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

MARCOS AURÉLIO VIEIRA

DETECÇÃO DE ANOMALIAS EM DADOS DA ADMINISTRAÇÃO PÚBLICA UTILIZANDO TÉCNICA DE APRENDIZADO DE MÁQUINA

TRABALHO DE CONCLUSÃO DE CURSO

MEDIANEIRA

2019

MARCOS AURÉLIO VIEIRA

DETECÇÃO DE ANOMALIAS EM DADOS DA ADMINISTRAÇÃO PÚBLICA UTILIZANDO TÉCNICA DE APRENDIZADO DE MÁQUINA

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 Computação".

Orientador: Prof. Dr. Evando Carlos Pessini

Co-orientador: Prof. Dr. Arnaldo Candido Junior

MEDIANEIRA

2019

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

DETECÇÃO DE ANOMALIAS EM DADOS DA ADMINISTRAÇÃO PÚBLICA UTILIZANDO TÉCNICA DE APRENDIZADO DE MÁQUINA

Por

MARCOS AURÉLIO VIEIRA

Este Trabalho de Conclusão de Curso foi apresentado às 11:10h do dia 8 de julho de 2019 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. Evando Carlos Pessini
UTFPR - Câmpus Medianeira

Prof. Dr. Alan Gavioli
UTFPR - Câmpus Medianeira

Prof. Msc. Fernando Schütz
UTFPR - Câmpus Medianeira

Prof. Msc. Jorge Aikes Junior
UTFPR - Câmpus Medianeira

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

RESUMO

VIEIRA, Marcos Aurélio. DETECÇÃO DE ANOMALIAS EM DADOS DA ADMINISTRAÇÃO PÚBLICA UTILIZANDO TÉCNICA DE APRENDIZADO DE MÁQUINA . 42 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2019.

O uso de aprendizado de máquina (AM) tem sido utilizado em diversas áreas da sociedade e organizações. Com isso diversas aplicações e descobertas tem surgido dessa área. O objetivo desse trabalho é encontrar anomalias em dados referente a gastos fornecidos pelas prefeituras utilizando técnicas de aprendizado de máquina e mineração de dados. Para o desenvolvimento desse trabalho, é utilizado a linguagem de programação Python,as bibliotecas, Pandas, NumPy, SciKit-learn. A primeira etapa realizada foi selecionar os dados que tenham algum tipo de correlação, assim optou-se por utilizar os dados do Sistema de Informação Contábil e Fiscal (SICONFI) do portal do Tesouro Nacional, mais especificamente os dados referente a gastos das prefeituras com educação, saúde, e segurança pública, entre outros atributos. Em seguida foram unificadas com dados do Índice Firjan de Gestão Fiscal (IFGF), Índice Firjan de Desenvolvimento Municipal (IFDM), dados do Departamento de Informática do Sistema Único de Saúde (DataSUS). Com essa nova proposta, novas anomalias de interesse foram encontradas. Outras cidades que, no experimento apenas com a base do SICONFI, eram consideradas como anomalias, nessa nova abordagem, passaram a não ser, e nos 10 casos analisados, o gasto dessas cidades eram justificados pelos índices propostos.

Palavras-chave: aprendizado de máquina, mineração de dados, descoberta do conhecimento em base de dados

ABSTRACT

VIEIRA, Marcos Aurélio. DETECTION OF ANOMALIES IN PUBLIC ADMINISTRATION DATA USING MACHINE LEARNING TECHNIQUE. 42 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2019.

The use of Machine Learning (ML) has been used in several areas of society and organizations. With this several applications and discoveries have arisen from this area. The objective of this work is to find data anomalies related to expenditures provided by municipalities using techniques of machine learning and data mining. For the development of this work, the Python programming language, libraries, Pandas, NumPy, SciKit-learn is used. The first step was to select the data that have some type of correlation, so we chose to use the data from the Accounting and Fiscal Information System (SICONFI) of the National Treasury portal, more specifically data on municipalities' spending on education , health, and public safety, among other attributes. They were then unified with data from the Firjan Index of Fiscal Management (IFGF), Firjan Municipal Development Index (IFDM), data from the Department of Information Technology of the Unified Health System (DataSUS). With this new proposal, new anomalies of interest were found. Other cities that, in the experiment with only the SICONFI base, were considered anomalies in this new approach, except in the ten cases analyzed, the expenditure of these cities was justified by the proposed indexes.

Keywords: machine learning, data mining, kdd

LISTA DE FIGURAS

LISTA DE SIGLAS

AM	Aprendizado de Máquina
CEAP	Cota para Exercício da Atividade Parlamentar
CID	Classificação Internacional de Doenças
DATASUS	Departamento de Informática do Sistema Único de Saúde
ENEM	Exame Nacional do Ensino Médio
FIRJAN	Federação das Indústrias do Estado do Rio de Janeiro
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatísticas
IDEB	Índice de Desenvolvimento da Educação Básica
IFDM	Índice FIRJAN de Desenvolvimento Municipal
IFGF	Índice FIRJAN de Gestão Fiscal
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IPC	Índice de Percepção da Corrupção
LOF	Local Outlier Factor
ONGs	Organizações não Governamentais
RNA	Rede Neural Artificial
SINCOFI	Sistema de Informações Contábeis e Fiscais

SUMÁRIO

1 INTRODUÇÃO

Conforme a Transparency International (2019), o Brasil caiu 9 posições no Índice de Percepção da Corrupção no ano de 2018 em comparação ao ano anterior, ocupando agora a 105^{a} colocação entre 180 países avaliados. A pontuação passou de 37 para 35 e este é o pior resultado desde 2012, quando os dados passaram a ser comparáveis ano a ano. O que representa a 3^{a} queda anual seguida. Esse ranking classifica como 100 totalmente íntegro e 0 (zero) totalmente corrupto. A Figura 1 apresenta a tendência de queda do Brasil desde 2012.

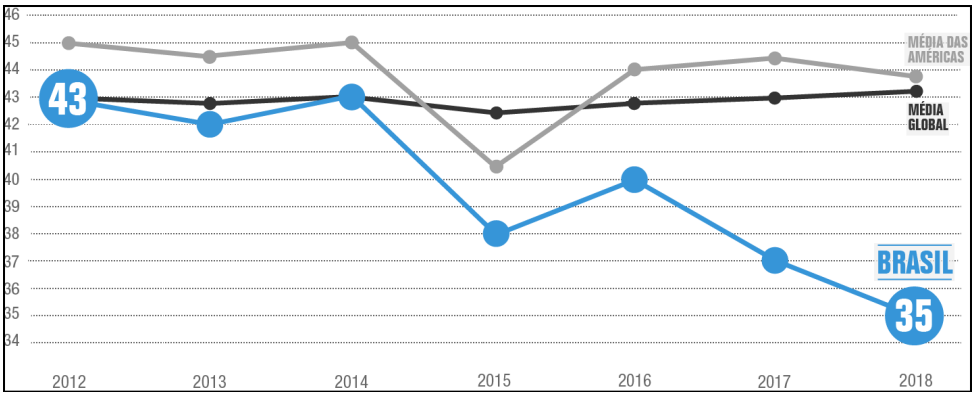


Figura 1 – Evolução do Brasil no ranking IPC

Fonte: (Transparency International, 2019)

Segundo Medeiros (2003), a falta de transparência no setor público é um dos elementos que pode contribuir para episódios de corrupção. Ainda em Medeiros (2003), "a corrupção é um fenômeno permanente, e o seu combate é, pois, uma tarefa sem fim".

Em 2009, foi criada a Lei Complementar 131 (LC 131/2009), conhecida como Lei da Transparência, onde toda a entidade pública, com prazo máximo de até 24 horas, deve lançar suas receitas e despesas publicamente na Internet. Complementar a isso, em 2011, foi promulgada a lei n^{o} 12.527/2011, chamada de Lei de Acesso à Informação (LAI), que regulamenta o direito constitucional de acesso às informações públicas.

Com essas ações, é possível a qualquer pessoa pesquisar os gastos públicos no Portal da Transparência[1]. A ideia é que a população em geral, com o uso dessa ferramenta, possa

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

fiscalizar os órgãos e agentes públicos, e coibir práticas de desvios de verba públicas. A principio, parece ser uma tarefa fácil efetuar essa pesquisa, porém, sem um mínimo de conhecimento em gestão pública, é muito difícil a uma pessoa, ou a um grupo, analisar, correlacionar e encontrar informações que sejam um indício de algum ato de corrupção ou desvio de verba ou mesmo má gestão.

No portal do Ministério do Planejamento, Desenvolvimento e Gestão (2018)[2] pode ser encontrado diversos aplicativos implementados nessa área. Como por exemplo uma Inteligência Artificial desenvolvida através do projeto Operação Serenata de Amor [3], que analisa dados referente à Cota para Exercício da Atividade Parlamentar (CEAP), e que já conseguiu encontrar dados relevantes, como pedido de reembolso de 13 refeições no mesmo dia, consumo de bebida alcoólica em Las Vegas, entre outras anomalias suspeitas. Outro aplicativo é o Monitora, Brasil! [4], que permite pesquisar e monitorar as atividades realizadas pelos Deputados Federais e Senadores, analisando seus projetos. Outro aplicativo é o Repasse [5], que verifica os repasses de verbas do governo federal, para os municípios.

A Mineração de Dados (*Data Mining*) é uma área da Ciência da Computação que pode ser usada para analisar grandes volumes de dados, na busca de padrões, previsões, erros, entre outros (AMARAL, 2016). A mineração de dados atua em conjunto com Aprendizado de Máquina, área da Inteligência Artificial que desenvolve algoritmos capazes de dar à máquina a capacidade de aprender a partir da experiência, ou seja, usando dados de eventos já ocorridos (AMARAL, 2016).

O aprendizado de máquina pode identificar padrões ou anomalias, que seriam difíceis de serem realizadas por uma mera análise visual, ou por técnicas tradicionais de análise de dados. A aplicação dessas técnicas nos dados fornecidos pelos principais portais de transparências, em conjunto com dados municipais de índices de ranqueamento, podem trazer informações relevantes sobre o uso do dinheiro pelo setor público e seus agentes.

[2] http://dados.gov.br/aplicativos
[3] https://serenata.ai/
[4] https://monitorabrasil.org/
[5] http://repasse.icmc.usp.br/

1.1 OBJETIVOS GERAL E ESPECÍFICOS

O objetivo geral é o uso de técnicas de Inteligência Artificial para encontrar anomalias no uso do dinheiro público, por meio de dados do portal da transparência em conjunto com índices referentes ao ranqueamento dos municípios disponibilizados por outros órgãos governamentais. Para atingir o objetivo geral foram divididos nos seguintes objetivos específicos:

- Unificação dados do Sistema de Informação Contábil e Fiscal (SICONFI) [6], índices da Federação das Indústrias do Estado do Rio de Janeiro (FIRJAN) [7] e óbitos da base do Departamento de Informática do Sistema Único de Saúde (DATASUS) [8];
- Aplicação do algorítimo LOF apenas nos dados do SICONFI nos anos de 2015 e 2016.
- Aplicação do algorítimo LOF nos dados unificados nos anos de 2015 e 2016;
- Comparação dos resultados, verificando se as anomalias com os dados unificados, são as de interesse, ou seja, referem-se a má gestão.

1.2 JUSTIFICATIVA

Diversas iniciativas de Organizações não Governamentais (ONGs) vem sendo realizadas com o intuito de monitorar o uso do dinheiro por órgãos públicos. Dentre elas, destaca-se o trabalho realizado pelo Observatório Social, onde, segundo reportagem de Lima (2017), estima-se que a fiscalização possibilitou a economia R$ 1,5 bilhões com gastos públicos no período de 2013 a 2016. Em 2018 são 134 observatórios sociais em 16 estados (Observatório Social do Brasil, 2018), o que demonstra o interesse da sociedade organizada em fiscalizar a utilização do dinheiro na busca por corrupção ou por má gestão pública. Uma ferramenta utilizando aprendizado de máquina para encontrar anomalias pode auxiliar sociedade e ONGs no seu trabalho de monitorar o setor público.

[6] https://siconfi.tesouro.gov.br/siconfi/index.jsf
[7] https://www.firjan.com.br/pagina-inicial.htm
[8] http://datasus.saude.gov.br/

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho é organizado como segue. O Capítulo 2 apresenta a fundamentação teórica abrangendo *Knowledge Discovery in Databases* (KDD), Mineração de Dados, Aprendizado de Máquina, Técnicas de Detecção de Anomalias e o algoritmo Local Outlier Factor (LOF). O Capítulo 3 descreve os materiais e métodos propostos, onde são descritas as etapas realizadas e as ferramentas utilizadas. O Capítulo 4 descreve os resultados de cada experimento realizado e uma discussão acerca das anomalias encontradas. O Capítulo 5 apresenta a conclusão do trabalho realizado, onde é explicitada as dificuldades encontradas, os resultados gerais alcançados e trabalhos futuros que podem ser implementados.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção é apresentado o estado da arte do tema escolhido. A Seção 2.1 apresenta o processo de Descoberta de Conhecimento em Base de Dados e como a Mineração de Dados é utilizada como uma etapa nesse processo. A Seção 2.2 contém uma introdução sobre Aprendizado de Máquina. Por fim, a Seção 2.3 descreve as técnicas utilizadas em detecção de anomalias.

2.1 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

Conforme Goldschmidt e Passos (2005), a Mineração de Dados é uma etapa da Descoberta de Conhecimento em Base de Dados (Knowledge Discovery in Databases - KDD). Para entender a diferença entre Mineração de Dados e KDD, Fayyad et al. (1996) propõe as etapas conforme Figura 2.

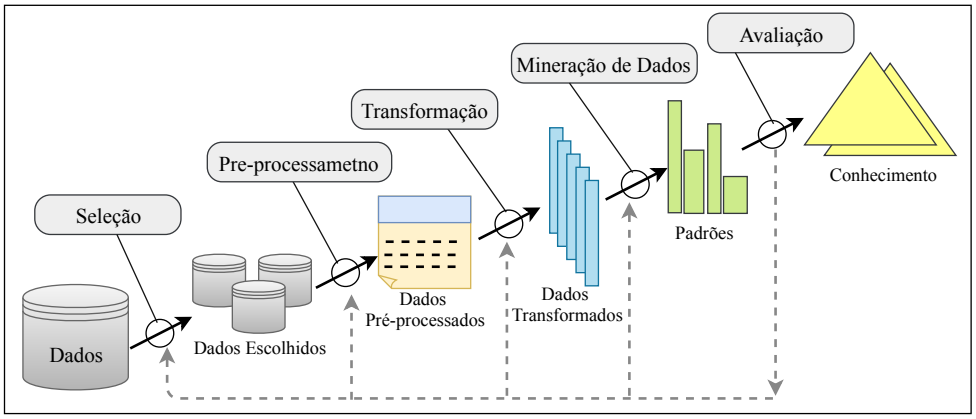


Figura 2 – Uma visão geral das etapas que compõem o processo do KDD

Fonte: Adaptado de Fayyad et al. (1996)

Assim, Fayyad et al. (1996) define 5 etapas que constituem esse processo:

1. Seleção dos Dados: define o domínio da aplicação bem como os objetivos que se deseja alcançar;
2. Pré-processamento: busca encontrar e eliminar dados inconsistentes;
3. Transformação: consiste em remover ou explicar os ruídos, decidir estratégia para eliminação ou não de dados ausentes, entre outros;
4. Mineração de Dados: é a etapa do KDD aplicada a um grande volume de dados que tem por objetivo a aplicação de algoritmos para a busca de padrões ou geração de novo conhecimento;
5. Avaliação: consiste em apresentar o conhecimento, que pode ser na forma de gráfico, tabela, planilhas, ou outras formas de visualização de dados.

Ambos autores Fayyad et al. (1996) e Goldschmidt e Passos (2005) ainda subdividem essas etapas, e ainda defendem que o processo é interativo, ou seja, requer a atuação humana no controle do processo, e iterativo, de forma que os processos podem ser repetidos seja integral ou parcialmente. Nas próximas seções é realizada uma breve explicação sobre cada etapa, sendo que a etapa de seleção e transformação é discutida em conjunto com pré-processamento.

2.1.1 Pré-Processamento

Esta etapa do KDD tem por finalidade definir os dados que são relevantes para fins da busca do conhecimento, ou seja verificar se o conjunto de dados faz parte do domínio a ser estudado. É possível que os dados estejam contidos em diversos formatos e com valores inconsistentes. Assim, após a seleção dos dados, é realizada a limpeza, codificação, enriquecimento e normalização dos dados. Será feita uma descrição da função de cada etapa de pré-processamento segundo Goldschmidt e Passos (2005).

A Seleção de Dados é uma importante subetapa do KDD que visa definir quais informações dentre as bases selecionadas que serão consideradas durante o processo. Aplica-se nessa etapa algumas técnicas, para que os dados estejam em uma única tabela bidimensional (ou matriz). Para que se tenha essa única tabela, pode se ocorrer dois cenários conforme Goldschmidt e Passos (2005):

- Junção direta: todos os dados são incluídos sem uma análise crítica para uma única tabela;
- Junção orientada: é selecionado apenas dados que tenha potencial para influenciar o processo de KDD.

Considerando que os dados já estejam em uma única tabela, a seleção de dados pode ter como objetivo a escolha de atributos ou a escolha de registros nos quais pode se ter redução de dados horizontal que visa definir quais registros serão analisados e redução de dados vertical que visa definir quais atributos serão utilizados.

Ainda segundo Goldschmidt e Passos (2005), o pré-processamento é composto pelas seguintes etapas: limpeza, codificação, enriquecimento e normalização.

A **limpeza** tem por finalidade corrigir os registros ausentes, inconsistentes e divergentes (*outliers*)(NETO; DINIZ, 2002). No caso de informações ausentes, pode-se optar por:

- Exclusão dos dados: consiste na exclusão do conjunto de dados o registro que tenha pelo menos um atributo sem informação;
- Preenchimento manual: inserção dados ausentes de forma manual, ou seja, atribuir uma informação, com base em dados existente, registro por registro. É um procedimento muitas vezes impraticável devido ao consumo de tempo, principalmente se o conjunto de dados for muito grande;
- Preenchimento com valores globais constantes: visa atribuir um valor global para todos os dados ausentes como ”desconhecido”ou ”null”. Técnica não recomendada de acordo com o algoritmo de mineração de dados, pois alguns podem interpretar esses dados como padrões e importantes para a análise;
- Preenchimento com medidas estáticas: consiste na utilização de medidas estáticas como média para atributos numéricos e moda para atributos categóricos;
- Preenchimento com métodos de mineração de dados: é a utilização de modelos preditivos para preenchimento de dados ausentes.

A **codificação** é o processo que define como os dados serão representados durante o processo de KDD. Os dados devem ser codificados de forma a atender as necessidades específicas dos algoritmos de mineração de dados, pois alguns aceitam apenas dados numéricos (quantitativos), enquanto outros apenas dados categóricos (qualitativos). A codificação segundo Boente et al. (2008) pode ser:

- Numérica - Categórica: mapeamento pode ser direto, que consiste em alterar valores numéricos para categóricos ou mapeamento em intervalos, também denominado de Discretização;
- Categórica - Numérica: substitui-se valores categóricos por valores numéricos.

O **enriquecimento** tem por finalidade encontrar informações extras que sejam

relevantes ao domínio do contexto que possam assim fornecer mais elementos para a descoberta do conhecimento. As operações mais utilizadas conforme Boente et al. (2008) são:

- Pesquisas: visa a busca de novas informações na fonte original que podem resultar na inclusão de novos atributos na tabela;
- Consulta a base de dados externa: consiste na incorporação de informações de outras bases de dados, como informações demográficas.

Por fim, a **normalização** é composta por técnicas que permitem atribuir uma nova escala a um atributo, de forma que os valores desse atributo estejam contidos nessa escala, como de -1.0 a 1.0 ou de 0 a 1 (ASSEISS, 2017). Podemos citar algumas técnicas, tais como: Normalização Linear, Normalização por Desvio Padrão, Normalização pela Soma dos Elementos, Normalização pelo Valor Máximo dos Elementos e Normalização por Escala Decimal (GOLDSCHMIDT; PASSOS, 2005).

2.1.2 Mineração de Dados

A mineração é uma importante etapa do KDD, conforme Camilo e Silva (2009), destacam-se três áreas de maior expressão no que diz respeito a conceituação de Mineração de Dados: Estatísticas, Aprendizado de Máquina e Banco de Dados. De acordo com Zhou (2003), destaca-se a análise comparativa feita nas três áreas ou perspectiva citadas:

- Perspectiva estatística: conforme Hand et al. (2001) mineração de dados "é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados";
- Perspectiva de banco de dados: segundo Cabena (1998) a mineração de dados "é um campo interdisciplinar que agrega técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados";
- Perspectiva do aprendizado de máquina: Fayyad et al. (1996) define que mineração de dados "é um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados".

Esta etapa, sobre a perspectiva do aprendizado de máquina, refere-se a capacidade

que o algoritmo a ser utilizado tem em aprender a partir de exemplos, conforme discutido na Seção 2.2.

2.2 APRENDIZADO DE MÁQUINA

Monard e Baranauskas (2003) definem Aprendizado de Máquina como uma área da Inteligência Artificial (IA) que tem como objetivo o desenvolvimento de técnicas computacionais sobre o aprendizado e a construção de sistemas capazes de adquirir conhecimento de forma automática.

Em mineração de dados, as principais abordagens referente ao aprendizado segundo Goldschmidt e Passos (2005) são:

- Aprendizado supervisionado: compreende a abstração de um modelo de conhecimento a partir dos dados apresentados na forma de pares ordenados (entrada, saída desejada), onde entrada refere-se ao conjunto de valores das variáveis de entrada, e saída desejada é o valor que se espera produzir sempre que são recebibos os valores especificados na entrada;
- Aprendizado não supervisionado: não existe a informação referente a saída desejada. Assim, o algoritmo utilizado, partindo dos dados informados, busca estabelecer uma relação entre os mesmos.

Russell e Norvig (2010) citam ainda o aprendizado por reforço, onde o aprendizado é baseado em uma série de comandos reforços que podem ser interpretados recompensa ou punição. Existe ainda conforme Chapelle et al. (2006) e Russell e Norvig (2010), o aprendizado semi-supervisionado que contém características tanto do supervisionado quanto do não supervisionado. Nesse caso, além dos dados não rotulados, são fornecidas algumas informações de supervisão.

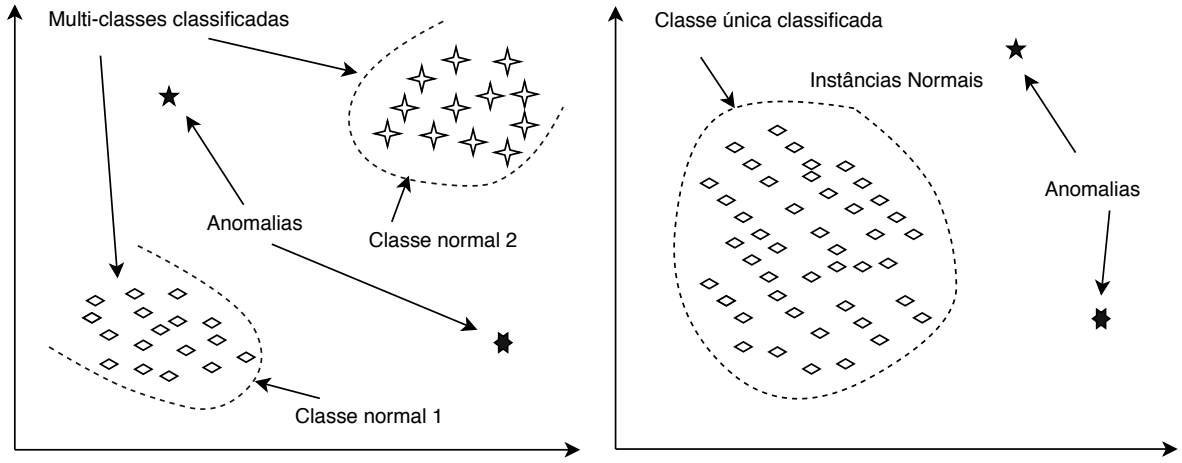
2.3 TÉCNICAS DE DETECÇÃO DE ANOMALIAS

Anomalia ou *Outliers* é uma observação que difere tanto das demais que levanta suspeitas se foi criada pelo mesmo mecanismo (HAWKINS, 1980). Objetos anômalos podem surgir nos dados por diversos motivos, como atividades maliciosas, fraudes ou falha no sistema de coleta. Estudos comparativos tendem a sub-categorizar técnicas de detecção de anomalias nos seguintes grupos (KINTOPP, 2017; CHANDOLA et al., 2009):

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

As técnicas baseadas em **classificação** são usadas para aprender um modelo ou função, que pode ser chamado de classificador, a partir de um conjunto de instâncias de dados rotuladas que pode ser entendido como treinamento e, em seguida, classificar uma instância de teste em uma das classes usando o modelo ou função aprendido (TAN et al., 2005; DUDA et al., 2000). Técnicas de detecção de anomalias baseadas em classificação atuam de maneira similar em duas fases. A fase de treinamento aprende um classificador usando os dados de treinamento rotulados disponíveis em seguida a fase de teste classifica uma instância de teste como normal ou anômala, usando o classificador (CHANDOLA et al., 2009).

Baseado nos rótulos disponíveis para a fase de treinamento, essas técnicas podem ser agrupadas em duas grandes categorias, técnicas de detecção de anomalias de múltiplas classes e de uma classe. As de multi-classe assumem que os dados de treinamento contêm instâncias rotuladas pertencentes a múltiplas classes normais (STEFANO et al., 2000; BARBARÁ et al., 2001). Essas técnicas de detecção de anomalia ensinam um classificador a distinguir entre cada classe normal e o resto das classes. Uma instância de teste é considerada anômala se não for classificada como normal por nenhum dos classificadores. Algumas técnicas nessa subcategoria associam um escore de confiança à previsão feita pelo classificador. Se nenhum dos classificadores estiver confiante em classificar a instância de teste como normal, a instância será declarada anômala, conforme Figura 3a. As técnicas de uma classe pressupõem que todas as instâncias de treinamento possuem apenas um rótulo de classe, conforme pode ser observado na Figura 3b.



(a) Anomalia multi-classe detectada (b) Anomalia classe única detectada

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

Especialmente as técnicas multi-classes, fazem uso de algoritmos eficazes que podem distinguir entre instâncias contidas em diferentes classes. A fase de testes das técnicas baseadas em classificação é rápida, uma vez que cada instância de teste precisa ser comparado com o modelo pré-computado. Segundo Chandola et al. (2009) essas duas características citadas, são consideradas as vantagens dessas técnicas. No que se refere as desvantagens, as técnicas multi-classes dependem da disponibilidade de rótulos precisos para várias classes normais, o que geralmente não é possível, e um segundo ponto, é que atribuem um rótulo a cada instância de teste, o que também pode ser prejudicial quando uma pontuação de anomalia significativa é desejada para as instâncias de teste. Existem alguns trabalhos que visam resolver esse problema (PLATT, 1999).

Existe uma variedade de técnicas de detecção de anomalias que usam algoritmos de classificação diferentes para construir classificadores, tais como:

- Baseado em Redes Neurais (*Neural Networks*): as redes neurais artificiais (RNA) são aplicadas à detecção de anomalias em configurações de várias classes e de uma única classe. Uma técnica básica de detecção de anomalias de múltiplas classes usando RNA opera em duas etapas. Primeiro, uma RNA é treinada nos dados normais de treinamento para aprender as diferentes classes normais. Em segundo lugar, cada instância de teste é fornecida como uma entrada para a rede neural. Se a RNA aceita a entrada de teste, é normal e se a RNA rejeita uma entrada de teste, é uma anomalia (CHANDOLA et al., 2009; STEFANO et al., 2000);
- Baseada em Redes Bayesianas (*Bayesian Networks*): têm sido usadas para detecção

de anomalias na configuração de várias classes. Uma técnica básica para um conjunto de dados categóricos uni-variados usando uma rede bayesiana simples, estima a probabilidade posterior de observar um rótulo de classe de um conjunto de rótulos de classe normal e o rótulo de classe de anomalia, considerando uma instância de dados de teste. O rótulo de classe com maior posterior é escolhido como a classe prevista para a instância de teste dada. A probabilidade de observar a instância de teste dada uma classe e a anterior nas probabilidades de classe é estimada a partir do conjunto de dados de treinamento (BABBAR, 2015). As probabilidades zero, especialmente para a classe de anomalias, são suavizadas usando o *Laplace Smoothing*(KIKUCHI et al., 2015). A técnica básica pode ser generalizada para conjuntos de dados categóricos multivariados, agregando as probabilidades posteriores por atributo para cada instância de teste e usando o valor agregado para atribuir um rótulo de classe à instância de teste (CARVALHO; CHIANN, 2013);

- Baseado em Máquina de Vetores de Suporte (*SVM - Support Vector Machines*): usam técnicas de aprendizado de uma classe para o SVM e aprendem uma região que contém as instâncias de dados de treinamento, ou seja, um limite ou fronteira (RÄTSCH et al., 2002). Para cada instância de teste, a técnica básica determina se a instância de teste está dentro da região aprendida. Se uma instância de teste estiver dentro da região aprendida, ela será declarada como normal, caso contrário, será declarada como anômala;
- Baseado em Regras (*Rules*): aprendem regras que capturam o comportamento normal de um sistema. Uma instância de teste que não é coberta por essa regra é considerada uma anomalia. Técnicas baseadas em regras foram aplicadas em configurações de várias classes e de uma única classe. Uma técnica básica baseada em regras de várias classes consiste em duas etapas. A primeira etapa é aprender regras a partir dos dados de treinamento usando um algoritmo de aprendizado de regras, como RIPPER (COHEN, 1995), Árvores de Decisão (QUINLAN, 1987) e assim por diante. Cada regra possui um valor de confiança associado que é proporcional à relação entre o número de instâncias de treinamento classificadas corretamente pela regra e o número total de instâncias de treinamento cobertas pela regra. A segunda etapa é encontrar, para cada instância de teste, a regra que melhor captura a instância de teste. O inverso da confiança associada à melhor regra é a pontuação de anomalia da instância de teste (KAO; HUANG, 2012).

As técnicas baseadas nos **vizinhos mais próximos** trabalham com a suposições de que instâncias de dados normais ocorrem em vizinhanças densas, enquanto anomalias ocorrem longe de seus vizinhos mais próximos. Essas técnicas requerem uma medida de distância ou similaridade definida entre duas instâncias de dados. A distância (ou similaridade) entre duas

instâncias de dados pode ser calculada de diferentes maneiras, onde para atributos contínuos pode ser usada a distância Euclidiana, Minkowski, Kullback-Leibler, Mahalanobis, Manhattan ou Hamming (WILLIAMS; LI, 2008). As técnicas de detecção de anomalias podem ser agrupadas em duas categorias:

1. técnicas que usam a distância de uma instância de dados para o seu vizinho mais próximo como a pontuação de anomalia;
2. técnicas que calculam a densidade relativa de cada instância de dados para calcular sua pontuação de anomalia.

Usando distância para *k*-ésimo vizinho mais próximo, técnica de detecção de anomalia de vizinho mais próximo baseia-se na seguinte definição: pontuação de anomalia de uma instância de dados é definida como sua distância ao *k*-ésimo vizinho mais próximo em um determinado conjunto de dados (CHANDOLA et al., 2009). Usando densidade relativa, estimam a densidade da vizinhança de cada instância de dados. Uma instância que fica em uma região com baixa densidade é declarada anômala, enquanto uma instância que se encontra em uma vizinhança densa é declarada normal. Para lidar com a questão das densidades variáveis no conjunto de dados, um conjunto de técnicas foi proposto para calcular a densidade de instâncias em relação à densidade de seus vizinhos. Breunig et al. (1999) e Breunig et al. (2000) atribui uma pontuação de anomalia a uma dada instância de dados, conhecida como *Local Outlier Factor* (LOF). Para qualquer instância de dados, a pontuação do LOF é igual à proporção da densidade média local dos *k* vizinhos mais próximos da instância e da densidade local da própria instância de dados. Para encontrar a densidade local para uma instância de dados, os autores primeiro encontram o raio da menor hiperesfera centrada na instância de dados, que contém seus *k* vizinhos mais próximos (UPADHYAYA; SINGH, 2012). O estudo referente a LOF será aprofundado da seção 2.4.

As técnicas baseadas em **agrupamento** ou *cluster* é usado para agrupar instâncias de dados em agrupamentos semelhantes. Agrupamento é basicamente uma técnica não supervisionada, embora tenha sido explorado também, o semi-supervisionado. Embora a detecção de agrupamento e anomalia pareça ser fundamentalmente diferente uma da outra, várias técnicas de detecção de anomalias baseadas em agrupamentos foram desenvolvidas. As técnicas podem ser agrupadas em três categorias:

1. As instâncias de dados normais pertencem a um agrupamento nos dados, enquanto as anomalias não pertencem a nenhum agrupamento. Aplicam um algoritmo baseado em agrupamento, como por exemplo *K-means* (KASTURE; GADGE, 2012), ao conjunto de dados conhecido e declaram qualquer instância de dados que não pertença a nenhum agrupamento como anômala. Uma desvantagem dessas técnicas é que elas não são

otimizadas para encontrar anomalias, uma vez que o principal objetivo do algoritmo é encontrar agrupamentos.

2. As instâncias de dados normais ficam próximas ao centroide de agrupamento mais próximo, enquanto as anomalias estão distantes de seu centroide de agrupamento mais próximo. Um exemplo de algoritmo é o CBLOF (AMER; GOLDSTEIN, 2012). É realizada em duas etapas. Na primeira etapa, os dados são agrupados usando um algoritmo de agrupamento. Na segunda etapa, para cada instância de dados, sua distância ao centroide de agrupamento mais próximo é calculada como sua pontuação de anomalia.
3. As instâncias de dados normais pertencem a agrupamentos grandes e densos, enquanto as anomalias pertencem a agrupamentos pequenos ou esparsos. Observando as categorias anteriores, se as anomalias nos conjuntos de dados se formarem por si mesmas, essas técnicas não poderão detectar tais anomalias. Por isso essa terceira categoria foi proposta. Nessa categoria declaram as instâncias pertencentes a agrupamentos cujo tamanho e/ou densidade estão abaixo de um limite, como anômalos.

As técnicas **estatísticas** ajustam um modelo, geralmente para o comportamento normal, aos dados fornecidos e, em seguida, aplicam um teste de inferência estatística para determinar se uma instância não visível pertence ou não a esse modelo. Instâncias com baixa probabilidade de serem geradas a partir do modelo aprendido, com base na estatística de teste aplicada, são declaradas como anomalias (CHANDOLA et al., 2009). As técnicas estatísticas podem ser agrupadas em paramétricas e não paramétricas.

As paramétricas são subdivididas ainda em:

- Modelo gaussiano (*Modelo gaussiano*): assumem que os dados são gerados de uma distribuição gaussiana. Os parâmetros são estimados usando Estimativas de Máxima Verossimilhança (MLE - *Maximum Likelihood Estimates*). A distância de uma instância de dados para a média estimada é a pontuação de anomalia para essa instância. Um limite é aplicado às pontuações de anomalia para determinar as anomalias;
- Modelo de regressão (*Regression Model*): é realizada em duas etapas, onde, primeiro um modelo de regressão é ajustado nos dados e em seguida para cada instância de teste, o residual da instância de teste é usado para determinar a pontuação da anomalia. O residual é a parte da instância que não é explicada pelo modelo de regressão (CHANDOLA et al., 2009);
- Mistura de distribuições paramétricas (*Mixture of Parametric Distributions*): nesta categoria podem ser agrupadas em duas subcategorias. A primeira subcategoria de técnicas modela as instâncias e anomalias normais como distribuições paramétricas separadas, enquanto a segunda subcategoria de técnicas modela apenas as instâncias

normais como uma mistura de distribuições paramétricas. Para a primeira subcategoria de técnicas, a fase de teste envolve determinar qual distribuição - normal ou anômala - pertence à instância de teste. Abraham e Box (1979) assumem que os dados normais são gerados a partir de uma distribuição gaussiana e as anomalias também são geradas a partir de uma distribuição gaussiana com a mesma média, mas com maior variância. Uma instância de teste é testada usando o teste de Grubb (GRUBBS, 1969) em ambas as distribuições e, consequentemente, rotulada como normal ou anômala.

As técnicas não paramétricas, usam modelos estatísticos não paramétricos, de tal forma que a estrutura do modelo não é definida como a priori, mas é determinada a partir de dados fornecidos. São divididas em duas subcategorias:

- Baseada em Histograma (*Histogram-Based*): consiste em duas etapas, onde, a primeira etapa envolve a construção de um histograma baseado nos diferentes valores obtidos por esse recurso nos dados de treinamento e, na segunda etapa, a técnica verifica se uma instância de teste está em qualquer um dos compartimentos do histograma. Em caso afirmativo, a instância de teste é normal, caso contrário, é anômala. Uma variante da é atribuir uma pontuação de anomalia a cada instância de teste com base na altura, ou seja, frequência, da caixa na qual ela se enquadra(GOLDSTEIN; DENGEL, 2012);
- Baseada na Função do Kernel (*Kernel Function-Based*): são semelhantes aos métodos paramétricos descritos anteriormente. A única diferença é a técnica de estimativa de densidade usada (LATECKI et al., 2007). Desforges et al. (1998) propuseram uma técnica estatística semi-supervisionada para detectar anomalias, que usa funções kernel para estimar a função de distribuição de probabilidade (PDF - Probability Distribution Function) para as instâncias normais. Uma nova instância, que se encontra na área de baixa probabilidade deste PDF, é declarada.

As técnicas da **teoria da informação** analisam o conteúdo da informação de um conjunto de dados usando diferentes medidas teóricas de informação, tais como Complexidade de Kolomogorov, entropia, entropia relativa, entre outras. Essas técnicas supõe que anomalias nos dados induzem irregularidades no conteúdo informativo do conjunto de dados. Complexidade de Kolmogorov é um ramo da teoria da informação que lida com a complexidade contida em um único objeto ou *string*. É outra medida de aleatoriedade, ao contrário da entropia, e não é baseada em probabilidade, mas considera o método ou algoritmo usado para calcular a string (BERNI, 2018). A interpretação típica da entropia é que ela especifica o número de bits necessários para codificar e transmitir a classificação de um item de dados. O valor de entropia é menor quando a distribuição de classe é distorcida, ou seja, quando os dados são “mais puros”. Por exemplo, se todos os itens de dados pertencerem a uma classe, a entropia será zero e, zero

bits precisará ser transmitido, pois o receptor sabe que existe apenas um resultado. O valor de entropia é maior quando a distribuição da classe é mais uniforme, ou seja, quando os dados são mais ”impuros”.

As técnicas **espectrais** tentam encontrar uma aproximação dos dados usando uma combinação de atributos que capturam a maior parte da variabilidade nos dados. Tem como premissa que os dados podem ser incorporados em um subespaço de menor dimensão no qual instâncias e anomalias normais aparecem significativamente diferentes.

2.4 LOCAL OUTLIER FACTOR

Breunig et al. (2000) aponta que a maioria dos estudos realizados no final da década de 90, tais como (KNORR; NG, 1998), (KNORR; NG, 1999), (ARNING et al., 1996) e (RUTS; ROUSSEEUW, 1996) consideram um *outlier* como uma propriedade binária, isto é, um objeto no conjunto de dados é um *outlier* ou não. Ocorre que para muitas aplicações, a situação pode ser mais complexa, tornando mais significativo atribuir a cada objeto um grau de ser um *outlier*. Assim Breunig et al. (2000), sugere a utilização de um fator local para determinar se o dado em questão é ou não uma anomalia. Esse fator está relacionado a distância ou isolamento que o objeto se encontra em relação ao seus vizinhos.

O LOF é baseado em densidade que depende da pesquisa de vizinhos mais próximos. O método marca cada ponto de dados calculando a proporção das densidades médias dos vizinhos do ponto com a densidade do próprio ponto. A densidade estimada de um ponto p é o número de vizinhos de p dividido pela soma das distâncias até os vizinhos do ponto.

Sendo:

- $N(p)$ o conjunto de vizinhos do ponto p;
- k o número de pontos desse conjunto;
- $d(p,x)$ a distância entre os pontos p e x.

A Equação 1 calcula a densidade estimada:

$$f(p) = \frac{k}{\sum_{x \in N(p)} d(p,x)} \tag{1}$$

E o fator local é calculado conforme a Equação 2:

$$LOF(p) = \frac{\frac{1}{k}\sum_{x \in N(p)} f(x)}{f(p)} \tag{2}$$

A Figura 4 mostra de forma intuitiva como o LOF se comporta. O ponto A tem um escore LOF alto porque sua densidade é baixa em relação às densidades dos vizinhos. Os círculos pontilhados indicam a distância até o terceiro vizinho mais próximo de cada ponto. Breunig et al. (2000) define que para objetos dentro do agrupamento (*cluster*) o valor de LOF é aproximadamente 1, e para os demais objetos é dado limites superiores e inferiores, valores esses definidos através de uma heurística de classificação de objetos por seu valor máximo de LOF dentro do intervalo selecionado.

Alguns trabalhos tem utilizado o LOF para detecção de anomalias em diversas áreas. Na detecção de lavagem de dinheiro podemos citar o trabalho de (ZENGAN, 2009), onde o objetivo é encontrar padrões comportamentais transacionais suspeitos de lavagem de dinheiro. Na detecção de anomalias em redes de computadores o trabalho de (AUSKALNIS et al., 2018), que visa criar um sistema de detecção de invasão de anomalias capaz de detectar ataques até então desconhecidos. Na área de saúde o trabalho de (CVETKOVIĆ; LUSTREK, 2012), que tem por objetivo, encontrar anomalias que possam indicar algum problema de saúde utilizando dados de prontuários de pacientes.

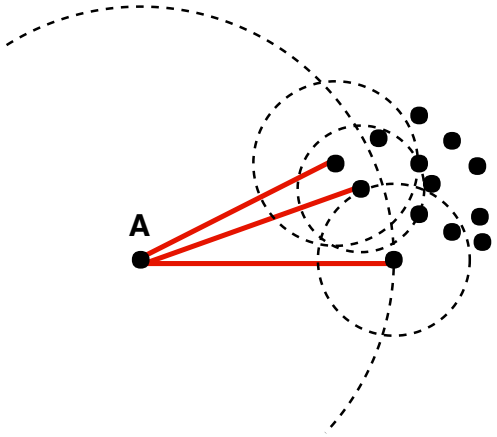


Figura 4 – Visão intuitiva do LOF

Fonte: Adaptado de (BREUNIG et al., 2000)

Na Seção 3.2 será apresentada a utilização do LOF no cenário de estudo deste trabalho utilizando a linguagem de programação Python. Explicado as bases utilizadas e seus atributos.

3 MATERIAL E MÉTODOS

Neste capítulo são descritos os materiais e métodos utilizados para o desenvolvimento deste projeto. São descritas as etapas do projeto e as principais tecnologias empregadas.

3.1 MATERIAIS

Optou-se por utilizar a linguagem de programação Python[1], devido a já comportar algumas ferramentas utilizadas em AM, tais como:

- Pandas[2]: biblioteca que fornece estrutura de dados e funções de alto nível projetada para trabalhar com estrutura de dados tabuladas de forma fácil, rápida e expressivas (MCKINNEY, 2017). Fornece funcionalidade de indexação sofisticada para facilitar a reformulação, a divisão, a agregação e a seleção de subconjuntos de dados;
- NumPy[3]: abreviação de *Numerical Python*, tem sido o alicerce da computação numérica em Python. Oferece a estrutura de dados como tensores, os algoritmos e um rol de bibliotecas necessários para a maioria dos aplicativos científicos que envolvem dados numéricos no Python (MCKINNEY, 2017);
- Scikit-learn[4]: kit de ferramentas de aprendizado de máquina de uso geral para programadores Python. Inclui sub-módulos para modelos como: classificação, regressão, agrupamento, redução de dimensionalidade, seleção de modelos e pré-processamento (MCKINNEY, 2017);
- Colaboratory: é uma ferramenta de pesquisa para educação e pesquisa em aprendizado de máquina. É um ambiente de notebook Jupyter - que não requer configuração para usar.
- Jupyter: é um projeto de código aberto do Projeto IPython em 2014, para suportar a

[1] https://www.python.org/
[2] https://pandas.pydata.org/
[3] http://www.numpy.org/
[4] http://scikit-learn.org/stable/

evolução da ciência de dados interativa e computação científica.

As bases utilizadas foram:

- SICONFI: base do Sistema de Informações Contábeis e que possui um banco de dados chamado Finbra, formado pelas informações das declarações recebidas pelo Tesouro Nacional por determinação da Lei Complementar 101/2000, a Lei de Responsabilidade Fiscal – LRF. Foram utilizados os atributos: Cod.IBGE, População, 01 - Legislativa, 04 - Administração, 08 - Assistência Social, 10 - Saúde, 12 - Educação, 13 - Cultura, 15 - Urbanismo, 17 - Saneamento, 18 - Gestão Ambiental, 20 - Agricultura, 26 - Transporte e 27 - Desporto e Lazer;
- DATASUS: Departamento de Informática do Sistema Único de Saúde disponibiliza informações na plataforma TABNET, que podem servir para subsidiar análises objetivas da situação sanitária. Utilizado os atributos: 110 - Agressões, 111 - Eventos (fatos) cuja a intenção é indeterminada e 112 - Intervenções legais e operações de guerra;
- IFDM: Índice FIRJAN de Desenvolvimento é o estudo do Sistema FIRJAN que acompanha anualmente o desenvolvimento socioeconômico dos municípios brasileiros. Abrange três áreas de atuação: Emprego & renda, Educação e Saúde;
- IFGF: Índice FIRJAN de Gestão Fiscal ferramenta de controle social com objetivo estimular a cultura da responsabilidade administrativa e aprimoramento da gestão fiscal dos municípios. É composto por cinco indicadores: Receita Própria, Gastos com Pessoal, Investimentos, Liquidez e Custo da Dívida.

Os dados coletados do DATASUS são referentes causa de mortalidade, conforme Classificação Internacional de Doenças (CID), mais especificamente a CID-BR-10, que refere-se a décima revisão. No TABNET são filtradas as causas: 110 - Agressões, 111 - Eventos (fatos) cuja a intenção é indeterminada e 112 - Intervenções legais e operações de guerra. Os anos são de 2015 e 2016, pois 2017 ainda não consta no TABNET.

Na Seção 3.2 será abordado de que forma e em que momento esses materiais serão utilizados nas etapas de KDD.

3.2 MÉTODO

A metodologia utilizada é baseada no KDD, método já descrito no Capítulo 2. Sendo assim, foi realizada as etapas conforme abaixo:

- Seleção dos dados;
- Pré-processamento, Transformação e Mineração de Anomalias;
- Apresentação, Analise e Comparação dos Resultados.

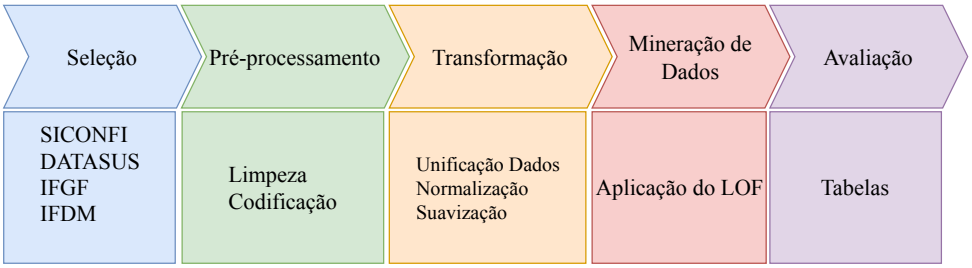


Figura 5 – Uma visão geral do processo realizado

Fonte: Autoria Própria

A figura 5 apresenta uma visão geral das etapas realizadas.

Nas próximas seções essas etapas são explicadas.

3.2.1 Seleção dos dados

As etapas de identificação e seleção de dados foram baseadas no trabalho de Kintopp (2017). Isso inclui usar como fonte para o aprendizado os dados do Sistema de Informações Contábeis e Fiscais (SICONFI)[5] do portal do Tesouro Nacional[6]. Dados esses referente a Despesa por Função, mas especificamente Despesas Empenhadas do triênio de 2013 a 2015 e como complemento serão utilizados os dados de 2016 e 2017.

Alem do SICONFI, foram analisadas as bases, conforme já discutidos na seção 3.1:

- DATASUS/TABNET [7];

[5] https://siconfi.tesouro.gov.br/siconfi/index.jsf

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

[7] http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10br.def

- IFGF[8];
- IFDM[9];
- IDEB[10];
- ENEM[11];

Os dados do Índice de Desenvolvimento da Educação Básica (IDEB) e Exame Nacional do Ensino Médio (ENEM) não foram utilizados, pois houve alteração na forma da disponibilização dos mesmos. Assim em um determinado período o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) utilizou uma metodologia, e em outro período, outra metodologia foi utilizada, prejudicando a comparação dos dados.

Para compor os dados, foi utilizado como chave entre as bases o código do município de acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE).

Após a coleta dos dados, foi necessário efetuar um série de transformações, que são descritos na subseção 3.2.2.

3.2.2 Pré-processamento, Transformação e Mineração de Dados

Foram reproduzidas as ações realizadas conforme Kintopp (2017), porém adaptadas às novas bases propostas com os indicadores do DATASUS e FIRJAN. Esse experimento e seus resultados é explicado na seção 4.1.

O pré-processamento foi executado conforme abaixo:

- Limpeza/Codificação: visando remover linhas e colunas desnecessárias. Pivotamento de linhas em colunas, ou seja, criação de atributos;
- Normalização: transformação dos valores dos atributos originais em um intervalo específico, por exemplo, o intervalo [0,1];
- Suavização: redução da escala dos dados por população, buscando utilizar valores próximos para cidades de tamanhos similares.

Após essa a etapa, foi feita transformação, onde os dados do SICONFI foram unificados com a base do DATASUS, onde o experimento e seus resultados são discutidos na seção 4.2.

[8] http://www.firjan.com.br/ifgf/
[9] https://www.firjan.com.br/ifdm/
[10] http://portal.inep.gov.br/ideb
[11] http://inep.gov.br/web/guest/microdados

Em seguida feita uma nova unificação, da base do SICONFI com IFDM e IFGF. O objetivo foi descobrir se com dados, como por exemplo IFDM, unificados com os dados do SICONFI referente a gastos com educação, o algoritmo por si só, já consegue identificar alguma anomalia, onde, por exemplo, uma cidade com alto gasto em educação, e um índice de IFDM de educação muito baixo, é marcado como anomalia, sem necessidade de recorrer a analise da tabela do IFDM. Os resultados e discussões desse experimento estão descritos na seção 4.3.

Para a mineração de dados foi utilizado o algoritmo LOF (seção 2.4). O valor do parâmetro *k*, foi definido como sendo 70, este parâmetro representa os *k-vizinhos* mais próximos. A distância euclidiana foi utilizada como medida de proximidade. Foram realizados experimentos com outros parâmetros, no intuito de verificar se com novos dados, um parâmetro mais adequado deveria ser utilizando, porém a única alteração foi no valor do LOF, sem alterar o número de anomalias encontradas.

Após a mineração de dados, foram exportadas tabelas no formato *Comma Separated Values* (CSV) e foram feitas as devidas análises, conforme é descrito na subseção 3.2.

3.2.3 Apresentação, Análise e Comparação dos Resultados

Para esta etapa foram implementadas funcionalidade para geração de tabelas para apresentar os dados. Foram realizadas analises das anomalias encontradas e sua relação com a instância original, verificando os gastos das dez cidades com maior valor de LOF.

Realizada a comparação do experimento 4.1 com o experimento 4.3 e os resultados e discussões são apresentados ,

4 RESULTADOS

Neste capítulo são apresentados os experimentos realizados, alguns apontamentos referente as dificuldades encontras, técnicas aplicadas e ao final uma discussão sobre as principais anomalias encontradas.

4.1 EXPERIMENTO LOF - ANÁLISE DOS DADOS

Realizado um pré-experimento apenas na base do SICONFI nos anos de 2013 a 2017, com o objetivo de analisar as ferramentas utilizadas conforme capítulo 3, bem como fazer um levantamento das dificuldades encontradas. A primeira etapa consistiu em definir os atributos a serem utilizados. Nesse experimento nenhuma instância foi removida, e para isso, foram selecionados apenas as contas que possuíam menos dados ausentes, e poderiam em um segundo momento ser relevante para a composição com outras bases referente a índices de ranqueamento. Logo foram selecionados como atributos as colunas: Cod.IBGE (apenas para manter uma chave de identificação), População, 01 - Legislativa, 04 - Administração, 08 - Assistência Social, 10 - Saúde, 12 - Educação, 13 - Cultura, 15 - Urbanismo, 17 - Saneamento, 18 - Gestão Ambiental, 20 - Agricultura, 26 - Transporte e 27 - Desporto e Lazer.

A próxima etapa foi realizar a normalização e suavização dos dados. Para isso, primeiro os valores das contas foram divididos pela população das cidades, assim os valores passaram a ser *per capita*. Mesmo assim, os valores dos atributos ficaram em escalas bem distintas. Assim, seguindo o método KDD, foi feita a normalização dos dados, colocando em uma escala de 0 a 1 utilizando a equação 3:

$$y = \frac{x - mínimo(\bar{x})}{máximo(\bar{x}) - mínimo(\bar{x})} \tag{3}$$

Em seguida foi aplicado o logaritmo decimal no valor da cidade. Em um primeiro momento foi utilizado a metodologia do IBGE, onde as cidades são agrupadas por População,

como por exemplo, de 0 a 10.000, cidade pequena, de 10.001 a 25.000, cidade pequena/média, de 25.001 a 50.000, cidade média, e assim por diante. Ocorre que não existe uma diferença tão grande entre uma cidade de 25.000 e 25.001 habitantes. Assim, aplicando o logaritmo decimal, em uma cidade de 10.000 habitantes o log é 4, já uma cidade de 25.000 o log é aproximadamente 4,30, já uma cidade de 100.000 habitantes o log é 5, assim no momento de utilizar a técnica de aprendizado, que utiliza a distância euclidiana, essa diferença entre as cidades vai ser suavizada. Após a normalização e suavização, foi aplicado o algoritmo de aprendizado de máquina LOF, utilizando como parâmetro *k* de vizinhos mais próximo igual a 70. Esse parâmetro é para definir o número de vizinhos para o cálculo da densidade.

Analisando o resultado ano a ano, ou seja, de 2013 a 2017, existe uma característica das cidades que aparecem com mais recorrência como anomalia. Se trata de cidades que tiveram gastos informados apenas em dois atributos (seção 3.1) do SICONFI. Existe a possibilidade de ser um erro de cadastramento, onde a cidade não fez o lançamento correto de seus gastos. Sendo assim, optou-se por apresentar apenas as cidades com gastos lançados em mais de dois atributos.

A tabela 1, mostra os dez maiores fatores LOF encontrados de 2013 a 2017. Os nomes reais de cidades foram omitidos seguindo a recomendação da assessoria jurídica da Universidade Tecnológica Federal do Paraná.

Tabela 1 – 10 maiores anomalias de 2013 a 2017 encontrados

2013			2014			2015			2016			2017	
Cidade	LOF		Cidade	LOF		Cidade	LOF		Cidade	LOF		Cidade	LOF
ZT7	10,50		YT2	6,83		HH4	5,57		HH4	8,25		HH4	6,20
HH4	5,20		HH4	5,37		DL5	4,27		DL5	5,25		DL5	5,38
D7	3,66		DN5	4,34		KK5	3,88		ZQ2	4,45		CE8	5,33
KK5	3,59		KU2	3,87		LP2	3,25		LM9	4,04		YR0	3,92
KC6	3,58		DL5	3,86		PP1	3,23		MM0	3,61		ER4	3,87
C9	3,41		KT5	3,65		MM0	3,21		KK5	3,55		LM9	3,57
NT2	3,30		BU4	3,49		BL2	3,02		RA2	2,87		OJ7	3,43
SH9	2,87		DN7	3,18		EU7	2,99		NT2	2,75		SA6	3,22
LL3	2,84		FM3	3,06		LA8	2,98		KU7	2,73		RN2	3,15
WA7	2,81		OT0	2,98		DN7	2,98		GI2	2,70		QK6	3,05

Fonte: Autoria Própria

A cidade YT2 de 2014, refere-se a uma cidade com 7 mil habitantes que teve gasto de R$ 21.506.365,42 com saneamento. Valor esse muito superior a média de cidades com número de habitantes similar. Apenas para comparação, foram destacadas dez cidades acima e dez cidades a abaixo da mesma, em termos de ordenação por população, e a média de gasto

com a conta saneamento dessas vinte cidades, foi de R$ 310.934,74. Encontrar essa anomalia foi importante, pois indica que mesmo removendo alguns atributos, foi possível encontrar a mesma anomalia encontrada por Kintopp (2017). Outra informação é que a cidade HH4 onde a particularidade é que em todos os anos não tem gastos lançados com educação, e um valor muito alto lançado em desportos e lazer. E isso também foi uma anomalia encontra por Kintopp (2017). Porém a informação mais relevante, e se refere ao tipo de anomalia que se deseja encontrar, é a cidade DL5, que aparece nos anos de 2014 a 2017, com gasto bem acima da média, em Saneamento, Transporte, Desportos & Lazer, entre outros. Essa cidade foi alvo de uma operação deflagrada pelo Ministério Público do seu estado. Uma nova anomalia encontra se refere a cidade CE8 de 2017, com 5949 habitantes, onde a mesma teve apontamento de R$ 6.992.941,85 com Gestão Ambiental. Outra anomalia desse ano é a cidade YR2 de 4552 habitantes que teve gasto com a conta Legislativa de R$ 8.145.081,42.

Após realizar essa pré-analise, e demonstrar que obteve-se resultados condizente com o que o trabalho se propõe a encontrar, optou-se nos próximos experimentos utilizar como base apenas os dados do Siconfi de 2015 e 2016. A justificativa principal é evitar o que em mineração de dados é chamado de explosão combinatória. Ou seja, como temos os dados do Siconfi de 2013 a 2017, temos 5 bases, que poderiam ser combinadas de diversas formas, como por exemplo Siconfi 2013 com Firjan 2013, Siconfi 2014 com Firjan 2014, e assim sucessivamente, até Siconfi 2017 com Firjan 2017, e ainda replicado para base do Datasus, e assim, gerando muitas combinações. Outro motivo é por que os dados do Datasus de 2017 ainda não estão no sistema Tabnet, e assim poder ter um comparativo, se a composição dos dados realmente ajuda em uma analise melhor.

4.2 EXPERIMENTO LOF - DADOS SICONFI E DATASUS

Os dados do DATASUS referem-se a óbitos cujo a causa seja homicídio. Ressaltando que no sistema TABNET essas mortes não são registradas como homicídio, mas em uma determinada categoria conforme definição da norma CID-BR-10. Isso foi discutido na seção 3.1.

Ao realizar esse experimento, ocorreu uma mudança significativa no período de 2015, devido a duas cidades. A cidade PD5 de 3978 habitantes do Rio Grande do Norte teve 3 óbitos registrados no TABNET, assim, a taxa de óbitos por 100 mil habitantes ficou em 175. O LOF

passou de 1,70, onde não figurava nem entre as 100 maiores anomalias, para quarta posição com LOF de 3,25.

Tabela 2 – 10 maiores anomalias de 2015 e 2016 - Experimento 4.2

2015			2016	
Cidade	LOF		Cidade	LOF
HH4	5,39		HH4	7,84
DL5	4,21		DL5	5,15
KK5	3,81		ZQ2	4,23
PD5	3,25		LM9	4,01
LP2	3,22		JD9	3,80
MM0	3,17		MM0	3,62
PP1	3,11		KK5	3,49
MO8	3,04		SO6	2,91
DN7	2,91		NT2	2,73
ML8	2,91		KU7	2,72

Fonte: Autoria Própria

A cidade MO8 com 2662 habitantes do estado do Paraná com 3 óbitos registrados no Tabnet, assim, a taxa de óbitos por 100 mil habitantes ficou em 187. O LOF passou de 1,42, onde não figurava nem entre as 380 maiores anomalias, e passou para 3,04, passando para sétima posição. Em ambos os casos, foram encontradas reportagens sobre casos homicídios nesse ano de 2015.

Esse resultado, se repetiu em 2016, com duas outras cidades (JD9 e SO6), também com poucos habitantes, e devido o número de óbitos, sua taxa de mortes por 100 mil habitantes ficou muito elevada. A tabela 2, mostra as dez maiores anomalias encontradas nesse experimento compondo a base do SICONFI com os dados do DATASUS.

Verificasse ainda que as três primeiras cidades, comparando a tabela 1 do experimento 4.1 com a tabela 2 desse experimento, permanecem as mesmas, tendo apenas uma variação do fator LOF.

4.3 EXPERIMENTO LOF - DADOS SICONFI E FIRJAN

Após realização do experimento 4.1, foi realizada composição com os índices do FIRJAN. Assim além dos atributos do SICONFI, foram acrescentados dez novos atributos. Os

mesmo se referem a:

- IFDM: Educação, Saúde, Emprego & Renda e Geral;
- IFGF: Custo da Dívida, Gestão com Pessoal, Investimento, Liquidez, Receita Própria e Geral.

Assim, nesse novo experimento no ano de 2015 três cidades entraram entre as dez maiores anomalias. A cidade LM9 teve gasto com Urbanismo de R$ 99.039.984,66 quase 16 vezes mais que a média das cidades de mesmo porte. A mesma cidade, se manteve no ano de 2016, subindo uma posição. No ano de 2016 foi apontado gasto com Urbanismo no valor de R$ 144.638.756,13 ou 20 vezes mais que a média das cidades semelhantes. E ainda em 2016, foi registrado gasto com Educação no valor de R$ 105.703.446,73 mais de 4 vezes a média de cidades com mesma características. Os índices FIRJAN dessa cidade são bem altos, e teoricamente os gastos se justificariam. Porém, ao pesquisar sobre a cidade, foi encontrada uma reportagem referente a esquema de caixa 2 e superfaturamento de licitação, praticado pelo prefeito e secretários da prefeitura.

A cidade ID4 teve um apontamento com Urbanismo no ano de 2015 de R$ 39.209.287,09 ou 26 vezes maior que a média das cidades de mesmo porte. Essa cidade foi alvo da operação do Ministério Público do Rio Grande do Norte, que apura fraudes em contratos públicos. Foi apontado o gasto com urbanismo, porém outros gastos foram bem altos, inclusive alguns citados na reportagem referente a operação que foi deflagrada.

Outra cidade ML8 tem gastos maiores que o normal em praticamente todas as despesas, e como parece ser recorrente, um valor que pode se citado é R$ 65.211.794,62 com Urbanismo.

Já no ano de 2016, quatro novas cidades subiram para as 10 maiores anomalias. Podemos citar a cidade P2 de pouco mais de 7 mil habitantes que teve gasto de R$ 30.129.904,70 aproximadamente 5 vezes mais do que cidades do mesmo porte que não são anômalas. E seu índice IFDM referente a educação foi de 0,6967, demonstrando que o gestão do dinheiro público não está refletindo na melhoria dos serviços prestados pela prefeitura. A tabela 3 apresenta as dez maiores anomalias encontradas no experimento 4.3.

Esse novo experimento se mostrou eficiente, se analisarmos o fato de que 229 cidades em 2015, que não eram consideradas anomalias no experimento 4.1, foram identificadas agora como anomalia. Em contrapartida, da mesma forma, 229 cidades que eram consideradas anomalias no experimento 4.1 passaram a não ser mais anômalas no 4.3. Isso se repetiu no ano de 2016 com 281 cidades.

A tabela 4 mostra as dez maiores anomalias do experimento 4.1 que deixaram de ser anomalias, e as dez maiores anomalias do experimento 4.3 que agora são consideradas anomalias nos anos de 2015 e 2016.

Tabela 3 – 10 maiores anomalias de 2015 e 2016 - Experimento 4.3

2015			2016	
Cidade	LOF		Cidade	LOF
HH4	3,25		HH4	4,60
DL5	2,98		DL5	3,76
KK5	2,70		LM9	2,87
LA8	2,37		KK5	2,74
LP2	2,37		MM0	2,55
MM0	2,32		DH7	2,48
LM9	2,31		ZQ2	2,38
ID4	2,23		P2	2,29
BL2	2,19		ZS2	2,27
ML8	2,18		EB7	2,25

Fonte: Autoria Própria

Tabela 4 – Comparativo Experimento 4.1 x Experimento 4.3

2015					2016			
Exp. 4.1		Exp. 4.3			Exp. 4.1		Exp. 4.3	
Cidade	LOF	Cidade	LOF		Cidade	LOF	Cidade	LOF
ZG5	2,38	K8	1,66		XG6	2,31	FT4	1,95
WO8	1,99	HK2	1,61		XC2	2,19	CT7	1,84
BP2	1,98	CS4	1,60		TK9	2,18	P5	1,81
DS9	1,92	FB5	1,59		DS9	2,04	JD5	1,75
JU1	1,88	JG6	1,59		JU1	1,98	CU3	1,73
CC5	1,85	MH4	1,59		LG1	1,97	Q7	1,73
ZA6	1,79	OJ9	1,58		NC9	1,92	CT6	1,72
CL0	1,78	OG0	1,57		QG2	1,91	JB9	1,70
UN6	1,76	K0	1,56		O2	1,86	AN0	1,68
MA1	1,76	BG3	1,55		LR6	1,86	G8	1,67

Fonte: Autoria Própria

Essa informação se torna importante, quando analisamos algumas dessas cidades. Exemplo ZG5 de 2015 com 45.938 habitantes, que antes era classificada como uma anomalia, e nesse novo experimento, devido aos dados do FIRJAN, os gastos da cidade foram justificados. Um gasto em particular foi com educação onde foi lançado R$ 41.344.103,44 e o índice IFDM referente a educação foi 0,7152.

Já como exemplo oposto, podemos citar a cidade K8, que não era considerada uma anomalia, e agora no novo experimento, foi considerada como anomalia. Para melhor compreensão, a tabela 5 faz um comparativo entre as cidades ZG5, que nesse experimento não é mais uma anomalia, com a cidade K8, que passou a ser uma anomalia. Como pode ser observado, a cidade ZG5 tem os índices FIRJAN, melhores do que os da cidade K8. Tornando injustificável gasto da cidade K8 como, por exemplo, educação que foi 2,5 vezes maior que a outra cidade, e seu IFDM referente a educação foi 0,5410 enquanto que da ZG5 0.6848. Ao pesquisar sobre a cidade K8, foi encontrada uma notícia que a corrupção virou rotina no município. Em outra notícia consta que o ex-prefeito foi condenado pelo Ministério Público do estado. Isso demonstra que ao compor os dados, a classificação da cidade como uma anomalia se tornou mais concisa.

4.4 RESULTADOS

Assim segue um resumo das principais anomalias encontradas e um breve descritivo de cada uma delas:

- Cidade HH4: aparece em todos os experimentos, pois se trata de uma cidade sem gasto com Educação, e um valor bem alto com Desportos & Lazer;
- Cidade DL5: cidade com 11300 habitantes que aparece em todos os experimentos a partir de 2014. No ano de 2016 teve gasto de R$ 52.215.000,00 com Saúde o que é 7 vezes mais do que a média de municípios de mesmo porte. Ainda no ano de 2016, gasto de mais de R$ 81 milhões com Educação, aproximadamente 8 vezes a mais que média de municípios de mesmo porte. Encontra uma reportagem, onde consta que cinco pessoas foram presas em operação deflagrada pelo Ministério Público do Espírito Santo. A reportagem aponta o envolvimento de agentes políticos e servidores, onde receberam propina de empresários dos ramos de limpeza pública e transporte coletivo;
- Cidade KK5: cidade de 10488 habitantes. No ano de 2015 teve lançamentos de

Tabela 5 – Comparativo ZG5 normal x K8 anômala - 2015

Situação	NORMAL	ANÔMALA
Cidade	ZG5	K8
População	45928	42439
01 - Legislativa	3.584.711,27	2.422.204,43
04 – Administração	6.117.980,87	9.872.621,83
08 – Assistência Social	2.711.595,67	2.997.331,04
10 - Saúde	39.335.669,04	10.710.964,49
12 – Educação	10.277.573,47	25.681.496,25
13 - Cultura	1.209.849,40	313.656,98
15 - Urbanismo	226.010,70	787.502,16
17 - Saneamento		71.710,00
18 – Gestão Ambiental		
20 - Agricultura		1.426.654,73
26 - Transporte	3.284.056,43	36.900,00
27 - Desporto e Lazer	10.442.188,08	692.466,52
28 - Encargos Especiais		3.657.441,30
ifdm-educacao	0,6848	0,5410
ifdm-saude	0,7472	0,3362
ifdm-emprego-renda	0,2490	0,3002
ifdm-geral	0,5604	0,3925
ifgf-geral	0,4630	0,2865
ifgf-custo-divida	0,9260	0,2036
ifgf-gestao-pessoal	0,4419	0,6230
ifgf-investimento	0,3550	0,0728
ifgf-liquidez	0,6407	0,4037
ifgf-receita-propria	0,2088	0,0832
LOF	1,26	1,66

Fonte: Autoria Própria

gasto muito superior a municípios de mesmo porte. Gasto de R$ 43.741.186,58 com administração, o que representa 13 vezes a mais que a média de cidade de mesmo porte. Gasto de R$ 20.116.902,07 com assistência social, ou 20 vezes a mais que a média de cidades semelhantes;

- Cidade LM9: cidade de 31599 habitantes. Gastou com Urbanismo em 2015 e 2016 respectivamente de R$ 99.039.984,66 e R$ 144.638.756,13. Encontrada uma reportagem referente a esquema de caixa 2 e superfaturamento de licitação, praticado pelo prefeito e secretários da prefeitura;
- Cidade MM0: cidade de 10102 habitantes. Gastou R$ 40.698.141,76 em urbanismo em 2015 e R$ 59.933.211,40 em 2016, respectivamente 15 e 18 vezes mais que a média de cidades de mesmos porte.

Algo que foi recorrente nesse experimento, são cidades litorâneas com gasto acima do normal com urbanismo. Seria interessante uma analise mais aprofundada nas contas dessas cidades, talvez solicitando junto ao município os comprovantes desses gastos e suas respectivas licitações, bem como fiscalizar as obras realizadas.

4.5 DISCUSSÃO

A técnica se mostrou promissora e robusta, principalmente devido ao fato de que 229 cidades, que na reprodução do experimento de Kintopp (2017) descrito na seçao 4.1, não foram consideradas anômalas e com a composição dos dados nessa nova abordagem foram apontadas como anomalias em 2015. O mesmo ocorreu em 2016 com 281 cidades apontadas como anomalias. E conforme tabela 5 da secção 4.3 é possível verificar que a cidade K8 é uma cidade que tem que ser considerada uma anomalia.

Assim os objetivos propostos foram atingidos e com composição dos dados do SICONFI com DATASUS, e do SICONFI com FIRJAN os resultados apontados denotam uma melhor precisão das anomalias encontradas, ou seja, menos falsos positivos.

Outra consideração importante é que das 4 cidades que se repetem entre as dez maiores anomalias de 2015 e 2016, DL5, KK5, LM9 e MM0, duas delas foram encontradas reportagem sobre o envolvimento de agentes públicos em esquema de corrupção (DL5 e LM9).

5 CONCLUSÃO

Esse trabalho aplicou o algoritmo LOF em uma nova abordagem, compondo os dados do SICONFI com outras bases de dados como DATASUS e índices disponibilizados pelo FIRJAN. Foi demonstrado assim, que com essa nova aplicação, torna a tarefa de fiscalização facilitada, sem a necessidade de se analisar os índices das cidades, permitindo focar as análises apenas nas distorções referente aos gastos.

5.1 TRABALHOS FUTUROS

O artigo referente a esse trabalho foi aceito e premiado no evento InterFORENSICS de 2019 realizado em São Paulo. Durante o evento surgiu por parte de agentes da Polícia Civil e Federal alguns questionamentos. Sendo assim, das dúvidas e ideias que foram relatadas, algumas podem ser colocadas como trabalhos futuros:

- Aplicar a técnica em dados brutos, ou seja, nas notas de gastos das prefeituras: Peritos da Polícia Civil e Federal tem acesso aos dados das prefeituras, ou seja, não são os dados consolidados, mas os fatos geradores.
- Criação de um *Dashboard* onde as pessoas possam pesquisar as anomalias encontras, e assim poder fiscalizar a sua cidade;
- Criação de um *bot* que envie a mensagem via rede social das anomalias encontradas;
- Aplicar a técnica em outras bases, que não sejam dados públicos, que vise encontrar fraudes (bancos, empreiteiras, cooperativas, entre outras).

REFERÊNCIAS

ABRAHAM, B.; BOX, G. Bayesian analysis of some outlier problems in time series. **Biometrika**, v. 66, p. 229–236, 08 1979.

AMARAL, F. **Aprenda Mineração de Dados: Teoria e Prática**. Rio de Janeiro: Alta Books, 2016. 240 p. ISBN 97887576089889.

AMER, M.; GOLDSTEIN, M. Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer. In: . [S.l.: s.n.], 2012.

ARNING, A.; AGRAWAL, R.; RAGHAVAN, P. A linear method for deviation detection in large databases. In: **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA**. [s.n.], 1996. p. 164–169. Disponível em: <http://www.aaai.org/Library/KDD/1996/kdd96-027.php>.

ASSEISS, M. S. G. Aplicação do processo de descoberta de conhecimento em banco de dados acadêmico utilizando as tarefas de agrupamento e classificação . In: **Universidade Estadual Paulista**. São Paulo: [s.n.], 2017. p. 115. Disponível em: <https://repositorio.unesp.br/handle/11449/151251>.

AUSKALNIS, J.; PAULAUSKAS, N.; BASKYS, A. Application of local outlier factor algorithm to detect anomalies in computer network. **Elektronika ir Elektrotechnika**, v. 24, n. 3, 2018. ISSN 2029-5731. Disponível em: <http://eejournal.ktu.lt/index.php/elt/article/view/20972>.

BABBAR, S. Detecting and describing non-trivial outliers using bayesian networks. In: **2015 International Conference on Cognitive Computing and Information Processing(CCIP)**. [S.l.: s.n.], 2015. p. 1–6.

BARBARÁ, D.; WU, N.; JAJODIA, S. Detecting Novel Network Intrusions Using Bayes Estimators. In: **Proceedings of the 2001 SIAM International Conference on Data Mining**. [S.l.: s.n.], 2001.

BERNI, A. **Survey of Kolmogorov Complexity and its Applications**. 2018. 7 p. Disponível em: <https://www.ece.uic.edu/˜devroye/courses/ECE534/project/project_Andrew_Berni.pdf>.

BOENTE, A. N. P.; GOLDSCHMIDT, R.; ESTRELA, V. Uma metodologia para apoio à realização do processo de descoberta de conhecimento em bases de dados. **Simpósio de Excelência em Gestão e Tecnologia**, SEGeT, p. 14, 08 2008.

BREUNIG, M. et al. Lof: Identifying density-based local outliers. In: **Proceedings Of The 2000 Acm Sigmod International Conference On Management Of Data**. [S.l.]: ACM, 2000. p. 93–104.

BREUNIG, M. M. et al. OPTICS-OF: identifying local outliers. In: **Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD '99, Prague, Czech Republic, September 15-18, 1999, Proceedings**. [s.n.], 1999. p. 262–270. Disponível em: <https://doi.org/10.1007/978-3-540-48247-5_28>.

CABENA, P. **Discovering data mining: from concept to implementation**. [S.l.]: Prentice Hall, 1998. (An IBM Press Book Series). ISBN 9780137439805.

CAMILO, C. O.; SILVA, J. C. d. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. Goiânia, 2009. 29 p.

CARVALHO, J. Vinícius de F.; CHIANN, C. Redes bayesianas: um método para avaliação de interdependência e contágio em séries temporais multivariadas. **Revista Brasileira de Economia**, v. 67, p. 201–217, 06 2013.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection. **ACM Computing Surveys**, 2009. ISSN 03600300.

CHAPELLE, O.; SCHÖLKOPF, B.; ZIEN, A. **Semi-Supervised Learning**. 1. ed. Cambridge, Massachusetts: The MIT Press, 2006. 524 p. ISBN 9780262033589.

COHEN, W. W. Fast effective rule induction. In: **In Proceedings of the Twelfth International Conference on Machine Learning**. [S.l.]: Morgan Kaufmann, 1995. p. 115–123.

CVETKOVIĆ, B.; LUSTREK, M. Risk assessment using local outlier factor algorithm. In: . [S.l.: s.n.], 2012.

DESFORGES, M. J.; JACOB, P. J.; COOPER, J. E. Applications of probability density estimation to the detection of abnormal conditions in engineering. **Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science**, v. 212, n. 8, p. 687–703, 1998. Disponível em: <https://doi.org/10.1243/0954406981521448>.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification (2Nd Edition)**. New York, NY, USA: Wiley-Interscience, 2000. ISBN 0471056693.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37, 1996. ISSN 0738-4602.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: um guia prático**. Rio de Janeiro: Elsevier, 2005. 261 p. ISBN 9788535218770.

GOLDSTEIN, M.; DENGEL, A. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. In: . [S.l.: s.n.], 2012.

GRUBBS, F. E. Procedures for detecting outlying observations in samples. **Technometrics**, Taylor Francis, v. 11, n. 1, p. 1–21, 1969.

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of Data Mining**. 1. ed. Massachusetts: A Bradford Book The MIT Press, 2001. 546 p. ISBN 026208290X.

HAWKINS, D. **Identification of Outliers**. [S.l.]: Chapman and Hall, 1980. (Monographs on applied probability and statistics). ISBN 9780412219009.

KAO, L.; HUANG, Y. Association rules based algorithm for identifying outlier transactions in data stream. In: **2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)**. [S.l.: s.n.], 2012. p. 3209–3214. ISSN 1062-922X.

KASTURE, P.; GADGE, J. Cluster based outlier detection. **International Journal of Computer Applications**, v. 58, 10 2012.

KIKUCHI, M. et al. Confidence interval of probability estimator of laplace smoothing. In: . [S.l.: s.n.], 2015. p. 1–6.

KINTOPP, P. M. **Aplicação de Técnicas de Aprendizado de Máquina em Dados Públicos para Detecção de Anomalias**. Medianeira: Universidade Tecnológica Federal do Paraná, 2017. 58 p.

KNORR, E. M.; NG, R. T. Algorithms for mining distance-based outliers in large datasets. In: **Proceedings of the 24rd International Conference on Very Large Data Bases**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. (VLDB '98), p. 392–403. ISBN 1-55860-566-5. Disponível em: <http://dl.acm.org/citation.cfm?id=645924.671334>.

KNORR, E. M.; NG, R. T. Finding intensional knowledge of distance-based outliers. In: **VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK**. [s.n.], 1999. p. 211–222. Disponível em: <http://www.vldb.org/conf/1999/P21.pdf>.

LATECKI, L. J.; LAZAREVIC, A.; POKRAJAC, D. Outlier detection with kernel density functions. In: **Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition**. Berlin, Heidelberg: Springer-Verlag, 2007. (MLDM '07), p. 61–75. ISBN 978-3-540-73498-7. Disponível em: <http://dx.doi.org/10.1007/978-3-540-73499-4_6>.

LIMA, L. De olho no prefeito. **Revista Época**, Brasil, v. 972, p. 78–81, 2017. ISSN 1415-5494. Disponível em: <https://epoca.globo.com/politica/noticia/2017/02/rede-de-cidadaos-voluntarios-que-fiscaliza-prefeitos-e-vereadores.html>.

MCKINNEY, W. **Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython**. O'Reilly Media, 2017. ISBN 9781491957615. Disponível em: <https://books.google.com.br/books?id=UiM3DwAAQBAJ>.

MEDEIROS, H. J. de. **O Papel do Ministério Público no Combate à Corrupção**. 32 p. Tese (Doutorado) — Ministério Público Federal, Brasilia, 2003. Disponível em: <http://www.mpf.mp.br/atuacao-tematica/ccr5/noticias-1/eventos/docs-monografias/monografia_2_lugar.pdf>.

Ministério do Planejamento, Desenvolvimento e Gestão. **Aplicativos, visualizações e infográficos produzidos com dados abertos**. 2018. Disponível em: <http://dados.gov.br/aplicativos>. Acesso em: 9 de agosto de 2018.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: **Sistemas Inteligentes Fundamentos e Aplicações**. 1. ed. Barueri-SP: Manole Ltda, 2003. p. 89–114. ISBN 85-204-168.

NETO, F.; DINIZ, C. **Técnicas estatísticas em data mining**. [S.l.]: IMCA, 2002. (Monografias del IMCA). ISBN 9789972899126.

Observatório Social do Brasil. **Observatórios pelo Brasil**. 2018. Disponível em: <http://osbrasil.org.br/observatorios-pelo-brasil/>. Acesso em: 19 de agosto de 2018.

PLATT, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: **ADVANCES IN LARGE MARGIN CLASSIFIERS**. [S.l.]: MIT Press, 1999. p. 61–74.

QUINLAN, J. R. Generating production rules from decision trees. In: **Proceedings of the 10th International Joint Conference on Artificial Intelligence - Volume 1**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1987. (IJCAI'87), p. 304–307. Disponível em: <http://dl.acm.org/citation.cfm?id=1625015.1625078>.

RÄTSCH, G. et al. Constructing boosting algorithms from svms: an application to one-class classification. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 9, p. 1184–1199, set. 2002.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. Third. Upper Saddle River, NJ: Prentice Hall, 2010. (Series in Artificial Intelligence). Disponível em: <http://aima.cs.berkeley.edu/>.

RUTS, I.; ROUSSEEUW, P. J. Computing depth contours of bivariate point clouds. **Comput. Stat. Data Anal.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 23, n. 1, p. 153–168, nov. 1996. ISSN 0167-9473. Disponível em: <http://dx.doi.org/10.1016/S0167-9473(96)00027-8>.

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**, v. 30, p. 84–94, 2000.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Us ed. [S.l.]: Addison Wesley, 2005. Hardcover. ISBN 0321321367.

Transparency International. **Índice de Percepção da Corrupção**. Berlin, Germany, 2019. 12 p. Disponível em: <www.transparency.org>.

UPADHYAYA, S.; SINGH, K. Nearest neighbour based outlier detection techniques. **International Journal of Computer Trends and Technology**, v. 3, p. 299–303, 01 2012.

WILLIAMS, J.; LI, Y. Comparative study of distance functions for nearest neighbors. In: . [S.l.: s.n.], 2008. p. 79–84.

ZENGAN, G. Application of cluster-based local outlier factor algorithm in anti-money laundering. In: **Proceedings - International Conference on Management and Service Science, MASS 2009**. [S.l.: s.n.], 2009. ISBN 9781424446391.

ZHOU, Z.-H. **Three perspectives of data mining**. [S.l.], 2003. v. 143, 139–146 p. Disponível em: <www.elsevier.com/locate/artint>.