Dissertação:

Um Sistema de Apoio à Detecção de Anomalias em Dados Governamentais usando Múltiplos Classificadores

Rafael Alexandrino Spíndola de Souza

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, 2021

Rafael Alexandrino Spíndola de Souza

Dissertação

Dissertação apresentada ao
Programa de Pós-Graduação em Informática - PPGI
do Centro de Informática, da Universidade Federal da Paraíba,
como requisito para a obtenção do grau de Mestre em Informática

Orientador: Tiago Maritan Ugulino de Araújo

Abril de 2021

Catalogação na publicação
Seção de Catalogação e Classificação

S729s Souza, Rafael Alexandrino Spíndola de.
Um sistema de apoio à detecção de anomalias utilizando múltiplos classificadores / Rafael Alexandrino Spíndola de Souza. - João Pessoa, 2021.
82 f. : il.

Orientação: Tiago Maritan Ugulino de Araújo Araújo.
Dissertação (Mestrado) - UFPB/CI.

1. Tratamento de dados. 2. Detecção de anomalias. 3. Detecção de outliers. 4. Aprendizagem supervisionada. 5. Aprendizagem não supervisionada. 6. Mineração de]dados. 7. Dados governamentais. I. Araújo, Tiago Maritan Ugulino de Araújo. II. Título.

UFPB/BC CDU 004.62

Elaborado por Gracilene Barbosa Figueiredo - CRB-15/794

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Dissertação para o Programa de Pós-Graduação em Informática - PPGI intitulado ***Um Sistema de Apoio à Detecção de Anomalias em Dados Governamentais usando Múltiplos Classificadores*** de autoria de Rafael Alexandrino Spíndola de Souza, submetido à banca examinadora constituída pelos seguintes professores:

Prof. Dr. Tiago Maritan Ugulino de Araújo
Universidade Federal da Paraíba - UFPB

Prof. Dr. Thais Gaudencio do Rego
Universidade Federal da Paraíba - UFPB

Prof. Dr. Telmo de Menezes e Silva Filho
Universidade Federal da Paraíba - UFPB

Prof. Dr. Rostand Edson Oliveira Costa
Universidade Federal da Paraíba - UFPB

Coordenador(a) do Programa de Pós-Graduação do Centro de Informática
Dr. Tiago Pereira do Nascimento
CI/UFPB

João Pessoa, 22 de abril de 2021

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

AGRADECIMENTOS

Primeiramente, agradeço a Deus e aos Mentores por concederem-me saúde e me contemplarem com o privilégio de poder estudar. Além Dele, sou infinita e especialmente grato à minha esposa, Luana, por toda a motivação e apoio que me ofertou ao longo de todo o trabalho desenvolvido nessa respeitada Universidade. Aos meus pais e irmãos, sempre próximos e com palavras de motivação e apoio, ressalto a minha profunda gratidão! Ao grande Nivaldo, por todas as conversas e ajudas que me proporcionou durante toda pesquisa. Ao Tribunal de Contas do Estado Paraíba, minha segunda família, agradeço por todo apoio, flexibilização e crédito ofertado, especialmente aos Chefes Carol e Luzemar, por sempre nos apontar o melhor caminho rumo ao desenvolvimento. Por fim, agradeço profundamente ao Professor Tiago Maritan, pelas diversas horas dedicadas, pelas valiosas orientações emitidas e pela abnegação demonstrada, apoiando e acreditando no desenvolvimento da pesquisa desde o primeiro momento.

RESUMO

Com quantidades cada vez maiores de dados para serem analisados e corretamente interpretados, a Detecção de Anomalias (ou *Outliers*) surge como uma das áreas de grande impacto no contexto da Mineração de Dados (MD). Suas aplicações estendem-se aos mais diversos campos da atuação humana, notadamente na medicina, administração, gestão de processos, ciência da informação, física, economia e em muitas outras atividades. Neste trabalho, propõe-se um Sistema não paramétrico de apoio à detecção de eventos aberrantes em bases de dados estacionárias, provenientes da Administração Pública e relacionadas aos Dados de Dispensas e Inexigibilidades de Licitações do Governo Federal entre 2014 e 2019, aos Dados Orçamentários do Fundo Municipal de Saúde de João Pessoa – PB, entre 2016 e 2020, e aos Dados relativos ao Gerenciamento de Frotas do Estado da Paraíba, entre 2017 e 2019. A soluçao proposta reúne múltiplos algoritmos de detecção supervisionada e não supervisionada (OCSVM, LOF, CBLOF, HBOS, KNN, Isolation Forest e Robust Covariance) para classificar os eventos como anomalias. Os resultados mostraram que, do total de eventos retornados pela solução, em média, 90,07% deles foram corretamente identificados como *outliers*. Portanto, há indicativos de que a solução proposta tem potencial de contribuir para as atividades de apoio a auditoria governamental, bem como para os processos de gerenciamento e tomada de decisão, estes decorrentes da interpretação dos fenômenos presentes nos dados.

Palavras-chave: Detecção de anomalias, Detecção de *outliers*, Aprendizagem supervisionada, Aprendizagem não supervisionada, Mineração de Dados, Dados Governamentais.

ABSTRACT

With increasing amounts of data to be analyzed and correctly interpreted, Anomaly Detection (or Outliers) appears as one of the areas of significant impact in the context of Data Mining (DM). Its applications extend to the most diverse human activity fields, such as medicine, administration, process management, information science, physics, economics, and many other activities. In this work, we propose a non-parametric system to support the detection of aberrant events in stationary databases. The database comes from the Public Administration and related to the Federal Government's Disbursement and Bidding Data between 2014 and 2019, to the Fund's Budget Data Municipal Health of João Pessoa - PB, between 2016 and 2020, and Data on the Fleet Management of the State of Paraíba between 2017 and 2019. The proposed solution combines some supervised and unsupervised detection algorithms (OCSVM, LOF, CBLOF, HBOS, KNN, Isolation Forest, and Robust Covariance) to classify events as anomalies. The results showed that the solution identifies an average of 90.07% correctly events as outliers. Therefore, there are indications that the proposed solution can contribute to government audit support activities and management and decision-making processes, these arising from the interpretation of the phenomena present in the data.

Key-words: Anomaly detection, outlier detection, supervised learning, unsupervised learning, Data Mining, Government Data.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

TCEPB - *Tribunal de Contas do Estado da Paraíba*

GerFrotas - *Gerenciamento de Frotas do Estado da Paraíba*

DIGF - *Dispensas e Inexigibilidade de Licitações realizadas pelo Governo Federal*

FMSJP - *Fundo Municipal de Saúde do Município de João Pessoa*

PBE - *Programa Brasileiro de Etiquetagem*

DM - *Data Mining*

KDD - *Knowledge Discovery in Databases*

AC - *Abordagem Convencional*

AP - *Abordagem Proposta*

SWOT - *Strengths, Weaknesses, Opportunities and Threats*

LOF – *Local Outlier Factor*

ABOD – *Angle-Based Outlier Detection*

AED - *Análise Exploratória de Dados*

kNN - *k-Nearest Neighbors*

OGD - *Open Governments Data*

e-GOV - *Governo Eletrônico*

TIC - *Tecnologia de Informação e Comunicação*

SVM - *Suport Vector Machine*

AFC - *Automated Fare Collection*

OCSVM - *One-Class Suport Vector Machine*

CBLOF - *Clustering Based Local Outlier Factor*

CB-LOF - *Cube-Based Local Outlier Factor*

ILOF - *Incremental Local Outlier Factor*

MLP - *Multi-Layer Perceptron*

LOCI - *Local Correlation Integral*

IForest - *Isolation Forest*

HBOS - *Histogram-Based Outlier Score*

RC - *Robust Covariance*

Sumário

1 Introdução

Atualmente, quantidades significativas de dados são geradas. Somente para o ano de 2020, a projeção do volume de dados adquirido atingiu o valor de 44 (quarenta e quatro) trilhões de gigabytes [1].

Ações das mais simples e variadas, como postar vídeos em plataformas de *streaming* ou a publicação de fotografias via Facebook ou Instagram, são destacadas como atividades geradoras de volumes imensos de dados. A rede social Facebook, por exemplo, gera diariamente um volume superior a 500 (quinhentos) terabytes de dados relacionados a seus usuários [2]. Enquanto isso, somente no Brasil, espera-se que o volume de dados em 2020 atinja cerca de 1,60 trilhão de gigabytes [3].

Apesar de a iniciativa privada gerar grandes quantidades de dados, os órgãos e entidades integrantes da Administração Pública também são responsáveis pela aquisição e disponibilização de volumes significativos de dados/informações de interesse social. Anualmente, no âmbito da atividade estatal, são realizadas inúmeras aquisições de bens e serviços, pagamentos a funcionários, celebrações de contratos, viabilização de obras e todas essas informações devem, por força do que determina o ordenamento jurídico, estar disponíveis em tempo real ao pleno conhecimento e acompanhamento da sociedade (artigo 48, § 10, inciso II da Lei Complementar 101, de 04 de maio de 2000).

Diante do cenário em que se vislumbra o aumento estrito nas quantidades de dados gerados, torna-se imprescindível a adoção de tecnologias computacionais capazes de, não somente processá-los com alto desempenho e disponibilidade, mas também implementar análises concomitantes das informações e tendências contidas nos dados adquiridos, de forma a possibilitar a extração de conhecimento e sua disseminação na sociedade.

Os dados colhidos pelo *Facebook*, por exemplo, congregam inúmeros registros de seus usuários que, se analisados, podem revelar informações importantes capazes de serem aplicadas para propósitos mais específicos ou abrangentes, como orientar empreendedores na tomada de decisões, orientar a produção de determinados ramos da indústria e descobrir tendências. De outro lado, a análise dos dados colhidos pela Administração Pública pode conduzir à otimização dos recursos da sociedade, promover o engajamento cívico no Governo Eletrônico (E-Gov) [4], além de identificar disparidades que exijam a adoção de novas políticas públicas voltadas às necessidades da população e que sejam amparadas por tendências gerais implícitas nos dados.

Ilustrando, somente a Administração Estadual Paraibana registra dados, para posterior conferência e fiscalização, referentes a cerca de 35 mil abastecimentos de veículos ao mês. Além disso, a Secretaria de Estado da Saúde, no período compreendido entre as datas de 23 de março a 02 de abril de 2020, lapso relacionado ao enfrentamento da

pandemia do SARS-CoV-2, registrou 3.054 empenhos de despesas associadas a aquisições de materiais de saúde, que totalizaram R$ 77.522.148,15. [5]. Na esfera municipal e no mesmo período, o Fundo de Saúde de João Pessoa, que é apenas um dos dezoito órgãos pertencentes à edilidade, registrou 217 empenhos de despesas relacionadas à promoção das ações de saúde pública voltadas à população, as quais totalizaram R$ 36.941.453,97.

Vale destacar que os dados, *per si*, representam apenas uma simples observação sobre o estado do mundo em um determinado instante. Constituem-se como um registro dos atributos de um ente, objeto ou fenômeno [6]. Assim, para que se possa extrair conhecimento desses registros, é necessário que eles sejam submetidos a análises criteriosas capazes de capturar informações dotadas de relevância e propósito.

Na área da Ciência da Informação (CI), há diversos estudos sendo realizados para avaliar os impactos sociais, econômicos e culturais da utilização de algoritmos na análise de dados e extração de informações [7][8][9].

Sabe-se que a análise manual para grandes bases de dados nem sempre é viável às empresas privadas e aos órgãos integrantes da administração pública, uma vez que recursos disponíveis, como tempo e mão de obra, são frequentemente escassos.

Diante dessas dificuldades, o presente trabalho procura contribuir fornecendo uma estrutura computacional que, fazendo uso de conceitos amplamente abordados pela literatura científica, seja capaz de oferecer uma resposta rápida e suficientemente objetiva aos usuários, particularmente aos auditores governamentais, permitindo-lhes que sejam selecionadas instâncias que constituam amostras consideradas anômalas, de forma a facilitar os processos de extração de amostras de auditoria e a verificação da conformidade das despesas públicas.

1.1 Motivação

Os dados digitalmente armazenados constituem importante fonte de pesquisa, principalmente para as atividades associadas ao plano estratégico e tático das corporações. Desde a primeira década do século XXI, quase toda a informação produzida passou a ser colocada diretamente no mundo de *bits* e *bytes*, e o que existe no meio físico, como em livros, revistas e jornais, para citar alguns exemplos, passou a ser transmutado para discos rígidos ou para memórias digitais [10].

Nas atividades desenvolvidas pelo setor público, a tendência de registros e disponibilização de dados em meios digitais também se faz presente, uma vez que com o advento da Constituição da República Federativa do Brasil de 1988, a transparência foi alçada à condição de princípio e passou a balizar todo o ordenamento jurídico, sendo, inclusive, reforçada pela Lei de Acesso à Informação (Lei 12.527, de 18 de novembro de 2011) e pela Lei Complementar 101, de 04 de maio de 2000, comumente conhecida como Lei de

Responsabilidade Fiscal.

Nessa esteira, diversas informações relacionadas às contas públicas e atos governamentais passaram a ser registradas em meios digitais e disponibilizados ao público em geral, com o intuito de se amplificar o controle social sobre as ações do governo, possibilitado pelo advento de diversas tecnologias de comunicação digital (Governo Eletrônico – E-Gov [4]).

Assim, a quantidade de informações registradas em meios digitais motivou a formulação da seguinte questão de pesquisa (QP):

QP1. É possível otimizar o processo de decisão dos gestores públicos e os planos de fiscalização através da implementação de técnicas computacionais aplicadas aos dados registrados pelos órgãos públicos?

Na literatura científica, identificou-se trabalhos que, embora não se relacionem diretamente com dados de origem pública, apresentavam técnicas de mineração de dados e detecção não supervisionada de *outliers* aplicadas à verificação de possíveis anomalias em conjuntos de dados numéricos [11].

Nessa linha, técnicas de detecção de anomalias em espaços multivariados, como o *Local Outliler Factor* (LOF) e o *Isolation Forest* (IForest), surgem como ferramentas capazes de otimizar tanto o processo fiscalizatório associados às despesas públicas como a própria ação voltada à tomada de decisões.

Diante disso, formulou-se uma nova questão de pesquisa, cujo objetivo é verificar se o uso de alguma das técnicas de detecção de anomalias pode permitir um maior controle sobre os gastos associados às despesas públicas.

QP2. Técnicas de detecção de anomalias aplicadas aos dados públicos são capazes de selecionar, na população disponível, uma amostra representativa de supostos candidatos a outliers?

Em estatística, *"Outliers"*, também conhecidos como valores aberrantes, valores atípicos ou anomalias, são medidas que apresentam grande afastamento das demais observações presentes na série [12]. Além disso, a existência desse tipo de observação na amostra deve conduzir os experimentadores a uma análise cautelosa, já que pode comprometer a interpretação dos resultados dos testes estatísticos a serem aplicados, em virtude de efeitos que levam a forte flutuação das medidas de tendência central.

Como se depreende, é de grande interesse que nas atividades de gestão, não somente associadas à área pública, sejam desenvolvidas soluções capazes de identificar possíveis anomalias, fornecendo uma maneira objetiva de se racionalizar os processos de acompanhamento dos resultados de determinada ação, bem como orientar na tomada de outras decisões. Aliás, a detecção de possíveis anomalias presentes no conjunto de dados poderia,

em alguma medida, auxiliar no fortalecimento dos controles administrativos da entidade, possibilitando a mitigação de erros e fraudes.

Entretanto, é necessário que as amostras consideradas anômalas pela aplicação da técnica correspondam a instâncias inconsistentes, o que requer o desenvolvimento de métodos de validação capazes de assegurar, com alguma margem de confiança, a característica destoante.

Com esse objetivo, propôs-se a última questão de pesquisa:

QP3. As instâncias consideradas anômalas pela técnica de detecção de outliers podem ser validadas por algum método ou técnica? Há como medir o contraste da técnica?

Na seara científica, a principal motivação é, a partir da abordagem das questões de pesquisa, propor uma solução computacional capaz de reduzir as dificuldades associadas ao processo de fiscalização e tomada de decisão dos órgãos públicos do Estado da Paraíba, através da seleção de amostras que contenham instâncias contrastantes com os padrões gerais observados. Sabe-se que as entidades integrantes do setor estatal frequentemente não dispõem de recursos humanos suficientes à fiscalização e ao acompanhamento das despesas e políticas públicas em execução, de forma que uma solução computacional responsiva, suficientemente simples e que permita a identificação de possíveis anomalias contribui, sobremaneira, principalmente no delineamento de trilhas de auditoria - escolha de amostras representativas que apresentem maior potencial de risco - e na construção de índices ou parâmetros orientadores da gestão.

Do ponto de vista tecnológico, a motivação é a implementação de uma estrutura de análise codificada capaz de extrair informações de bases de dados de despesas públicas. Em outras palavras, propõe-se uma ferramenta de análise voltada à seara governamental e focada na extração de amostras consideradas anômalas oriundas dos gastos estatais relacionados aos dados de Gerenciamento de Abastecimentos da Frota Estadual (GerFrotas) [13], à execução orçamentária do Fundo Municipal de Saúde de João Pessoa (FMSJP) [5] e aos dados de dispensas e inexigibilidades de licitação do Governo Federal (DIGF) [14].

De outro lado, há motivação social, uma vez em que se fornece à sociedade, particularmente à Administração Pública, uma maneira segura, ágil e científica para se otimizar os processos relacionados à tomada de decisão e fiscalização dos recursos aplicados, os quais comumente atingem montas elevadas.

1.2 Objetivos - Gerais e Específicos

Os principais objetivos deste trabalho são a investigação e o desenvolvimento de uma solução computacional codificada capaz de realizar, para bases de dados de gastos públicos relacionadas ao gerenciamento de abastecimentos da frota Estadual, à execução

orçamentária do Fundo Municipal de Saúde de João Pessoa e aos dados de dispensas e inexigibilidades de licitação do Governo Federal, a extração de amostras de eventos que contenham alta propensão de constituírem-se como *outliers*.

Nesse contexto, como objetivos específicos da pesquisa, estabeleceu-se a investigação das principais técnicas de detecção de anomalias em bases de dados e o desenvolvimento de um sistema que combina essas técnicas de forma a classificar as instâncias de acordo com o perfil de anormalidade apresentado.

Dessa forma, o objetivo geral deste trabalho está associado à detecção de potenciais anomalias presentes nos dados, estas que poderão ser selecionadas e classificadas, fornecendo-se uma amostra de instâncias que, por se afastarem do padrão geral observado, deverão sofrer fiscalização prioritária, contribuindo para o processo de seleção dos gastos públicos que representam os maiores riscos ao Erário.

Por se tratar de uma solução baseada na detecção supervisionada e não supervisionada de anomalias, cujo propósito é a avaliação dos gastos públicos relacionado às três bases aqui tratadas (GerFrotas, DIGF e FMSJP), sua utilização poderá se mostrar valiosa em diversos nichos, como na constatação de fraudes, avaliação das tendências de uma determinada política pública, análise de risco, entre outras aplicações.

1.3 Hipóteses

A proposta a ser defendida nesse trabalho é a de que é possível otimizar o processo de tomada de decisão dos gestores públicos e os planos de fiscalização a partir da estratégia de implementação de uma estrutura computacional capaz de identificar, por técnicas específicas de mineração de dados, possíveis anomalias nas informações relacionadas à execução orçamentária e financeira dos órgãos e entidades públicas, de forma a se permitir a seleção de amostras que racionalizem os processos de acompanhamento e fiscalização das despesas. Tendo-se em vista as questões de pesquisas levantadas, informalmente, pôde-se definir as seguintes hipóteses: 1) Técnicas de detecção de anomalias aplicadas aos dados públicos são capazes de selecionar amostras representativas de candidatos a *outliers*; e 2) As instâncias consideradas anômalas pela técnica de detecção de *outliers* podem ser validadas e o contraste da técnica pode ser aferido.

Formalmente, as hipóteses podem ser apresentadas da seguinte maneira:

1) – ***Hipótese nula H0***: Técnicas de detecção de anomalias aplicadas aos dados públicos não permitem selecionar amostras representativas de candidatos a *outliers*. - ***Hipótese alternativa H1***: Técnicas de detecção de anomalias aplicadas aos dados públicos são capazes de selecionar amostras representativas de candidatos a *outliers*.

2) – ***Hipótese nula H0***: As instâncias consideradas anômalas pela técnica de

detecção de *outliers* não podem ser seguramente validadas, bem como o contraste da técnica não pôde ser aferido. - ***Hipótese alternativa H1***: As instâncias consideradas anômalas pela técnica de detecção de *outliers* podem ser validadas e o contraste da técnica pode ser aferido.

Na Seção 1.4, será apresentada sucintamente a metodologia utilizada, de forma que sua descrição geral será discutida no Capítulo 4, momento em que será apresentado o plano de experimentos realizado a fim de testar as hipóteses gerais apresentadas.

1.4 Metodologia

No presente trabalho, a metodologia balizou-se pela investigação de técnicas e abordagens relacionadas à detecção de *outliers*, as quais pudessem ser utilizadas, no contexto da Administração Pública, para subsidiar tanto os processos relacionados à auditoria quanto àqueles que demandam conhecimento dos gestores públicos para a tomada de decisão. Com esse intuito, e através da sistematização da pesquisa apresentada no Anexo A, procurou-se fazer o levantamento das técnicas e abordagens de detecção de anomalias disponíveis, de maneira a se entender os desafios enfrentados pela área, assim como onde essas abordagens vêm sendo aplicadas e os resultados alcançados. A apresentação dos trabalhos relacionados mais relevantes é realizada no Capítulo 3.

Outrossim, a metodologia consistiu na organização automática do conjunto de dados submetido à solução, de forma a se buscar, simultaneamente, a melhor estrutura de dados para os atributos das chaves primárias e a preservação da maior quantidade de informações relevantes associadas aos atributos das instâncias disponíveis. Para tanto, utilizou-se a linguagem de programação *Python* para aplicar estratégias, baseadas em diversos métodos, capazes de dar a cada um dos atributos a melhor estrutura para interpretação dos dados (*"string", "int", "float", "category",* entre outros).

Na última fase do pré-processamento, o conjunto de dados reestruturado passa pelo processo de padronização dos dados, fazendo com que seus atributos sejam expressos no mesmo perfil de variação de escala.

Passada a fase de estruturação, o conjunto de dados remanescente é submetido ao módulo desenvolvido para detecção de anomalias. Este, por sua vez, é integrado por sete abordagens distintas voltadas à detecção de *outliers*, sendo elas o *Local Outlier Factor (LOF)*, *Isolation Forest (Iforest)*, *Clustering Based Local Outlier Factor - CBLOF*, *Histogram-based outlier score - HBOS*, *One-Class Suport Vector Machine - OCSVM*, *k-Nearest Neighbors Detector - kNN* e *Robust Covariance - RC*, que usa o método *EllipticEnvelope*.

Dessa forma, no módulo de detecção de anomalias, cada instância da base de dados é avaliada pelos estimadores visando a constituição de uma amostra de eventos que

possuam alto grau de anormalidade.

No Capítulo 2, será apresentada a fundamentação teórica para este trabalho e no Capítulo 4, a metodologia da solução proposta será discutida.

1.5 Escopo

Nas seções 1.2 e 1.4, discorreu-se sobre os objetivos gerais e a metodologia utilizada nesta pesquisa, destacando-se que o escopo do trabalho relaciona-se à obtenção de amostras de eventos com alto potencial de constituírem-se como instâncias anômalas.

Entretanto, para que o sistema proposto possa classificar instâncias como *outliers*, é essencial que a base de dados sob análise contenha atributos numéricos. Essa limitação deriva das estratégias de detecção de anomalias utilizadas pela ferramenta proposta, as quais necessitam de dados do tipo numérico para classificar os eventos. Dessa forma, a solução proposta não pode ser aplicada a bases de dados onde não existam atributos numéricos.

Além disso, o trabalho avaliou a solução proposta apenas sobre as bases de dados públicas mencionadas na Seção 1.2, razão pela qual não é possível afirmar se a solução proposta pode ser generalizada para outras bases de dados, inclusive de natureza pública.

1.6 Estrutura da Dissertação

Esta dissertação está organizada em seis capítulos. Neste, apresentam-se as motivações, introduzindo-se conceitos relacionados à análise de dados e a sua interpretação. Ademais, são evidenciadas as questões de pesquisas, os objetivos gerais e específicos e as hipóteses que serão posteriormente testadas.

O Capítulo 2, por sua vez, busca apresentar as fundamentações teóricas do trabalho, desenvolvendo conceitos gerais e definições encontradas na literatura, bem como a contextualização geral de como os conceitos abordados estão sendo aplicados na sociedade. No terceiro capítulo, objetivou-se evidenciar os trabalhos relacionados à pesquisa, particularmente no que se refere às técnicas de detecção de *outliers* e os desafios existentes na área.

No Capítulo 4, é trazida a descrição detalhada da metodologia e da solução proposta, especificando-se os módulos da solução (pré-processamento e detecção de anomalias), as tarefas que são executadas por eles, bem como a estratégia adotada para se extrair uma amostra de possíveis instâncias anômalas.

Por fim, os Capítulos 5 e 6, respectivamente, apresentam os resultados finais obtidos da análise das bases de dados utilizadas neste trabalho e as conclusões decorrentes da

referida análise, sendo que no Capítulo 6 há, ainda, a reapresentação geral de toda a pesquisa e os resultados obtidos.

2 Fundamentação Teórica

O presente capítulo trata dos principais conceitos que fundamentam este trabalho. Inicialmente, serão expostos os principais desafios e oportunidades relacionados ao volume de dados atualmente disponível e o processo de extração de informações dessa matéria prima.

Em seguida, serão apresentados os conceitos relacionados à área de descoberta de conhecimento em bases de dados, à mineração de dados, além de uma visão geral sobre os algoritmos utilizados na tarefa de detecção de anomalias.

Por fim, serão discutidas as possibilidades de aplicação dessas técnicas no cenário real e como elas podem contribuir para a otimização do controle das transações financeiras, das atividades de fiscalização e dos processos de tomada de decisão.

2.1 Era Digital e a Obtenção de Conhecimento

Como discutido no Capítulo 1, a quantidade de informações armazenadas sob a forma digital aumenta a cada dia. Dispositivos como sensores, relógios, eletrodomésticos, câmeras, entre outros, captam dados que refletem o estado de mundo em determinado instante.

Dados podem ser definidos como o elemento puro, quantificável sobre um determinado evento e que pode ser processado [15]. Em outras palavras, são registros soltos que não sofreram nenhuma tipo de análise.

Para que os dados possam ser úteis, devem ser submetidos a um processo de interpretação sistemática que, através da consideração de padrões, associações ou relações, possa extrair informações capazes de subsidiar o processo de formação do conhecimento, este que pode ser entendido como a descoberta de características que possibilitam, em alguma medida, a previsão de fatos futuros [15].

Como os dados despontam como uma fonte riquíssima de informações, o desenvolvimento de técnicas e ferramentas capazes de extraí-las, assim como os seus padrões e relações, tornou-se imprescindível.

Nessa linha, a descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases – KDD*) é o processo que pode ser definido como a identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados [16]. Em suma, *KDD* refere-se às atividades de encontrar conhecimento oculto e relevante em grandes repositórios de dados, utilizando-se de aplicações de alto nível e, em particular, técnicas de mineração de dados.

A *KDD* é de grande interesse para pesquisadores de diversas áreas, como aprendi-

zagem de máquina, reconhecimento de padrões, banco de dados, estatística, aquisição de conhecimentos para sistemas especialistas e visualização de dados. Até mesmo em áreas como a administração, técnicas de *KDD* são utilizadas para descobrir regras, identificar fatores e tendências-chave, descobrir padrões e relacionamentos ocultos em grandes bancos de dados, tudo isso para auxiliar os processos de tomada de decisões sobre estratégias e vantagens competitivas [17] [18] [19].

Entre as técnicas de *KDD* destinadas à exploração de grandes massas de dados, ressalta-se a mineração de dados que, através da busca de padrões consistentes, detecta relacionamentos sistemáticos entre variáveis que são de difícil visualização pelos especialistas. Sendo assim, sua aplicação pode facilitar a extração de conhecimento útil de um conjunto de dados [20].

De forma geral, mineração de dados (também conhecida como prospecção de dados ou *"data mining" - DM*) pode ser definida como o processo de extração de padrões interessantes e potencialmente úteis ou conhecimentos de uma grande quantidade de dados [21]. Pode ser dividida nas seguintes etapas: conhecimento do domínio; pré-processamento; extração de padrões; pós-processamento e utilização do conhecimento.

Na etapa de conhecimento do domínio, reúnem-se as atividades de identificação do problema, definição dos objetivos e metas. No pré-processamento, por sua vez, os esforços são dirigidos à formatação e transformação dos dados, redução de seu volume, seleção e extração de dados. Nesta fase, introduzem-se técnicas que buscam, entre outras finalidades, verificar correlações entre os diversos atributos da base de dados e eliminar aqueles que representam informações redundantes. A fim de ilustrar, suponha dois atributos numéricos que possuam correlação estatística de Pearson muito próxima de 1. Nesse contexto, a interpretação estatística é a de que os atributos comportam-se de maneira muito semelhante, sendo que uma variação aplicada a um deles gera um comportamento similar no outro. Assim, pode-se eliminar da análise um dos atributos sem ocorrer perda de informação relevante.

Logo, percebe-se que a aplicação de estratégias dessa natureza, não somente sobre os atributos numéricos, colabora para a redução do volume de dados, diminuindo, em muitos casos, a própria complexidade do algoritmo.

Outra atividade que pode ser executada sobre o conjunto de dados durante o pré-processamento é a padronização (ou normalização) dos dados. Nessa estratégia, o que se busca é a uniformização da escala de apresentação dos atributos, de forma que todos eles possam ser expressos no mesmo perfil de variação. A principal vantagem do uso da normalização é a possibilidade de reduzir o risco de viés, na medida em que variáveis com escalas muito distintas podem distorcer o resultado retornado pelo algoritmo. Contudo, principalmente em tarefas associadas à detecção de *outliers*, o uso da normalização deve

ser analisado, já que pode suavizar os efeitos das anomalias.

Essas são algumas das vantagens relacionadas à etapa do pré-processamento em mineração de dados. Embora não sejam exatamente iguais, o pré-processamento possui uma correlação alta com a análise exploratória de dados (AED), esta que é uma maneira de observar as características principais do conjunto de dados estudado e representá-las de uma forma que seja fácil de visualizar. Na AED, podem ser utilizadas diversas técnicas como diagrama de caixa, histograma, diagrama de Pareto, gráficos de dispersão, diagrama de ramos e folhas, coordenadas paralelas, razão de possibilidades, redução de dimensionalidade [22].

Nesse cenário, a AED pode ser encarada como uma parte essencial do pré - processamento, que congrega operações ainda mais abrangentes, como verificação de combinações inconsistentes nos dados (animal: "cão", asa: "sim"), atributos não totalmente preenchidos (*missing values*), entre outras.

Na fase subsequente, a de extração de padrões, determina-se a tarefa a ser realizada, que a depender dos objetivos, pode variar entre classificação, clusterização, regressão, regras de associação, detecção de anomalias, etc. Definida a tarefa, elege-se o algoritmo que irá executá-la.

Especificamente em relação às tarefas de detecção de anomalias, muitas técnicas podem servir de base à escolha do algoritmo. Entre elas, pode-se citar as baseadas em densidade, redes Bayesianas e análise de *clusters*.

Neste trabalho, o objetivo principal associa-se ao processo de identificar, para as três bases de dados de despesas públicas utilizadas (GerFrotas, DIGF e FMSJP), pré-processadas e no espaço n-dimensional, amostras de instâncias que apresentam alta propensão de constituírem-se como *outliers*.

Com esse intuito, utilizou-se os seguintes algoritmos para a detecção das anomalias: *Local Outlier Factor (LOF)* [23], *Clustering Based Local Outlier Factor (CBLOF)* [24], *Histogram-based outlier score (HBOS)* [25]; *One-Class Suport Vector Machine (OCSVM)* [26], *k-Nearest Neighbors Detector (kNN)* [27], *Isolation Forest (IForest)* [28] e *Robust Covariance (RC)* que usa o método *EllipticEnvelope* [29].

A escolha dos algoritmos (ou estimadores) considerou os potenciais de utilização e os resultados alcançados em diversos trabalhos relacionados na literatura, como em [30],[31], [32] e em [33]. O funcionamento de cada estimador é detalhadamente tratado nos trabalhos indicados nas referências [23], [24], [25], [26], [27], [28] e [29].

A última fase do processo de mineração de dados é o pós-processamento, oportunidade em que haverá análise dos padrões descobertos e a verificação se as soluções obtidas são capazes de explicar o problema. Se a solução for satisfatória, considera-se que houve

obtenção de conhecimento novo e útil, pronto para ser utilizado. Caso contrário, à luz do método científico, revisitam-se as hipóteses, realizam-se os ajustes necessários e repete-se o experimento.

Na Seção 2.2, apresenta-se uma percepção geral sobre as tarefas relacionadas à detecção de anomalias.

2.2 Visão Geral sobre Detecção de Anomalias

As aplicações de *KDD* focam seus esforços na obtenção de padrões comuns presentes nos dados. Entretanto, para diversas atividades, como as que envolvem a ocorrência de atividades criminosas, pode ser muito mais interessante encontrar eventos raros, desvios consideráveis do padrão geral observado ou casos excepcionais [23].

Vê-se, então, que a própria definição de *KDD* já deixa claro que os esforços são direcionados às tarefas de detecção de padrões e não ao tratamento adequado das anomalias. Além disso, mesmo na área de detecção de *outliers*, pode-se encontrar na literatura trabalhos que encaram a classificação de uma instância como *outlier* sob o ponto de vista puramente binário [23]. Dito de outra forma, a estratégia de classificação trazida nesses trabalhos considera apenas duas possibilidades para a instância em estudo, isto é, ou ela é ou não é *"outlier"*. A desvantagem marcante dessa interpretação é que, para aplicações reais, frequentemente a situação é bem mais complexa.

Por exemplo, sob o ponto de vista em que a classificação é binária, a avaliação do grau de anormalidade do comportamento das instâncias muitas vezes depende de ferramentas auxiliares, como a representação gráfica do conjunto de dados (que pode ser inviável a depender do número instâncias e da dimensionalidade da base de dados) ou mesmo a definição de novos critérios capazes de mensurar o grau de aberração de um determinado evento. Além disso, há dificuldades para se identificar instâncias que se encontram em regiões limítrofes (região de borda) para a classificação do comportamento como normal ou anormal.

A interpretação em que a anormalidade da instância é vista como um parâmetro contínuo mostra-se mais adequada do que o ponto de vista binário, justamente por permitir uma avaliação mais abrangente do evento, distinguindo instâncias de forma qualitativa (*outliers* e *inliers*), mas, também, mensurando o valor do desvio observado (verificação quantitativa - grau da anormalidade).

As anomalias podem, ainda, ser interpretadas pelo ponto de vista da *clusterização*. Nesse contexto, os *outliers* são eventos que não se encontram suficientemente próximos das regiões com alta densidade de eventos (*clusters*), obtidas através do conjunto de dados.

A partir dessa perspectiva, os *outliers* obtidos passam a ser considerados como

ruídos provenientes dos algoritmos de clusterização. Isto, por sua vez, materializa a grande desvantagem desse ponto de vista, uma vez que o ruído depende fortemente dos parâmetros utilizados na execução do algoritmo [23].

Na abordagem da *clusterização*, embora haja técnicas focadas na detecção de *outliers*, elas os consideram sobre uma perspectiva global, o que as tornam incapazes de classificar, como *outliers*, instâncias que fazem parte de estruturas mais complexas, comumente encontradas em conjuntos de dados reais.

No trabalho que propôs o LOF como uma estratégia para detecção de anomalias [23], os autores trazem um exemplo que evidencia a falha da interpretação global dos *outliers*, assim como o problema da interpretação binária. Suponha um conjunto de 502 instâncias distribuídas em dois *clusters*, C_1 e C_2, gerados por um algoritmo qualquer. Assuma que 400 objetos pertençam ao C_1 e 100 ao C_2. Além disso, considere a existência de dois pontos adicionais, o_1 e o_2.

Conforme pode ser observado na Figura 1 [23], o aglomerado C_1 é menos denso do que o *cluster* C_2.

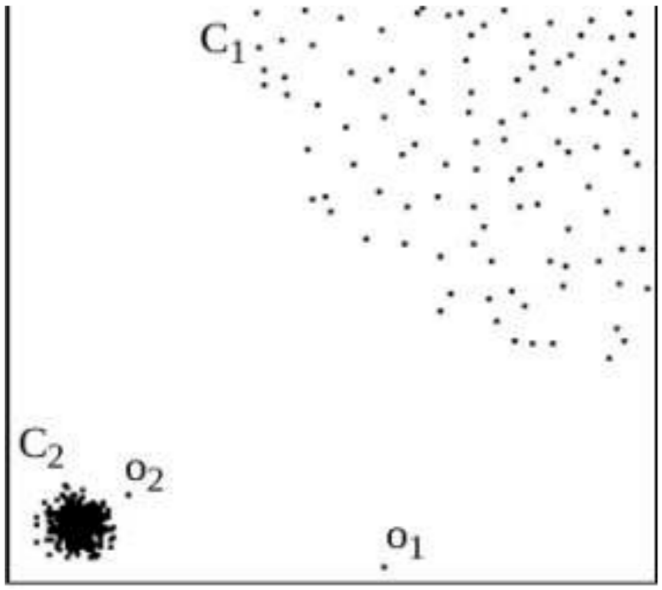

Figura 1: Aglomerados C_1 e C_2 e seus *Outliers* o_1 e o_2.

Pela definição intuitiva dada por Hawkins [23], a qual "*outlier* é uma observação que, por se desviar significativamente dos outros eventos, leva-nos a suspeitar que o seu mecanismo de geração seja distinto do mecanismo que criou os outros eventos", os objetos o_1 e o_2 são considerados *outliers* e os outros, por pertencerem ao C_1 ou ao C_2, não seriam considerados como anomalias.

Contudo, sem a noção de *outlier* baseada em distância, o o_2 poderia não ser considerado como *outlier*, na medida em que se encontra arbitrariamente próximo do aden-

samento C_2. Nessa linha, vale relembrar que os *outliers*, na perspectiva da *clusterização*, são altamente dependentes dos parâmetros utilizados no algoritmo que os retornou, ressaltando a desvantagem desse ponto de vista.

A introdução da perspectiva do grau da anormalidade, dada por um fator que mede a tendência de uma determinada instância ser uma anomalia, mostra-se útil, uma vez que permite avaliar instâncias que se encontram próximas da região de fronteira para a classificação. Assim, para o exemplo trazido, o grau de discrepância para o evento o_1 identificaria uma tendência maior desse objeto pertencer ao conjunto de *outliers* do que o grau obtido para a observação o_2, afastando a falha promovida pela interpretação binária da anomalia, bem como oferecendo uma forma para avaliar a discrepância local do evento.

Nas Figuras 2 e 3, pode ser verificado como o uso do grau de discrepância pode fornecer uma avaliação mais ampla das observações decorrentes dos dados.

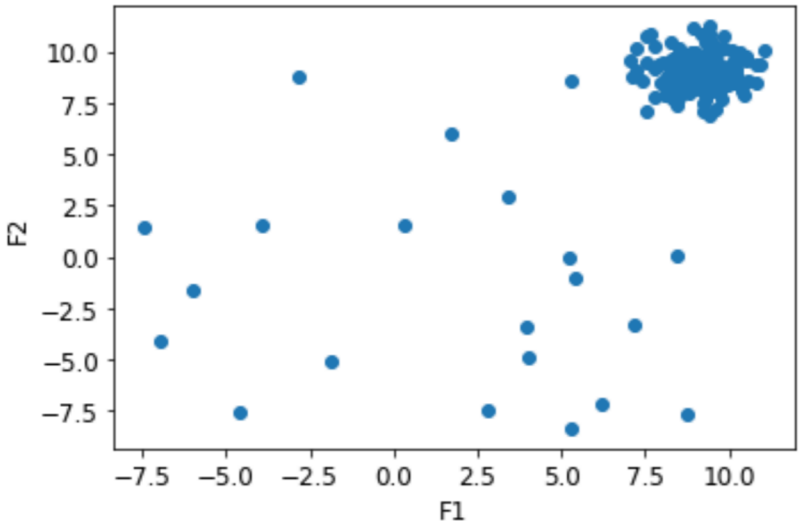

Figura 2: Base de dados gerada aleatoriamente para ilustração.

Após o uso de diferentes técnicas para detecção de anomalias sobre o conjunto da Figura 2, o resultado obtido pode ser verificado na Figura 3:

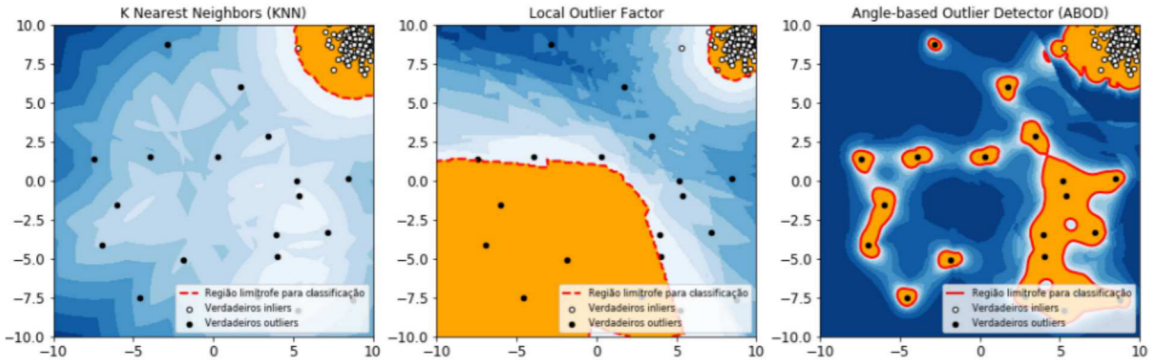

Figura 3: Detecção de anomalias por três técnicas distintas.

A estratégia baseada em *clusterização* (kNN), embora tenha classificado diversas instâncias como anomalias, por partir da premissa de que a anomalia é uma característica binária, não identifica regiões, externas ao *cluster*, onde as instâncias apresentam o maior grau de discrepância. De outro lado, tanto o *Local Outlier Factor* quanto o *Angle-based Outlier Detector (ABOD)* são capazes de identificar, mesmo fora dos aglomerados principais, regiões onde o grau de anomalia dos eventos tornam-se mais severos.

Dessa forma, explorar métodos que avaliem o grau de anormalidade de uma determinada instância são mais adequados do que aqueles que interpretam a anomalia como uma propriedade puramente binária.

Apesar de existirem diversas estratégias capazes de identificar anomalias em conjuntos de dados, neste trabalho, optou-se por utilizar as estratégias de detecção de *outliers* citadas na Seção 2.1, em razão dos resultados alcançados e descritos na literatura especializada (Capítulo 3).

Na Seção 2.3, promove-se uma visão geral sobre o funcionamento das técnicas de detecção escolhidas, apresentando as principais ideias envolvidas nas estratégias por elas exploradas.

2.3 Estratégias utilizadas para a tarefa de Detecção de *Outliers*

Como mencionado na Seção 2.1, para a execução da tarefa de detecção de anomalias, foram utilizados sete estimadores distintos: o *Local Outlier Factor* (LOF), *Isolation Forest* (Iforest), *Clustering Based Local Outlier Factor* - CBLOF, *Histogram-based outlier score* - HBOS, *One-Class Suport Vector Machine* - OCSVM, *k-Nearest Neighbors Detector* - kNN e *Robust Covariance* - RC.

Os estimadores utilizados na detecção de anomalias podem ser classificados em três grandes grupos, sendo que um deles refere-se às estratégias baseadas em uma abordagem de aprendizagem de máquina não supervisionada; o segundo, em uma abordagem

supervisionada; e, por fim, o grupo referente à aprendizagem por reforço [34].

Os algoritmos baseados em aprendizagem de máquina supervisionada requerem a utilização de uma amostra do conjunto de dados para a etapa de treinamento do algoritmo. Em outras palavras, o treinamento pode ser resumido como a etapa em que o sistema objetiva aprender uma regra geral implícita, através do mapeamento das entradas para as saídas, partindo-se da base de dados de treinamento.

Por sua vez, na aprendizagem não supervisionada, nenhum tipo de premissa é fornecida ao algoritmo, deixando-o livre para encontrar estruturas implícitas nas bases de dados de entrada (treinamento ocorre sem informação a priori sobre as saídas desejadas).

É importante destacar que, entre as aprendizagens supervisionada e não supervisionada, existe a semi-supervisionada, que embora requeira uma etapa para treinamento, o conjunto fornecido para essa tarefa contém saídas ausentes (às vezes, muitas saídas faltantes). Nesse caso, o conjunto de dados de entrada é totalmente conhecido no momento da aprendizagem, mas com ausência de valores esperados para diversos objetivos.

Por último, no aprendizado por reforço, o algoritmo interage com um sistema dinâmico, recebendo uma premiação ou uma punição, a depender do alcance de um determinado objetivo.

Neste trabalho, os estimados de detecção de anomalias utilizados relacionam-se às abordagens não supervisionadas (LOF, CBLOF, HBOS, IForest e RC), semi-supervisionada (OCSVM) e supervisionada (kNN).

É importante, também, fazer distinção entre as tarefas de *Novelty Detection* e *Outlier Detection*. Embora as duas estejam intimamente relacionadas com a detecção de *outliers*, há uma diferença conceitual entre as duas modalidades. Na *Outlier Detection*, os estimadores tentam ajustar o modelo ignorando, no conjunto de treinamento, instâncias que se desviam consideravelmente do padrão geral dos dados. Por outro lado, na *Novelty Detection*, os estimadores treinam sobre um conjunto de dados que não contém anomalias, de forma que, após o ajuste do modelo, novas instâncias (que não estavam no conjunto de treinamento) serão submetidas ao estimador a fim de comparar se elas desviam (*outlier*), ou não (*inlier*), do modelo previamente ajustado [35].

O *Local Outlier Factor* (LOF) é um estimador para a tarefa de detecção de anomalias que se baseia na ideia de vizinhança mais próxima (k - instâncias mais próximas ao ponto observado). Nesse tipo de estratégia, a densidade local do objeto observado (sua vizinhança) é avaliada e comparada à vizinhança dos objetos vizinhos. Na execução do algoritmo, é avaliada a quantidade de observações localizadas em uma dada região do espaço, de forma que, quanto maior o número de observações nessa região, maior será sua densidade populacional. [23]

Dessa forma, as observações que habitarem regiões com baixa densidade, isto é, com quantidade de vizinhos significativamente menor do que as vizinhanças adjacentes, serão consideradas como *outliers* pelo LOF. Além disso, como o LOF não interpreta a anomalia sob o ponto de vista puramente binário, as observações mais discrepantes recebem um grau para a anormalidade, sendo arbitrariamente maior para as instâncias localizadas em regiões com baixa densidade.

A Figura 4 traz a ilustração da estratégia do algoritmo:

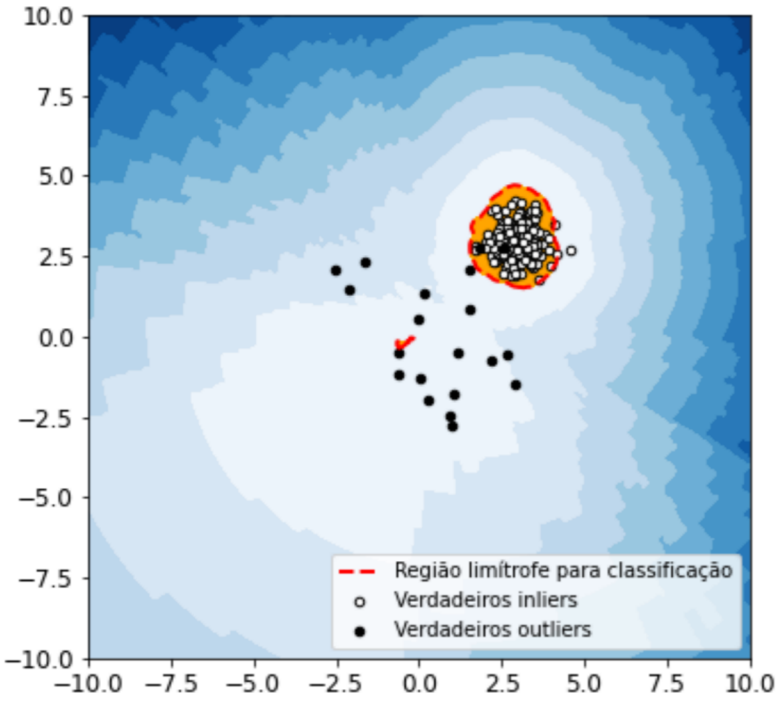

Figura 4: Atuação do LOF sobre uma base de dados aleatória
LOF interpretação: *Região fora do aglomerado (linha tracejada) com baixo adensamento - Anomalias.*

O CBLOF, por sua vez, é um algoritmo que reúne estratégias de vizinhança mais próxima (*k-Nearest Neighbor*), como o LOF, e de clusterização. Ao invés de procurar objetos que habitam regiões de baixa densidade através do cálculo das distâncias envolvidas entre o objeto observado e sua vizinhança, o CBLOF utiliza um algoritmo de *clusterização* para definir os adensamentos existentes no espaço, sendo utilizado o algoritmo de clusterização *Squeezer* no trabalho que o propôs [24]. Após a identificação dos aglomerados existentes no conjunto de dados analisado, o CBLOF procura instâncias que 1) não integram nenhum dos aglomerados ou que 2) estão contidas em adensamentos que são muito menores ou esparsos quando comparados aos aglomerados principais. Assim, a

pontuação da anomalia leva em consideração o tamanho do agrupamento a que pertence o objeto estudado, o tamanho do agrupamento mais próximo e a distância ao centroide.

De acordo com os autores [24], o CBLOF foi motivado, basicamente, pela observação do custo computacional envolvido na execução de algoritmos baseados em vizinhança mais próxima (como o LOF) e pela tendência de algoritmos de aglomeração (*clustering*) tratarem as anomalias apenas como ruído.

Na Figura 5, é apresentada uma execução do CBLOF sobre um conjunto de dados aleatoriamente gerado. Quanto maior o raio da circunferência (gradiente em azul), maior a distância das anomalias frente ao Grupo principal.

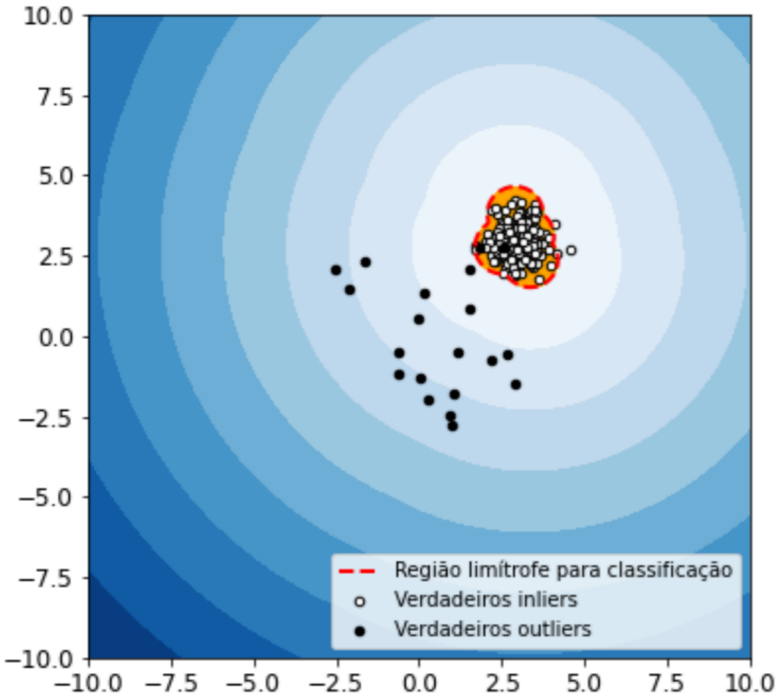

Figura 5: Exemplo de Atuação do CBLOF.
Interpretação CBLOF: *Instâncias externas ao grupo principal (delimitado pela linha tracejada) representam anomalias que não estão contidas em nenhum adensamento ou pertencem a grupos pequeno e/ou esparsos.*

Outro estimador utilizado é o *Isolation Forest* (*IForest*). A estratégia principal deste algoritmo é isolar as anomalias presentes no conjunto de dados, diferenciando-o de outros algoritmos que buscam, frequentemente, estabelecer um perfil para os pontos normais. O funcionamento do *IForest*, conforme indicado em [28], parte de duas premissas acerca das anomalias presentes no conjunto de dados. A primeira delas é a de que, por

serem eventos anômalos, a quantidade deles é muito pequena frente à cardinalidade do conjunto de entrada. Por essa premissa, os autores assumem a raridade dos eventos anômalos. A segunda, por sua vez, está associada às informações trazidas pelos atributos das instâncias, que são significativamente distintas das normais.

Partindo dessas premissas, a detecção das anomalias é realizada através do particionamento do espaço, usando a estratégia conhecida como "árvores isoladas"(*isolation tree - iTree*).

Uma das vantagens na utilização deste método é a baixa complexidade no tempo e a baixa requisição de memória[28].

A Figura 6 apresenta um caso onde o *IForest* foi utilizado sobre um conjunto de dados gerado aleatoriamente:

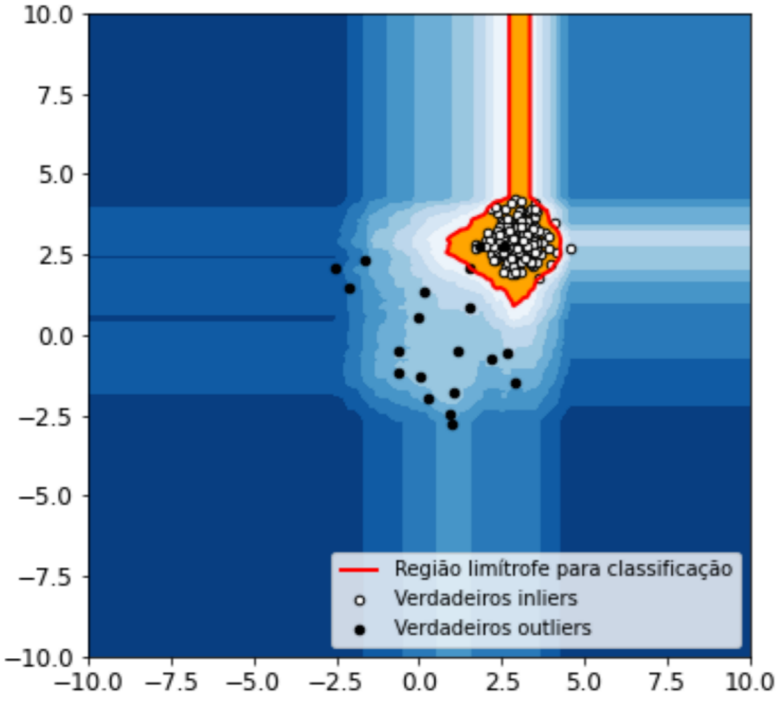

Figura 6: Exemplo de atuação do *IForest*.
Interpretação Isolation Forest: *Linhas (verticais e horizontais - cores mais claras) indicam regiões fora do Aglomerado onde Anomalias foram identificadas pelo particionamento do espaço.*

Ainda na apresentação das estratégias relacionadas à aprendizagem não supervisionada, o HBOS, proposto em [25], é um algoritmo baseado em análise de histogramas. Ele assume a independência dos atributos e é uma combinação de métodos de análise

unidimensional. Nesse sentido, o algoritmo analisa cada atributo do conjunto de dados isoladamente e, através da combinação de diversas técnicas para construção de histogramas (categórica, dinâmica e estática), obtém a altura de cada um dos histogramas, que representa sua densidade estimada. Após a normalização dos histogramas (que garante o peso de cada atributo) e a composição dos pesos dos atributos, é extraído um fator de anormalidade para a instância observada.

A Figura 7, apresenta uma execução do HBOS sobre a base de dados representada na Figura 2:

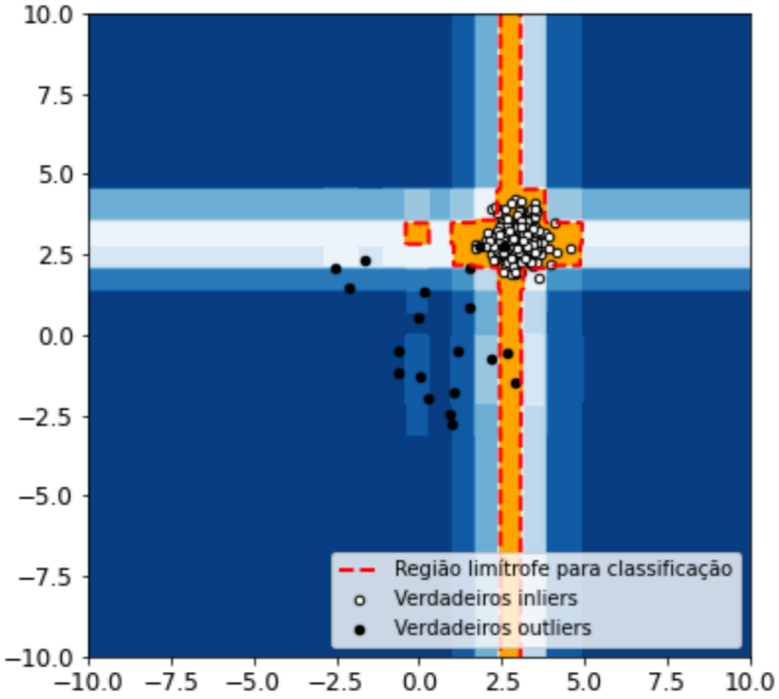

Figura 7: Anomalias identificadas com o uso do HBOS.

Em [25], os autores do HBOS explicam que o algoritmo apresenta bons resultados para identificar *outliers* globais, mas falha na classificação de anomalias locais. Além disso, informam que o tempo de execução do algoritmo é linear, o que pode ser uma vantagem em aplicações cujos dados possuem muitas instâncias ou elevada dimensionalidade.

O *Robust Covariance* (RC) é um estimador altamente robusto de localização multivariada e dispersão, baseado no método do mínimo determinante da matriz covariância (MCD). Classifica-se como uma abordagem não supervisionada e sua estratégia assume que os dados são normalmente distribuídos e os ajusta a uma elipse (região delimitada pela linha tracejada na Figura 8), como região limítrofe para a classificação de uma instância como anômala.

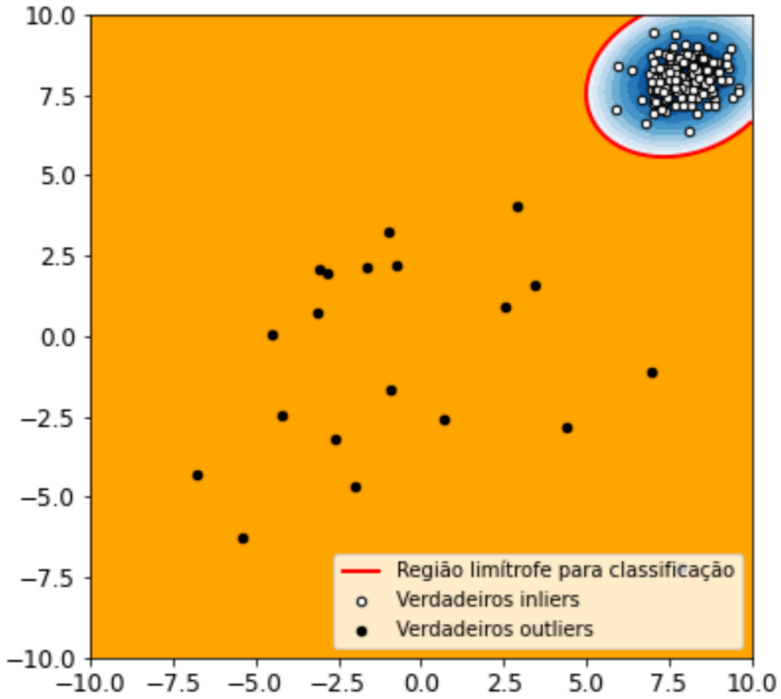

Figura 8: Aplicação do *Robust Covariance*.
Estimador aplicado ao conjunto de dados da Figura 2.

Com o ajuste realizado, duas regiões no espaço são criadas: uma que comporta as instâncias que se ajustaram ao modelo e possuem características similares (*inliers* - região interna à Elipse da Figura 8) e a outra região, onde os eventos têm comportamento que destoa do esperado pelo modelo (*outliers* - região em amarelo na Figura 8).

O OCSVM é um algoritmo de aprendizagem semi-supervisionada. A estratégia de classificação que ele utiliza consiste em, através do treinamento de uma amostra da base de dados, identificar duas classes de instâncias (*inliers* e *outliers*). A separação dessas duas classes é realizada através de um hiperplano, que pode ser entendido como a região de fronteira para a classificação das instâncias. Dessa forma, cada nova instância (conjunto de teste) é comparada ao modelo ajustado, de maneira que seja possível sua inclusão em uma das classes definidas.

Em outras palavras, os eventos que não apresentam o comportamento previsto pelo modelo são consideradas como *outliers*. A Figura 9 apresenta a divisão da base de dados (usada na Figura 2) em duas classes distintas de instâncias.

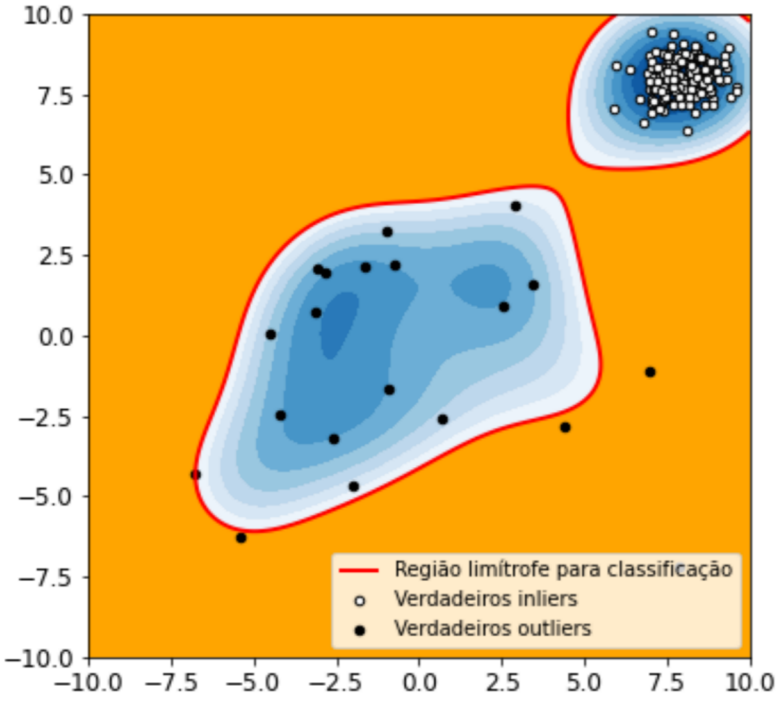

Figura 9: Atuação do OCSVM sobre uma base de dados aleatória
Interpretação OCSVM: *Regiões delimitadas pela linha em vermelho indicam as duas classificações possíveis realizadas pelo estimador.*

Finalmente, o kNN deriva de uma família de algoritmos baseados em distância, sendo um dos métodos mais comuns de aprendizagem supervisionada. A premissa básica da estratégia do kNN é que observações normais estão próximas uma das outras, enquanto que os *outliers* são observações solitárias e afastadas do aglomerado de instâncias normais (*inliers*).

Para determinar a distância entre a instância observada e sua vizinhança mais próxima (delimitada por k elementos - vizinhança k), o algoritmo pode utilizar uma métrica específica, como a Euclidiana, a de Minkowski ou a de Hamming. Esse parâmetro é que será utilizado para determinar se a instância é, ou não, uma anomalia.

Nesse contexto, quanto maior for a distância da instância observada a sua vizinhança mais próxima (com k elementos), mais isolada estará a instância no espaço. Essa é a ideia principal que ampara a classificação de anomalias por esse estimador (ver Figura 10).

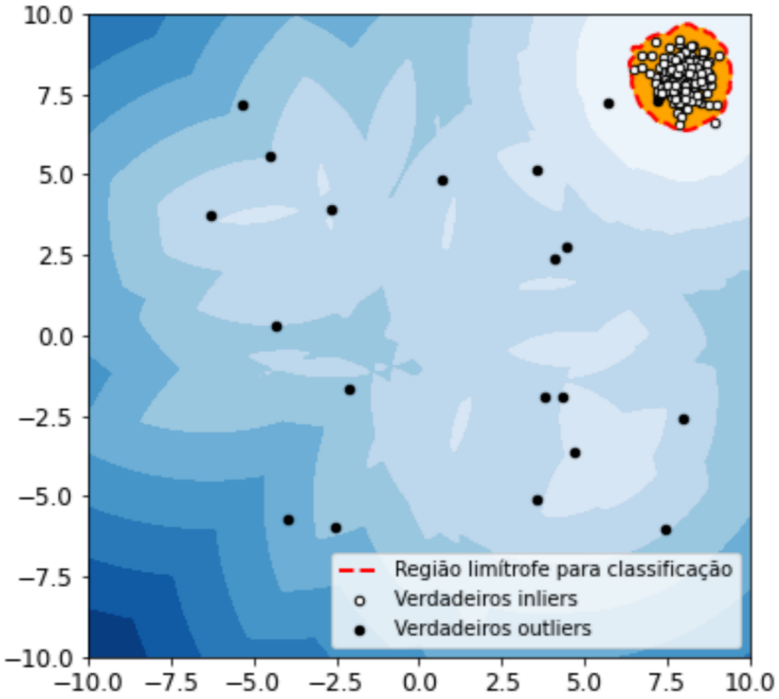

Figura 10: Anomalias detectadas pelo kNN
Interpretação kNN: *Instâncias externas ao Grupo principal (linha tracejada) possuem distância, em relação à vizinhança k, significativamente maior do que as do Grupo principal, sendo, portanto, consideradas anômalas pelo estimador.*

Após a apresentação das ideias básicas que envolvem o funcionamento dos esti-

madores, a Seção 2.4 traz algumas aplicações das técnicas de detecção de anomalias. A descrição detalhada sobre o funcionamento dos estimadores pode ser obtida através das referências [23] a [29].

2.4 Aplicações

Nas seções 2.1 a 2.3, discutiu-se a necessidade de interpretação do enorme volume de dados atualmente disponível. Além disso, apresentou-se, brevemente, que as atividades relacionadas à descoberta de conhecimento nessas bases (KDD), como a mineração de dados, podem promover uma melhora na taxa de interpretação e extração de informações úteis capazes de influir no cotidiano, particularmente nas atividades que requeiram conhecimento para a tomada de decisão.

Do ponto de vista do conhecimento novo e útil, verificou-se que a aplicação de técnicas, como o *Local Outliers Factor* (LOF) e o *Isolation Forest* (IForest), promove a evidenciação de padrões destoantes de difícil visualização pelos especialistas, fomentando a identificação de possíveis eventos candidatos a anomalias.

Preliminarmente, essa identificação otimiza o processo associado à seleção de amostras a serem submetidas à fiscalização, pois evidencia eventos que se apresentam como distorções relevantes, fornecendo uma amostra que deverá ser auditada prioritariamente, além de tornar objetiva a seleção das amostras de auditoria.

Na Auditoria Governamental, que é o conjunto de técnicas que visa avaliar a gestão pública, pelos processos e resultados gerenciais, e a aplicação de recursos públicos por entidades de direito público e privado, mediante o confronto entre uma situação encontrada e a esperada, à luz de um determinado critério técnico, operacional ou legal [36], o objetivo primordial é garantir resultados operacionais na gestão da coisa pública. Isso é conseguido através da perseguição do objetivo primordial das atividades de auditoria que, conforme as Normas Brasileiras de Contabilidade [37], é aumentar o grau de confiança nas demonstrações contábeis por parte dos usuários, através da emissão de uma opinião, pelo auditor, sobre se as demonstrações contábeis foram elaboradas, em todos os aspectos relevantes, em conformidade com uma estrutura de relatório financeiro aplicável.

No contexto específico da Administração Pública, tendo-se em vista os princípios que a norteiam, o escopo do trabalho de auditoria não se resume à análise de demonstrações contábeis, mas alcança, também, a avaliação dos resultados decorrentes da implantação de políticas públicas e atos de governo, particularmente daqueles que impliquem a assunção de obrigações com credores e, consequentemente, o uso de recursos públicos. Mais do que isso, os "usuários" não se limitam àqueles que necessitam de informações relevantes e confiáveis para avaliar a saúde financeira de determinada entidade, mas estende-se a todo público que, de alguma forma, se relaciona com a Administração. Assim, toda a

sociedade pode ser considerada como *stakeholder*, isto é, público-alvo e interessado nos resultados e na asseguração fornecida pelas atividades de auditoria.

É nessa esteira que os dados coletados pela Administração, principalmente em meio digital, podem auxiliar no aumento da capacidade de controle pelos órgãos competentes e pela própria sociedade, uma vez que através das técnicas computacionais, particularmente de detecção de anomalias, eventos destoantes podem ser facilmente identificados, inseridos e analisados na amostra de auditoria.

A literatura vem demonstrando que diversos países vêm adotando o conceito de dados governamentais abertos (*Open Governments Data – OGD*) com o objetivo de fomentar a transparência e o engajamento social através da disponibilização de informações precisas, tempestivas e úteis aos interessados. Além disso, aponta que a transparência pode integrar o processo de gestão dos recursos públicos ao planejamento, programação e operações de controle interno, auditoria e avaliação das despesas. Por fim, define o OGD como as atividades relacionadas à disponibilização de dados e informações em formatos e meios que permitam o livre acesso, o uso, a distribuição e a exploração de dados" [38].

Portanto, como há uma tendência mundial no sentido da disponibilização de dados relevantes sobre a gestão pública, observa-se que o uso de técnicas computacionais capazes de interpretar e extrair padrões ou anomalias possui elevado potencial de contribuição nos cenários de avaliação dos gastos, geração de valor público e apoio nas tomadas de decisões dos gestores.

No Brasil, o Governo Eletrônico (e-Gov) é entendido como o uso de tecnologias de informação e comunicação (TIC) para a promoção de um melhor governo [39]. Ele é um indicativo dos esforços perpetrados na busca da otimização da aplicação dos recursos de origem pública.

Embora haja um amplo espaço para a aplicação de técnicas de detecção de anomalias e extração de conhecimento no setor público, não é somente neste em que elas podem ser utilizadas. Na literatura há diversas aplicações relacionadas à iniciativa privada, mostrando que muitos setores apropriaram-se dessas técnicas com o intuito de aumentar a capacidade de tomada de decisão. Como exemplo, pode-se citar o trabalho de K. Chitra and B. Subashini [40], o qual traz aplicações de técnicas e algoritmos de mineração de dados na área bancária, nas tarefas de detecção de fraudes, operações financeiras não usuais e gerenciamento de risco.

Em outro trabalho [20], são apresentadas aplicações das técnicas com o objetivo de analisar o risco associado à concessão de empréstimos através de mecanismos de pontuação (*scoring*). O estudo ainda descreve, em detalhes, os algoritmos baseados em árvores de decisão, classificação de Bayes, *Suport Vector Machine* (SVM) e *Booting Random Forest*, sendo que todos servem como exemplos de técnicas de mineração de dados.

Também foram encontradas aplicações das técnicas no setor da agropecuária [41], visando a obtenção de padrões presentes nos dados capazes de auxiliar no processo de tomada de decisão, particularmente no que se refere à predição de medições importantes para o setor, como preços de produção e de alimentação.

Em suma, a literatura é farta no que se refere às aplicações das técnicas de mineração de dados e, em especial, das que remetem à detecção de anomalias.

No Capítulo 3, como resultado da execução do protocolo de revisão da literatura, serão trazidas mais aplicações referente às diversas técnicas de detecção de anomalias, promovendo-se uma complementação da visão geral sobre a área e de sua importância para a mineração de dados.

3 Trabalhos Relacionados

Neste capítulo, os principais trabalhos relacionados à pesquisa serão apresentados. Para tal, foi executado o protocolo de revisão da literatura proposto neste trabalho, o qual selecionou 123 artigos mais relevantes da área. O protocolo de revisão sistemática é apresentado no Anexo A deste documento.

Como mencionado anteriormente, a detecção de anomalias é uma tarefa crítica e desafiadora para a mineração de dados, sendo considerada, também, um dos problemas principais associados às aplicações de descoberta de conhecimento – KDD, no aperfeiçoamento da robustez do processo de aprendizagem de máquina. Seu objetivo principal é a busca de objetos que apresentem padrões significativamente aberrantes quando comparados aos demais.

Os chamados *outliers*, ou anomalias, podem ser gerados por diversos mecanismos, desde erros manuais, erros de sistemas, falhas mecânicas decorrentes do processo de captura de dados de fontes distintas e, até mesmo, podem ser provenientes de padrões anormais gerados por atividades ilícitas, como fraudes.

Detecção de anomalias é um tema recorrente na literatura, sendo aplicada às mais diversas áreas como detecção de intrusão em redes [42] [43], identificação de fraudes em cartões de crédito, testes clínicos e muitas outras. Por ser um problema relevante para a mineração de dados, a área tem recebido muita atenção dos pesquisadores que envidam grandes esforços para obter melhorias e desenvolver novas abordagens para a identificação de padrões anômalos.

A detecção de anomalias pode ser dividida em diferentes grupos. Em Salehi, Mahsa et. al (ver Figura 11) [44], os autores trazem uma classificação sobre as abordagens não supervisionadas existentes, dividindo-as em técnicas envolvendo grafos e dados contínuos (*data streams*). Para este último, subclassifica as abordagens em baseadas em estatística, agrupamento (clusterização) e baseadas em vizinhos mais próximos (*nearest neigbors based*). Naquelas que envolvem clusterização, há estratégias que consideram a distância do objeto ao centroide ou as que se baseiam no tamanho ou densidade do *cluster*. Para as baseadas em vizinhança mais próxima, as estratégias dividem-se, no geral, em baseadas em distância ou densidade das regiões adjacentes.

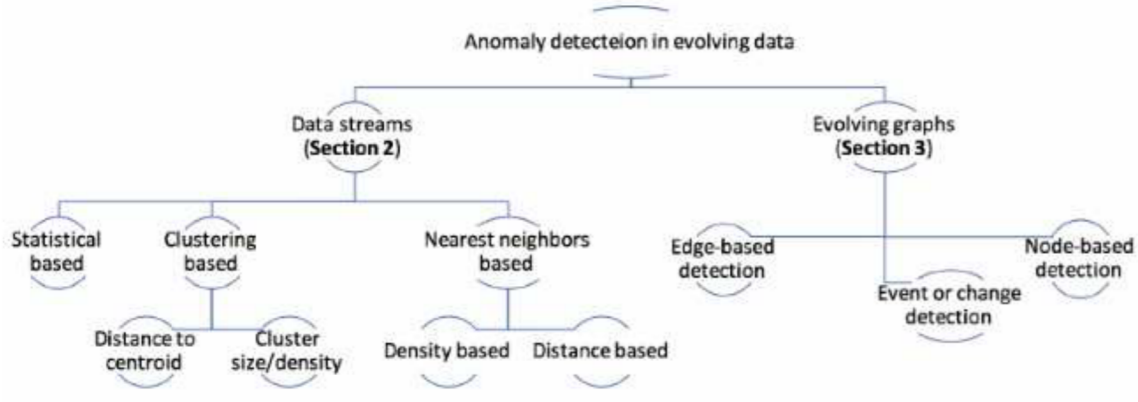

Figura 11: Classificação das abordagens de detecção de anomalias envolvendo dados.

Entre as abordagens não supervisionadas baseadas em densidade, a mais famosa é a *Local Outlier Factor (LOF)*, onde o grau da anormalidade de um objeto é determinado considerando-se a estrutura de clusterização delimitada por uma região, em função da vizinhança k considerada.

Essa técnica foi proposta por Breunig et al. [23] sendo que no trabalho de avaliação da efetividade das técnicas de detecção de anomalias [32], publicado em 2016, os pesquisadores concluíram que o LOF, junto aos métodos KNN e KNNW, permaneceu como o estado da arte, mesmo após o advento de abordagens mais recentes, como LDF e KDEOS.

O LOF é uma abordagem que possui várias aplicações na área de mineração de dados. Em Bao et al. [45], o LOF foi utilizado para detectar anomalias associadas a dados de trajetória, isto é, dados provenientes do movimento de determinado objeto. O resultado do trabalho foi a proposição de um novo algoritmo, chamado *TRAOD*, que combinou estratégias não supervisionadas baseadas em distância e em densidade, mostrando ser capaz de identificar desvios em relação às trajetórias reais estudadas.

Na área econômica, o LOF foi aplicado para medir a saturação de mercados em Beijing [46], uma vez que a decisão sobre onde estabelecer um negócio é fundamental às atividades comerciais. Dessa forma, a aplicação do LOF reduziu a dependência entre a decisão a ser tomada e a experiência humana, bem como outros fatores majoritariamente subjetivos.

Além disso, diante do cenário atual de coleta automática e massiva de dados - AFC (*Automated Fare Collection*), há estudos focados na análise de padrões da mobilidade coletiva. Nesse contexto, em Du et al.[30], os pesquisadores propuseram um sistema de vigilância e detecção de atividades anômalas para identificar suspeitos de serem "ba-

tedores de carteiras", através de seus registros de deslocamentos diários. Os resultados experimentais demonstraram que o método é efetivo, principalmente quando se associam técnicas de detecção de anomalias supervisionadas e não supervisionadas. Quando o LOF foi utilizado junto ao OCSVM (*One- Class Suport Vector Machine* – técnica semi-supervisionada), os melhores resultados foram obtidos.

Em Shan et al.[47], os pesquisadores deram uma nova aplicação ao LOF: detecção de faturamentos duvidosos em gastos médicos. Esse estudo foi conduzido no contexto de avaliação da conformidade do gerenciamento de saúde pública. Como resultado, o estudo sugere que métodos de detecção de anomalias baseados em densidade são efetivos para identificar padrões de faturamento inapropriados, constituindo-se como uma ferramenta para a avaliação no monitoramento do faturamento médico na promoção dos serviços de saúde.

Em [31], os pesquisadores aplicaram uma versão otimizada do LOF (*Cluster-Based LOF*) na área de engenharia de petróleo e geofísica, obtendo, de forma confiável, medidas em tempo real dos fluxos de gás, óleo e água, sem a necessidade de separação de fases. Além do LOF, eles exploraram a estratégia levada a cabo pelo *Isolation Forest*, cujo grau da anomalia é definido em função do particionamento do espaço.

Em [48], quando comparado às técnicas RRS (Ramaswamy, Rastogi and Shim) e FastVOA, o LOF apresentou o melhor desempenho quando aplicado a atividades relacionadas à lavagem de dinheiro. Esta, por sua vez, é a ação que visa dar natureza lícita à renda que foi obtida através de métodos ilícitos. No referido trabalho, o LOF identificou *outliers* relacionados às atividades de empresas "de fachada", comumente utilizadas para a "lavagem de dinheiro".

Abordagens baseadas em densidade estão entre os mais populares métodos de detecção de *outliers*. Entretanto, esses métodos sofrem com baixa performance quando os problemas apresentam padrões associados à baixa densidade.

Assim, muito embora o LOF seja uma técnica consagrada e com muitas aplicações na área de detecção de anomalias, ela possui desvantagens que devem ser consideradas na avaliação do problema a ser examinado. Entre essas desvantagens, a literatura aponta as relacionadas ao elevado uso de memória e a longa sequência de *outliers* obtidos.

Tendo em vista essas limitações, há trabalhos que propõem variações das técnicas clássicas com o objetivo de resolver ou, ao menos, mitigar as desvantagens existentes. Entre essas variações, há as baseadas em densidade relativa (RDOS), as quais introduzem uma nova maneira para se medir a densidade da vizinhança de um objeto [49].

O algoritmo DILOF, que também é uma variação do LOF, foi proposto para tratar bases de dados não estacionárias com eficiência no uso de memória e mantendo-se a efetividade das atividades relacionadas à detecção de *outliers* [50]. Também em [51]

propõe-se uma variação do LOF que indicou melhorar sua eficiência.

De forma geral, diversas técnicas derivaram do LOF, como o ILOF, CB-LOF, SLOF, COF. Essas variações surgiram, basicamente, para superar desafios impostos pelos grandes conjuntos de dados aos algoritmos convencionais, principalmente no que se refere à eficiência no tempo e no espaço.

Por exemplo, o *Cube-Based Local Outlier Factor (CB-LOF)* [52] foi proposto para aliviar os problemas de eficiência apresentados pelo *Incremental Local Outlier Factor (ILOF)*. Este, por sua vez, que é usado em problemas que tratam de dados gerados continuamente (*data streamming*) por fontes diversas não estacionárias, atribui um grau de anormalidade a cada objeto representado. No entanto, como ele atualiza os parâmetros de cada ponto dinamicamente, sua eficiência é facilmente comprometida quando o fluxo de dados é massivo e contínuo, devido à geração de regiões com altas densidades locais, as quais conduzem a um elevado consumo de tempo e espaço nas tarefas de atualização a gravação das novas informações dos pontos.

Quanto à acurácia, em Ahmad et al. [53] foram comparados diversos métodos supervisionados e não supervisionados de detecção de anomalias, sendo que as melhores acurácias foram obtidas pela rede neural *Multi-Layer Perceptron* (MLP), pelo modelo *"Decision Tree"* e pelo *Linear Suport Vector Machine*, os quais atingiram, respectivamente, 100%, 98% e 97% de acurácia. No entanto, esse estudo não abordou a LOF.

Há ainda algoritmos de aprendizagem não supervisionada que se baseiam na análise unidimensional da base de dados, buscando reduzir o custo computacional no tempo para a execução da tarefa. O HBOS [25] assume a independência das variáveis e constrói, para cada uma delas, histogramas para avaliar a frequência relativa de ocorrência de determinando evento no atributo, promovendo uma medida indireta sobre a densidade da vizinhança. Ao final, em bases multivariadas, os histogramas obtidos, que são estimadores de densidade frequentemente utilizados em abordagens semi-supervisionadas, são combinados de forma a retornarem uma pontuação (*scoring*) para cada instância - que define se ela é ou não *outlier*. Em [25], é demonstrado que esse algoritmo tem tempo de execução linear, sendo indicado como um método eficiente para detecção de *outliers* globais. Apesar de possuir baixa precisão para identificar *outliers* locais, o HBOS teve excelentes resultados quando comparado ao LOF e ao KNN, principalmente quanto ao tempo de execução.

Além do uso de histogramas como estimadores de densidades, existem estratégias que usam o método de Rousseewn como um estimador altamente robusto para a detecção de anomalias em dados multivariados [29]. Essa abordagem, que possui aplicações no campo da Astronomia (grandes bases de dados), utiliza o método da mínima covariância do determinante (MCD - *Mininum Covariance Determinant*) para se derivar observações

candidatas a anomalias.

De outro lado, existem trabalhos que propuseram novos *frameworks* para a resolução de algum tipo de problema. Em Jia Guo et al. [54], os autores propuseram um *framework* de detecção não supervisionada, chamado AEKNN, que objetiva incorporar as vantagens da representação da aprendizagem automática promovida pelas redes neurais profundas com a eficiência dos algoritmos de detecção de anomalia. Os resultados experimentais obtidos demonstraram que o AEKNN atingiu maior acurácia que o LOF e o *Isolation Forest*.

Em [55], é proposto um algoritmo de detecção de anomalias, não supervisionado e não paramétrico (SECODA) para *datasets* com atributos contínuos e categóricos. O método garante a identificação de diversos tipos de anomalias, como *outliers* extremos, anomalias de classe esparsa, entre outras. O estudo mostrou que o algoritmo tem grau de confiança superior a 90%, podendo ser utilizado em diversas aplicações reais.

O *Angle-Based Outlier Detection* (ABOD), proposto em Kriegel et. al. 2008 [56] é outra abordagem que, similarmente ao LOF e a suas variações, oferece uma medida do grau de anormalidade de determinado objeto, não reduzindo a anormalidade a uma característica puramente binária, como é realizado por algumas estratégias de clusterização.

Em [57], o ABOD é utilizado na área de inovação tecnológica. A ideia principal é aplicar a técnica de detecção de *outlier* sobre dados de patentes industriais, com o objetivo de identificar ideias inovadoras (nesse caso, *outliers*) que pudessem indicar uma nova oportunidade de negócio às empresas.

Em [58] é proposto um algoritmo baseado em ângulo (VAOF) que é comparado ao ABOD e sua variação mais eficiente, o FastABOD. Os resultados experimentais evidenciaram que, para bases de alta dimensionalidade, o novo algoritmo possui resultados competitivos com o ABOD, embora este, de forma geral, tenha se mantido como o algoritmo mais preciso. O ABOD também vem sendo aplicado na borda de redes, com o objetivo de capturar dados de treinamento do modelo de limpeza de dados, como pode ser visto em [59].

Anomalias, apesar de corresponderem a uma pequena fração da base de dados, podem apresentar alto potencial de conduzir a significativas perdas econômicas, sociais e de recursos públicos. Nessa linha, é muito comum encontrar na literatura aplicações de detecção de anomalias direcionadas a minimizar a probabilidade de perdas.

Com o intuito de identificar comportamentos anômalos ou fraudulentos, na área da administração, há trabalhos que aplicam a detecção de anomalias para gerenciar processos de negócios. Em Tavares et al. [60], o algoritmo proposto foi capaz de identificar mais de 98% dessas anomalias de forma *online*.

Na área médica, também, a detecção de *outliers* se mostra relevante. Em [61] concluiu-se que técnicas de detecção de anomalias (BIRCH) podem identificar *outliers* mesmo quando estes não são descobertos pelos métodos tradicionais de análise de dados, tais como as análises de regressões. No trabalho, o uso da detecção de anomalia foi capaz de identificar pessoas com suspeita de depressão, quando as análises de regressão não foram capazes de identificar nenhum *outlier*, mesmo quando a base de dados continha 50 pessoas diagnosticadas com depressão. De outro lado, ao se analisar um conjunto com 576 admissões iatrogênicas, o número de comedicações obtido não foi significativo usando um preditor log-linear. Entretanto, usando-se técnicas de detecção de *outliers*, obteve-se um *cluster* de anomalias que continha 174 pacientes, representando 30% da amostra estudada.

Outra aplicação interessante de detecção de *outliers* está relacionada à verificação de fraudes associadas a planos de saúde. O modelo desenvolvido em Soleymani et al. [62] foi capaz de identificar prescrições médicas com suspeita de fraudes.

Além desse, em [63], os autores propuseram utilizar a *Local Correlation Integral – LOCI*, que é uma técnica de detecção de anomalias não supervisionada, para detectar fraudes no sistema de seguros de saúde australiano (MEDICARE). Os resultados obtidos, apesar de não claramente apresentados, indicaram a efetividade do modelo.

Há, ainda, aplicações que, usando aprendizagem de máquina não supervisionada, procuram identificar possíveis fraudes fiscais decorrentes da subnotificação de valores nas declarações tributárias [64].

Diante do exposto, é possível afirmar que existem inúmeros exemplos de aplicações de detecção de *outliers* na literatura. Mais do que isso e ao contrário do que se poderia imaginar, essas aplicações não se restringem à área da computação, pois impactam fortemente outras como as associadas à administração, economia, engenharia, geofísica, inovação e gestão pública.

Apesar disso, embora se tenha observado o amplo espectro das aplicações, na revisão da literatura realizada neste trabalho, não foram encontradas aplicações que explorassem o potencial dessas técnicas no âmbito dos dados gerados pelos órgãos públicos.

Vale dizer que as técnicas de detecção de anomalias, pelos resultados gerais dos diversos trabalhos brevemente apresentados, podem conduzir a melhorias na gestão da coisa pública, bem como dos processos relacionados à atividade estatal, uma vez que a otimização da eficiência contribui para o aumento da efetividade dos recursos e das políticas públicas a serem desenvolvidas pelos gestores.

Dessa forma, existe a possibilidade de que haja uma "lacuna" na área de detecção de *outliers*, justamente por não se ter encontrado na literatura, aplicações e implementações engajadas no tratamento de dados oriundos de órgãos públicos, mesmo havendo

uma tendência mundial de maior prestação de contas por parte dos governos (e-Gov) e de maior controle social.

Tendo isso em vista, o presente trabalho propõe uma solução baseada em técnicas de detecção de anomalias para o tratamento de dados públicos. No Capítulo 4, será apresentada a solução proposta, bem como a metodologia utilizada.

4 Metodologia e Solução Proposta

Este Capítulo destina-se à apresentação da arquitetura da solução proposta. O objetivo principal é fornecer uma visão esquemática da solução e delinear o funcionamento dos seus módulos, que são os módulos de pré-processamento e detecção de anomalias.

Como mencionado em Capítulos 1 a 3, foram apontadas aplicações das técnicas de detecção de *outliers* a diversos setores como o agronegócio, bancário, análise de risco, fraude financeira e *marketing*. Entretanto, na literatura, não foram encontrados usos mais invasivos dessas técnicas no tocante aos dados armazenados por órgãos ou entidades governamentais. Tendo isso em vista, a solução busca o fornecimento de uma ferramenta capaz de processar e extrair, de dados relacionados a gastos públicos, uma amostra que contenha instâncias com alta propensão de constituírem-se como *outliers*.

No escopo desta pesquisa, a solução proposta foi desenvolvida para ser aplicada a bases de dados públicas (bases FMSJP [5], GerFrotas [13] e DIGF [14]) que registram aquisições de combustíveis para a frota Estadual; insumos médicos necessários às atividades desenvolvidas pela Fundo Municipal de Saúde de João Pessoa; e despesas decorrentes de dispensas e inegixibilidades de licitação registradas no Portal da Transparência do Governo Federal. O objetivo é que a seleção de amostras com alto potencial de anormalidade contribua com as tarefas em que se necessite selecionar eventos/transações governamentais que representam maior risco ao Erário, possibilitando uma forma ágil e científica para a definição dos eventos que deverão, prioritariamente, sofrer diligências por parte das entidades de fiscalização. Dessa forma, com a identificação de instâncias que contrastam significativamente com o padrão geral observado, é possível aos órgãos responsáveis focar seus esforços na avaliação desses eventos.

A Figura 12 apresenta uma visão esquemática da solução proposta, evidenciando a atuação dos módulos pré-processador e detector de anomalias. A base de dados inicial é recebida e tratada pelo módulo de pré-processamento, sendo posteriormente submetida ao detector de anomalias.

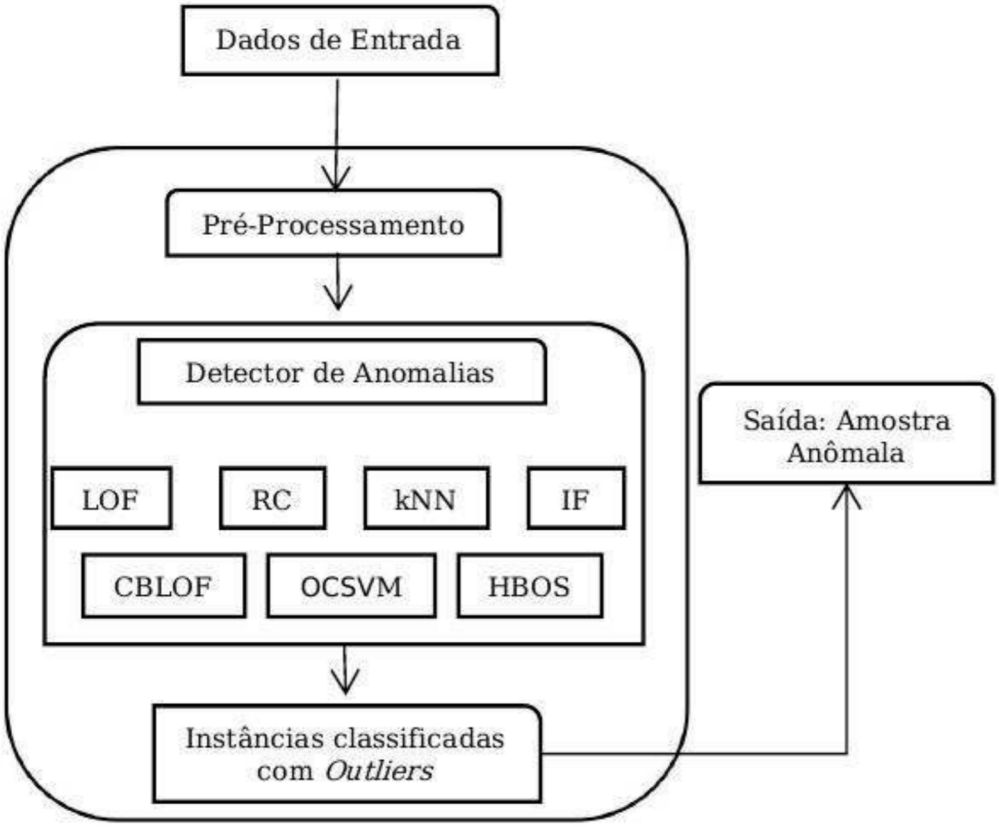

Figura 12: Visão esquemática da Solução Proposta.

4.1 Pré-Processador

A solução proposta foi desenvolvida visando sua aplicação sobre as bases de dados públicas objeto deste trabalho (GerFrotas, DIGF e FMSJP). Seus módulos foram criados em linguagem de programação Python, que lê a base de dados a ser processada através de um arquivo de entrada *".csv"*.

Em linhas gerais, o Pré-Processador é um módulo criado com o objetivo de realizar, de forma automatizada, para as bases utilizadas (GerFrota, DIGF e FMSJP), as tarefas relacionadas ao pré-processamento dos dados, essencial à mineração de dados. Em suas diversas fases, conforme será visto a seguir, o módulo executará verificações sobre repetições de valores nos atributos, existência de valores faltantes (*missing values*), correlação entre atributos numéricos e entre estes e categóricos, além de diversas outras tarefas. A Figura 13 fornece uma visão esquemática sobre o módulo de pré-processamento:

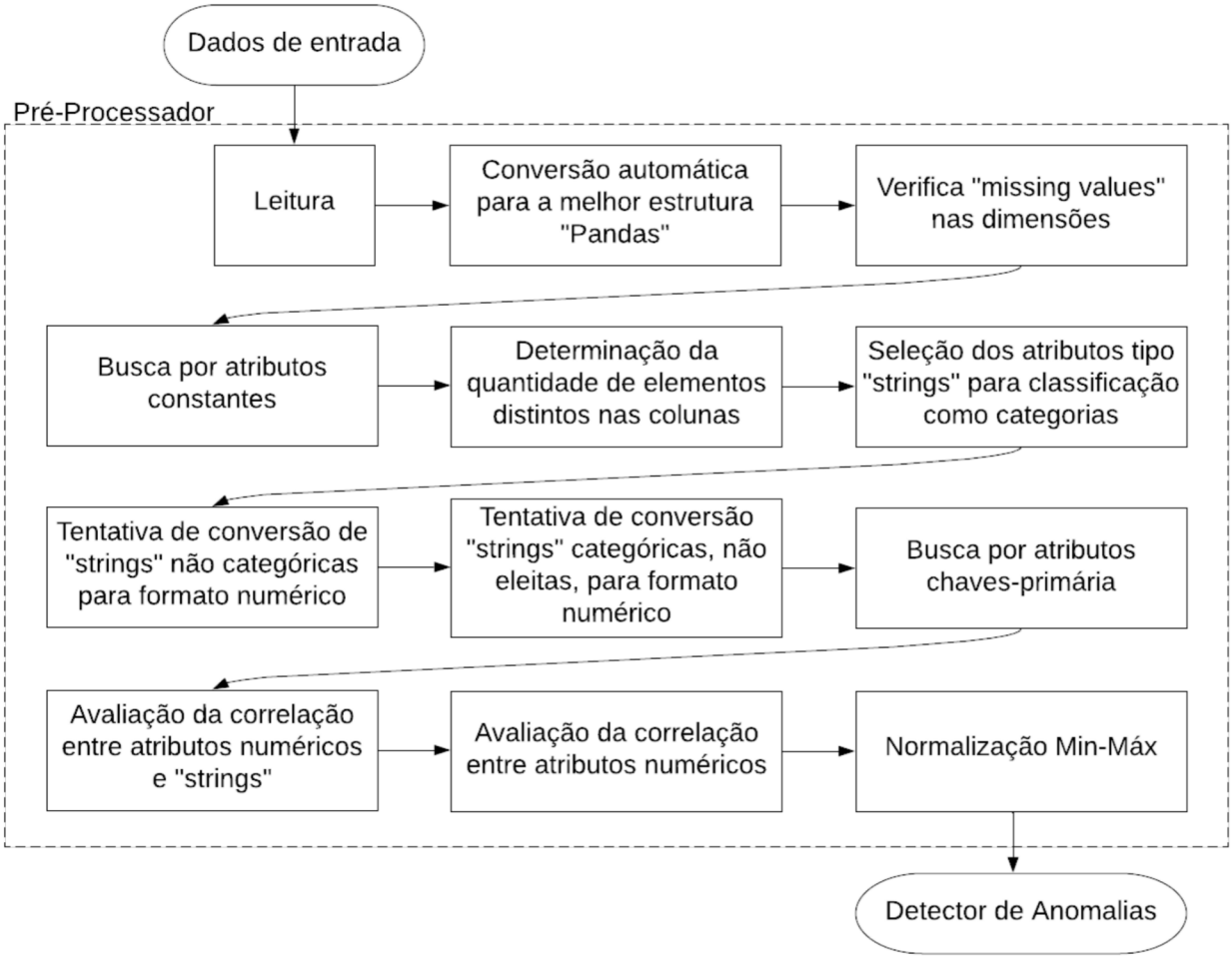

Figura 13: Visão geral sobre a atuação do Pré-Processador.

Passada a fase inicial da leitura da base de dados, o pré-processador busca conformar o conjunto a melhor estrutura de dados possível. Essa conformação é essencial, pois, como não foi estabelecido uma estrutura fixa para os conjuntos de dados a serem submetidos à solução, o pré-processador precisa reconhecer o tipo de dados associados a cada um dos atributos.

Nesse sentido, o pré-processador reúne funções que lhe permite identificar, através da leitura do arquivo de texto, os tipos de dados dos atributos que integram a base de dados (inteiro, real, *string*, *datetime* e objeto).

Após a definição automática da melhor estrutura de dados para a base, no passo seguinte do pré-processamento verifica-se se há instâncias com atributos que não estejam totalmente preenchidos. Nessa tarefa, caso sejam detectadas instâncias que não estejam totalmente preenchidas (dimensões com valores faltantes), o algoritmo as exclui da base de dados.

Após a remoção das instâncias com dados incompletos (valores não preenchidos), o pré-processador procura por atributos que sejam constantes no conjunto de dados, isto é, busca atributos que possuem a mesma informação para todas as instâncias, caracterizando-o como um valor fixo e, portanto, sem valor à análise estatística.

A analogia vetorial para o atributo constante é que ele representa uma componente espacial de valor fixo "a" em um espaço d-dimensional. Dessa forma, como todos os vetores do espaço, dado pela base de dados, possuem um mesmo valor na componente analisada, a projeção deles sobre a componente se concentrará sobre o valor "a", não havendo, portanto, desvios capazes de serem considerados como anomalias.

Superadas as etapas de leitura, formatação da estrutura de dados e frequência de ocorrência dos valores presentes em cada atributo, o algoritmo seleciona apenas os atributos que foram reclassificados para o formato *string*.

Essa seleção, por sua vez, procura analisar quais desses atributos podem ser encarados como categóricos, isto é, quais podem ser utilizados como agregadores ou subconjuntos do mesmo atributo. Suponha que, no conjunto de dados estudado, haja um atributo que se refira a marcas de veículos fabricados no Brasil. Nesse contexto, embora o conjunto de dados possa conter instâncias representando eventos diversos, o número de elementos distintos presentes no mencionado atributo é fixo. Ilustrando, considere que no Brasil somente existam cinco fabricantes de veículos: VolksWagen, Fiat, Volvo, Honda e Hyundai. Por mais que o conjunto possua mais de 10^6 instâncias, o atributo "Fabricantes" terá, no máximo, cinco elementos distintos em sua composição. Assim, o atributo em comento é considerado uma categoria, ou dimensão categórica, uma vez que o número de elementos distintos que o compõe é muito menor do que número de instâncias presentes na população estudada.

O critério estabelecido para a definição de um atributo *string* como categórico considerou a quantidade de elementos distintos presentes. Caso a quantidade de elementos distintos seja menor ou igual a X% do número total de instâncias, onde X é um parâmetro configurado no pré-processador, então a coluna é atributo *string* categórica. Atendendo à restrição imposta, o atributo é reclassificado como tipo "categórico", deixando de ser interpretado como tipo *string*.

Essa etapa é importante porque evidencia para o usuário a existência de atributos que podem ser utilizados para a obtenção de determinadas características, inclusive estatísticas, dos subconjuntos presentes nos dados, além de permitir a reorganização da base de dados.

No exemplo fornecido anteriormente sobre os fabricantes de veículos no Brasil, a classificação de atributos como categóricos pode fazer com que os usuários reconfigurem o conjunto de dados a fim de obterem um novo conjunto mais adequado aos seus propósitos.

Uma situação que ocorre, a depender dos critérios de classificação, é que poderá haver atributos *strings* não considerados como categorias. Nesses casos, o pré-processador tentará convertê-los ao formato numérico, com o objetivo de preservar a maior quantidade de informações numéricas no conjunto de dados. Isto tem relevância porque a estratégia de detecção de anomalias a ser aplicada sobre os dados requer que as colunas estejam representadas numericamente (*int* ou *float*). Assim, tentar converter as *strings* não categóricas em inteiros ou reais pode preservar informações relevantes ao processo de extração de observações discrepantes.

Na etapa precedente, foram realizadas tarefas que buscavam a conversão dos atributos *strings* não categóricos em números, na tentativa de maximizar a quantidade de informações disponíveis na fase de detecção de anomalias. Entretanto, as *strings* categóricas também podem ser passíveis de conversão numérica. Assim, o algoritmo executa a mesma tarefa para as dimensões categóricas restantes – aquelas que não foram eleitas pelo usuário no processo de reindexação. Portanto, ao final dessas tarefas, a estrutura da base de dados possui a maior quantidade possível de atributos numéricos, maximizando a dimensionalidade do conjunto submetido ao módulo de detecção de *outliers* – constituído pelos estimadores apresentados no Capítulo 2.

Após essas alterações sobre a estrutura dos dados, o módulo de pré-processamento procura identificar, dentre os atributos existentes, aqueles que se enquadram no conceito de chave primária. Chaves primárias referem-se a informações que nunca se repetem em uma dada tabela e, por isso, podem ser exploradas como índices de referência na indexação do conjunto estudado. Em outras palavras, o que essa tarefa realiza é a procura por atributos que, por possuírem alto grau de variabilidade frente a quantidade de instâncias, podem ser utilizados como índice no conjunto de dados pré-processado.

Com esse objetivo, definiu-se que, se a quantidade de elementos distintos em uma determinada coluna for maior ou igual a Y% da quantidade de linhas da tabela (instâncias), onde Y é outro parâmetro utilizado no pré-processador, ele poderá ser classificado como uma possível chave primária. Caso não haja atributos que atendam à regra exposta, então, o conjunto de dados manterá a indexação original.

O módulo de pré-processamento ocupou-se em preservar a maior quantidade de informações no conjunto de dados, mantendo atributos que foram completamente preenchidos, os categóricos, não categóricos e os numéricos. Entretanto, embora seja desejável a manutenção da maior dimensionalidade possível do conjunto de dados, não se deve manter informações que sejam redundantes, pois isto aumenta o custo computacional e afeta a eficiência do algoritmo, sem trazer melhora em sua eficácia. Tendo isso em vista, dotou-se o pré-processador de uma estratégia que visa a busca por atributos que estejam fortemente correlacionados, de maneira a se permitir a exclusão de um desses atributos do conjunto de dados.

Nesse sentido, colunas altamente correlacionadas representam informações sobre um mesmo fenômeno presente nos dados. Assim, torna-se desnecessário aplicar a técnica de detecção de anomalias a dimensões que possuam essa natureza, uma vez que os atributos com essa propriedade reagem, essencialmente, da mesma forma quando são submetidos a variações.

Com esse intuito, o algoritmo realiza dois tipos de correlações: a primeira, sobre os atributos numéricos; e a segunda, compara a relação entre os numéricos e aqueles classificados como categóricos.

Em relação à primeira, realizada sobre os atributos numéricos, considera a correlação estatística de Spearman como parâmetro para avaliar o grau da associação entre os atributos. Optou-se pela correlação de Spearman em virtude desta avaliar relações monótonas, tanto lineares quanto não lineares (sendo uma vantagem sobre a correlação de Pearson, que avalia apenas relações lineares). Por padrão, definiu-se que se o parâmetro for igual ou superior a 0,95 – caracterizando a correlação de Spearman como muita alta - o tipo de informação contida nesses atributos é idêntica. Logo, se as informações são iguais do ponto de vista estatístico, preserva-se apenas uma das dimensões correlacionadas e procede-se à exclusão das restantes.

No que se refere à correlação entre atributos categóricos (não numéricos) e os numéricos, a associação foi medida por meio da procura de mapeamentos bijetivos entre esses dois conjuntos. Assim, se determinado atributo numérico possui a mesma quantidade de elementos distintos quando comparado a um categórico, significa que há uma relação unívoca entre os atributos mencionados, eliminando-se o categórico do conjunto de dados e mantendo-se apenas o numérico, uma vez que este é o necessário ao módulo de detecção de anomalias.

Através desses procedimentos, o método de estruturação, então, reduz a dimensionalidade do sistema, mas preserva a qualidade das informações, haja vista que apenas as consideradas redundantes foram excluídas da base de dados a ser submetida ao detetor de anomalias.

Na Figura 14, é apresentado um exemplo comparando a estrutura inicial de um conjunto de testes com a estrutura após a saída do módulo de pré-processamento.

Atributos	Tipo de dado	
	Entrada	Saída
CODIGO	int64	Int64
cliente	object	Eliminado por Correlação
FROTA	object	category
ID_TRANSACAO	int64	Int64
HODOMETRO	int64	Int64
NOME	object	category
CPF	object	category
Ineficiência 2 R$	float64	Eliminado por Correlação
STATUS	object	category
litros consumidos	float64	float64
RBASE_CREDENC	int64	Int64
NOME_FANTASIA	object	category
CIDADE	object	category
ESTADO	object	category
DESCRICAO	object	category
QUANTIDADE	float64	Eliminado por Correlação
UNITARIO	float64	float64
Ineficiência 1 R$	float64	float64
PREFIXO	object	Eliminado por Correlação

Atributos	Tipo de dado	
	Entrada	Saída
VALOR	float64	Eliminado por Correlação
DATAHORA_TRANSACAO	object	int64
NUM_CARTAO	int64	Int64
PLACA	object	category
TIPO VEICULO	object	category
MARCA VEICULO	object	category
MODELO VEICULO	object	category
TIPO	object	category
litros esperados	float64	Eliminado por Correlação
ANO_FABRICACAO	int64	Eliminado por Correlação
ANO_MODELO	int64	Int64
DESLOCAMENTO	int64	Int64
CONSUMO	float64	float64
ID_CENTRO_CUSTO	float64	Eliminado por Correlação
CENTRO DE CUSTO VEICU	object	Eliminado por Correlação
CENTRO DE CUSTO FUNCI	object	Eliminado por Correlação
Referência km/l	float64	float64
Desvio Referência km/l	float64	float64

Figura 14: Amostra do conjunto de dados pré-processado, mas não normalizado.

Comparação entre os tipos de dados do conjunto de entrada e saída. Dimensões: Entrada =37 atributos e Saída=27 atributos. Houve preservação do número linhas: 159.650.

A última fase do pré-processamento relaciona-se à padronização (normalização) das escalas de representação dos dados numéricos. Essa etapa é essencial a diversas atividades de mineração de dados, em especial àquelas que procuram identificar discrepâncias, pois pode reduzir o viés associado a dimensões representadas em ordens de grandeza muito maiores do que as outras. Na Figura 15, apresenta-se um exemplo com as cinco primeiras instâncias dos oito primeiros atributos do conjunto de dados (base GerFrotas), antes de ser submetido à padronização da escala:

CODIGO	ID_TRANSACAO	HODOMETRO	RBASE_CREDENCIADO	UNITARIO	DATAHORA_TRANSACAO	NUM_CARTAO	ANO_MODELO
33630	217698	89763	18483	4.39	1559345700000000000	63936206013402432	2016
33630	217687	23910	25650	4.72	1559345400000000000	63936206013366242	2019
33630	217681	43802	2096	4.39	1559345340000000000	63936206013275162	2018
33630	217677	90000	2291	4.60	1559345040000000000	63936206013049842	2018
33630	217675	84721	54854	4167.00	1559345040000000000	63936206013076892	2017

Figura 15: Amostra do conjunto de dados pré-processado, mas não normalizado.

De acordo com a Figura 15, é possível observar a diferença entre as ordens de grandeza das escalas. Enquanto o atributo "UNITARIO" possui, em geral, ordem de grandeza 10^0, o atributo "DATAHORA_TRANSACAO", possui escala na ordem de 10^{18}. Endereçado a esse problema, os dados foram normalizados usando uma estratégia de *minmax*, que compara o valor de determinado atributo com seus valores máximo e mínimo para a padronização dos dados (ver Figura 16) e não parte da premissa que os dados obedecem à distribuição normal.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Figura 16: Estratégia de normalização MinMax.

A Figura 17 evidencia uma amostra do conjunto de dados após sofrer o procedimento de normalização:

CODIGO	ID_TRANSACAO	HODOMETRO	RBASE_CREDENCIADO	UNITARIO	DATAHORA_TRANSACAO	NUM_CARTAO	ANO_MODELO
0.0	0.217697	0.084542	0.299441	0.000073	0.610737	0.999981	0.9956
0.0	0.217686	0.022519	0.429036	0.000080	0.610730	0.999975	0.9989
0.0	0.217680	0.041254	0.003128	0.000073	0.610728	0.999960	0.9978
0.0	0.217676	0.084765	0.006654	0.000078	0.610721	0.999922	0.9978
0.0	0.217674	0.079793	0.957109	0.089837	0.610721	0.999927	0.9967

Figura 17: Amostra da base de dados pré-processada e normalizada. Perfil de variação da escala [0,1].

Pontua-se, inicialmente, que a estrutura do conjunto de dados foi alterada de forma a melhor representar as informações presentes em seus atributos. Dados do tipo numérico mantiveram sua estrutura. Contudo, aqueles que foram classificados preliminarmente como "objetos", tiveram seus tipos de dados alterados, conforme suas características, para (*string* não categórica, *string* categórica, *datetime*).

Especificamente no que tange ao tipo de dado "booleano", por sua própria natureza (dois elementos distintos no atributo), acabaram por ser encarados como atributos *strings* categóricos. Por sua vez, os *datetime*, que foram classificados no processo de leitura como "objetos", durante a execução da rotina, foram convertidos ao tipo de dado numérico, através de um mapeamento bijetivo.

Por fim, o módulo de pré-processamento promoveu a normalização das escalas dos atributos do conjunto de dados, deixando-o pronto para atuação do módulo subsequente, o de detecção de *outliers*.

4.2 Módulo de Detecção de *Outliers*

Na Seção 4.1, foram apresentadas as principais tarefas executadas pelo módulo de pré-processamento da solução. Detalharam-se os objetivos gerais relacionados à organização inicial dos dados, evidenciando que o conjunto de dados de saída, entendido como aquele entregue ao final do primeiro módulo, continha a maior quantidade de informações relevantes, inclusive não numéricas.

O módulo de detecção de *outliers*, por sua vez, é aquele responsável por executar a abordagem de detecção de anomalias, a qual, neste trabalho, baseia-se nas estratégias

materializadas pelos estimadores apresentados no Capítulo 2.

O módulo de detecção atua sobre o conjunto pré-processado e, através das informações numéricas disponíveis na base de dados, de maneira não supervisionada, busca identificar instâncias presentes nas regiões no espaço d- dimensional que possuam características capazes de indicar seu comportamento anormal (eventos em regiões com baixa densidade, isolados, incompatíveis com o ajuste realizado com uso de histogramas, entre outros).

Dessa forma, o módulo identifica eventos presentes na base de dados que destoam, significativamente, da distribuição geral observada, colaborando para a consecução do principal objetivo da solução proposta, que é a obtenção de uma amostra integrada por eventos com alta propensão de constituírem-se como anomalias.

Operacionalmente, o pré-processador entrega, ao detector de anomalias, o conjunto de dados numérico, pré-processado e padronizado, sendo o conjunto uma das entradas da função de detecção de *outliers*.

O módulo de detecção, por sua vez, ao receber a base de dados pré-processada, a submete às estratégias de deteção de anomalias que integram o módulo, de forma que cada instância do conjunto de dados é avaliada por todos os estimadores presentes no módulo (*Local Outlier Factor* - LOF [23]; *Clustering Based Local Outlier Factor* - CBLOF [24]; *Histogram-based outlier score* - HBOS [25]; *One- Class Suport Vector Machine* - OCSVM [26]; *k-Nearest Neighbors Detector* - kNN [27]; *Isolation Forest* - IF [28] e *Robust Covariance* [29]).

Conforme detalhado no Capítulo 2, uma das vantagens de se utilizar técnicas que enxergam a anormalidade como uma característica contínua da instância (em oposição à visão puramente binária) é justamente a possibilidade de se determinar uma pontuação para a intensidade da anormalidade observada. Essa pontuação, fornecida por cada um dos estimadores a cada instância, é o que define se o evento é ou não uma anomalia perante a estratégia de detecção de determinado estimador.

Após o processamento do conjunto de dados pelo detector de anomalias, cada estimador terá classificado as instâncias em *inliers* ou *outliers*, a depender das características do evento e da intensidade da anomalia, permitindo que o detector de anomalias reúna as instâncias que foram classificadas como *outliers* pelos estimadores. Com a lista de todas as instâncias da base de dados, o detector seleciona uma amostra contendo todas as instâncias que, simultaneamente, foram consideradas anômalas por três ou mais estimadores.

Dessa forma, o módulo define o grau de anomalia de cada instância, entendido como o número de vezes em que cada uma delas foi considerada como *outlier* pelos estimadores do módulo. A propensão de anormalidade dos elementos que integram a amostra é dada,

justamente, pelo grau das anomalias, uma vez que quanto maior o número de vezes em que uma determinada instância é reprovada (considerada *outlier* por diversos critérios), maior torna-se a chance de que ela represente um *outlier* verdadeiro.

Portanto, como saída da solução proposta, o detector de anomalias entrega uma tabela contendo uma amostra das instâncias da base de dados de entrada contendo eventos que foram considerados *outliers* por, no mínimo, três estimadores, remetendo a amostra a um grau de anormalidade ≥ 3.

No Capítulo 5, serão apresentados os resultados obtidos com o uso da solução proposta sobre as bases de dados reais, relativas aos dados provenientes da execução orçamentária do Fundo Municipal de Saúde de João Pessoa (FMSJP) [5], à aquisição de combustíveis pela Administração Pública do Estado da Paraíba (GerFrotas) [13] e a dispensas e inexigibilidades de licitação do Governo Federal (DIGF) [14].

5 Resultados e Discussões

No Capítulo 4, apresentou-se a metodologia utilizada no desenvolvimento da solução proposta. De forma detalhada, discorreu-se sobre a visão esquemática da solução e sobre cada módulo que a integra, evidenciando as tarefas executadas durante o pré-processamento automatizado e delineando a abordagem operacional do módulo de detecção de *outliers*.

Além disso, destacou-se que a probabilidade da característica anômala das instâncias contidas na amostra decorre da quantidade de vezes que cada uma delas foi considerada como *outlier* pelos estimadores do módulo de detecção.

Neste capítulo, serão trazidos os resultados obtidos com o uso da solução sobre três conjuntos de dados reais, provenientes da Administração Pública. Esses dados referem-se a aquisições realizadas pelo Estado em suas diferentes esferas de atuação (Governo Federal, Governo Estadual e Municipal).

Na Seção 5.1, será fornecida uma visão geral sobre as bases de dados utilizadas.

5.1 Bases de Dados utilizadas nos Experimentos

Foram utilizadas, neste trabalho, três bases de dados distintas provenientes da Administração Pública Brasileira, a saber: dados orçamentários do Fundo Municipal de Saúde de João Pessoa – PB entre 2016 a 2020 (FMSJP) [5]; dados relativos ao Gerenciamento do Abastecimento da Frotas do Estado da Paraíba entre 2017 a 2019 (GerFrotas) [13]; e dados de Dispensas e Inexigibilidades de Licitações do Governo Federal entre 2014 e 2019 (DIGF) [14].

As características gerais dessas bases são apresentadas na Tabela 1:

Tabela 1: Visão geral das Bases de Dados utilizadas

Bases de Dados	Instâncias	Atributos
FMSJP	20.389	23
DIGF	816.138	12
GerFrotas	930.524	27

As bases FMSJP e DIGF contêm atributos relacionados à identificação das unidades orçamentárias responsáveis por cada evento registrado. Entre os atributos existentes estavam o valor da nota de empenho, sua numeração, a data da emissão, identificação do credor, valor pago ou empenhado e unidade gestora responsável pelo pagamento.

Por outro lado, a GerFrotas contém atributos relacionados a todos os abastecimentos dos veículos e equipamentos integrantes da Frota do Estado da Paraíba. As

informações disponíveis revelam os valores totais e unitários (R$/l) por abastecimento; o tipo de combustível adquirido (gasolina, álcool, diesel); os rendimentos efetivo, médio e mediano (em km/l) para o equipamento abastecido; sua marca e modelo; além de outras informações.

As bases de dados da Tabela 1 continham instâncias que, usando um método convencional, foram classificadas como "normais" ou "anômalas" a depender da satisfação cumulativa da restrição de compatibilidade imposta pelo seguinte critério estatístico:

$$\textbf{\textit{Restrição:}}\ \text{Se} \quad v_0 - v_{\text{médio}} > 3.\,\sigma_m \quad \textbf{\textit{e}} \quad v_0 - v_{\text{mediano}} > 3.\text{mad}, \tag{1}$$

então v_0 é um ***outlier***,

onde v_0 representa o valor da instância em estudo (observada) pertencente à base de dados; $v_{\text{médio}}$ é o valor médio do atributo de interesse; σ_m é o seu desvio padrão da média; v_{mediano} representa o valor mediano do atributo de interesse; e, por fim, o *mad* (*Median Absolute Deviation*) é desvio mediano absoluto correspondente.

Para a GerFrotas, a análise estatística convencional confrontou o rendimento efetivamente apresentado pelo equipamento (v_0) com os rendimentos médio ($v_{\text{médio}}$) e o mediano (v_{mediano}) esperados. Assim, nos casos em que a restrição foi atendida, a análise convencional classificou a instância como anômala. Já para a base FMSJP, a análise de compatibilidade confrontou o gasto observado (v_0) e os gastos médio ($v_{\text{médio}}$) e mediano (v_{mediano}) com determinado credor no período a que se refere o conjunto de dados. Por fim, para a base DIGF, o confronto foi realizado entre os gastos observado, médio e mediano das diversas unidades orçamentárias da União. Dessa forma, nos casos em que houve incompatibilidade entre o valor empenhado e os valores médio e mediano esperados para as respectivas unidades orçamentárias no período, a análise convencional classificou os eventos como aberrantes.

A execução da análise estatística convencional sobre as bases de dados da Tabela 1 considerou como instâncias anômalas todas aquelas que atenderam às exigências da restrição apresentada na Equação 1, culminando na identificação de *outliers*, nas quantidades e proporções apresentadas na Tabela 2:

Tabela 2: Presença de *Outliers* identificados pela análise estatística convencional sobre as bases de dados utilizadas

Bases de Dados	QTDE. *Outliers*	Proporção
FMSJP	1.156	5,67%
DIGF	27.572	3,38%
GerFrotas	20.292	2,18%

As classificações realizadas pela análise estatística convencional, expressas na Tabela 2, ampararam o processo de validação dos resultados obtidos pelo sistema de apoio proposto, uma vez que permitiu a comparação entre as instâncias indicadas como *outliers* pelo sistema e aquelas identificadas pela metodologia estatística convencional.

Na Seção 5.2, serão evidenciados os resultados obtidos pela solução objeto deste estudo.

5.2 Resultados obtidos com o uso do Sistema de Apoio à Detecção de Anomalias

Esta seção apresentará os resultados obtidos com a aplicação da solução proposta aos conjunto de dados utilizados neste trabalho.

Após a fase do pré-processamento, o módulo de detecção de anomalias atuou sobre os atributos numéricos das bases de dados pré-processadas, de forma que o módulo de detecção foi executado oito vezes, sendo que a cada execução, a quantidade máxima de *outliers* a serem obtidos foi variada (150, 300, 600, 900, 1200, 1500, 2000 e 2500). O aumento progressivo dessa quantidade permitiu a ampliação da cardinalidade da amostra retornada pelo algoritmo (instâncias candidatas a *outliers*) e uma avaliação mais abrangente sobre se os eventos dessa amostra constituíam, de fato, instâncias cujo comportamento é anômalo.

Como mencionado na Seção 5.1, os resultados do algoritmo proposto foram comparados àqueles obtidos com o uso da metodologia estatística convencional. Assim, cada instância presente nas amostras devolvidas pelo Sistema de Apoio à Detecção de Anomalias foi comparada à classificação feita pela metodologia convencional, considerando-se como *outliers* verdadeiros todos os eventos das amostras que, simultaneamente, foram classificados como *outliers* pelas duas metodologias.

Considerando os procedimentos adotados, preliminarmente obteve-se os seguintes resultados para cada base de dados:

Tabela 3: Resultados parciais para a base de dados FMSJP

Qtde. Max. Retornada	**Outliers Grau≥3**	**AC ∩ AP**	Outliers - AP	Taxa AC ∩ AP
150	95	67	28	70,53%
300	231	139	92	60,17%
600	387	184	203	47,55%
900	563	224	339	39,79%
1200	794	271	523	34,13%
1500	1024	313	711	30,57%
2000	1465	392	1073	26,76%
2500	2009	455	1554	22,65%

Tabela 4: Resultados parciais para a base de dados DIGF

Qtde. Max. Retornada	**Outliers Grau≥3**	**AC ∩ AP**	Outliers - AP	Taxa AC ∩ AP
150	58	55	3	94,83%
300	162	147	15	90,74%
600	320	298	22	93,12%
900	464	425	39	91,59%
1200	638	572	66	89,65%
1500	827	726	101	87,79%
2000	1114	948	166	85,10%
2500	1542	1198	344	77,69%

Tabela 5: Resultados parciais para a base de dados GerFrotas

Qtde. Max. Retornada	**Outliers Grau≥3**	**AC ∩ AP**	Outliers - AP	Taxa AC ∩ AP
150	72	71	1	98,61%
300	90	89	1	98,89%
600	133	127	6	95,49%
900	176	166	10	94,32%
1200	255	231	24	90,59%
1500	375	315	60	84,00%
2000	558	441	117	79,03%
2500	739	549	190	74,29%

O atributo "Qtde. Max. Retornada"é um parâmetro de execução definido pelo usuário e utilizado para definir a quantidade máxima de *outliers* a ser retornado pelo sistema (*Outliers* com grau ≥1), de forma a se ajustar a quantidade obtida com a capacidade efetiva de análise - perante os usuários que poderão analisar a amostra. Já o atributo "*Outliers* Grau≥ 3"representa o somatório das colunas AC ∩ AP e *Outliers* - AP. Estas, por sua vez e respectivamente, indicam a i) intersecção entre os *outliers* identificados pela abordagem convencional (AC) e as anomalias identificadas pela abordagem proposta (AP) e ii) os *outliers* de grau superior a três identificados apenas pela AP. Finalmente, a "Taxa AC ∩ AP" indica percentualmente as instâncias identificadas como *outliers* pelas duas abordagens, considerando o total de anomalias obtidas pela AP ("*Outliers* Grau≥ 3").

Da análise das informações, verificou-se que, apesar da quantidade de *outliers* identificados pela abordagem convencional (AC) representar 2,18% e 3,38%, respectivamente, das bases GerFrotas e DIGF, o sistema proposto foi capaz de identificar, com acurácia média de 89,40% +/- 3,28% e 88,81% +/- 1,91%, instâncias que cumulativamente foram classificadas como anômalas pela AC e pela AP. Rememora-se que as classificações da AP possuem como característica grau igual ou superior a três, isto é, representam eventos que foram considerados como *outliers* por, no mínimo, três estimadores de detecção de anomalias distintos. Para a base FMSJP, embora o percentual de *outliers* identificados pela AC tenha sido de 5,67% e, portanto, maior do que para as bases DIGF e GerFrotas, a acurácia média atingida foi de 41,52% +/- 5,94%.

Independentemente da base utilizada (FMSJP, DIGF ou GerFrotas), verificou-se que, ao aumentar a quantidade máxima de *outliers* a serem obtidos (Qtde. Max. Retornada), havia aumento na quantidade de instâncias classificadas como *outliers* apenas pela AP. A Figura 18 apresenta esse comportamento:

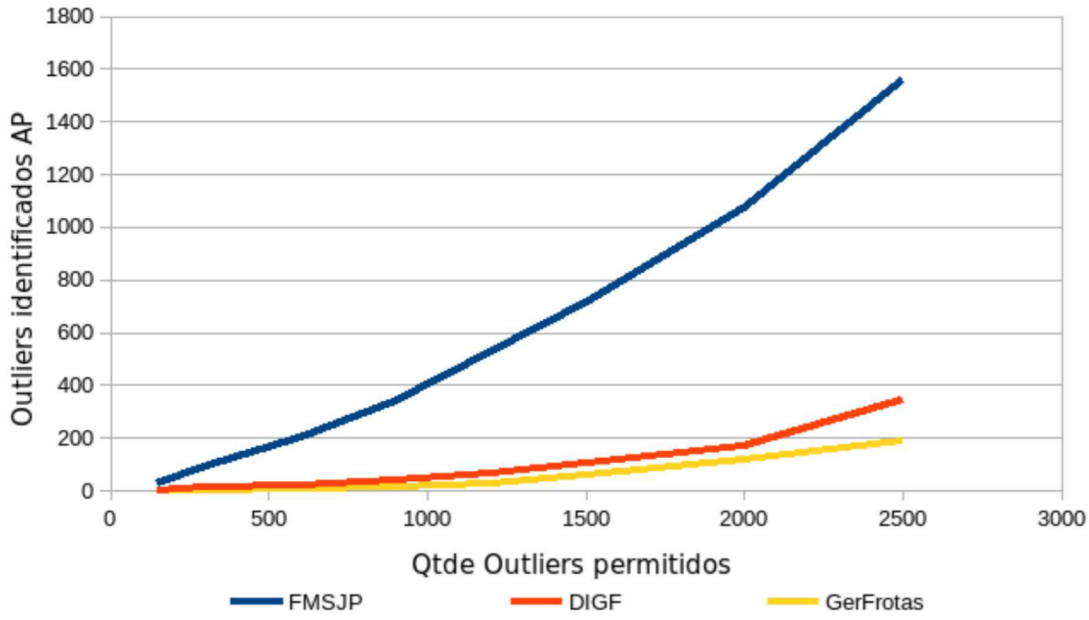

Figura 18: Quantidade de *outliers* identificados apenas pela AP em função da quantidade máxima retornada (Qtde. Max. Retornada).

Tendo em vista a tendência representada pela Figura 18, tornou-se necessária a realização de uma análise complementar endereçada a verificar se as instâncias identificadas apenas pela AP eram classificações equivocadas do algoritmo proposto ou se representavam *outliers* não identificados pela abordagem convencional - AC.

Para a realização desse procedimento adicional de validação do sistema proposto, solicitou-se apoio de três especialistas, Auditores Governamentais integrantes dos quadros funcionais do Tribunal de Contas do Estado da Paraíba, para que, através dos critérios utilizados em auditoria das contas públicas, pudessem opinar sobre a regularidade/legitimidade dos eventos classificados como anomalias apenas pela AP (colunas *Outliers* - AP, presentes nas Tabelas 3, 4 e 5).

Dessa forma, a cada um dos especialistas, foram enviadas três relações dos eventos considerados como *outliers* pela AP, sendo uma relação por base de dados. Nessas relações, as informações sobre cada evento, indicado como *outlier* pela AP, foram disponibilizadas de forma que os especialistas tivessem acesso a todas as informações constantes do banco de dados original (conjunto de entrada - bases DIGF, FMSJP e GerFrotas antes do pré-processamento) e, assim, pudessem avaliar as instâncias com as mesmas informações presentes no conjunto original.

A análise complementar procedida pelos especialistas buscou avaliar cada um dos eventos identificados como anomalias pela abordagem proposta (AP), no intuito de verificar se o evento em questão possuía características capazes de o incluir em uma amostra diligenciável pela Auditoria. Como item diligenciável, entende-se todo evento que, por

possuir um determinado comportamento, não pode ser considerado regular sem que sejam realizadas fiscalizações mais aprofundadas, como a solicitação de documentos e esclarecimentos adicionais, fiscalização *in loco*, circularização e outros procedimentos de auditoria.

Como resposta aos procedimentos realizados, cada um dos especialistas devolveu três bases de dados (DIGF, FMSJP e GerFrotas) contendo seus pareceres sobre a regularidade dos eventos apresentados.

Após a feitura das análises, com a emissão dos pareceres dos especialistas, as informações fornecidas foram consolidadas tendo-se em vista três critérios distintos. No primeiro, definiu-se como *outlier* verdadeiro toda instância em que houve concordância unânime entre os especialistas sobre o seu perfil de anormalidade, remetendo a um cenário de pior caso. O segundo critério, por sua vez, definiu-se como *outlier* verdadeiro toda instância que foi considerada anômala por, pelo menos, um dos especialistas - cenário mais flexível (melhor caso). Por fim, em um cenário intermediário, considerou-se como *outlier* verdadeiro instâncias que tiveram o perfil anômalo identificado por, pelo menos, dois dos especialistas. Nos outros casos, isto é, para as instâncias não classificadas como *outliers* verdadeiros, embora o algoritmo proposto tenha indicado seu perfil de anormalidade, tal comportamento não foi ratificado pelos especialistas.

A realização da análise complementar pelos especialistas permitiu verificar que, embora não identificadas pela abordagem convencional (AC), diversas instâncias classificadas como *outliers* apenas pela abordagem proposta (AP) representavam inconsistências decorrentes de registros (identificados pelo id[1]) i) que possuíam erro, a exemplo dos rendimentos veiculares negativos apresentados na Tabela 8; ii) que indicavam equipamentos abastecidos e não utilizados - rendimento nulo - Tabela 8; iii) de aquisição de combustíveis com preço unitário irreal - Tabela 8; iv) que, apesar de possuírem valores empenhados, não houve registro dos pagamentos correspondentes - Tabelas 6 e 7; e v) que continham eventos incompatíveis com a média $(v_0 – v_{\text{médio}})/3.\sigma_m > 1$ (Desvio[2]) - Tabelas 6 e 7.

As Tabelas 6, 7 e 8 apresentam alguns exemplos dessas instâncias inconsistentes identificadas apenas pela AP:

[1]id: identificador da transação.
[2]Desvio: $(v_0 - v_{\text{médio}})/3.\sigma_m$.

Tabela 6: Exemplos de *outliers* identificados apenas pela AP com comportamento anômalo confirmado pelos especialistas - base FMSJP

id	Valor Empenhado	Valor Pago	Desvio
27	R$ 385.736,39	0	11,56
25	R$ 330.000,00	0	8,70
16404	R$ 403.090,80	0	5,66
6642	R$ 382.014,15	0	5,32
2191	R$ 107.250,00	0	4,94
4748	R$ 349.461,02	0	4,79
11420	R$ 194.600,00	0	3,85

Tabela 7: Exemplos de *outliers* identificados apenas pela AP com comportamento anômalo confirmado pelos especialistas - base DIGF

id	co_uasg	Valor Estimado	Desvio
25000507000532	250005	R$ 46.728.660,00	17,19
25000506000942	250005	R$ 45.966.453,00	16,85
25000506004782	250005	R$ 44.835.625,00	16,33
97400306000062	974003	R$ 0,00	2,46
97400306000102	974003	R$ 0,00	2,46
97400306000072	974003	R$ 0,00	2,46

Tabela 8: Exemplos de *outliers* identificados apenas pela AP com comportamento anômalo confirmado pelos especialistas - base GerFrotas

id	Placa	Valor Pago	Valor R$/litro	Rendimento Efetivo km/l
752660	MOK9302	R$ 300,00	3,09	-7439,98
907663	MOH2236	R$ 364,00	3,64	-494,98
473727	FUM8051	R$ 57,90	3,86	-86,67
904486	OFG2242	R$ 307,00	3,07	0
904487	OFG2342	R$ 307,00	3,07	0
505454	QFY0520	R$ 36,00	4749	65,04

id	Placa	Valor Pago	Valor R$/litro	Rendimento Efetivo km/l
466153	MOV1230	R$ 28,44	4989	111,4
487472	NQJ3777	R$ 28,00	4746	111,53
575070	QFV6696	R$ 24,97	4729	136,36
562634	MOS3749	R$ 20,00	4796	136,93
653315	NQJ9959	R$ 9,50	4634	181,95

Finda a análise complementar, as instâncias que tiveram o comportamento anômalo confirmado pelos especialistas, em virtude de características semelhantes às descritas em i) a v), foram inseridas no rol daquelas que representam *outliers* verdadeiros classificados pela AP. Já as instâncias em que não se pôde confirmar o comportamento aberrante, foram excluídas do grupo de possíveis *outliers*, uma vez que podem representar classificações errôneas do método proposto ou anomalias de difícil visualização.

As Tabelas 9, 10 e 11 indicam os resultados obtidos, tendo-se em vista o cenário de pior caso, isto é, a situação em que a instância é considerada *outlier* verdadeiro apenas se há unanimidade entre os especialistas sobre a anormalidade do evento:

Tabela 9: Resultado para FMSJP considerando os *Outliers*-AP confirmados no cenário de Pior Caso.

Qtde. Máx.	Outliers Grau≥3	Outliers Verdadeiros	Outliers não confirmados	AC∩AP	Total AP	Taxa AP
150	95	23	5	67	90	94,74%
300	231	69	23	139	208	90,04%
600	387	113	90	184	297	76,74%
900	563	164	175	224	388	68,92%
1200	794	225	298	271	496	62,47%
1500	1024	268	443	313	581	56,74%
2000	1465	355	718	392	747	50,99%
2500	2009	440	1114	455	895	44,55%

Tabela 10: Resultado final para DIGF considerando os *Outliers*-AP confirmados no cenário de Pior Caso.

Qtde. Máx.	Outliers Grau≥3	Outliers Verdadeiros	Outliers não confirmados	AC∩AP	Total AP	Taxa AP
150	58	0	3	55	55	94,83%
300	162	4	11	147	151	93,21%
600	320	6	16	298	304	95,00%
900	464	14	25	425	439	94,61%
1200	638	31	35	572	603	94,51%
1500	827	49	52	726	775	93,71%
2000	1114	87	79	948	1035	92,91%
2500	1542	150	194	1198	1348	87,42%

Tabela 11: Resultado para GerFrotas considerando os *Outliers*-AP confirmados no cenário de Pior Caso.

Qtde. Máx.	Outliers Grau≥3	Outliers Verdadeiros	Outliers não confirmados	AC∩AP	Total AP	Taxa AP
150	72	1	0	71	72	100,00%
300	90	1	0	89	90	100,00%
600	133	3	3	127	130	97,74%
900	176	5	5	166	171	97,16%
1200	255	17	7	231	248	97,25%
1500	375	50	10	315	365	97,33%
2000	558	101	16	441	542	97,13%
2500	739	157	33	549	706	95,53%

Dessa forma, considerando o cenário de pior caso, a acurácia da solução proposta para as bases FMSJP, DIGF e GerFrotas atingiu, respectivamente, os percentuais de 68,15% +/- 6,37%, 93,28% +/- 0,88% e 97,77% +/- 0,54%, culminando na acurácia média de 86,40% +/- 3,41% para a solução no pior caso.

Já pelo critério de melhor caso (ao menos um especialista identificou o comportamento da instância), o desempenho da solução proposta pode ser obtido através da verificação das Tabelas 12, 13 e 14:

Tabela 12: Resultado final para FMSJP considerando os *Outliers*-AP confirmados no cenário de Melhor Caso.

Qtde. Máx.	Outliers Grau≥3	Outliers Verdadeiros	Outliers não confirmados	AC∩AP	Total AP	Taxa AP
150	95	28	0	67	95	100,00%
300	231	91	1	139	230	99,57%
600	387	168	35	184	352	90,96%
900	563	254	85	224	478	84,90%
1200	794	357	166	271	628	79,09%
1500	1024	454	257	313	767	74,90%
2000	1465	639	434	392	1031	70,38%
2500	2009	870	684	455	1325	65,95%

Tabela 13: Resultado final para DIGF considerando os *Outliers*-AP confirmados no cenário de Melhor Caso.

Qtde. Máx.	Outliers Grau≥3	Outliers Verdadeiros	Outliers não confirmados	AC∩AP	Total AP	Taxa AP
150	58	3	0	55	58	100,00%
300	162	11	4	147	158	97,53%
600	320	18	4	298	316	98,75%
900	464	30	9	425	455	98,06%
1200	638	51	15	572	623	97,65%
1500	827	78	23	726	804	97,22%
2000	1114	120	46	948	1068	95,87%
2500	1542	235	109	1198	1433	92,93%

Tabela 14: Resultado final para GerFrotas considerando os *Outliers*-AP Confirmados no cenário de Melhor Caso.

Qtde. Máx.	Outliers Grau≥3	Outliers Verdadeiros	Outliers não confirmados	AC∩AP	Total AP	Taxa AP
150	72	1	0	71	72	100,00%
300	90	1	0	89	90	100,00%
600	133	5	1	127	132	99,25%
900	176	7	3	166	173	98,30%
1200	255	20	4	231	251	98,43%
1500	375	56	4	315	371	98,93%
2000	558	111	6	441	552	98,92%
2500	739	179	11	549	728	98,51%

Logo, considerando o cenário de melhor caso, a acurácia da solução proposta para as bases FMSJP, DIGF e GerFrotas atingiu, respectivamente, os percentuais de 83,22% +/- 4,55%, 97,25% +/- 0,75% e 99,04% +/- 0,24%, culminando na acurácia média de 93,17% +/- 2,08% para a solução no critério de melhor caso.

Por fim, as Tabelas 15, 16 e 17 indicam os resultados obtidos no caso intermediário, isto é, situação que considera como *outliers* verdadeiros as instâncias classificadas como anômalas pela AP e que tiveram confirmação do comportamento aberrante por, pelo menos, dois dos especialistas:

Tabela 15: Resultado final para FMSJP considerando os *Outliers*-AP confirmados no cenário Intermediário.

Qtde. Máx.	Outliers Grau≥3	Outliers Verdadeiros	Outliers não confirmados	AC∩AP	Total AP	Taxa AP
150	95	28	0	67	95	100,00%
300	231	91	1	139	230	99,57%
600	387	161	42	184	345	89,15%
900	563	241	98	224	465	82,59%
1200	794	331	192	271	602	75,82%
1500	1024	407	304	313	720	70,31%
2000	1465	561	512	392	953	65,05%
2500	2009	740	814	455	1195	59,48%

Tabela 16: Resultado final para DIGF considerando os *Outliers*-AP Confirmados no cenário Intermediário.

Qtde. Máx.	Outliers Grau≥3	Outliers Verdadeiros	Outliers não confirmados	AC∩AP	Total AP	Taxa AP
150	58	0	3	55	55	94,83%
300	162	4	11	147	151	93,21%
600	320	6	16	298	304	95,00%
900	464	14	25	425	439	94,61%
1200	638	33	33	572	605	94,83%
1500	827	51	50	726	777	93,95%
2000	1114	90	76	948	1038	93,18%
2500	1542	177	167	1198	1375	89,17%

Tabela 17: Resultado final para GerFrotas considerando os *Outliers*-AP Confirmados no cenário Intermediário.

Qtde. Máx.	Outliers Grau≥3	Outliers Verdadeiros	Outliers não confirmados	AC∩AP	Total AP	Taxa AP
150	72	1	0	71	72	100,00%
300	90	1	0	89	90	100,00%
600	133	3	3	127	130	97,74%
900	176	5	5	166	171	97,16%
1200	255	17	7	231	248	97,25%
1500	375	52	8	315	367	97,87%
2000	558	105	12	441	546	97,85%
2500	739	163	27	549	712	96,35%

Para o caso médio (cenário intermediário), a acurácia do método proposto para as bases FMSJP, DIGF e GerFrotas atingiu, respectivamente, os percentuais de 80,25% +/- 5,39%, 93,60% +/- 0,68% e 98,03% +/- 0,46%, fazendo com que a acurácia média para o caso intermediário atingisse 90,62% +/- 2,35%.

Portanto, a análise complementar sobre o comportamento das instâncias classificadas como *outliers* apenas por AP (vide Figura 18) possibilitou o aumento da acurácia do Sistema de Apoio proposto, pois confirmou o comportamento anômalo para diversas daquelas instâncias. Além disso, o uso de diferentes critérios, aplicados na interpretação das análises realizadas pelos especialistas, permitiu a obtenção dos resultados da solução para os cenários de melhor caso, pior caso e caso intermediário. A Tabela 18 consolida o resultado obtido com a aplicação de cada um dos critérios:

Tabela 18: Resultados obtidos nos 3 cenários avaliados.

Cenário	Acurácia
Pior Caso	86,40% +/- 3,41%
Melhor Caso	93,17% +/- 2,08%
Caso Médio	90,64% +/- 2,35%
Resultado do método	90,07% +/- 1,98%

Dessa forma, considerando a média dos resultados obtidos nos três cenários avaliados, a acurácia geral do método proposto para as bases FMSJP, DIGF e GerFrotas atingiu 90,07% +/- 1,98% para as bases testadas.

O resultado do sistema ainda pode ser avaliado por base de dados, através da consolidação da performance da solução nos três cenários em cada conjunto de dados. As Tabelas 19, 20 e 21 ilustram a acurácia média da solução por base de dados testada:

Tabela 19: Resultados obtidos nos 3 cenários avaliados para FMSJP.

Cenário	Acurácia
Pior Caso	68,15% +/- 6,37%
Melhor Caso	83,22% +/- 4,55%
Caso Médio	80,25% +/- 5,39%
Média para a FMSJP	74,20% +/- 3,32%

Tabela 20: Resultados obtidos nos 3 cenários avaliados para DIGF.

Cenário	Acurácia
Pior Caso	93,28% +/- 0,88%
Melhor Caso	97,25% +/- 0,75%
Caso Médio	93,60% +/- 0,68%
Média para a DIGF	94,72% +/- 0.58%

Tabela 21: Resultados obtidos nos 3 cenários avaliados para GerFrotas.

Cenário	Acurácia
Pior Caso	97,77% +/- 0,54%
Melhor Caso	99,04% +/- 0,24%
Caso Médio	98,03% +/- 0,46%
Média para a GerFrotas	98,28% +/- 0,26%

Considerando-se a visão por base, a solução apresentou acurácia média de 74,20% +/- 3,32%, para a base FMSJP; 94,72% +/- 0.58%, para a DIGF; e de 98,28% +/- 0,26%, para a base de dados GerFrotas. Assim, embora o resultado geral do método seja dado por 90,07% +/- 1,98% (Tabela 18), nas aplicações sobre as bases DIGF e GerFrotas, a solução proposta mostrou performance superior à acurácia média geral.

6 Considerações Finais

No presente trabalho, propôs-se uma solução que fosse capaz de extrair conhecimento de três bases de dados públicas, relacionadas às aquisições de combustíveis para atender à frota estadual paraibana (base GerFrotas), às dispensas e inexigibilidades de licitação realizadas pelo Governo Federal (base DIGF) e às compras de materiais de consumo pelo Fundo Municipal de Saúde de João Pessoa (base FMSJP), de forma a subsidiar os processos associados à tomada de decisão e à fiscalização das despesas, tendo-se em vista que a quantidade de informações armazenadas digitalmente vem crescendo e, por sua vez, análises manuais não são mais adequadas para responder à demanda de interpretação desses dados.

Nesse contexto, o trabalho procurou verificar se é possível otimizar o processo de decisão dos gestores públicos e os planos de fiscalização, através da implementação de técnicas computacionais relacionadas à detecção de anomalias aplicadas aos dados registrados pelos órgãos públicos, de forma a se selecionar amostras de eventos com alta propensão de representarem *outliers*.

A qualidade (grau) dos *outliers* obtidos está diretamente associada às atividades voltadas à fiscalização, onde torna-se imprescindível a otimização dos recursos disponíveis, direcionando-os a amostras de eventos com maior potencial de risco, que podem ser representados pelas anomalias de maior grau.

Quanto à definição do conceito de grau da anomalia, neste trabalho, foi considerado como o número de vezes em que cada instância foi classificada como *outlier* pelos estimadores. Dessa forma, se após a atuação do algoritmo, uma determinada instância tiver sido classificada como anômala por apenas um dos estimadores, o grau da anomalia do evento será 1 (um).

No desenvolvimento do trabalho, com o intuito de constituir a amostra de saída do Sistema de Apoio à Detecção de Anomalias, foram considerados somente os eventos que possuíam grau mínimo 3 (três). Em outras palavras, as instâncias da amostra retornada foram, simultaneamente, classificadas como *outliers* por, pelo menos, três estimadores diferentes.

A definição de um grau mínimo para as instâncias da amostra foi necessária para garantir que ela fosse integrada por apenas eventos que possuíssem alta propensão à anormalidade. A definição do grau mínimo três foi feita de maneira empírica, mas amparada pelo entendimento de que instâncias reprovadas por três ou mais estimadores diferentes dão indícios suficientes de seu comportamento dissonante frente ao padrão geral observado na base.

Nesse cenário, a abordagem proposta foi capaz de retornar amostras de instâncias

que foram corretamente classificadas como *outliers* em 86,40% +/- 3,41% para o critério de pior caso; 93,17% +/- 2,08% para o melhor caso; e 90,64% +/- 2,35% para o caso intermediário. De maneira consolidada em relação às bases de dados testadas, a acurácia alcançada pela solução foi, em média, de 90,07% +/- 1,98%.

Verificou-se, também, que na visão que considera a performance da solução por base de dados, a solução apresentou resultados superiores à acurácia média obtida para as bases DIGF e GerFrotas, cujos resultados alcançaram, respectivamente, 94,72% +/- 0,58% e 98,28% +/- 0,26% (vide Tabelas 20 e 21).

Tendo-se em vista os resultados alcançados com a ferramenta proposta, foi possível responder às questões de pesquisas que motivaram o trabalho (Seção 1.1), bem como avaliar as hipóteses formalmente apresentadas na Seção 1.3.

Quanto à primeira questão de pesquisa (**QP1**), a qual indagava sobre a possibilidade de se otimizar o processo de decisão dos gestores públicos e os planos de fiscalização através da implementação de técnicas computacionais aplicadas aos dados registrados pelos órgãos públicos, o Sistema de Apoio proposto indica haver benefícios para a fiscalização e processos de tomada de decisão. Para a fiscalização, a vantagem está associada à obtenção de amostras que contêm instâncias com alta propensão de serem anômalas, facilitando o procedimento de amostragem, através de técnicas de aprendizagem de máquina. Além disso, a seleção da amostra feita pelo algoritmo proposto reduz a subjetividade envolvida em sua escolha, na medida em que passa a não depender da experiência de um dado profissional para a seleção das instâncias da amostra. Para o processo de tomada de decisão, pode-se afirmar que a solução também promove benefícios, pois, conhecendo-se os eventos com características aberrantes, o gestor pode se dedicar à investigação das causas que conduziram a tais comportamentos, podendo corrigir procedimentos e processos relacionados ao objeto avaliado. Nesse sentido, inclusive, o gestor pode identificar as maiores fragilidades e ameaças ao seu negócio de maneira a se proteger de fatores negativos, provenientes de acontecimentos tanto externos como internos (análise *SWOT - Strengths, Weaknesses, Opportunities and Threats*).

A hipótese nula *H0* proveniente da **QP1** afirmava que "os processos de decisão de gestão e planos de fiscalização não podem ser melhorados com o uso da estrutura computacional proposta". Contudo, os resultados alcançados pela Solução Proposta foram capazes de refutar tal hipótese, tendo-se em vista que a Solução fornece aos interessados uma amostra com elementos cujo comportamento anômalo foi confirmado em cerca de 90% dos casos, grau de confiança que favorece a facilitação dos processos tratados na hipótese.

Em relação à segunda questão de pesquisa (**QP2**), a qual buscava verificar se técnicas de detecção de anomalias aplicadas aos dados públicos são capazes de selecionar

amostras representativas de supostos candidatos a *outliers*, os resultados obtidos indicam que a solução é capaz de oferecer, para as bases de dados utilizadas neste trabalho, amostras de instâncias com alta propensão à anormalidade. Mesmo com bases de dados com poucos *outliers* identificados pela abordagem estatística convencional (AC) - Tabela 2, o algoritmo proposto foi capaz de identificar, com acurácia média de 90,07% +/- 1,98%, instâncias que cumulativamente foram classificadas como *outliers* pela AC e pela validação realizada pelos especialistas.

Assim, a hipótese nula (*H0*) decorrente da **QP2**, a qual afirmava que "técnicas de detecção de anomalias aplicadas aos dados públicos não permitem selecionar amostras representativas de candidatos a *outliers*"pode ser refutada, uma vez que, do total de instâncias retornadas pelo algoritmo proposto através da amostra, em média, 90,07% +/- 1,98% delas representavam eventos cujo comportamento anômalo foi confirmado (*outliers*) na validação.

Vale dizer, também, que a acurácia do algoritmo proposto atingiu valores superiores à média para as bases GerFrotas (98,28% +/- 0,26%) e DIGF (94,72% +/- 0,58%), o que permite concluir que o método tem potencial para contribuir na fiscalização desses tipos de despesas. Já para a base FMSJP, apesar de a acurácia ter sido menor do que a média (74,20% +/- 3,32%), em virtude da elevada quantidade de atributos não-numéricos na base (que não foram considerados pelo módulo de detecção de anomalias), o algoritmo mostrou bons resultados na seleção de amostras de *outliers* presentes nos dados.

Por fim, em relação à última questão de pesquisa (**QP3**), que indagava sobre se era possível validar a ferramenta proposta, bem como realizar a avaliação do contraste da técnica, foi verificado que o uso de critérios estatísticos de amostragem (testes de compatibilidade), aliado à avaliação procedida por especialistas da área de despesas públicas, puderam ser combinados para validar o sistema de apoio proposto. Inclusive, o contraste da técnica, representado pela acurácia média geral alcançada, pôde ser determinado. Então, a hipótese nula (*H0*) associada a esta questão, a qual afirmava que "as instâncias consideradas anômalas pela técnica de detecção de *outliers* não podem ser seguramente validadas, bem como o contraste da técnica não pôde ser aferido"deve ser refutada, uma vez que o método pôde ser validado e, além disso, permitiu a extração de seu contraste - acurácia.

Portanto, é possível concluir que o Sistema de Apoio à Detecção de Anomalias proposto demonstrou potencial de contribuir para o objetivo central deste trabalho, que é subsidiar os processos relacionados ao controle e à fiscalização dos gastos públicos, através da seleção de amostras que contenham eventos com alta propensão de constituírem-se como *outliers*. Além disso, a abordagem proposta pode contribuir para os processos relacionados à tomada de decisão pelos gestores, uma vez que, conhecendo-se os eventos anômalos nas bases de dados, passa a ser possível a investigação mais aprofundada de

suas causas, com vistas à otimização da gestão da coisa pública.

Em relação às limitações impostas à pesquisa, a principal decorreu dos escassos recursos computacionais disponíveis, os quais impediram que fossem realizados testes com quantidades retornadas de *outliers* superiores a 2500, pois acima desta quantidade, a execução do algoritmo em computadores de uso doméstico, por base de dados, consumiu cerca de 15h.

Em relação a trabalhos futuros, deseja-se 1) agregar novos estimadores ao módulos de detecção de anomalias, com estratégias distintas de detecção (baseadas, inclusive, em redes neurais); 2) desenvolver uma interface *on-line* para que o sistema possa ser executado em nuvem; e 3) realizar testes em outras bases de dados públicas, com o intuito de verificar se o sistema proposto pode ter sua aplicação estendida.

REFERÊNCIAS

[1] NEOWAY. Disponível em: https://www.neoway.com.br/o-que-e-big-data/. Acesso em: 15/06/2020.

[2] SLEASH GEAR. Disponível em https://www.slashgear.com/facebook-data-grows-by-over-500-tb-daily-23243691. Acesso em: 17/06/2020.

[3] Jornal do Comércio. Disponível em: https://www.jornaldocomercio.com/site/noticia.php?codn=159347. Acesso em: 20/06/2020.

[4] A. A. Rodrigues; D. A. B. Neves; G. A. Dias. Dados Abertos Governamentais e Apropriação Tecnológica: Acesso e Engajamento Cívico. **Rev. FSA, Edição, Teresina PI**, v. 17, n. 5, art. 2, p. 26-41, mai. 2020.

[5] Tribunal de Contas do Estado da Paraíba. Disponível em: https://sagres.tce.pb.gov.br.

[6] Gomes; Pimenta; Schneider.Data Mining in Information Science Research: Challenges And Opportunities».

[7] PIMENTA, R. M. Big data e controle da informação na era digital: tecnogênese de uma memória a serviço do mercado e do estado. Tendências da Pesquisa Brasileira em Ciência da Informação, v.6, n. 2, 2013.

[8] BEZERRA, Arthur Coelho. Vigilância e filtragem de conteúdo nas redes digitais: desafios para a competência crítica em informação. In: Encontro Nacional de Pesquisa em Ciência da Informação, 16., 2015, João Pessoa, UFPB, 2015.

[9] CALDAS, C. O. L.; CALDAS, P. N. L. Estado, democracia e tecnologia: conflitos políticos no contexto do big-data, das fakenews e das shitstorms. Perspectivas em Ciência da Informação, v. 24, n. 2, p. 196–220, 2019.

[10] LIMA JÚNIOR, Walter Teixeira. Jornalismo inteligente (JI) na era do data mining. In: Anais do II SBPJor (CD-ROM). Salvador-BA/Brasil, Ano: 2004.

[11] Zimek, A., Schubert, and Kriegel, H.-P. Unsupervised Outlier Detection in High-Dimensional Numerical Data, Published online 27 August 2012 in Wiley Online Library.

[12] Gladwell, Malcolm.***Fora de Série***. ISBN 9788575424483.

[13] Banco de dados do Tribunal de Contas do Estado da Paraíba. Disponível através de solicitação.

[14] Portal da Transparência do Governo Federal. Disponível em: http://www.portaltransparencia.gov.br.

[15] REZENDE, S. O., PUGLIESI, J. B., MELANDA, E. A., PAULA M. F., 2003, Mineração de Dados., Sistemas Inteligentes – Fundamentos e Aplicações. Manole.

[16] FAYYAD, U. M., PIATESKY-SHAPIRO, G.; SMYTH, P., 1996 "From Data Mining to Knowledge Discovery: An Overview". In: Advances in Knowledge Discovery and Data Mining, AAAI Press.

[17] O'brien, James A. (2005). Sistemas de Informação e as decisões gerenciais na era da internet 2º ed. São Paulo: Saraiva. p. 143. ISBN 9788502044074.

[18] Laudon, Kenneth. Laudon, Jane (2011). Sistemas de Informações Gerenciais: Fundamentos da inteligência de negócios: gestão da informação e de banco de dados. 9º ed. São Paulo: ABDR. p. 159.

[19] O'brien, James A.; Marakas, George M. (2007). Administração de Sistemas de Informação: uma introdução 13º ed. São Paulo: McGraw-Hill. p. 171. ISBN 8586804770.

[20] Abhijit A. Sawant and P. M. Chavan, "Study of Data Mining Techniques used for Financial Data Analysis", Volume 2, Issue 3, May 2013.

[21] Dinesh J. Prajapati, Jagruti H. Prajapati, "Handling Missing Values: Application to University Data Set", Issue 1, Vol.1, August-2011, ISSN 2249-6149 e Minakshi, Dr. Rajan Vohra and Gimpy, "Missing Value Imputation in Multi Attribute Data Set", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4). Ano: 2014

[22] Martin., Theus, (2009). Interactive graphics for data analysis : principles and examples. Boca Raton: CRC Press. ISBN 9781584885948. OCLC 245023938; C., Toit, S. H.; H., Stumpf, R. (1986). Graphical Exploratory Data Analysis. New York, NY: Springer New York. ISBN 9781461293712. OCLC 840279850; 1936-, Inselberg, Alfred, (2009). Parallel coordinates : visual multidimensional geometry and its applications. Dordrecht: Springer. ISBN 9780387686288. OCLC 656399247.

[23] Breunig, Markus M., Kriegel, Hans-Peter , T. Ng, Raymond, Sander, Jörg. LOF: Identifying Density-Based Local Outliers. Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX. Ano: 2000.

[24] Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. Pattern Recognition Letters, 24(9-10):1641–1650. Ano: 2003.

[25] Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): a fast unsupervised anomaly detection algorithm. KI-2012: Poster and Demo Track, pages 59–63. Ano: 2012.

[26] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. Neural computation, 13(7):1443–1471. Ano: 2001.

[27] Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In European Conference on Principles of Data Mining and Knowledge Discovery, 15–27. Springer. Ano 2002.

[28] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, 413–422. IEEE. Ano: 2008.

[29] Rousseeuw, P.J., Van Driessen, K. "A fast algorithm for the minimum covariance determinant estimator" Technometrics 41(3), 212 (1999).

[30] Du, Bowen; Liu, Chuanren; Zhou, Wenjun; Hou, Zhenshan e Xiong, Hui. Catch me if you can detecting pickpocket suspects from large scale transit records. Ano: 2016.

[31] Tomaso Barbariol, Enrico Feltresi, Gian Antonio Susto. Disponível em: https://www-sciencedirect.ez15.periodicos.capes.gov.br.

[32] Campos, Guilherme o.; Zimek, Arthur; Sander, Jorg; Campello, Ricardo J. G. B.; Micenkova, Barbora; Schubert, Erich; Assent, Ira and Houle, Michael E. On the evaluation of unsupervised outlier detection measures datasets and an empirical study. Ano: 2016. Disponível em: https://link-springer-com.ez15.periodicos.capes.gov.br/content/pdf/10.1007/s10618-015-0444-8.pdf.

[33] Schubert, E.; Zimek, A.; Kriegel, H. -P. (2012). "Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection". Data Mining and Knowledge Discovery. 28: 190–237. doi:10.1007/s10618-012-0300-z.

[34] Russell, Stuart; Norvig, Peter (2003). Artificial Intelligence: A Modern Approach (em inglês) 2 ed. [S.l.]: Prentice Hall. ISBN 978-0137903955.

[35] Scikit-Learn. Disponível em: https://scikit-learn.org/stable/modules/outlierdetection. Acesso em: 20/06/2020.

[36] Instrução Normativa 01/2001 da Controladoria-Geral da União – CGU. Disponível em: https://www.legisweb.com.br/legislacao/?id=75181. Acesso em: 10/06/2020.

[37] Normas Brasileiras de Contabilidade NBC - TA: Objetivos Gerais do Auditor Independente e a Condução da Auditoria em Conformidade com Normas de Auditoria. Disponível em: https://cfc.org.br/tecnica/normas-brasileiras-de-contabilidade/nbc-ta-de-auditoria-independente/. Acesso em: 10/06/2020.

[38] C.G Reddick, A. T. Chatfield, G. Purón, Online Budget Transparency Innovation in Government: A Case Study of the U.S. State Governments, Cid. 2017. dgo Proceedings Paper. In Proceedings of the 17th Annual International Digital Government Research Conference, New York City, NY, June 2017.

[39] OECD. 2014. Recommendation of the Council on Digital Government Strategies. Disponível em http://www.oecd.org/gov/digital-government/Recommendation-digital-government-strategies.pdf. Acesso em: 11/06/2020.

[40] K. Chitra and B. Subashini, "Data Mining Techniques and its Applications in Banking Sector", Volume 3, Issue 8, August 2013.

[41] A. McCarren, S. McCarthy, C. O'Sullivan and M. Roantree, "Anomaly detection in agri warehouse construction", Proceedings of the ACSW, pp. 1-17. Ano: 2017.

[42] Hu, Xue Li; Zhang, Lian Cheng e Wang, Zhen Xing. An adaptive smartphone anomaly detection model based on data mining. Ano: 2018. Disponível em: https://jwcn-eurasipjournals.springeropen.com/track/pdf/10.1186/s13638-018-1158-6.

[43] Zanero, Stefano e Savaresi, Sergio M. Unsupervised learning techniques for an intrusion detection system. Ano: 2004 https://doi-org.ez15.periodicos.capes.gov.br/10.1145/967900.967988.

[44] Salehi, Mahsa and Rashidi, lida. A Survey on Anomaly Detection in Evolving data with application to forest fire risk prediction. Ano: 2018. Disponível em: https://doi-org.ez15.periodicos.capes.gov.br/10.1145/3229329.3229332.

[45] Bao, Liang; Wu, Shanshan; Chen, Weizhao; Zhu, Zisheng and Yi. Fan trajectory outlier detection based on partition and detection framework. Ano: 2017. Disponível em: https://ieeexplo re-ieee-org.ez15.periodicos.capes.gov.br/stamp/stamp.jsp?.

[46] Fu, D. and He, S. New combination algorithms in commercial area data mining and clustering. Ano: 2016. Disponível em: https://ieeexplore-ieee-org.ez15.periodicos.capes.gov.br/stamp/stamp.jsp?arnumber=7509786.

[47] Shan, Yin; Murray, D. Wayne and Sutinen, Alison. Discovering inappropriate billings with local density based outlier detection method. Ano: 2009. Disponível em: https://dl.acm. org/doi/10.5555/2449360.2449380.

[48] Devendra Kumar Luna, Girish Keshav Palshikar, Manoj Apte e Arnab Bhattacharya. Finding Shell Company Accountsusing Anomaly Detection. 2018. https://doi-org.ez15.periodicos.capes.gov.br/10.1145/3152494.3152519.

[49] Ning, Jin; Chen, Leiting and Chen, Junwei. Relative density based outlier detection algorithm. Ano: 2018. Disponível em: https://doi-org.ez15. periodicos.capes.gov.br/10.1145/3297156.3297236.

[50] Na, Gyoung S.; Kim, Donghyun and Yu, Hwanjo. Dilof effective and memory efficient local outlier detection in data streams. Ano: 2018. Disponível em: https://doi-org.ez15.periodicos.capes.gov.br.

[51] Fan Gaofeng; Chen Hongmei; Ouyang Zhiping and Wang Lizhen. Density based top k outlier detection on uncertain objects. Ano: 2011. Disponível em: https://ieeexplore-ieee-org.ez15.periodicos.capes.gov.br.

[52] Jianhua Gao, Weixing Ji, Lulu Zhang, Anmin Li, Yizhuo Wang, Zongyu Zhang. Cube-based incremental outlier detection for streaming computing.

[53] Ahmad, U.; Asim, H.; Hassan, M. T. and Naseer, S. Analysis of Classification Techniques for Intrusion Detection. Disponível em: https://ieeexplore-ieee-org.ez15.periodicos.capes.gov.br/stamp/stamp.jsp?arnumber=8966675.

[54] Jia Guo, Guannan Liu, Yuan Zuo, Junjie Wu. An Anomaly Detection Framework Based on Autoencoder and Nearest Neighbor. Ano: 2018. Disponível em: https://ieeexplore-ieee-org.ez15.periodicos.capes.gov.br.

[55] Foorthuis, R. secoda segmentation and combination based detection of anomalies. Ano: 2017. Disponível em: https://ieeexplore-ieee-org.ez15.periodicos.capes.gov.br.

[56] Kriegel, Hans-Peter; Schubert, Matthias and Zimek, Arthur. Angle-based outlier detection in high dimensional data, 2008, https://doi-org.ez15.periodicos.capes.gov.br/10.1145/1401890.1401946.

[57] Wang, Juite and Chen, Yi-Jing. Technological opportunity analysis for the telehealth industry. Ano: 2017. Disponível em: https://ieeexplore-ieee-org.ez15.periodicos.capes.gov.br.

[58] Liu, Wenting. Outlier detection based on variance of angle in high dimensional data. Ano: 2015. Disponível em: https://www-spiedigitallibrary.ez15.periodicos.capes.gov.br/conference-proceedings-of-spie/9794/979407/Outlier-detection-based-on-variance-of-angle-in-high-dimensional.

[59] Tian Wang , Haoxiong Ke, Xi Zheng , Kun Wang , Arun Kumar Sangaiah , Member, IEEE, and Anfeng Liu. Big Data Cleaning Based on Mobile Edge Computing in Industrial Sensor-Cloud.

[60] Tavares, Gabriel Marques; da Costa, Victor G. Turrisi; Martins, Vinicius Eiji; Ceravolo, Paolo and Barbon, Sylvio. Anomaly detection in business process based on data stream mining. Ano: 2018. Disponível em: https://doi-org.ez15.periodicos.capes.gov.br.

[61] Cleophas, Ton j. Machine learning in therapeutic research the hard work of outlier detection in large data. Ano: 2016. Disponível em: https://ovidsp.dc2ovid.ez15.periodicos.capes.gov.br.

[62] Soleymani, Mohammad Haddad; Yaseri, Mehdi; Farzadfar, Farshad; Mohammadpour, Adel; Sharifi, Farshad and Kabir, Mohammad Javad. Detecting medical prescriptions suspected of fraud using an unsupervised data mining algorithm. Ano: 2018. Disponível em: https://www-ncbi-nlm-nih.ez15.periodicos.capes.gov.br.

[63] Tang, Mingjian; Mendis, B. Sumudu. U.; Murray, D. Wayne; Hu, Yingsong and Sutinen, Alison. Unsupervised fraud detection in medicare australia. Ano: 2011. Disponível em: https://dl.acm.org/doi/10.5555/2483628.2483641.

[64] De Roux, Daniel , Boris Pérez, Moreno, Andrés , Maria del Pilar Villamil and César Figueroa. tax fraud detection for under reporting declarations using an unsupervised machine learning approach. Ano: 2018. Disponível em: https://doi-org.ez15.periodicos.capes.gov.br.

[65] INMETRO. Metodologia Consumo Veicular. Disponível em: http://www.inmetro.gov.br/consumidor/pbe/Metodologia_Consumo_Veicular.pdf. Acesso em: 20 de junho de 2020.

ANEXO A - Protocolo de Revisão da Literatura

Conforme discutido no Capítulo 3, o levantamento da literatura foi realizado tendo por amparo um protocolo de revisão. Nesse protocolo, procurou-se obter os trabalhos mais relevantes sobre a área, como aqueles que descrevem, propõem e aplicam a detecção de anomalias aos mais diversos campos de atuação.

Com esse objetivo, o processo de busca por trabalhos relacionados pautou-se pelas seguintes questões de pesquisa:

QP1: Quais estratégias para detecção de outliers não-supervisionada já foram propostas na literatura?

QP2: Em quais contextos elas vêm sendo utilizadas?

QP3: Essas técnicas podem auxiliar no processo de tomada de decisão?

QP4: Quais as contribuições trazidas pelo uso dessas técnicas nos setores alvos?

De forma a executar a pesquisa pelos trabalhos relacionados na área, foi feita uma exploração automática utilizando-se os principais engenhos de busca de publicações de artigos científicos e periódicos. Nesse protocolo foram utilizados os seguintes engenhos de busca:

ACM Digital Library (http://dl.acm.org)

IEEE Xplore (http://ieeexplore.ieee.org)

ScienceDirect (http://sciencedirect.com)

Web of Science (http://webofknowledge.com)

Com o objetivo de realizar a pesquisa nos portais e engenhos de busca selecionados, foi necessário a definição das palavras-chave que seriam empregadas nessas pesquisas.

As palavras-chave são usadas pelos engenhos de busca para retornar, como resultado das consultas, os trabalhos que contenham esses termos, de acordo com as opções definidas no motor de busca.

Nesses motores de busca, os filtros foram considerados de maneira a retornar os trabalhos que contivessem os termos no título, no resumo ou nas palavras-chave. Na Tabela 22, é possível visualizar as palavras-chave, os sinônimos e os termos relacionados que foram aplicados nesse protocolo:

Tabela 22: Palavras-chave e Termos Relacionados

Palavras-chave	Sinônimos e/ou Termos relacionados
Data Mining	database mining, DM
Outlier Detection	anomaly detection
Unsupervised	
Angle-based	
Density-based	

Os termos e palavras-chave da Tabela 22 serviram de base para a definição de uma "*string* de busca" para a pesquisa. A *string*, por sua vez, é um texto utilizado para consultas avançadas nos engenhos de busca, de modo a se retornar os trabalhos relacionados.

Após organizar as palavras-chave, elaborou-se a seguinte *string* de busca para este protocolo:

("database mining"OR "data mining"OR "dm") AND ("outlier detection"OR "anomaly detection") AND ("unsupervised"OR "angle-based"OR "density-based")

Para cada um dos engenhos de busca selecionados, foi realizada a adaptação da *string*, de forma a conformá-la à sintaxe de pesquisa empregada nesses motores. Na Tabela 23, são apresentados os engenhos e a "*string* de busca" adaptada:

Tabela 23: Engenhos de Busca e *String* de Busca Adaptados

Engenho de Busca	String de Busca Adaptada
ACM Digital Library	Title:(("database mining"OR "data mining"OR "dm") AND ("outlier detection"OR "anomaly detection") AND ("unsupervised"OR "angle-based"OR "density-based")) OR Abstract:(("database mining"OR "data mining"OR "dm") AND ("outlier detection"OR "anomaly detection") AND ("unsupervised"OR "angle-based"OR "density-based")) OR Keyword:(("database mining"OR "data mining"OR "dm") AND ("outlier detection"OR "anomaly detection") AND ("unsupervised"OR "angle-based"OR "density-based"))

Engenho de Busca	String de Busca Adaptada
IEEE Xplore	("Document Title":"database mining"OR "Document Title":"data mining"OR "Document Title":"dm") AND ("Document Title":"outlier detection"OR "Document Title":"anomaly detection") AND ("Document Title":"unsupervised"OR "Document Title":"angle-based"OR "Document Title":"density-based") OR ("Abstract":"database mining"OR "Abstract":"data mining"OR "Abstract":"dm") AND ("Abstract":"outlier detection"OR "Abstract":"anomaly detection") AND ("Abstract":"unsupervised"OR "Abstract":"angle-based"OR "Abstract":"density-based") OR ("Author Keywords":"database mining"OR "Author Keywords":"data mining"OR "Author Keywords":"dm") AND ("Author Keywords":"outlier detection"OR "Author Keywords":"anomaly detection") AND ("Author Keywords":"unsupervised"OR "Author Keywords":"angle-based"OR "Author Keywords":"density-based")
ScienceDirect	title, abstract, keywords: ("database mining"OR "data mining"OR "dm") AND ("outlier detection"OR "anomaly detection") AND ("unsupervised"OR "angle-based"OR "density-based"

Engenho de Busca	String de Busca Adaptada
Web of Science	TS=((”database mining”OR ”data mining”OR ”dm”) AND (”outlier detection”OR ”anomaly detection”) AND (”unsupervised”OR ”angle-based”OR ”density-based”)) OR AB=((”database mining”OR ”data mining”OR ”dm”) AND (”outlier detection”OR ”anomaly detection”) AND (”unsupervised”OR ”angle-based”OR ”density-based”)) OR AK=((”database mining”OR ”data mining”OR ”dm”) AND (”outlier detection”OR ”anomaly detection”) AND (”unsupervised”OR ”angle-based”OR ”density-based”))

Após a pesquisa pelos trabalhos nos engenhos de busca, é necessário realizar um processo de seleção dos documentos relevantes para a revisão. Os documentos retornados foram filtrados com base nos seguintes critérios de inclusão (*CI*) e exclusão (*CE*):

- ***CI1****: Serão incluídos artigos completos relacionados ao tema, nos quais tragam no título ou resumo informações suficientes sobre o trabalho desenvolvido;*

- ***CE1****: Serão desconsiderados todos os documentos que não sejam artigos científicos;*

- ***CE2****: Serão desconsiderados todos os trabalhos que não estiverem em vernáculo ou em inglês;*

- ***CE3****: Caso existam artigos repetidos, serão mantidos, apenas, os mais recentes;*

- ***CE4****: Serão desconsiderados artigos em que o acesso completo não esteja disponível, os quais apenas o resumo esteja disponível;*

- ***CE5****: Serão desconsiderados artigos que foram retratados ou “despublicados”, se houver; e*

- ***CE6****: Serão desconsiderados artigos que não tenham aplicação em detecção de outliers/anomalias não supervisionada, após a leitura do título, resumo e palavras-chave.*

Após uso dos critérios de seleção sobre os artigos, estes foram analisados e, as aplicações e proposições mais relevantes, integraram o capítulo de trabalhos relacionados.