolhido para todos os experimentos foi 70 devido a esse valor ser próximo a raiz quadrada do total de instâncias desse conjunto de dados.

## 3.4 EXIBIÇÃO, ANÁLISE E COMPARAÇÃO DOS RESULTADOS

Nessa etapa, foram criados métodos na linguagem Java para tratar os dados de saída do ELKI e gerar um possível explicador das anomalias encontradas. O explicador foi elaborado calculando as diferenças dos valores das despesas de cada instância com o valor do centroide de 20 cidades de população parecida. Com isso, cada categoria de gasto recebeu uma porcentagem representando sua influência no resultado de acordo com a diferença do valor dessa categoria com o centroide das cidades parecidas. Portanto, categorias com valor muito acima ou abaixo da média das cidades parecidas receberam porcentagens maiores, indicando a possível razão dessa instância ter sido considerada uma anomalia.

Ainda nessa etapa, foram feitas comparações entre algumas das cidades com maiores pontuações de anomalias atribuidas pelo algoritmo LOF e seus respectivos Índices FIRJAN

de Gestão Fiscal[4] (IFGF). O IFGF, apresenta de maneira simples como anda a gestão dos municípios Brasileiros (FIRJAN, 2016). Ou seja, como é possível ver na Figura 8, o IFGF é um índice que leva em consideração vários fatores da gestão de uma cidade tais como: Arrecadação, investimentos, gastos com pessoal, suficiência de caixa e o custo da dívida do município. Com isso, uma pontuação é atribuída a cada prefeitura, indicando a qualidade da gestão dessa prefeitura para um determinado ano. Essa pontuação vai de 0,00 até 1,00, sendo que o valores acima de 0,8 significa uma gestão de excelência, entre 0,6 e 0,8 é considerada uma boa gestão, entre 0,4 e 0,6 a gestão passa ser considerada em dificuldade e se ficar com valor inferior a 0,4 é considera gestão crítica. Portanto, ao comparar as anomalias encontradas com as suas respectivas situações no quesito gestão fiscal, de acordo com o IFGF, foi possível interpretar e trazer discussões mais interessantes a respeito dos gastos efetuados por algumas prefeituras.

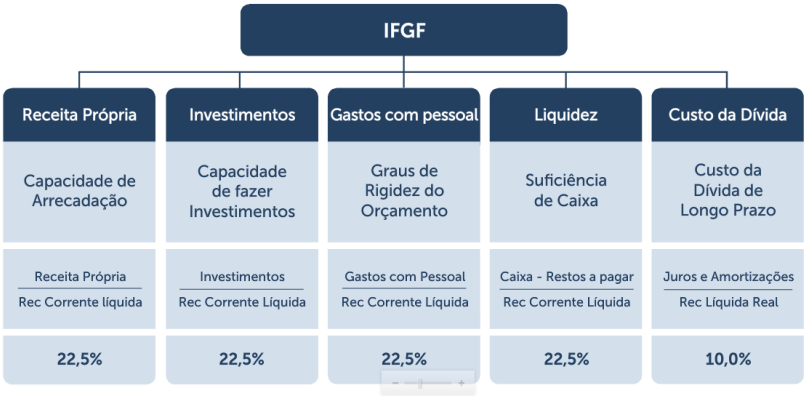



**Figura 8 – Composição da pontuação IFGF**

**Fonte: (FIRJAN, 2016)**

A última fase dessa etapa de análise dos resultados, foi comparar os resultados obtidos pelo algoritmo LOF com os resultados dos experimentos em que foram aplicados outros algoritmos de AM. O objetivo disso foi verificar se esses algoritmos trariam os mesmo resultados que o LOF ao classificar as anomalias. Portanto, alguns algoritmos com abordagens um pouco diferentes do LOF foram utilizados, e as maiores anomalias que esses algoritmos encontraram foram confrontadas entre-si e entre as maiores anomalias encontradas pelo LOF.

[4]http://www.firjan.com.br/ifgf/

# 4 EXPERIMENTO PRELIMINAR

Neste capítulo são apresentados os objetivos, procedimentos e resultados da execução de um experimento preliminar aplicado em uma amostra de dados reais.

## 4.1 OBJETIVOS

A falta de estudos e publicações de trabalhos similares a proposta apresentada por este trabalho tornou a elaboração de seu planejamento um processo mais complexo do que o esperado. A execução de um experimento pode ajudar de maneira geral na preparação para a aplicação real do trabalho. Logo, os principais motivos para a execução desse experimento foram:

1. Melhorar a identificação dos materiais e métodos a serem utilizados;
2. Entender melhor o funcionamento da ferramenta de extração de conhecimento ELKI;
3. Analisar o comportamento e resultado do algoritmo LOF aplicado em dados de gastos reais;
4. Verificar a viabilidade, assim como a complexidade, da identificação de anomalias em dados públicos abertos.

No experimento, foi possível aplicar as etapas discutidas na seção de materiais e métodos, podendo assim identificar os passos que estão de acordo ou não com o plano. Outro fator relevante desta prática será compreender melhor a ferramenta ELKI por meio de testes executados durante esse experimento. O programa ELKI possui a característica de ser muito promissor na aplicação de técnicas de AM não supervisionadas, assim como na extração de conhecimento utilizando as mais diversas técnicas. Entretanto, essa ferramenta não possui uma documentação adequada, o que acaba dificultando um melhor entendimento sobre seus parâmetros e seu funcionamento. Por fim, analisar a eficiência e dificuldades encontradas para aplicar um algoritmo de detecção de anomalias em dados reais é de grande importância para os

resultados futuros deste trabalho.

## 4.2 PROCEDIMENTOS

Para a execução deste experimento, inicialmente foi feita uma busca seguida da coleta de dados reais. Os dados escolhidos e baixados foram os dados abertos sobre os gastos parlamentares registrados na Câmara dos Deputados referente ao período de Janeiro até Outubro de 2016. Esses dados estão disponibilizados no formato XML e podem ser encontrados e baixados por qualquer pessoa no portal da Câmara[1]. Devido ao grande tamanho do arquivo XML, ele foi convertido para o formato CSV utilizando a ferramenta *MoorXmlToCsvConverter*. O novo arquivo CSV gerado foi então importado e aberto em uma planilha do *Excel*. Na planilha contendo os dados originais dos gastos parlamentares, foi inserida uma nova coluna para ser utilizada futuramente como um rótulo de identificação. Portanto, para cada linha desta nova coluna foi inserido um número único. Em seguida toda a planilha foi copiada em uma nova planilha. Nessa nova planilha foram excluídas todas as colunas não relevantes, restando somente três colunas: a coluna identificadora (numérica), a coluna da categoria ao qual o gasto pertence (nominal) e a coluna do valor líquido (numérica). Por fim, essa planilha foi exportada pela ferramenta *Excel* para o formato CSV, podendo assim ser utilizada como dados de entrada da ferramenta ELKI.

Após vários testes executando os dados gerados anteriormente como entrada no ELKI foi constatada a necessidade de separar os dados de cada categoria do tipo de gastos em arquivos diferentes. Ao todo, são 18 valores possíveis para a categoria do tipo do gasto. Portanto, seria necessário gerar 18 novos arquivos. Entretanto, devido esta prática ser um experimento e para agilizar o processo, nem todas categorias foram utilizadas. Então, vários arquivos foram criados utilizando filtros do próprio *Excel* e exportando os arquivos separadamente. Em seguida, com os arquivos gerados, foi aplicado o algoritmo LOF por meio da ferramenta ELKI, sendo que o principal parâmetro utilizado foi a entrada como sendo cada um dos arquivos gerados anteriormente. O parâmetro *dbc.parser* foi configurado como *CategoricalDataAsNumberVectorParser* para converter o atributo da categoria nominal para numérico, podendo assim exibir o resultado em duas dimensões (categoria e valor). O parâmetro

[1] http://www2.camara.leg.br/transparencia/cota-para-exercicio-da-atividade-parlamentar/dados-abertos-cota-parlamentar

*parser.labelIndices* foi escolhido como 0 para rotular a primeira coluna que é a do identificador, evitando assim que ela influencie durante aplicação do algoritmo. Foi também aplicado uma técnica de normalização dos dados disponível no parâmetro *dbc.filter*. No parâmetro *algotithm* foi selecionado o algoritmo LOF. A distância escolhida para ser utilizada foi a euclidiana. O valor do parâmetro $k$ do algoritmo LOF foi escolhido de acordo com o tamanho $n$ dos dados. Onde, para os arquivos com menos de 2000 instâncias foi estipulado um valor de $k$ igual a $\sqrt{n}$ , e para conjuntos maiores que isso $k$ foram utilizados principalmentes valores maiores, por exemplo 300, 400 e 500. Já para o parâmetro de saída foram escolhidas as opções de gerar visualização gráfica e criar arquivo de texto ao mesmo tempo. Por fim, utilizando essas configurações, o algoritmo LOF foi então executado várias vezes para todos arquivos contendo gastos de diversas categorias.

## 4.3 RESULTADOS

Os resultados do experimento mostraram tantos aspectos positivos como negativos da aplicação do algoritmo LOF para identificar anomalias em dados reais. Dentre os pontos negativos percebidos durante o experimento, o que maior se destacou foi a existência de grandes conjuntos de dados esparsos. Um motivo disso é que a categoria a qual esses dados pertencem é muito genérica, isto é, os valores das despesas que se enquadram nessa categoria podem variar muito. Por exemplo, como mostra a Figura 9, as despesas de consultoria ficaram distribuídas entre vários agrupamentos em sequência, sendo alguns muito grande e outros nem tanto, tendo como intervalo de valores em torno de 0 até 30 mil. Os diversos valores esparsos existentes nesse intervalo fizeram o algoritmo identificar um alto número de instâncias falsas positivas, ou seja, classificar dados possivelmente normais como anomalias. Além disso, o algoritmo atribuiu a uma grande quantidade de instâncias um *score* LOF considerado infinito, isto é, maior do que 1000. Apesar disso, o algoritmo também foi capaz de identificar possíveis instâncias positivas. Isto é, em meio a *outliers* falsos também foram reconhecidas anomalias reais, tais como a instância localizada no extremo do eixo $x$ da Figura 9, com valor de R$ 50.000,00 e que recebeu um *score outlier* de 7.9, o que já pode ser considerado alto.

Para categorias de gastos mais específicas, isto é, que possuem um certo padrão na distribuição dos valores, o algoritmo LOF mostrou um rendimento promissor na identificação de valores anormais. Como mostram as Figura 10 e 11, tanto para gastos com alimentação

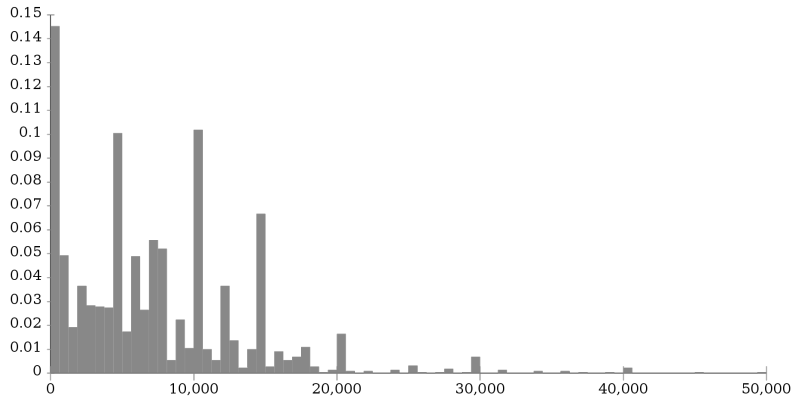



**Figura 9 – Distribuição de gastos dos parlamentares com consultoria**

**Fonte: autoria própria**

como despesas com passagens aéreas, foram obtidos elevados *scores* de anormalidade. Em alimentação o maior *outlier* recebeu 103.8256 como pontuação de anormalidade. Rastreando essa instância pelo rótulo identificador, foi possível constatar que ela é referente a um gasto de R$ 5.142,25 o que é um valor muito acima das outras instâncias encontradas nesse conjunto. Já na categoria de gastos com passagens aéreas, a maior anomalia identificada pelo algoritmo LOF recebeu *score* de 20.6327. Esse *outlier* também obteve uma pontuação alta em consequência de seu valor elevado se comparado ao restos dos dados. Essa instância trata-se de uma passagem aérea internacional que custou o total de R$ 16.137,2. Além dessas instâncias, várias outras também receberam pontuações de anormalidades elevadas como pode-se observar nas Figuras 10 e 11. Na execução real do trabalho, tais instâncias poderão ser analisadas melhor para assim ser elaborada uma conclusão mais concreta.

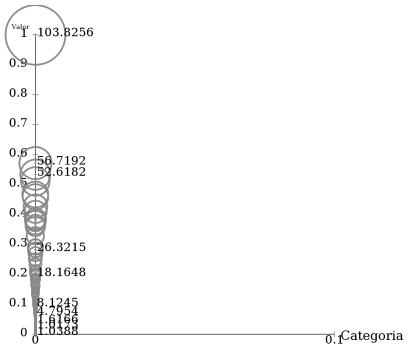



**Figura 10 – Pontuação LOF para gastos com alimentação**

**Fonte: Autoria própria**

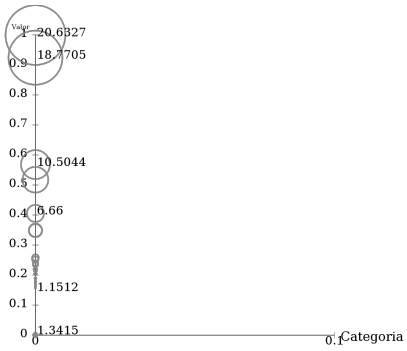



**Figura 11 – Pontuação LOF para gastos com passagens**
**Fonte: Autoria própria**

De maneira geral, este experimento teve sucesso em satisfazer seus objetivos. Foi possível identificar os principais materiais e métodos que podem ser utilizados, assim como visualizar o funcionamento da ferramenta e do algoritmo utilizado. Os testes auxiliaram também no enriquecimento do conhecimento empírico, para por exemplo, uma melhor utilização dos parâmetros tal como o parâmetro *k* utilizado pelo LOF. Também forem levantadas considerações a respeito do uso de normalização nos dados por uso de logaritmo ou raiz quadrada o que possivelmente pode ajudar com dados muito dispersos.

# 5 RESULTADOS

Neste capitulo são apresentados todos os resultados dos experimentos executados durante a etapa de aplicação das técnicas utilizadas. Para cada experimento, foram feitas discussões relevantes para auxiliar no entendimento da abordagem utilizada, assim como para interpretar as anomalias encontradas.

## 5.1 EXPERIMENTOS LOF DE NORMALIZAÇÃO E SUAVIZAÇÃO

Boa parte dos resultados deste trabalho tiveram como base a execução de vários experimentos com variações na etapa de pré-processamento. Logo, para cada experimento foram aplicadas alterações nos parâmetros e técnicas de pré-processamento dos dados. De maneira geral, as principais alterações no tratamento foram: a utilização de valores absolutos dos atributos em comparação com valores relativos à população, assim como a utilização dos dados suavizados via logaritmo decimal (isto é, com intervalo de variação reduzido) em comparação com dados normalizados por média e desvio padrão (também com variação reduzida) e por fim também normalizados e suavizados.

Com esses experimentos, foi possível observar a importância de técnicas de pré-processamentos tais como a normalização e suavização dos dados. Graças a elas foi possível identificar diferentes tipos de anomalias, sendo estas: anomalias originadas de cidades com população muito elevada, anomalias devido ao lançamento de valores de maneira errada para uma ou mais categorias de gastos, e por fim as anomalias relacionadas a gastos suspeitos nas categorias quando comparados com instâncias de população similar.

Além disso, ao término desses experimentos, foram também aplicadas diferentes técnicas para o pós-processamento dos dados gerados. O objetivo disto foi tentar facilitar a análise dos resultados, verificando quais técnicas permitem explicar melhor as anomalias encontradas. Sendo que técnica mais utilizada para auxiliar no entendimento de anomalias

relevantes a este trabalho foi a comparação com cidades de tamanho populacional similar. Isto é, para cada atributo de despesa, foi calculado o valor do centroide das 20 cidades com população próxima da cidade considerada anômala, e em seguida foi atribuído um grau de porcentagem para cada uma dessas categorias. Essa porcentagem é calculada de acordo com o valor da diferença entre a cidade anômala e o centroide das cidades vizinhas. No contexto deste trabalho, o módulo responsável por esse pós-processamento foi denominado módulo Explicador, pois possui o papel de tentar apresentar a razão pela qual uma cidade é considerada anômala, a partir de cidades semelhantes (com populações parecidas).

Os experimentos foram executados em dados públicos referentes às despesas das prefeituras no período de um ano. Esses dados utilizados foram acessados no portal do tesouro nacional na seção de contas anuais. Esse arquivo de entrada utilizado foi categorizado como "despesas por função empenhadas" do ano de 2014, onde o ano foi escolhido de maneira aleatória.

Em seguida, foram executadas diferentes técnicas de pré processamento resultando em diferentes experimentos. Entretanto, todos utilizaram o mesmo algoritmo de detecção de anomalias com os mesmo parâmetros. Isto é, o algoritmo utilizado nesses experimentos foi o LOF, sendo o parâmetro $k$ escolhido como o valor da raiz quadrada do total de instâncias. Logo, o valor utilizado foi 70, uma vez que o total de instâncias era próximo de 5000. A utilização da raiz quadrada para escolha do parâmetro $k$ em técnicas baseadas nos vizinhos mais próximos é uma boa prática, principalmente quando a metodologia utilizada tem caráter exploratório (GOLDSTEIN; DENGEL, 2012). Outro parâmetro que manteve-se o mesmo para todos experimentos foi referente a medida de similaridade, sendo esta definida como a distância euclidiana.

Apesar desses experimentos partilharem esses parâmetros durante a etapa de execução, diferentes resultados eram esperados, uma vez que, diferentes estratégias de pré-processamentos foram aplicadas. Logo, os experimentos executados nesta etapa foram:

- Experimento 1.1: LOF absoluto;
- Experimento 1.2: LOF relativo;
- Experimento 1.3: LOF suavizado ;
- Experimento 1.4: LOF relativo e normalizado;
- Experimento 1.5: LOF suavizado e normalizado.

No experimento 1.1, os valores absolutos dos atributos não foram alterados. Isto é, o algoritmo LOF foi aplicado para aprender diretamente dos valores reais dos atributos referentes aos gastos por categoria e inclusive a população real de cada cidade. Logo, neste experimento, muitas anomalias relacionadas ao tamanho da população foram encontradas. É

possível observar isto na coluna do Experimento 1.1 presente na Tabela 1, onde são mostrados os 10 maiores escores e onde observa-se que cidades grandes receberam escores elevados de anormalidade, sendo as cidades A1 e B1 as duas maiores pontuações com valores de 46,34 e 23,56, respectivamente. Considerando que os valores das despesas e da população foram mantidos de maneira absoluta, esse tipo de anomalia era esperado uma vez que o número de cidades grandes, que naturalmente possuem gastos em elevada proporção, é bem inferior ao de cidades médias e pequenas que, normalmente possuem despesas inferiores.

Por outro lado, anomalias devido a ausência de valores também receberam escores elevados. Por exemplo a instância referente a Cidade D1, que para o ano de 2014, declarou apenas valores em uma das 29 categorias existentes. Esse tipo de anomalia é mais difícil de ser tratado durante o pré-processamento, e eventualmente estará presente dentre todos experimentos como é possível observar na Tabela 1 e mesmo na tabela 2. Portanto, apesar de que essas anomalias não sejam o foco deste trabalho, elas contribuíram para avaliar e demonstrar a eficácia do algoritmo de AM utilizado.

**Tabela 1 – As dez maiores anomalias encontradas pelos experimentos 1.1, 1.2 e 1.3**

| | **Experimento 1.1** | **Experimento 1.2** | **Experimento 1.3** |
|---|---|---|---|
| 1 | 46,34 Cidade A1 | 10,03 Cidade D1 | 1,87 Cidade D1 |
| 2 | 23,56 Cidade B1 | 06,88 Cidade H1 | 1,72 Cidade P1 |
| 3 | 09,12 Cidade C1 | 06,55 Cidade J1 | 1,71 Cidade Q1 |
| 4 | 08,40 Cidade D1 | 05,85 Cidade K1 | 1,67 Cidade R1 |
| 5 | 06,30 Cidade E1 | 04,93 Cidade G1 | 1,67 Cidade S1 |
| 6 | 06,26 Cidade F1 | 04,66 Cidade L1 | 1,67 Cidade O1 |
| 7 | 06,07 Cidade G1 | 03,97 Cidade M1 | 1,67 Cidade M1 |
| 8 | 06,00 Cidade H1 | 03,77 Cidade E1 | 1,67 Cidade T1 |
| 9 | 05,23 Cidade I1 | 03,74 Cidade N1 | 1,67 Cidade L1 |
| 10 | 05,16 Cidade J1 | 03,73 Cidade O1 | 1,66 Cidade U1 |

No experimento 1.2, os valores dos atributos relativos as despesas foram divididos pela população da instância. Dessa maneira, esses atributos foram alterados para o valor do gasto por habitante. Por exemplo, se uma cidade gastou R$ 2.000.000,00 de reais na categoria saúde e possui um total de 20.000 habitantes, o gasto por habitante é de R$100,00. Logo, para este experimento não foi utilizado o atributo com o valor da população total, uma vez que esse atributo está implícito em cada um dos atributos relativos a despesas.

Portanto, os resultados desse experimento (Tabela 1, Experimento 1.2), trouxeram com enfase maior, instancias de cidades pequenas. Sendo parte dessas cidades possivelmente consideradas anormais devido a ausência ou declaração de valores em uma única categoria, tais como as cidades: D1, L1, M1 e O1. Entretanto, anomalias relevantes de uma análise mais detalhada também receberam escores elevados. Por exemplo a segunda maior anomalia desse

experimento (Cidade H1), que gastou aproximadamente 97 vezes a mais em Saneamento do que cidades com população similar. Este tipo de anomalia, condiz com o que este trabalho propôs encontrar. Assim, em casos semelhantes, seria interessante averiguar melhor a utilização do dinheiro público. Isto é, verificar se por exemplo, o saneamento básico foi apropriadamente implantado nessa cidade e se era realmente necessário um valor tao elevado para efetuar essa ação.

No experimento 1.3, tanto a população quanto os valores absolutos das despesas foram normalizados por meio do uso de logaritmo decimal. O objetivo disso, foi agrupar os valores de maneira a aproximá-los, de modo que cidades pequenas fiquem próximas umas das outras e o mesmo se aplique para cidades médias e grandes. Esse processo visa diminuir casos de anomalias devido ao tamanho da população. Portanto, como é possível ver na Tabela 1, na coluna do experimento 1.3, as anomalias devido ao tamanho da população ficaram ausentes dentre os 10 maiores escores. Por outro lado, as anomalias devido a ausência de valores predominaram. Sendo assim, após uma análise manual, verificou-se que apenas a segunda e a terceira cidades da coluna Experimento 1.3 parecem ter declarados corretamente. Entretanto, mesmos essas instâncias não declaram valores gastos na categoria saúde, e tiveram valores muito abaixo de cidades similares na categoria educação. Logo, seria necessário verificar se realmente essas cidades erraram ao declarar ou se a educação e saúde nesses municípios andavam tão bem que não foram necessários investimentos nessas áreas.

Os experimentos 1.4 e 1.5 foram elaborados a partir dos experimentos 1.2 e 1.3, respectivamente, onde foi feita a inclusão de técnica de normalização pelo uso da média zero e desvio padrão unitário. O objetivo desta técnica foi mitigar a dominância de atributos numericamente maiores sobre os demais. Por exemplo, evitar que o algoritmo de AM utilizado desse maior relevância ao atributo “Administração”, que geralmente possui gastos elevados, do que ao atributo “Defesa pública”, que geralmente não possui valores elevados.

Portanto, quando comparado os resultados desses experimentos (Tabela 2) com os resultados dos experimentos anteriores (Tabela 1), é possível observar novas instâncias. Boa parte dessas instâncias são anomalias relevantes a este trabalho e possivelmente foram desconsideradas nos outros experimentos devido a dominância de atributos. Muitas dessas instâncias receberam escores elevados por possuírem valores distantes de seus vizinhos em categorias que geralmente não possuem gastos muito elevados.

Por exemplo, ao analisar a instância referente a Cidade V1, que é a maior anomalia tanto no experimento 1.4 como no 1.5, observa-se que essa cidade gastou aproximadamente R$ 991.533,92 reais na categoria relações exteriores. Logo, considerando que esta cidade não pertence a nenhuma fronteira internacional, tratando-se de uma cidade localizada no interior do

Rio Grande do Norte e com cerca de 7 mil habitantes, está quantia é elevada para esta categoria e não parece justificada, carecendo de melhor investigação para apurar o gasto. Além disso, essa prefeitura também gastou relativamente menos em saúde e urbanismo quando comparado com cidades de população similar. E ainda não teve ou não declarou gasto algum na categoria referente ao transporte, o que também acarretou seu alto escore de anormalidade.

Ainda na Tabela 2, é possível observar a existência de instancias que também ocorreram nos experimentos da Tabela 1. Por exemplo, a instância referente a prefeitura da Cidade G1, que possivelmente foi classificada como anormal devido a possuir ausência de despesas em educação, ao mesmo tempo que, possui um elevado valor na categoria de desporto e lazer (R$ 10.110.193,70) se comparado ao valor que cidades de mesmo tamanho gastam em média (R$ 221.671.01).

**Tabela 2 – As dez maiores anomalias encontradas nos experimentos 1.4 e 1.5**

| | **Experimento 1.4** | **Experimento 1.5** |
|---|---|---|
| 1 | 61,16 Cidade V1 | 8,51 Cidade V1 |
| 2 | 13,96 Cidade X1 | 4,04 Cidade D2 |
| 3 | 12,75 Cidade W1 | 3,96 Cidade E2 |
| 4 | 11,79 Cidade Y1 | 3,87 Cidade D1 |
| 5 | 10,26 Cidade Z1 | 3,64 Cidade F2 |
| 6 | 10,17 Cidade A2 | 3,63 Cidade G2 |
| 7 | 08,73 Cidade B2 | 3,54 Cidade Q1 |
| 8 | 08,41 Cidade C2 | 3,54 Cidade G1 |
| 9 | 07,84 Cidade D1 | 3,49 Cidade H2 |
| 10 | 07,10 Cidade H1 | 3,39 Cidade I2 |

Ao verificar a qualidade da gestão fiscal dessas cidades presentes como maiores anomalias nos resultados dos experimentos 1.4 e 1.5, é possível observar que a grande maioria encontrava-se em situação de gestão fiscal com dificuldade ou crítica. Isto é, ao analisar as pontuações IFGF que as cidades presentes na Tabela 2, para o ano de 2014, constata-se que apenas as cidades D2, H2, I2 e F2 conseguiram pontuações iguais ou acima de 0,6 (Gestão boa). Entre o resto dessas prefeituras, a grande maioria conseguiu a pontuação entre 0,4 e 0,6 (Gestão em dificuldade). E algumas, tais como a Cidade V1, estavam em situação critica (IFGM menor que 0,4).

Como pode ser observado nesses dois últimos experimentos, a normalização pelo uso da média zero e desvio padrão unitário trouxe novas anomalias interessantes para este trabalho. Com isso, as técnicas de pré-processamento usadas nesses experimentos foram seguidas para todos os experimentos seguintes.

## 5.2 EXPERIMENTOS LOF PARA OS ANOS 2013 E 2015

Com base nos experimentos 1.4 e 1.5, novos experimentos foram feitos. Esses experimentos foram aplicados aos dados das despesas empenhadas das prefeituras para os anos de 2013 e 2015. Na Tabela 3 é possível observar os dez maiores escores atribuídos pelo algoritmo LOF para o ano de 2013, tanto na abordagem relativa (Experimento 2.1) quanto na suavizada (Experimento 2.2). Em ambos experimentos a Cidade V1 recebeu as maiores pontuações de anormalidade. Essa é mesma cidade que foi a maior anomalia nos experimentos anteriores (Tabela 2), porém referentes ao ano de 2014. Para o ano de 2013, o motivo dessa prefeitura receber os escores elevados foi novamente devido a gastos com relações exteriores, sendo dessa vez a quantia de R$ 471.169,43. Além disso, essa instância teve gastos inferiores de maneira geral nas outras categorias ao comparar com cidades de porte similar. Por exemplo, os gastos em saúde foram R$ 3.118.205,90, enquanto a média das 20 cidades com população igual ou aproximada foi de R$ 3.605.014,30.

Um caso de gasto anormal, porém possivelmente justificável, é instância da linha 3 no Experimento 2.1, referente a Cidade J2. O motivo da pontuação elevada (12,31) é o gasto de R$ 2.576.722,60 na categoria relacionada a defesa nacional. Entretanto, ao fazer um levantamento mais aprofundado das informações, conclui-se que essa cidade, localizada no litoral nordestino, é uma ilha e esse gasto efetuado, especificamente em defesa naval, parece ser justificado.

**Tabela 3 – As dez maiores anomalias encontradas para o ano 2013**

| | **Experimento 2.1** | **Experimento 2.2** |
|---|---|---|
| 1 | 47,52 Cidade V1 | 6,94 Cidade V1 |
| 2 | 27,10 Cidade X1 | 5,60 Cidade M2 |
| 3 | 12,31 Cidade J2 | 5,60 Cidade P2 |
| 4 | 11,93 Cidade K2 | 5,46 Cidade Q2 |
| 5 | 11,15 Cidade Z1 | 4,63 Cidade R2 |
| 6 | 10,89 Cidade L2 | 3,96 Cidade A1 |
| 7 | 10,55 Cidade M2 | 3,73 Cidade S2 |
| 8 | 10,41 Cidade N2 | 3,71 Cidade T2 |
| 9 | 09,51 Cidade O2 | 3,48 Cidade E2 |
| 10 | 09,23 Cidade P2 | 3,43 Cidade I2 |

Dentre os maiores escores do experimento 2.2 da Tabela 3, é possível identificar anomalias relativas ao tamanho da cidade. Por exemplo, as cidades A1 (linha 6), D2 (linha 9) e I2 (linha 10). Além disso, um novo tipo de anomalia foi identificado durante esse

experimento. Esse novo tipo de anomalia refere-se a cidades que tiveram gastos negativos. Isto é, possivelmente fizeram devolução de verba pública para uma ou mais categorias. Isso foi identificado devido a instância referente a Cidade T2, presente na linha 8 da coluna referente ao Experimento 2.2, que declarou ter gasto R$ 1.097.168,81 negativos em assistência social.

Verificando a situação da gestão fiscal das cidades presentes na Tabela 3 para o respectivo ano (2013), é possível observar que apenas as cidades A1, L2, R2, D2 e I2 encontravam-se em situação fiscal considerada boa (IFGF entre 0,6 e 0,8). As outras cidades receberam pontuações relativas a gestão em dificuldade ou crítica. Portanto, os gastos em que essas cidades saíram muito do padrão, não parecem ser justificáveis.

A Tabela 4, mostra os resultados dos experimentos executados para o ano de 2015. Ou seja, nessa tabela são exibidas as cidades com maior pontuação de anormalidade segundo o LOF. A primeira coluna, referente ao Experimento 2.3 mostra as maiores anomalias identificadas ao seguir o tratamento de dados de maneira relativa à população. A segunda coluna, nomeada como Experimento 2.4, trata-se dos resultados ao utilizar a abordagem de suavização dos dados. Ao observar essa tabela nota-se que ambas abordagens resultaram na Cidade U2, como a maior anomalia. Ao verificar melhor os gastos e informações dessa cidade, constatou-se que para uma cidade de aproximadamente 11.500 habitantes, os gastos com previdência social foram de R$ 1.378.211,16 a mais que cidades de mesmo tamanho populacional. Além disso, essa prefeitura gastou R$ 780.986,79 em relações exteriores, o que possivelmente acarretou o escore de anormalidade tão elevado. Considerando a localização geográfica dessa cidade, assim como a baixa população, essa anomalia é mais um caso que seria interessante em ser apurado para verificar a justificativa desses gastos.

Outro exemplo de anomalia que instiga uma averiguação do uso das verbas públicas é a instância referente a Cidade G1, presente na oitava linha da coluna do experimento 2.4. Essa prefeitura já havia aparecido, em experimentos do ano de 2014, entre as dez maiores anomalias devido a gastos elevados em desporto e lazer. Novamente, um gasto elevado nessa função (R$ 13.115.620,14), se comparado ao gasto médio de cidades com tamanho próximo (R$ 195.391,39), foi a possível causa dessa instância estar dentre as dez maiores anomalias. Entretanto, ao levar em consideração a elevada pontuação IFGF que essa cidade recebeu durante esse ano, esses gastos elevados podem ser de interesse para a população dessa cidade.

Ainda na Tabela 4, é possível observar alguns casos de anomalias devido a maneira errada de declarar os dados. Esses casos são refentes as instâncias das cidades D1 , F3 e D3 . A prefeitura da Cidade D1 declarou apenas gastos como despesas intraorçamentarias, a Cidade F3 declarou todos gastos na categoria referente a administração e a Cidade D3 declarou gastos em apenas duas funções, sendo estas: despesas intraorçamentarias e saúde. Entre as outras

**Tabela 4 – As dez maiores anomalias encontradas para o ano 2015**

| | Experimento 2.3 | Experimento 2.4 |
|---|---|---|
| 1 | 40,01 Cidade U2 | 5,79 Cidade U2 |
| 2 | 17,68 Cidade V2 | 4,71 Cidade A3 |
| 3 | 12,97 Cidade W2 | 4,35 Cidade B3 |
| 4 | 10,39 Cidade W1 | 4,33 Cidade C3 |
| 5 | 10,12 Cidade X1 | 4,06 Cidade D1 |
| 6 | 09,33 Cidade Z1 | 3,87 Cidade D3 |
| 7 | 08,11 Cidade Y2 | 3,47 Cidade E3 |
| 8 | 07,61 Cidade X2 | 3,44 Cidade G1 |
| 9 | 07,59 Cidade D1 | 3,40 Cidade E2 |
| 10 | 07,51 Cidade Z2 | 3,23 Cidade F3 |

cidades presente nessa tabela, que declaram os gastos de maneira correta, apenas três receberam o conceito B ao verificar seus índices FIRJAN de gestão fiscal para o ano de 2015. Isto é, apenas as prefeituras das cidades Z2, G1 e D2 receberam uma pontuação suficiente no IFGF para terem sua gestão considerada boa. Grande parte das outras instâncias ficaram com conceito C, que as qualifica como gestão em dificuldade, e as prefeituras de C3 e W1 foram consideras em gestão crítica devido aos baixos índices FIRJAN que receberam.

## 5.3 EXPERIMENTOS LOF BIÊNIOS E TRIÊNIO

Os resultados dos experimentos discutidos nas seções 5.1 e 5.2 demonstraram a aplicação do algoritmo de aprendizado de máquina LOF em detectar anomalias nos gastos das prefeituras durante o período de um ano. Alternativamente, uma abordagem que pode ser tomada é aplicar esse algoritmo para identificar anomalias nos gastos das prefeituras ao longo de 2 ou mais anos seguidos. Com isso, é esperado identificar como anomalias quaisquer prefeituras que desviarem-se dos padrões da série de tempo criada. Para executar essa abordagem, basta trazer todas as categorias de todos anos que farão parte da entrada como atributos das instâncias. Com isso, a dimensionalidade dos dados é aumentada em $n$ vezes, conforme o número de anos a serem considerado.

Os experimentos foram executados nos biênios referentes a 2013-2014 , 2014-2015 e no triênio 2013-2014-2015. Os resultados das maiores pontuações LOF para o primeiro biênio mencionado podem ser observados na Tabela 5. Já os resultados do segundo biênio (2014-2015) pode ser analisados na Tabela 6. E os resultados das dez maiores anomalias durante o

triênio estão presentes na Tabela 7. Para cada uma dessas tabelas, as colunas dos experimentos à esquerda, trazem as dez maiores anomalias ao utilizar os gastos das prefeituras relativo ao atributo população (gasto por pessoa). As colunas da direita referem-se as dez maiores anomalias ao utilizar a transformação de dados de maneira a suavizar esses dados e manter a população como um atributo próprio e não relativo aos gastos.

**Tabela 5 – Resultados LOF Biênio 2013-2014**

| | **Experimento 3.1** | **Experimento 3.2** |
|---|---|---|
| 1 | 49,95 Cidade V1 | 6,86 Cidade V1 |
| 2 | 19,25 Cidade X1 | 4,48 Cidade Q2 |
| 3 | 09,64 Cidade K2 | 4,19 Cidade M2 |
| 4 | 09,49 Cidade W1 | 3,50 Cidade E2 |
| 5 | 08,84 Cidade J2 | 3,39 Cidade A1 |
| 6 | 08,16 Cidade A2 | 3,31 Cidade P2 |
| 7 | 08,11 Cidade P2 | 3,22 Cidade I2 |
| 8 | 08,08 Cidade L2 | 3,22 Cidade R2 |
| 9 | 08,08 Cidade Y1 | 3,09 Cidade H2 |
| 10 | 08,02 Cidade M2 | 3,09 Cidade Q1 |

Ao comparar os resultados da Tabela 5 (2013-2014) com os resultados das Tabelas 3 (2013) e 2 (2014), que referem-se aos mesmos anos, é possível perceber que as maiores anomalias trazidas pelos experimentos no biênio estão presentes, não necessariamente na mesma ordem, nos experimentos executados nesses respectivos anos. O melhor exemplo disso, é a anomalia de maior pontuação compartilhada por todas essas tabelas. Essa anomalia refere-se a instância da prefeitura da Cidade V1. Essa cidade teve gastos elevados e contínuos em relações exteriores durante esse biênio. Com isso, ela garantiu sua colocação em primeiro lugar como a maior pontuação de anormalidade nesse experimento, referente a série temporal desses anos.

De maneira similar aos resultados obtidos na Tabela 5, os resultados da Tabela 6, referentes ao biênio 2014-2015, compartilharam boa parte dos resultados obtidos nos experimentos aplicados aos anos 2014 (Tabela 2) e 2015 (Tabela 4). A instância com maior pontuação de anormalidade em ambos experimentos aplicados nesse biênio foi a Cidade U2. Os motivos disso foram gastos elevados em relações exteriores e previdência social no ano de 2015, discutidos nos resultados da Tabela 4 na seção 5.2. Além disso, essa prefeitura também teve gastos elevados (R$1.865.309,64) em despesas intraorçamentarias no ano de 2014, quando comparado a média (R$72.018,69) de cidades de tamanho similar.

O último experimento executado nessa abordagem utilizando séries temporais, foi a aplicação do algoritmo LOF nos dados referentes ao triênio 2013 a 2015. De maneira análoga aos experimentos anteriores, os resultados esperados desse experimento deveriam refletir os

**Tabela 6 – Resultados LOF Biênio 2014-2015**

| | Experimento 3.3 | Experimento 3.4 |
|---|---|---|
| 1 | 25,59 Cidade U2 | 4,35 Cidade U2 |
| 2 | 13,02 Cidade V2 | 4,08 Cidade D3 |
| 3 | 11,02 Cidade X1 | 3,96 Cidade D1 |
| 4 | 10,74 Cidade W1 | 3,57 Cidade A3 |
| 5 | 09,04 Cidade Z1 | 3,55 Cidade E2 |
| 6 | 08,53 Cidade Y1 | 3,4 Cidade C3 |
| 7 | 08,19 Cidade W2 | 3,33 Cidade B3 |
| 8 | 07,73 Cidade A2 | 3,31 Cidade H3 |
| 9 | 07,47 Cidade D1 | 3,28 Cidade G1 |
| 10 | 06,79 Cidade G3 | 3,26 Cidade G2 |

resultados dos anos que compõem esse conjunto. De fato, ao observar a Tabela 7, que mostra as dez maiores anomalias para ambos experimentos executados nesse ano, é possível identificar anomalias que estão presente entre as tabelas referentes a esses mesmos anos (individualmente). Por exemplo, a instância de Cidade X1, presente na coluna do experimento 3.5, linha 2. Essa prefeitura apareceu tanto nos dez primeiros resultados do experimento 3.3 (Tabela 6), como nos resultados do experimento 3.1 (Tabela 5). Ao analisar os gastos dessa prefeitura ao longo desse triênio, nota-se que ela gastou aproximadamente R$3.670.00,00 a mais na categoria urbanismo do que cidades com população aproximada. Além disso, de maneira geral essa cidade ainda teve vários gastos relativamente baixos ao longo desse período, sendo as categorias que mostraram as maiores diferenças: saúde, transporte, assistência social e administração.

**Tabela 7 – Resultados LOF Triênio 2013 à 2015**

| | Experimento 3.5 | Experimento 3.6 |
|---|---|---|
| 1 | 20,70 Cidade U2 | 3,70 Cidade Q2 |
| 2 | 15,05 Cidade X1 | 3,53 Cidade U2 |
| 3 | 11,84 Cidade V2 | 3,52 Cidade M2 |
| 4 | 09,32 Cidade W1 | 3,49 Cidade D3 |
| 5 | 08,32 Cidade K2 | 3,40 Cidade E2 |
| 6 | 07,84 Cidade L2 | 3,31 Cidade A1 |
| 7 | 07,75 Cidade J2 | 3,09 Cidade H2 |
| 8 | 07,18 Cidade M2 | 3,09 Cidade I2 |
| 9 | 07,11 Cidade Y1 | 3,07 Cidade H3 |
| 10 | 06,86 Cidade W2 | 3,02 Cidade I3 |

Apesar dos experimentos utilizando um período de tempo maior que um ano terem mostrado resultados equivalentes aos da abordagem inicial (único ano), novas anomalias não foram encontradas, ou pelo menos não receberam escores elevados o suficiente para serem destacadas. Além disso, dependendo da maneira com que os dados forem tratados, algumas anomalias deixaram de existir. Por exemplo, durante esses experimentos, prefeituras que não

haviam declarado os gastos para um dos anos pertencentes ao biênio ou triênio, não foram consideradas. Em consequência disso, a maior anomalia presente nos resultados dos anos 2013 e 2014 (Cidade V1), sequer apareceu entre os resultados do biênio e triênio em que o ano 2015 foi levado em consideração. O motivo disso, é que essa prefeitura não havia declarado os gastos referentes ao ano de 2015. Outra desvantagem ao utilizar essa abordagem é o crescimento da dimensionalidade do problema. Isto é, ao trabalhar com um único ano existem 29 atributos referentes as categorias de gastos, enquanto ao processar um biênio e um triênio, o numero desses atributos será 58 e 87 respectivamente. Com isso, o processo de análise e verificação de quais atributos influenciaram mais no escore de anormalidade elevado da instância passa a ser bem mais trabalhoso.

## 5.4 EXPERIMENTOS COM OUTROS ALGORITMOS

O algoritmo de aprendizado de máquina LOF, mostrou-se eficiente para ao identificar diferentes tipos de anomalias como pode ser visto nos resultados das seções anteriores. Entretanto, com o objetivo de validar, verificar e comparar os resultados obtidos pelo LOF com outros algoritmos de AM, novos experimentos foram executados utilizando com finalidade de identificar anomalias. Alguns critérios para escolha destes algoritmos foram considerados. A principal consideração foi tentar trazer pelo menos um algoritmo de AM de cada uma das categorias das técnicas de detecção de anomalias detalhadas no Capítulo 2. Além disso, a disponibilidade desses algoritmos na ferramenta KDD utilizada (ELKI) e a capacidade desses algoritmos em atribuir escores de anormalidade também foram fatores relevantes para a escolha. Dessa maneira, foi possível comparar as dez maiores anomalias de cada um desses algoritmos entre si e com o LOF. Portanto, os algoritmos de AM utilizados para esses experimentos foram:

- Modelo gaussiano: técnica estatística;
- Support Vector Machine (SVM): técnica de classificação;
- Correlation Outlier Probability (COP): técnica espectral;
- Ordering Points to Identify the Clustering Structure with Outlier Factor (OPTICSOF): técnica de agrupamento.

Esse algoritmos foram aplicados aos dados referentes ao gastos das prefeituras para o ano de 2014. Esses mesmos dados, foram utilizados pelos experimentos discutidos na seção 5.1. Com isso, seus resultados podem ser comparados ao resultados já obtidos pelo LOF. Dentre

esses algoritmos, o OPTICSOF e o COP, que fazem uso do parâmetro *k* e de uma função de distância para cálculo de um escore, utilizaram *k* com valor igual a 70 e a distância euclidiana como função de dissimilaridade. Além disso, o algoritmo COP utilizou o PCARunner como o parâmetro de PCA, e o SVM fez uso de RBF como *kernel*.

A Tabela 8, mostra as dez cidades que cada um desses algoritmos atribui como maiores anomalias no experimento utilizando os dados de 2014 com os gastos relativos por pessoa. Todos os resultados estão ordenados pela maior pontuação de anormalidade, mas observa-se que no algoritmo COP, há muitos empates, de modo que todas as cidades listadas compartilham a primeira posição (escore 1,00) .

Ao observar essa tabela e ao mesmo tempo comparar os resultados desses algoritmos entre si e com os resultados do LOF, exibidos na Tabela 2 (coluna do experimento 1.4), é possível perceber uma boa concordância entre esses algoritmos quando se diz respeito as dez maiores anomalias encontradas. Alguns desses algoritmos concordam até mesmo em boa parte da ordem dos resultados, por exemplo ao comparar os resultados do LOF (Tabela 2) com os do OPTICS-OF (Tabela 8). A ordem de resultados desses dois algoritmos é a mesma para os 10 resultados. Uma possível explicação para isso, é a similaridade na abordagem de aprendizado dessas técnicas, pois ambas utilizam *k* vizinhos mais próximos, uma medida de similaridade (distância euclidiana), e funções de densidade para atribuir um fator local.

**Tabela 8 – Dez maiores anomalias com os outros algoritmos de AM usando abordagem relativa**

| | **Modelo Gaussiano** | **SVM** | **COP** | **OPTICS-OF** |
|---|---|---|---|---|
| 1 | Cidade V1 | Cidade V1 | Cidade D1 | Cidade V1 |
| 2 | Cidade Z1 | Cidade W1 | Cidade P1 | Cidade X1 |
| 3 | Cidade X1 | Cidade X1 | Cidade E2 | Cidade W1 |
| 4 | Cidade J3 | Cidade Z1 | Cidade M3 | Cidade Y1 |
| 5 | Cidade A2 | Cidade A2 | Cidade V1 | Cidade Z1 |
| 6 | Cidade W1 | Cidade G3 | Cidade G1 | Cidade A2 |
| 7 | Cidade G3 | Cidade J3 | Cidade H2 | Cidade B2 |
| 8 | Cidade K1 | Cidade K3 | Cidade N3 | Cidade C2 |
| 9 | Cidade D1 | Cidade L3 | Cidade D2 | Cidade D1 |
| 10 | Cidade Y1 | Cidade Y1 | Cidade A1 | Cidade H1 |

A Tabela 9 mostra as dez maiores anomalias encontradas por esses algoritmos nos mesmo dados, porém, com o pré-processamento feito de maneira a manter a população como atributo e suavizar todos valores dos atributos pelo logaritmo decimal, antes de normalizar pela média e desvio padrão. Portanto, os resultados das colunas dessa tabela também podem ser comparados entre si, assim como podem ser comparados com os resultados do LOF na coluna do Experimento 1.5 (Tabela 2). Mesmo com a suavização dos dados, devido ao limite de pontuação máximo do algoritmo COP (1,00), muitas prefeituras também alcançaram nesse

experimento essa pontuação. Portanto, os resultados da coluna desse algoritmo, novamente, representam dez das maiores anomalias sem a necessariamente representar uma ordem de pontuação. Já os resultados das outras colunas, apesar de terem concordado menos na ordem das anomalias quando comparado ao experimento anterior, tiveram um alto grau de coincidência entre as dez maiores anomalias. Por exemplo, as instâncias referentes as cidades D1, D2, V1 e H2 estão presentes entre as dez maiores anomalias em todos algoritmos mostrados na Tabela 9, assim como na coluna referente ao experimento 1.5, com os resultados do LOF nessa mesma abordagem, da Tabela 2.

**Tabela 9 – Dez maiores anomalias com os outros algoritmos de AM usando suavização**

| | **Modelo Gaussiano** | **SVM** | **COP** | **OPTICS-OF** |
|---|---|---|---|---|
| 1 | Cidade D1 | Cidade E2 | Cidade D1 | Cidade V1 |
| 2 | Cidade G2 | Cidade H2 | Cidade P1 | Cidade D2 |
| 3 | Cidade P1 | Cidade A1 | Cidade E2 | Cidade E2 |
| 4 | Cidade E2 | Cidade V1 | Cidade M3 | Cidade D1 |
| 5 | Cidade M3 | Cidade D1 | Cidade V1 | Cidade F2 |
| 6 | Cidade V1 | Cidade G2 | Cidade G1 | Cidade G2 |
| 7 | Cidade G1 | Cidade O3 | Cidade H2 | Cidade G1 |
| 8 | Cidade H2 | Cidade I2 | Cidade N3 | Cidade Q1 |
| 9 | Cidade N3 | Cidade O1 | Cidade D2 | Cidade H2 |
| 10 | Cidade D2 | Cidade L1 | Cidade A1 | Cidade I2 |

Com os resultados desses experimentos, ficou claro que as diferenças na identificação de anomalias por parte de diferentes algoritmos são poucas. Além disso, entre os algoritmos utilizados, todos partilharam em atribuir os escores que caracterizavam anomalias em suas abordagens as mesmas instâncias que os outros algoritmos também consideraram anômalas. Isto é, todos os algoritmos utilizados, aparentemente, identificaram as mesmas anomalias, entretanto com ordenações um pouco variadas. Com isso, para este domínio de aplicação, fazer o uso de algoritmos baseados nos vizinhos mais próximos, tais como o LOF, parece ser vantajoso, uma vez que esse tipo de técnica trabalha bem de maneira não supervisionada e é considerada fácil de adaptar aos dados (CHANDOLA et al., 2009).

Portanto, o fator que acabou mais influenciando na diferenciação da detecção de anomalias, como pode ser visto em praticamente todas tabelas dos resultados, foi a abordagem de pré-processamento utilizada. Ambas abordagens, relativa a população com normalização e suavizada com normalização, trouxeram resultados interessantes para este trabalho. Entretanto, a utilização dos valores relativos população destacou mais os casos de anomalias com gastos suspeitos, enquanto a técnica da suavização dos dados deu mais evidência as anomalias relativas as cidades com grande população ou devido a declaração dos gastos de maneira incorreta.

# 6 CONSIDERAÇÕES FINAIS

Nesse capítulo são descritas as conclusões finais a respeito do trabalho e de sua utilização. Além disso, é evidenciado a maneira com que este trabalho pode contribuir no domínio em que foi aplicado. Por fim, são propostas recomendações para qualquer um que queira dar continuidade a este trabalho.

## 6.1 CONCLUSÕES

Este trabalho propôs uma abordagem, utilizando técnicas do Aprendizado de Máquina e da extração de conhecimentos, para identificar possíveis anomalias nos gastos das prefeituras. Isto é, tentar identificar o uso do dinheiro público de maneira suspeita, ou fora do padrão. Com os resultados, foi possível perceber diferentes tipos de anomalias, algumas não tão relevantes ou fáceis de interpretar, já outras, bem interessantes de se fazer uma análise mais aprofundada. Com isso, é possível observar o grande potencial que o aprendizado de máquina pode oferecer nessa área de aplicação. O algoritmo utilizado (LOF), mostrou-se eficiente em identificar anomalias nos gastos do dinheiro público. Isto implica que todo esse processo em identificar gastos fora do padrão pode ser automatizado, e que quaisquer dados públicos podem gerar informações úteis, considerando análises adicionais para validar suspeitas.

O método proposto é, portanto, útil em vários cenários. Por exemplo, é possível verificar se uma cidade com gastos baixos em educação está apresentando índices satisfatórios no Exame Nacional do Ensino Médio (ENEM). Caso isso não aconteça, a cidade não está investindo adequadamente em educação. De modo análogo, indicadores de saúde podem ser analisados para cidades anômalas com baixo investimento em saúde. Convém observar que o outro lado do espectro também permite análises úteis. Por exemplo, uma cidade anômala com investimento em educação relativamente baixo, mas com alto desempenho no ENEM. Candidatos podem utilizar essa informação de gestões anteriores como um indicador de gestão

eficiente em suas campanhas. Dessa maneira, essa abordagem pode se tornar uma ferramenta poderosa para auxiliar tanto a auditorias como a qualquer cidadão que queira fiscalizar o uso do dinheiro público.

## 6.2 TRABALHOS FUTUROS

Para trabalhos futuros recomenda-se:

- Aplicar o método proposto nos dados referentes aos gastos das prefeituras nos anos de 2004 à 2012;
- Aplicar o método proposto em outros dados públicos, tais como: gastos dos parlamentares, despesas da União e dos estados;
- Publicar os resultados em um portal Web de maneira similar as pontuações de gestão do IFGF;
- Incluir nas instâncias novos atributos que possam ser relevantes ao algoritmo de AM, tais como o Índice de Desenvolvimento Humano(IDH), o IFGF, entre outros dados relevantes;
- Fazer o uso de outros algoritmos de AM, e comparar os resultados já obtidos;

# REFERÊNCIAS


ACHTERT, E.; KRIEGEL, H.-P.; ZIMEK, A. Elki: a software system for evaluation of subspace clustering algorithms. In: SPRINGER. **International Conference on Scientific and Statistical Database Management**. [S.l.], 2008. p. 580–585.

AGGARWAL, C. C.; YU, P. S. Outlier detection for high dimensional data. In: ACM. **ACM Sigmod Record**. [S.l.], 2001. v. 30, n. 2, p. 37–46.

AGOVIC, A.; BANERJEE, A.; GANGULY, A. R.; PROTOPOPESCU, V. 6 anomaly detection in transportation corridors using manifold embedding. **Knowledge Discovery from Sensor Data**, CRC Press, p. 81–105, 2008.

AGYEMANG, M.; BARKER, K.; ALHAJJ, R. A comprehensive survey of numeric and symbolic outlier mining techniques. **Intelligent Data Analysis**, IOS Press, v. 10, n. 6, p. 521–538, 2006.

AMER, M.; GOLDSTEIN, M.; ABDENNADHER, S. Enhancing one-class support vector machines for unsupervised anomaly detection. In: ACM. **Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description**. [S.l.], 2013. p. 8–15.

ANSCOMBE, F. J. Rejection of outliers. **Technometrics**, Taylor & Francis Group, v. 2, n. 2, p. 123–146, 1960.

BAMNETT, V.; LEWIS, T. Outliers in statistical data. JSTOR, 1994.

BERRY, M. J.; LINOFF, G. **Data mining techniques: for marketing, sales, and customer support**. [S.l.]: John Wiley & Sons, Inc., 1997.

BREUNIG, M. M.; KRIEGEL, H.-P.; NG, R. T.; SANDER, J. Optics-of: Identifying local outliers. In: SPRINGER. **European Conference on Principles of Data Mining and Knowledge Discovery**. [S.l.], 1999. p. 262–270.

BREUNIG, M. M.; KRIEGEL, H.-P.; NG, R. T.; SANDER, J. Lof: identifying density-based local outliers. In: ACM. **ACM sigmod record**. [S.l.], 2000. v. 29, n. 2, p. 93–104.

BRIGADE, D. S. **Operação Serenata de Amor**. 2017. Disponível em: <https://serenatadeamor.org/>.

BYERS, S.; RAFTERY, A. E. Nearest-neighbor clutter removal for estimating features in spatial point processes. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 93, n. 442, p. 577–584, 1998.

CARVALHO, J. M. D. **Cidadania no Brasil**. [S.l.]: Civilização Brasileira, 2001.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM computing surveys (CSUR)**, 2009.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. [S.l.]: John Wiley & Sons, 2012.

DUTTA, H.; GIANNELLA, C.; BORNE, K. D.; KARGUPTA, H. Distributed top-k outlier detection from astronomy catalogs using the demac system. In: SIAM. **SDM**. [S.l.], 2007. p. 473–478.

ELKI. **ELKI: Environment for Developing KDD-Applications Supported by Index-Structures**. 2017. Disponível em: <https://elki-project.github.io/>.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Kdd**. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

FIRJAN. do rio de janeiro (2016). **IFGF–Índice FIRJAN de Gestão Fiscal**, 2016.

GOLDSTEIN, M.; DENGEL, A. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. **KI-2012: Poster and Demo Track**, Citeseer, p. 59–63, 2012.

GOLUB, G. H.; LOAN, C. F. V. **Matrix computations**. [S.l.]: JHU Press, 2012.

GUHA, S.; MISHRA, N.; MOTWANI, R.; O'CALLAGHAN, L. Clustering data streams. In: IEEE. **Foundations of computer science, 2000. proceedings. 41st annual symposium on**. [S.l.], 2000. p. 359–366.

GUTTORMSSON, S. E.; MARKS, R.; EL-SHARKAWI, M.; KERSZENBAUM, I. Elliptical novelty grouping for on-line short-turn detection of excited running rotors. **IEEE Transactions on Energy Conversion**, IEEE, v. 14, n. 1, p. 16–22, 1999.

HAWKINS, D. M. **Identification of outliers**. [S.l.]: Springer, 1980.

HE, Z.; XU, X.; DENG, S. Discovering cluster-based local outliers. **Pattern Recognition Letters**, Elsevier, v. 24, n. 9, p. 1641–1650, 2003.

HODGE, V. J.; AUSTIN, J. A survey of outlier detection methodologies. **Artificial intelligence review**, Springer, v. 22, n. 2, p. 85–126, 2004.

IDÉ, T.; KASHIMA, H. Eigenspace-based anomaly detection in computer systems. In: ACM. **Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2004. p. 440–449.

KOHAVI, R.; PROVOST, F. Glossary of terms. **Machine Learning**, v. 30, n. 2-3, p. 271–274, 1998.

KRIEGEL, H.-P.; KROGER, P.; SCHUBERT, E.; ZIMEK, A. Outlier detection in arbitrarily oriented subspaces. In: IEEE. **Data Mining (ICDM), 2012 IEEE 12th International Conference on**. [S.l.], 2012. p. 379–388.

LAZAREVIC, A.; ERTÖZ, L.; KUMAR, V.; OZGUR, A.; SRIVASTAVA, J. A comparative study of anomaly detection schemes in network intrusion detection. In: SIAM. **SDM**. [S.l.], 2003. p. 25–36.

LEE, W.; XIANG, D. Information-theoretic measures for anomaly detection. In: IEEE. **Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on**. [S.l.], 2001. p. 130–143.

LORENA, A. C.; CARVALHO, A. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007.

MITCHELL, T. M. Machine learning. 1997. **Burr Ridge, IL: McGraw Hill**, v. 45, p. 37, 1997.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas Inteligentes-Fundamentos e Aplicações**, v. 1, n. 1, 2003.

PATCHA, A.; PARK, J.-M. An overview of anomaly detection techniques: Existing solutions and latest technological trends. **Computer networks**, Elsevier, v. 51, n. 12, p. 3448–3470, 2007.

RAMASWAMY, S.; RASTOGI, R.; SHIM, K. Efficient algorithms for mining outliers from large data sets. In: ACM. **ACM SIGMOD Record**. [S.l.], 2000. v. 29, n. 2, p. 427–438.

RUSSELL, S. J.; NORVIG, P.; CANNY, J. F.; MALIK, J. M.; EDWARDS, D. D. **Artificial intelligence: a modern approach**. [S.l.]: Prentice hall Upper Saddle River, 2003.

SHANNON, C. E. A mathematical theory of communication. **ACM SIGMOBILE Mobile Computing and Communications Review**, ACM, v. 5, n. 1, p. 3–55, 2001.

SHEWHART, W. A. **Economic control of quality of manufactured product**. [S.l.]: ASQ Quality Press, 1931.

SILVA, K. da; FLACH, L. Ranking de corrupção e fraudes ocorridas no brasil entre 1999 e 2012. 2013.

SMITH, R.; BIVENS, A.; EMBRECHTS, M.; PALAGIRI, C.; SZYMANSKI, B. Clustering approaches for anomaly based intrusion detection. **Proceedings of intelligent engineering systems through artificial neural networks**, p. 579–584, 2002.

STEFANO, C. D.; SANSONE, C.; VENTO, M. To reject or not to reject: that is the question-an answer in case of neural classifiers. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, IEEE, v. 30, n. 1, p. 84–94, 2000.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Data mining cluster analysis: Basic concepts and algorithms**. 2013.

Transparency International. **Corruption Perceptions Index 2012**. [S.l.], 2013.

Transparency International. **Corruption Perceptions Index 2015**. [S.l.], 2016.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2016.

YU, D.; SHEIKHOLESLAMI, G.; ZHANG, A. Findout: finding outliers in very large datasets. **Knowledge and Information Systems**, Springer, v. 4, n. 4, p. 387–412, 2002.