TESOURONACIONAL

REVISTA

CADERNOS DE FINANÇAS PÚBLICAS

2022

EDIÇÃO ESPECIAL

SUBSÍDIO ÀS FISCALIZAÇÕES PÚBLICAS: Identificação dos Municípios com gastos discrepantes na Educação Básica

Renata Guanaes Machado
Auditor Federal de Finanças e Controle - Controladoria-Geral da União

RESUMO

Os gastos públicos devem ser constantemente monitorados pelos órgãos governamentais. Neste contexto, torna-se primordial a aplicação de tecnologia para a produção de informações estratégicas que apoiem ações de combate à corrupção e à má gestão dos recursos públicos. Com a disponibilização do Sistema de Informações sobre Orçamentos Públicos em Educação (SIOPE), o presente trabalho emprega técnicas de mineração de dados para a detecção de despesas atípicas no Ensino Fundamental, realizadas pelos municípios em 2018 – que podem constituir eventos ocasionais (como obras em escolas) ou representar indícios de irregularidades. Aplicou-se clusterização de municípios e algoritmos de detecção de anomalias em um grupo de municípios semelhantes. Os resultados alcançados (se o município é anômalo e seu grau de anomalia) podem ser agregados ao planejamento das ações de controle e, ainda, subsidiar a adoção de providências cabíveis por parte de demais instâncias, como o Ministério da Educação e conselhos de controle social.

Palavras-chave: Clusterização de municípios. Detecção de anomalias. Despesas públicas. Educação Básica.

Classificação JEL: C38. H52. C65.

SUMÁRIO

RESUMO 2

1 INTRODUÇÃO 5
1.1 MOTIVAÇÃO 5
1.2 PROBLEMA E JUSTIFICATIVA 5
1.3 OBJETIVOS GERAIS E ESPECÍFICOS 6

2 METODOLOGIA UTILIZADA 6

3 FASE DE ENTENDIMENTO DO NEGÓCIO 7
3.1 O SIOPE PARA MONITORAMENTO DOS GASTOS NA EDUCAÇÃO 8
3.2 DESPESAS COM O FUNDEB 9
3.3 O PAPEL DA CONTROLADORIA-GERAL DA UNIÃO 11
3.4 IDENTIFICAÇÃO DE PROBLEMAS OU DESAFIOS 12
3.5 OBJETIVOS DE NEGÓCIO E DA MINERAÇÃO DE DADOS 13

4 FASE DE ENTENDIMENTO E PREPARAÇÃO DOS DADOS 13
4.1 ENTENDIMENTO DOS DADOS DO SIOPE MUNICIPAL 14
4.1.1. Programas Vinculados 14
4.1.2. Grupos de Despesas 14
4.1.3. Subfunções da Educação estruturadas em Pastas e SubPastas 15
4.1.4. Contas Contábeis (Natureza e Elemento da Despesa) 17
4.2 PREPARAÇÃO DOS DADOS – DATAFRAME DE DESPESAS 17
4.2.1. Coleta e limpeza dos dados 17
4.2.2. Inclusão de colunas: “Classificação” e “Tipo de Gasto” 18
4.2.3. Inclusão de dados adicionais: dados econômicos e demográficos 19
4.2.4. Filtro dos dados para contexto ao Ensino Fundamental 20
4.2.5. Resumo do dataframe de despesas 21
4.3 PREPARAÇÃO DOS DADOS – DATAFRAME DE MUNICÍPIOS 23
4.3.1. Pivoteamento dos dados 24
4.3.2. Consolidação de Contas Contábeis 24
4.3.3. Resumo do dataframe de municípios 25

5 FASE DE MODELAGEM 26
5.1 ANÁLISE EXPLORATÓRIA DE DADOS 26
5.2 CLUSTERIZAÇÃO DE MUNICÍPIOS SEMELHANTES 27
5.2.1. Objetivos da clusterização de municípios 27
5.2.2. Decisão sobre o escalonamento dos dados 27

5.2.3. Clusterização k-Means 30
5.2.4. Clusterização DBSCAN 33
5.2.5. Clusterização Hierárquica – Agglomerative Clustering 34
5.2.6. Validação dos algoritmos de clusterização 37
5.3 DETECÇÃO DE ANOMALIAS 39
5.3.1. Delimitação das estratégias para detecção de outliers 39
5.3.2. A biblioteca Python Outlier Detection (PyOD) 40
5.3.3. Escolha de Cluster para ser submetido aos algoritmos 42
5.3.4. Resultados da detecção de anomalias 42
5.3.5. Validação dos modelos de detecção de anomalias 43

6 FASE DE AVALIAÇÃO E IMPLANTAÇÃO 49

7 CONCLUSÃO 51

REFERÊNCIAS 53

1. INTRODUÇÃO

1.1. MOTIVAÇÃO

A corrupção pública custa, todos os anos, muitos milhões aos governos em todo o mundo, sendo o seu combate um desafio para o setor público e para a sociedade (THESING, 2019). Para alguns autores, a má gestão do dinheiro público pode acarretar perdas ainda piores do que os atos de corrupção (ANGELICO, 2012; AGUIAR, 2019). Nesse contexto, o uso conjunto de tecnologias de ponta, de métodos de inteligência e de ferramentas de análise de dados, nos últimos anos, provou-se ser um poderoso e efetivo aliado não só para enfrentar a corrupção, mas também para minimizar ou evitar o desperdício dos recursos públicos.

Diversos órgãos de controle têm, entre seus principais objetivos, o monitoramento de gastos públicos e a produção de informações estratégicas como apoio à tomada de decisão no controle interno. Tal monitoramento pode se dar em variados contextos, como: compras e licitações públicas; programas de governo; benefícios sociais; convênios e transferências; entre outros. Entre as diversas ferramentas e métodos que apoiam o controle interno, pode-se mencionar: processamento analítico de dados, técnicas estatísticas, big data, business intelligence, inteligência artificial, mineração de dados e aprendizagem de máquina (machine learning). Todo esse arcabouço ferramental permite a descoberta de conhecimento de alto valor agregado a partir de grandes bases de dados governamentais, a potencialização da capacidade de análise e a modernização do controle, seja interno ou externo.

Nesse momento, torna-se fundamental a definição de alguns termos. A Ciência de Dados é um conjunto de princípios fundamentais que norteiam a extração de conhecimento a partir de dados; Mineração de Dados é a extração de conhecimento propriamente dita, por meio de tecnologias que incorporam tais princípios (PROVOST e FAWCETT, 2016). Assim, a aprendizagem de máquina é uma dessas tecnologias, na forma de algoritmos de indução (afirmar uma verdade generalizada a partir da observação de alguns elementos).

1.2. PROBLEMA E JUSTIFICATIVA

Este trabalho foi motivado pela necessidade de se realizar uma análise sistemática nos dados do Sistema de Informações sobre Orçamentos Públicos em Educação (SIOPE), sob a gestão do Fundo Nacional de Desenvolvimento da Educação (FNDE). Neste sistema, os estados e municípios registram as receitas e despesas realizadas com a educação pública – as quais são, posteriormente, transmitidas às demais instâncias de controle para fins de validação das informações. Em seguida, o FNDE gera os indicadores de gestão e de conformidade às disposições legais - como, por exemplo, o cumprimento da aplicação de 25% das receitas na Manutenção e Desenvolvimento do Ensino (MDE).

Não obstante, alguns relatórios produzidos por órgãos de controle, como a Controladoria-Geral da União (CGU) e o Tribunal de Contas da União (TCU), apontaram falhas nos relatórios produzidos pelo SIOPE, que não evitaram os desvios de recursos públicos dedicados à educação em diversos municípios. Estes problemas se encontram mais detalhados no item 3.4 do presente trabalho, após a contextualização sobre o SIOPE, sobre o Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (FUNDEB) e sobre as atribuições específicas da CGU enquanto instância de controle.

Diante desses fatos – a disponibilidade dos dados do SIOPE e a existência concreta de desvios dos recursos destinados à educação – surgiu o seguinte problema de pesquisa: a partir dos registros no SIOPE, como identificar possíveis indícios de irregularidades nos gastos públicos dos municípios com a Educação Básica?

Por fim, a justificativa do presente trabalho é a possibilidade de gerar conhecimento de valor estratégico no tema de monitoramento de gastos públicos na Educação, o qual poderá conferir uma maior qualidade de dados ao SIOPE, bem como propiciar subsídios complementares aos trabalhos de auditoria e fiscalização desempenhados pelas instâncias de controle.

1.3. OBJETIVOS GERAIS E ESPECÍFICOS

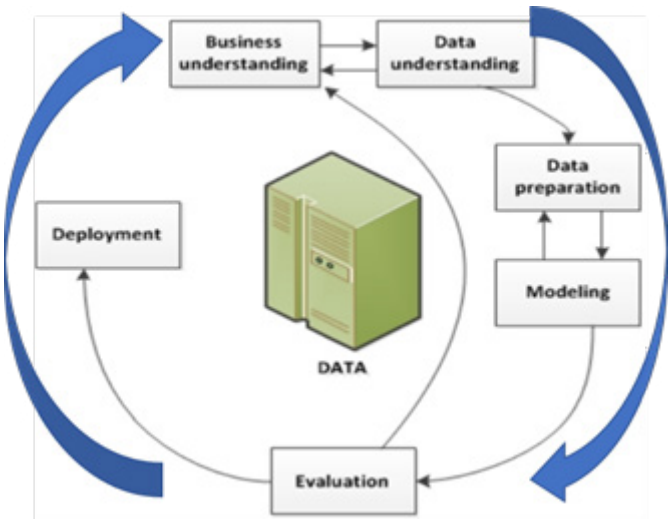
Tendo em vista a oportunidade de explorar os dados do SIOPE, definiu-se o seguinte objetivo geral: "Propor mecanismos para a identificação de discrepâncias nos gastos públicos realizados pelos municípios, especificamente no Ensino Fundamental em 2018, com o intuito de produzir um relatório com informações de valor agregado para as instâncias de controle", o qual proporciona os objetivos específicos abaixo:

- contextualização acerca da legislação sobre as políticas e os investimentos na educação;
- entendimento do sistema SIOPE;
- entendimento sobre a atuação das instâncias de controle;
- leitura de relatórios de auditoria e fiscalização dos órgãos de controle, para entender as falhas do SIOPE e os desvios na aplicação dos recursos federais em Educação Básica;
- realizar análises exploratórias nos dados do SIOPE, a fim de levantar hipóteses iniciais sobre os dados;
- escolher as técnicas de mineração de dados a serem utilizadas para a identificação de discrepâncias nos gastos públicos.

2. METODOLOGIA UTILIZADA

O presente trabalho foi realizado seguindo-se a metodologia de referência Cross-Industry Standard Process for Data Mining (CRISP-DM), comumente utilizada para projetos de mineração de dados (CHAPMAN, 2000). Nesta metodologia, o ciclo de vida de um projeto de mineração de dados, conforme apresentado abaixo, é composto das seguintes fases: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação.

Figura 1 – O ciclo de vida do CRISP-DM



Fonte: IBM (2019).

A fase de entendimento do negócio consiste em identificar o contexto institucional, entender o problema de negócio que a organização espera resolver e levantar as necessidades mais prioritárias. Em seguida, deve-se determinar os objetivos de negócio e seus critérios de sucesso; verificar os recursos disponíveis (pessoal, dados, hardware e software); determinar os objetivos da mineração de dados a partir do objetivo de negócio; e, finalmente, criar o plano de projeto.

A fase de entendimento dos dados envolve a coleta, a verificação da qualidade e a exploração dos dados com estatística descritiva, de forma a realizar descobertas e detectar possíveis correlações.

A fase de preparação dos dados realiza tratamentos sobre os dados, a fim de torná-los adequados para a aplicação dos algoritmos de mineração na próxima fase. Tais tratamentos incluem: limpeza dos dados, substituição de valores omissos, seleção de atributos relevantes, redução de dimensionalidade, criação de atributos derivados, integração de dados externos, entre outros.

A fase de modelagem compreende a seleção e aplicação de técnicas para criar modelos e encontrar padrões ou descobertas de conhecimento – exemplos dessas técnicas são classificação, agrupamento e regressão. Procedimentos de teste devem ser executados (como, por exemplo, matriz de confusão e medidas de acurácia) para verificar a qualidade e validade dos resultados.

A fase de avaliação objetiva analisar se os resultados do modelo atendem aos objetivos de negócio e aos critérios de sucesso definidos anteriormente, bem como se o modelo pode ser implantado ou se é necessário aprimorá-lo através de novas iterações das fases anteriores.

A fase de implantação resume-se em colocar o modelo obtido em efetiva produção, incluindo, quando aplicável, a confecção de relatório final com os resultados alcançados.

3. FASE DE ENTENDIMENTO DO NEGÓCIO

Esta fase consiste em levantar problemas existentes para, em seguida, estabelecer os objetivos de negócio, seus critérios de sucesso e os objetivos da mineração de dados. Desta forma, faz-se necessário uma prévia contextualização sobre o sistema SIOPE; sobre o FUNDEB como a principal fonte de financiamento federal para a Educação Básica; e sobre o papel fiscalizador da CGU sobre os recursos federais repassados a estados e municípios.

3.1. O SIOPE PARA MONITORAMENTO DOS GASTOS NA EDUCAÇÃO

O SIOPE, instituído pela Portaria Ministerial MEC nº. 06, de 20 de junho de 2006, e operacionalizado pelo FNDE, surgiu como resposta a uma demanda do então Ministro da Educação, Cristovam Buarque (2003), que almejava identificar o quanto se investia na educação pública brasileira. Conforme publicado no sítio do FNDE:

> *"O SIOPE é uma ferramenta eletrônica instituída para coleta, processamento, disseminação e acesso público às informações referentes aos orçamentos de educação da União, dos estados, do Distrito Federal e dos municípios, sem prejuízo das atribuições próprias dos Poderes Legislativos e dos Tribunais de Contas." (BRASIL, FNDE, 2019).*

Esse sistema possibilita que entes federativos registrem, bimestralmente, as receitas e despesas realizadas com a educação pública, incluindo a remuneração de profissionais do magistério e as despesas custeadas com recursos de programas federais relacionados, como o Programa Nacional de Alimentação Escolar (PNAE) e Programa Nacional de Apoio ao Transporte do Escolar (PNATE). Embora os dados sejam de natureza declaratória, há uma série de regras de integridade, embutidas no sistema, que checam os dados lançados antes da transmissão ao Conselho de Acompanhamento e Controle Social do FUNDEB (CACS) e ao FNDE. Por exemplo, o sistema informa se há omissão de alguma receita, proveniente de impostos, que se encontra lançada nos sistemas da STN; ou quando há alunos na Educação Infantil, declarados no Censo Escolar (INEP), sem o registro de despesas nesta mesma modalidade de ensino.

Após a validação dos dados, o sistema indica se cada ente federativo atinge os percentuais legalmente estabelecidos – tendo em vista que o artigo 212 da Constituição Federal (BRASIL, 1988) decreta que os estados e municípios devem aplicar o mínimo de 25% da receita de impostos e transferências no MDE. Ainda, a Lei no 9.394/96 (BRASIL, 1996) – Lei de Diretrizes e Bases da Educação Nacional (LDB) – especifica, em seu artigo 70, as despesas elegíveis com recursos do MDE:

> *Art. 70. Considerar-se-ão como de manutenção e desenvolvimento do ensino as despesas [...]:*
>
> *I - remuneração e aperfeiçoamento do pessoal docente e demais profissionais da educação;*
>
> *II - aquisição, manutenção, construção e conservação de instalações e equipamentos necessários ao ensino;*
>
> *III – uso e manutenção de bens e serviços vinculados ao ensino;*
>
> *IV - levantamentos estatísticos, estudos e pesquisas visando precipuamente ao aprimoramento da qualidade e à expansão do ensino;*
>
> *V - realização de atividades-meio necessárias ao funcionamento dos sistemas de ensino;*
>
> *VI - concessão de bolsas de estudo a alunos de escolas públicas e privadas; [...]*
>
> *VIII - aquisição de material didático-escolar e manutenção de programas de transporte escolar. (BRASIL, 1996)*

Há penalidades impostas às unidades que não utilizem o SIOPE. A inserção, atualização e transmissão dos dados aos CACS e ao FNDE são obrigatórias e com prazos definidos, sob pena de bloqueio de repasses dos recursos financeiros (transferências voluntárias e de convênios com o Governo Federal) e de situação irregular no Serviço Auxiliar de Informações para Transferências Voluntárias (CAUC).

Finalmente, o sistema permite a geração periódica de indicadores da gestão educacional, subsidiando a definição, implementação e monitoramento de políticas públicas educacionais por diversos atores, conforme mostra a figura abaixo.

Figura 2 – Rede de Parceiros do SIOPE

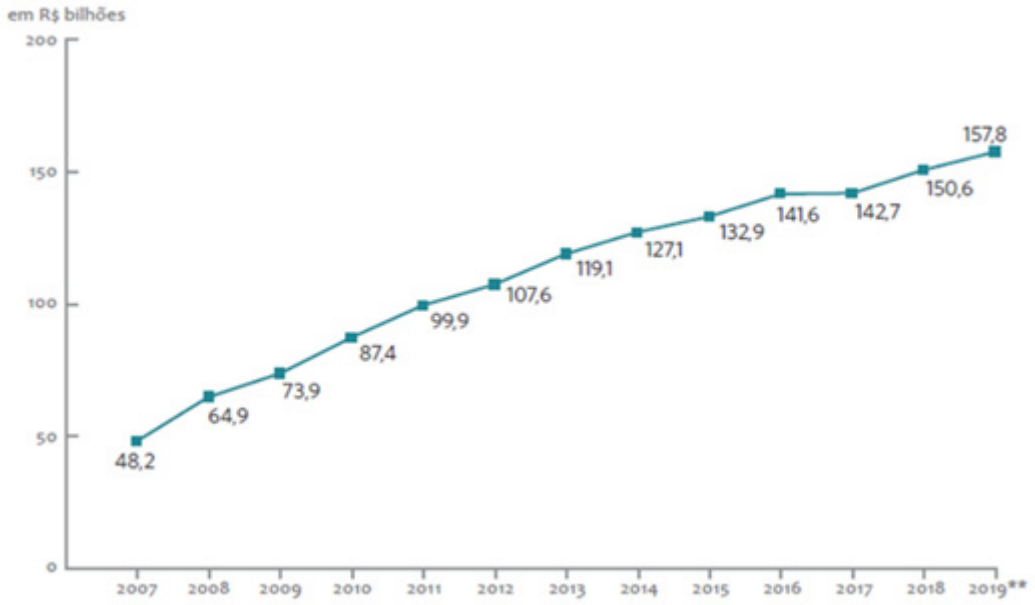
Fonte: BRASIL, MEC (2018).

De todo modo, o SIOPE é uma importante ferramenta para o acompanhamento e a avaliação dos gastos públicos em educação – tanto pelo Ministério da Educação/FNDE e órgãos de controle, como pelos gestores educacionais e CACS. Ademais, as informações registradas, em conjunto com indicadores e relatórios consolidados, são disponibilizadas pelo FNDE em página na internet para acesso pelos cidadãos, assegurando a transparência e o controle social dos recursos públicos destinados à educação.

3.2. DESPESAS COM O FUNDEB

Há no SIOPE o registro de uma importante fonte de recursos: trata-se do FUNDEB. Esse fundo foi criado em 2006 por meio da Emenda Constitucional nº. 53/2006, em substituição ao Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério (FUNDEF), e encontra-se regulamentado pela Lei no 11.494/2007 (BRASIL, 2007). Todos os recursos do FUNDEB devem ser aplicados, exclusivamente, na Educação Básica (particularmente, na valorização do magistério) – esse montante atingiu, conforme a figura abaixo, o valor de R$ 150 bilhões em 2018. Sua vigência encerrou-se em 2020, mas um projeto de lei foi criado para torná-lo permanente.

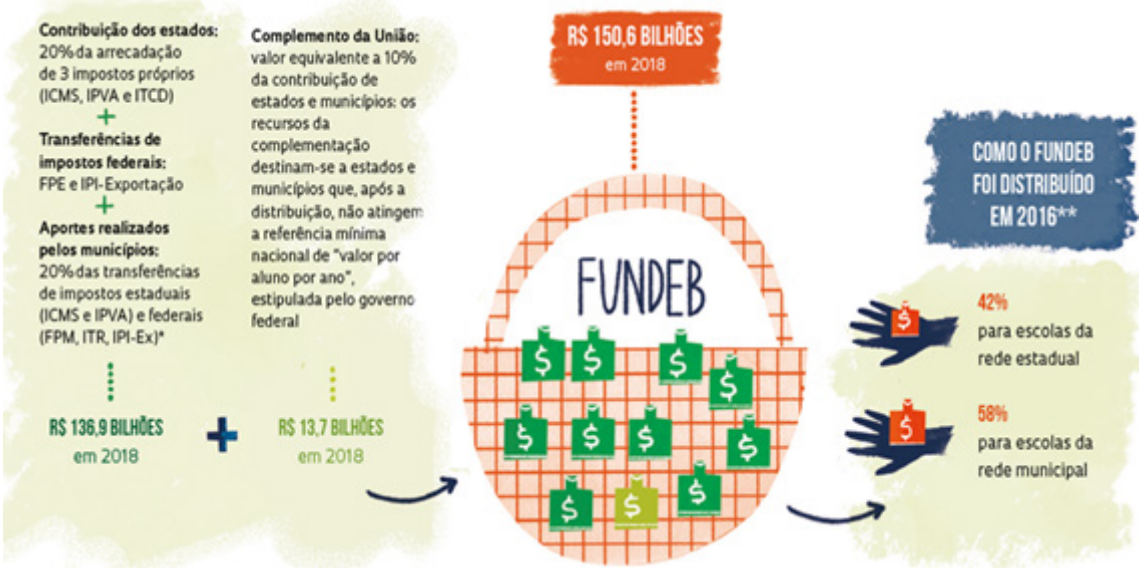
Figura 3 – Evolução dos Investimentos do FUNDEB



Fonte: JEDUCA (2019).

O FUNDEB é um fundo especial e redistributivo, de natureza contábil e de âmbito estadual. Conforme a figura seguinte, é formado por 20% dos recursos provenientes dos impostos e transferências dos estados e municípios, e por parcela financeira de recursos federais (a título de complementação). Posteriormente, valores são recalculados (com base no custo por aluno, calculado pelo FNDE, e o número de alunos matriculados, levantado pelo INEP) e redistribuídos aos entes, de forma proporcional e igualitária.

Figura 4 – Composição e redistribuição do FUNDEB



Fonte: QUEIROZ (2020).

Essa redistribuição procura garantir o valor mínimo nacional por aluno/ano a cada ente federativo, contribuindo para diminuir as desigualdades de recursos entre as redes de ensino e universalizar a oferta de ensino a todos os cidadãos.

3.3. O PAPEL DA CONTROLADORIA-GERAL DA UNIÃO

Na esfera federal, a CGU é o órgão central do Sistema de Controle Interno (SCI), do Sistema de Correição e do Sistema de Ouvidoria do Poder Executivo (BRASIL, CGU, 2020), e a ela compete desenvolver funções de defesa do patrimônio público, de controle interno, de auditoria pública, de correição e de ouvidoria, além de ações para a promoção da transparência; para o incentivo à integridade pública; e para a prevenção e combate à corrupção.

O Regimento Interno da CGU (BRASIL, CGU, 2020b) dispõe sobre a sua estrutura organizacional, cabendo à Secretaria Federal de Controle Interno (SFC) exercer as competências de órgão central do SCI do Poder Executivo Federal e, ainda:

> *XIV - fiscalizar e avaliar a execução dos programas de governo [...];*
>
> *XVI - realizar auditorias sobre a gestão dos recursos públicos federais [...];*
>
> *XVIII - apurar atos ou fatos ilegais ou irregulares praticados por agentes públicos ou privados na utilização de recursos públicos federais [...].*

Entre as atividades desempenhadas no controle interno, está a Avaliação dos Programas de Governo (AEPG) e a Fiscalização em Entes Federativos (FEF), que verificam a regularidade da aplicação de recursos públicos federais repassados aos estados e municípios. Estes instrumentos passam por constantes melhorias, como a criação da Matriz de Vulnerabilidade em 2015, uma ferramenta de análise de risco composta por doze indicadores – esta gera um ranking de municípios com maiores fragilidades na aplicação dos recursos, e do qual são selecionados aqueles a serem fiscalizados.

Com relação aos recursos federais provenientes do FUNDEB para os municípios, a CGU realiza ações de fiscalização, em campo, nos municípios escolhidos pela Matriz de Vulnerabilidade ou por sorteios públicos. Essas ações resultam em Relatórios de Avaliação, com análises detalhadas sobre atos de execução operacional e financeira do FUNDEB - como, por exemplo, sobre a adequação da folha de pagamentos dos profissionais do magistério ou sobre a correta execução de processos de contratação de fornecimento de bens e serviços destinados às escolas.

Alguns relatórios – em especial, os resultantes da FEF em dois municípios do estado do Piauí, produzidos em 2016 (BRASIL, CGU, 2019a e 2019b); e em um município do estado do Maranhão, produzido em 2019 (BRASIL, CGU, 2020a) – apontaram para irregularidades na aplicação dos recursos FUNDEB, repassados às prefeituras municipais, entre elas:

- *Realização de despesas inelegíveis com recursos do FUNDEB, ou despesas incompatíveis com o objetivo do FUNDEB;*
- *Pagamento de profissionais sem comprovada atuação na docência;*
- *Realização de gastos sem apresentação de documentos comprobatórios;*
- *Pagamentos efetuados por serviços não recebidos;*
- *Superfaturamento por quantidade na aquisição de bens, móveis, equipamentos e materiais permanentes com recursos do FUNDEB;*
- *Superfaturamento por quantidade na execução de obras de construção, reforma e ampliação de escolas com recursos do FUNDEB.*

Pode-se afirmar que metodologias e instrumentos para o controle interno auxiliam no processo de seleção dos municípios e possibilitam, quando em campo, a detecção das irregularidades

existentes. Não obstante, com a disponibilidade das informações de receitas e despesas no SIOPE (no período de 2005 a 2019) e com o uso concomitante de cruzamento de bases de dados, de técnicas estatísticas e de algoritmos de mineração de dados, é possível agregar valor aos métodos existentes – como a Matriz de Vulnerabilidade – no sentido de indicar perspectivas ainda não exploradas.

3.4. IDENTIFICAÇÃO DE PROBLEMAS OU DESAFIOS

O artigo 212 da Constituição Federal (BRASIL, 1988) e a LDB (BRASIL, 1996) fomentam a descentralização da Educação Básica e a municipalização do ensino fundamental, concedendo autonomia para que estados e municípios organizem e mantenham as instituições oficiais de seus sistemas de ensino. Entretanto, essa descentralização administrativa, em um país de grande extensão territorial, gera um sistema educacional de proporções colossais – mais de 180 mil escolas de Educação Básica e 48 milhões de matrículas, conforme o Censo Escolar de 2019 (BRASIL, INEP, 2020). Ainda, o substancial volume de recursos financeiros (em bilhões de reais), investido pelo Governo Federal, resulta em um modelo de financiamento de difícil monitoramento e, consequentemente, em uma maior complexidade dos mecanismos de controle interno.

Além da CGU como órgão de controle interno, há ainda a atuação do FNDE, que realiza o monitoramento das aplicações dos recursos do FUNDEB em estados e municípios por meio do SIOPE (BRASIL, MEC, 2018); e a atuação dos CACS, que são incumbidos de desempenhar o acompanhamento e controle social sobre a distribuição, o planejamento e a aplicação dos recursos do FUNDEB (BRASIL, CGU, 2019d), bem como de analisar as contas informadas no SIOPE pelos entes federativos. Esclarece-se que "monitorar" não tem o mesmo sentido de "fiscalizar", conforme nota emitida pelo MEC:

> *"A fiscalização e o controle quanto à aplicação dos recursos do Fundeb [...] competem aos tribunais de contas locais e ao Ministério Público dos estados, resguardada a competência do Ministério Público Federal, para os estados que recebem o aporte federal de recursos. Ao MEC, por meio do FNDE, compete o monitoramento quanto à aplicação, que é feito por meio do SIOPE [...]. Porém, o Siope é um sistema de monitoramento cuja base é declaratória. Significa que a fiscalização e o controle só são exercidos diretamente para fins de realização de auditoria, inspeção e eventual punição, pelos tribunais de contas locais e pelo Ministério Público." (ARCOVERDE e TOLEDO, 2019)*

Todavia, o Relatório de Auditoria Anual de Contas (AAC), sobre a prestação de contas do FNDE como unidade auditada (BRASIL, CGU, 2019c), apontou para as seguintes constatações sobre o SIOPE como sistema de controle da aplicação dos recursos do FUNDEB:

a. a atuação das diversas instâncias de controle adicionais (MEC, FNDE, CACS) não são suficientes para evitar perdas, desvios e/ou fraudes nas aplicações dos recursos do FUNDEB, detectadas e documentadas em diversos relatórios de auditoria elaborados pela CGU (BRASIL, CGU, 2019a, 2019b, 2020a);

b. a apresentação dos relatórios gerados pelo SIOPE em seu sítio: *"não permitem perceber a situação da educação nos entes, exigindo trabalho especializado para tratamento dos dados, compreensão das informações e criação de parâmetros que dêem significado aos números. É a comparação dos gastos e dos resultados entre entes federativos semelhantes que permite qualificar a atuação do controle" (BRASIL, CGU, 2019c).*

c. os CACS não dispõem das informações gerenciais adequadas para a sua atuação como instância de controle.

Em levantamento realizado pela CGU/CGEBC sobre os Acórdãos da Corte de Contas, verificou-se que o TCU, no Acordão nº 618/2014 – Plenário (BRASIL, TCU, 2020), avaliou o SIOPE quanto à confiabilidade das informações sobre os gastos com MDE e listou as seguintes constatações, dentre outras:

a. impossibilidade de atestar que as informações prestadas pelos entes federados no SIOPE refletem os gastos em MDE, e que a despesa de pessoal é fidedigna;

b. validação de informações antes da transmissão dos dados ocorre somente em itens de receitas – itens relativos às despesas não possuem o mesmo nível de controle.

Diante dos fatos mencionados, os principais problemas identificados são: difícil monitoramento dos investimentos em educação; alta complexidade dos mecanismos de controle interno; relatórios gerenciais do SIOPE inadequados para uso pelos CACS; e ausência de validação prévia de informações no SIOPE para itens de despesa.

3.5. OBJETIVOS DE NEGÓCIO E DA MINERAÇÃO DE DADOS

Dado todo o contexto até o momento, deu-se prioridade ao problema: "Relatórios gerenciais do SIOPE inadequados para uso pelos CACS" e formulou-se o seguinte objetivo de negócio: realização de técnicas de análise exploratória e de mineração de dados nas despesas vinculadas à Educação Básica, com o intuito de produzir informações de valor agregado para as instâncias de controle. Como consequência, estas informações poderão fornecer subsídios complementares à Matriz de Vulnerabilidade para os trabalhos da CGU, bem como poderá melhorar o acompanhamento da aplicação dos recursos pelo FNDE e CACS.

Neste trabalho, foi estabelecido o escopo inicial de indicar os municípios com discrepâncias nos seus gastos com o Ensino Fundamental em 2018 – ou seja, com despesas inconsistentes em comparação a um determinado padrão, podendo estas serem indícios de falhas ou de irregularidades nos gastos públicos. Como critério de sucesso, espera-se obter ao menos 1% de entes federativos com discrepâncias nos seus gastos educacionais declarados, e que alguns entes sejam identificados como anômalos em todos os algoritmos de detecção a serem utilizados.

Com relação à mineração de dados, foram estabelecidos os seguintes objetivos:

a. Realizar análises exploratórias nos dados para gerar novos conhecimentos;

b. Utilizar ao menos três técnicas de clusterização para a criação de grupos de municípios semelhantes;

c. Em um cluster de municípios semelhantes, utilizar ao menos cinco algoritmos de detecção de anomalias nas despesas desses municípios.

4. FASE DE ENTENDIMENTO E PREPARAÇÃO DOS DADOS

Essas fases envolvem a coleta, exploração e tratamentos sobre os dados; e ocorreram simultaneamente ao longo da execução deste trabalho. Ao final do capítulo, tem-se os seguintes dataframes: dataframe de despesas (dados tabulares) para uso por análises exploratórias, no qual cada registro

representa dados de uma despesa; e dataframe de municípios para uso pelas técnicas de clusterização, no qual cada registro representa um município com o seu vetor de características. Ambos os dataframes contêm apenas as despesas municipais.

4.1. ENTENDIMENTO DOS DADOS DO SIOPE MUNICIPAL

São descritos os componentes existentes no SIOPE para familiarização com o domínio da aplicação: grupos de despesas, programas, subfunções da educação e contas contábeis.

4.1.1. Programas Vinculados

Trata-se de recursos do FNDE repassados aos entes federativos e, naturalmente, este campo é preenchido somente para os registros do grupo de Despesas Vinculadas. No SIOPE Municipal de 208 para o Ensino Fundamental, estão registrados os seguintes programas:

- PNAE, PNATE, PDDE;
- Vinculadas a Contribuição Social do Salário-Educação;
- Outras Transferências de Recursos do FNDE;
- Transferências de Convênios – Educação;
- Outros Recursos Destinados à Educação;
- Ação Judicial FUNDEF – Precatórios.

4.1.2. Grupos de Despesas

As despesas são classificadas em grupos, descritos na tabela abaixo.

Tabela 1 - Descrição de cada Grupo de Despesa

Despesas próprias custeadas com impostos e transferências	São despesas vinculadas aos recursos próprios de cada ente, ou seja, tem como fonte o Tesouro do Estado, do Distrito Federal ou do Município.
	Não poderão ser consideradas as despesas com convênios, recursos transferidos pelo FNDE, royalties de petróleo e indenizações.
	Algumas despesas devem cumprir o percentual mínimo de 25% no MDE.
Despesas efetuadas com os recursos do FUNDEB	Fundo especial formado por parcela de recursos federais e por recursos provenientes dos impostos e transferências dos entes federativos.
	Devem ser empregados exclusivamente em ações de manutenção e desenvolvimento da Educação Básica pública. No caso de municípios, somente são disponibilizadas as modalidades de ensino Educação Infantil (creche e pré-escola) e Ensino Fundamental.
	Mínimo de 60% destinados à remuneração dos profissionais do magistério.
	Máximo de 40% destinados nas demais ações de MDE.
	Dos 25% das receitas de impostos e transferências destinadas ao MDE, 20% de algumas comporão o FUNDEB.

Despesas custeadas com recursos vinculados	Despesas custeadas com recursos de programas federais de educação como: PNAE, PNATE, PDDE, entre outros.
	Recursos legalmente vinculados a uma finalidade específica.
	Recursos recebidos provenientes de transferências constitucionais, co-mo: Salário-Educação, alimentação (merenda), transporte, Programa Dinheiro Direto na Escola, de transferências voluntárias (convênios) firmados com as unidades federativas, recursos provenientes de royalti-es do petróleo, etc.
	Não contam como MDE.

Fonte: Elaborada pelo autor (2020).

4.1.3. Subfunções da Educação estruturadas em Pastas e SubPastas

A inclusão de despesas no SIOPE atende ao modelo orçamentário brasileiro, utilizando a classificação funcional e programática. Conforme a Portaria no 42, de 14 de abril de 1999, do Ministério do Planejamento, a função representa o maior nível de agregação das diversas áreas de despesas que competem ao setor público (BRASIL, MPDG, 1999). Diz respeito, portanto, à área de ação do governo (educação, saúde, previdência social, etc.). Por tratar da função educação, a tabela abaixo lista as subfunções da educação cadastradas no sistema.

Tabela 2 - Subfunções da Função Educação no SIOPE Municipal.

Subfunção típica da educação modalidade de ensino	361 - Ensino Fundamental (EF) 362 – Ensino Médio (EM) 363 – Ensino Profissional (Qualif. para o Trabalho) 364 – Ensino Superior (ES) 365 – Educação Infantil (EI) 366 – Educação de Jovens e Adultos (EJA) 367 – Educação Especial (EE)	Pode ser Pasta Pai/ Pasta
Subfunção de apoio administrativo (de infraestrutura)	121 – Planejamento e Orçamento 122 – Administração Geral 123 – Administração Financeira 125 – Normatização e Fiscalização 126 – Tecnologia da Informação 128 – Formação de Recursos Humanos 131 – Comunicação Social	Apenas Pasta
Subfunção Outras considerada no cálculo do MDE	331 – Proteção e Benefícios ao Trabalhador (*) 722 – Telecomunicações (Educação a Distância) 782 – Transporte Escolar (*) 841 – Refinanciamento da Dívida Interna (*) 842 – Refinanciamento da Dívida Externa (*) 843 – Serviço da Dívida Interna (*) 844 – Serviço da Dívida Externa (*) 846 – Outros Encargos Especiais (*)	Apenas Pasta

Subfunção Outras NÃO considerada no cálculo do MDE	242 – Assistência ao Portador de Deficiência 243 – Assistência à Criança e ao Adolescente 271 – Previdência Básica 272 – Previdência do Regime Estatutário 273 – Previdência Complementar 274 – Previdência Especial 306 – Alimentação e Nutrição - Merenda Escolar (*) 392 – Difusão Cultural 695 - Turismo 812 – Desporto Comunitário 813 – Lazer	Apenas Pasta

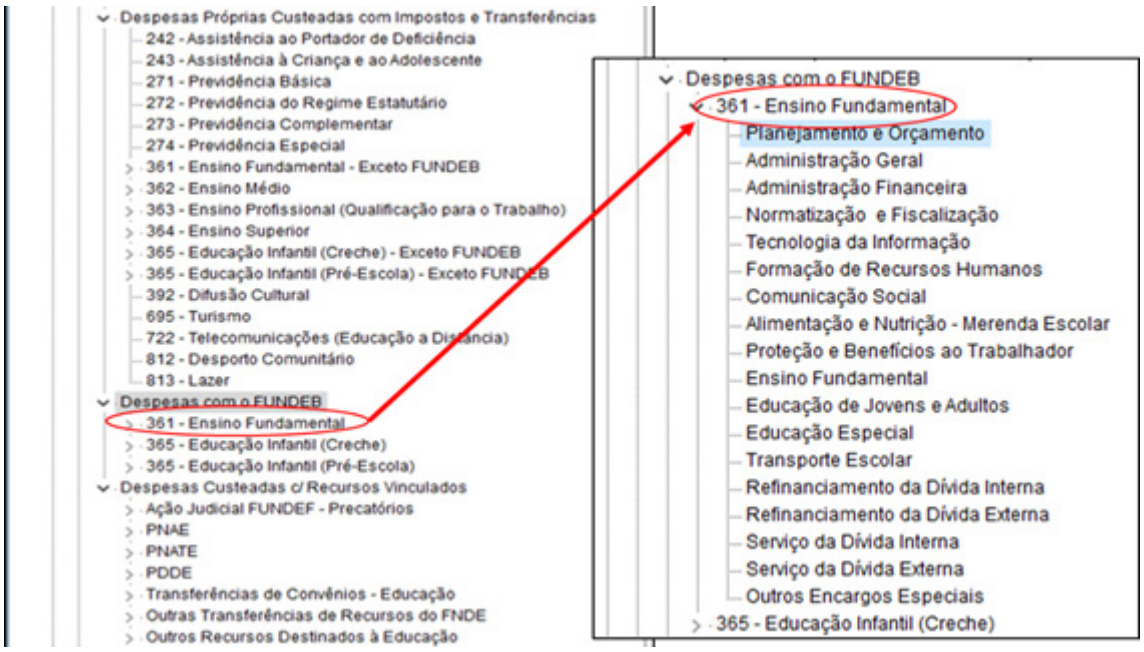
(*) São também consideradas subfunções de apoio administrativo e se localizam como subpastas abaixo de alguma modalidade de ensino (Pasta Pai de numeração 361 a 365).

Fonte: Elaborada pelo autor (2020), adaptado de BRASIL, MEC (2018).

Na alocação de uma dada despesa, após a escolha do grupo de despesa, seleciona-se uma ou mais subfunções que a representem, conforme demonstrado na Figura 5. Por tratar-se de uma hierarquia de subfunções, convencionou-se pelo uso dos campos Pasta_Pai e Pasta - a Pasta_Pai trata-se de uma subfunção típica (a modalidade de ensino) ou do item "Despesas Próprias Custeadas com Impostos e Transferência" – referentes às subfunções 242, 243, etc. que são despesas da escola que independem de uma modalidade de ensino. A Pasta é somente uma outra subfunção, subordinada à Pasta_Pai.

Na figura seguinte, abaixo de subfunções típicas de numeração 361 a 365 (modalidades de ensino) estão desdobradas as subfunções de apoio administrativo. Estas são atividades-meio que favorecem o desenvolvimento das atividades escolares e influenciam, indiretamente, para a execução das subfunções típicas – naturalmente, são despesas rateadas pelo número de matrículas de cada modalidade de ensino, subsidiando os cálculos do custo por aluno e facilitando a apuração, em nível nacional, do quanto se gasta em cada uma destas áreas (BRASIL, MEC, 2018).

Figura 5 - Hierarquia de Subfunções utilizadas para classificar uma despesa



Fonte: BRASIL, MEC (2018).

4.1.4. Contas Contábeis (Natureza e Elemento da Despesa)

O Plano de Contas Contábeis no SIOPE segue as determinações do Manual Técnico de Orçamento (MTO), base para a elaboração dos Orçamentos Fiscal e da Seguridade Social da União (BRASIL, MPDG, 1999). Conforme a Figura 6, após a seleção do grupo de despesa e das subfunções (no lado esquerdo), são apresentadas, no lado direito, as contas contábeis hierarquizadas em analíticas (na cor em branco) e sintéticas (na cor em azul, desabilitados para preenchimento). Escolhe-se, enfim, a conta mais relacionada à despesa sendo registrada. A cor amarela é a conta que está, no momento, selecionada pelo usuário na aplicação.

Figura 6 - Uso de conta contábil para classificar uma despesa

Fonte: BRASIL, MEC (2018).

4.2. PREPARAÇÃO DOS DADOS – DATAFRAME DE DESPESAS

4.2.1. Coleta e limpeza dos dados

O escopo da seleção dos dados do SIOPE Municipal se restringiu às despesas dos municípios no ano de 2018, com os atributos mais relevantes conforme listados na tabela abaixo. Procedimentos de limpeza estão resumidos na Tabela 4.

Tabela 3 - Campos do modelo de dados do SIOPE relevantes para o trabalho

CodUF	Código da UF
NomeUF	Nome da UF
SigUF	Sigla da UF
CodMunicipio	Código do município
NomeMunicipio	Nome do município
Classif_Pasta	Equivale ao Grupo de Despesa (Próprias, FUNDEB, Vinculadas).
NomePrograma	Nome do Programa do FNDE (somente para despesas vinculadas).
CodPasta_Pai	Código da Pasta Pai (representa a modalidade de ensino).
NomePasta_Pai	Nome da Pasta Pai (representa a modalidade de ensino).
CodPasta	Código da Pasta (subfunção da educação).
NomePasta	Nome da Pasta (subfunção da educação).
CodCC	Número da conta contábil, sem pontos. Ex. 34490523400.
Cod_CC_f	Número da conta contábil com pontos que separam grupos de dois dígitos. Ex. 3.44.90.52.34.00
NomeCC	Nome da conta contábil. Ex. Máquinas, Utensílios e Equip. Diversos.

Cod_NomeCC_f	Campo que junta o Nome ao Código da Conta Contábil - necessário porque há contas contábeis diferentes que apresentam o mesmo nome.
Dotação Atualizada (DA)	Dotação prevista no Orçamento (mais as suplementações, menos as anulações registradas).
Desp. Empenhadas (DE)	Despesa originária de ato emanado de autoridade competente que cria para o Estado uma obrigação de pagamento.
Desp. Liquidadas (DL)	Verificação do direito adquirido pelo credor, com base em documentos comprobatórios da entrega do material ou da prestação de serviço.
Desp. Pagas (DP)	Consiste na quitação do bem adquirido ou do serviço contratado.

Fonte: Elaborada pelo autor (2020).

Estes dados foram acessados no ambiente Jupyter Notebook (PROJECT JUPYTER, 2019) e consolidados em diferentes dataframes, por meio do emprego de linguagem Python (PYTHON, 2001) e de uma diversidade de bibliotecas, entre elas: Pandas para análise de dados (THE PANDAS PROJECT, 2019); PYODBC para conexão ao banco de dados; e SEABORN (SEABORN, 2019) para a geração de gráficos. Para uso dos algoritmos de mineração de dados (clusterização), foram utilizados os módulos da biblioteca de aprendizagem de máquina de código aberto SCIKIT-LEARN (PEDREGOSA, 2011).

Tabela 4 - Procedimentos de limpeza dos dados

DA, DE, DL, DP	Campos nulos foram preenchidos com valor ZERO (se deve a alguns tipos de despesa não preenchidos pelos entes federativos – consequência do pivoteamento na coluna "Tipo de Despesa").
Nome Programa	Este campo se aplica somente para Despesas Vinculadas. Demais registros (despesas próprias e FUNDEB) com valor nulo foram atualizados para "Não se aplica".
CodPasta Pai	Campos nulos foram atualizados para "Não se aplica", pois são referentes às subfunções que não se encaixam em alguma modalidade de ensino. Embora nulo, o campo NomePasta_Pai está preenchido com o valor "Despesas Próprias Custeadas com Impostos e Transferências".

Fonte: Elaborada pelo autor (2020).

4.2.2. Inclusão de colunas: "Classificação" e "Tipo de Gasto"

A coleta de dados lista todas as contas contábeis sintéticas (agrupadas) e analíticas (nas quais valores são efetivamente inseridos). Desta forma, tornou-se necessário indicar a classificação das contas em sintéticas e analíticas, para que o dataframe de despesas contivesse somente as analíticas que seriam submetidas à análise de dados.

Figura 7 – Quantitativo de registros de contas analíticas e contas sintéticas

```
Lista de Grupos de Despesa:  3 valores
 ['Desp proprias - não EF', 'Desp FUNDEF', 'Desp com Recursos Vinculados']

Contagem de contas analíticas (unique):  476
Contagem de contas sintéticas (unique):  40

Considerando-se todos os grupos de Despesas dos Municípios:
Total de registros de despesas dos Municipios:  2026697 despesas
Total de registros de despesas Analíticas dos Municipios:  923778 despesas
Total de registros de despesas Sintéticas dos Municipios:  1102919 despesas
```

Fonte: Elaborada pelo autor (2020).

Ademais, em virtude de um grande número de contas contábeis (mais de 300 contas), cada conta contábil analítica foi classificada, manualmente, em uma categoria de gasto mais genérico, com o intuito de consolidar as contas em 9 grandes grupos de tipo de gasto: Remuneração, Formação, Didático, Alimentação, Transporte, Manutenção, Investimentos, Conveniadas e Outros. Essa classificação foi realizada pela equipe da CGU, sendo validada por um gestor do SIOPE no FNDE. A figura abaixo exibe a contagem de despesas analíticas em cada grupo de Tipo de Gasto.

Figura 8 – Contagem de despesas analíticas em cada Tipo de Gasto

index	Total registros	Total de registros em %
6. Manutenção	409755	44.36
1. Remuneração	264333	28.61
7. Investimentos	99349	10.75
4. Alimentação	47067	5.10
5. Transporte	39084	4.23
3. Didático	33816	3.66
9. Outros	18912	2.05
8. Conveniadas	6763	0.73
2. Formação	4699	0.51

Fonte: Elaborada pelo autor (2020).

4.2.3. Inclusão de dados adicionais: dados econômicos e demográficos

As tentativas iniciais de modelagem nas despesas não produziram bons resultados, pois não é recomendável comparar todas as despesas sem o devido cuidado com as particularidades de cada município (municípios maiores, com mais professores e alunos, gastam muito mais e vice-versa). Em vista disso, houve um retorno à fase de preparação de dados para a inclusão de dados externos que caracterizassem os municípios para fins de clusterização. A tabela abaixo resume os dados externos adicionados ao dataframe de despesas.

Tabela 5 – Dados de fontes externas adicionados ao estudo das despesas municipais

Dados do IBGE	Código do município IBGE; região, mesoregião e microregião; população nos 5.570 municípios brasileiros (BRASIL, IBGE, 2019).
Dados do SIOPE e INEP	Dados referentes ao contexto educacional de cada município: 1) quantitativos de matrículas, por modalidade de ensino (subfunções 361 a 367), para cada município (dados do INEP previamente cadastrados no SIOPE); 2) quantitativos de escolas e de professores (dados do INEP); 3) Índice de Desenvolvimento da Educação Básica (IDEB); e 4) Taxa de evasão (2014/2015) para dependência administrativa pública.
Dados do PNUD	Dados referentes ao Índice de Desenvolvimento Humano Municipal (IDHM), para cada município: IDHM Educação, IDHM Longevidade e IDHM Renda.

Fonte: Elaborada pelo autor (2020).

Os quantitativos de matrículas, por modalidade de ensino e para cada município, são dados fornecidos pelo INEP e cadastrados no SIOPE. Com relação aos quantitativos de escolas e professores, as condições de seleção foram obtidas do documento "Filtros da Educação Básica" do INEP, que trata de instruções para a utilização dos Microdados do Censo da Educação Básica - 2018 (BRASIL, INEP,

2019e), e executadas na base de dados do INEP.

O Índice de Desenvolvimento da Educação Básica (IDEB) (BRASIL, INEP, 2019a), elaborado pelo MEC, é um indicador de qualidade dos ensinos fundamental e médio, abrangendo as redes pública e privada, sendo resultado do cruzamento do desempenho (Prova Brasil e Avaliação Nacional da Educação Básica - ANEB) com o rendimento escolar (aprovação). Foram acrescentados ao dataframe de despesas os seguintes indicadores:

- IDEB Anos Iniciais (IDEB_AI) e IDEB Anos Finais (IDEB_AF): se referem às notas IDEB – Ensino Fundamental (EF) - Anos Iniciais e Anos Finais (referência 2017) de cada município (apenas a nota das escolas urbanas da rede municipal); e
- IDEB Ensino Médio (IDEB_EM): se refere às notas IDEB - Ensino Médio (referência 2017) de cada município (apenas a nota das escolas da rede estadual).

A taxa de evasão representa a proporção de alunos que, em 2014 estavam matriculados na série k (etapa de ensino seriada do ensino fundamental ou médio), e em 2015 não estavam matriculados. No momento em que se buscou estes dados, a taxa mais recente disponibilizada era com relação ao ano de 2014 para o ano de 2015.

Com relação ao Índice de Desenvolvimento Humano Municipal (IDHM), gerado pelo Programa das Nações Unidas para o Desenvolvimento (PNUD) – trata-se de medida composta de indicadores de três dimensões do desenvolvimento humano: longevidade, educação e renda, para avaliar o desenvolvimento dos municípios brasileiros (BRASIL PNUD, 2019).

- IDHM Educação (IDHM_E): média geométrica do subíndice de frequência de crianças e jovens à escola, com peso de 2/3, e do subíndice de escolaridade da população adulta, com peso de 1/3.
- IDHM Longevidade (IDHM_L): obtido a partir do indicador Esperança de vida ao nascer (valores mínimo e máximo são 25 e 85 anos, respectivamente).
- IDHM Renda (IDHM_R): obtido a partir do indicador Renda per capita (valores mínimo e máximo são 8,00 e 4.033,00).

4.2.4. Filtro dos dados para contexto ao Ensino Fundamental

Conforme escopo delimitado para o presente trabalho, manteve-se no dataframe de despesas apenas os registros do Ensino Fundamental.

Figura 9 – Quantidade de registros de despesas analíticas em cada modalidade de ensino

	GrupoDespesa	CodPasta_Pai	NomePasta_Pai	Total Registros
0	Desp FUNDEF	361	Ensino Fundamental	131431
3	Desp com Recursos Vinculados	361	Ensino Fundamental	171823
9	Desp proprias - não EF	361	Ensino Fundamental - Exceto FUNDEB	273025
4	Desp com Recursos Vinculados	362	Ensino Médio	7214
10	Desp proprias - não EF	362	Ensino Médio	5155
5	Desp com Recursos Vinculados	363	Ensino Profissional	1668
11	Desp proprias - não EF	363	Ensino Profissional (Qualificação para o Trabalho)	3276
6	Desp com Recursos Vinculados	364	Ensino Superior	3983
12	Desp proprias - não EF	364	Ensino Superior	5705
1	Desp FUNDEF	365	Educação Infantil (Creche)	39364
2	Desp FUNDEF	365	Educação Infantil (Pré-Escola)	45082
7	Desp com Recursos Vinculados	365	Educação Infantil (Creche)	39159
8	Desp com Recursos Vinculados	365	Educação Infantil (Pré-Escola)	38131
13	Desp proprias - não EF	365	Educação Infantil (Creche) - Exceto FUNDEB	74784
14	Desp proprias - não EF	365	Educação Infantil (Pré-Escola) - Exceto FUNDEB	74960
15	Desp proprias - não EF	Não se aplica	Despesas Próprias Custeadas com Impostos e Transferências	9018

Fonte: Elaborada pelo autor (2020).

Entende-se que cada modalidade de ensino utiliza subfunções e contas contábeis diferenciadas. É provável que o Ensino Médio, por exemplo, tenha muito mais gastos com a infraestrutura de laboratórios (Física, Química e Biologia) do que o Ensino Fundamental e a Educação Infantil (que não terá esse tipo de gasto). Consequentemente, os estudos devem ser realizados, de forma separada, por modalidade de ensino - e o presente trabalho se limitou aos estudos das despesas aplicadas no Ensino Fundamental (registros de campo CodPasta_Pai igual ao valor 361) – lembrando-se que são incluídas as modalidades Ensino de Jovens e Adultos e Educação Especial no contexto do Ensino Fundamental.

4.2.5. Resumo do dataframe de despesas

Após a realização de todos os tratamentos, tem-se um dataframe de despesas municipais executadas para o Ensino Fundamental, no ano de 2018 – suas principais características estão resumidas na figura abaixo. Alguns comentários se fazem necessários:

- Há o total de 4.989 municípios, pois nem todos os municípios havia entregue a declaração das contas no momento do recebimento do sistema SIOPE pela CGU;
- Alguns campos inseridos inicialmente no dataframe ("Custo da educação por aluno" e "Despesa da educação com Professor, por Aluno") foram descartados por apresentarem inconsistências que prejudicariam a eficácia dos algoritmos de mineração de dados;
- Os valores zero existentes para as variáveis "IDHM dos Municípios", "Número de matrículas no EF", "IDEB – EF Anos Iniciais", "IDEB – EF Anos Finais" e "Custo da educação por Aluno" significam a ausência de dados (não houve valores registrados dessas variáveis para determinados municípios);
- Os valores zero existentes para a variável Taxa de Evasão significam que realmente não houve evasão de alunos.

Figura 10 – Resumo dos atributos do dataframe de despesas municipais

```
Total de registros de despesas dos Municipios:  576279 despesas
Total de Municipios:  4989 Municípios

Total de colunas do dataframe:  50 colunas

Colunas do dataframe de despesas:
 ['CodUF' 'NomeUF' 'SigUF' 'CodMun' 'CodIBGE' 'CodIBGE_Completo'
 'NomeMunicipio' 'Regiao' 'MesoRegiao' 'NomeMesoRegiao' 'MicroRegiao'
 'NomeMicroRegiao' 'NomePrograma' 'GrupoDespesa' 'Tipo de Gasto'
 'CodPasta_Pai' 'NomePasta_Pai' 'CodPasta' 'NomePasta' 'CodCC' 'CodCC_f'
 'NomeCC' 'Cod_NomeCC' 'Cod_NomeCC_f' 'DA' 'DE' 'DL' 'DP' 'Vlr_FUNDEB_STN'
 'Pop_estimada' 'IDHM' 'IDHM_E' 'IDHM_L' 'IDHM_R' 'QtdEscolas'
 'QtdDocentes' 'NUM_MATR_361' 'NUM_MATR_362' 'NUM_MATR_363' 'NUM_MATR_365'
 'NUM_MATR_365_1' 'NUM_MATR_365_2' 'NUM_MATR_366' 'NUM_MATR_367' 'IDEB_AI'
 'IDEB_AF' 'IDEB_EM' 'TxEvasao_EF' 'CustoAluno' 'DespesaProf']

 Regiões cujos Municipios entregaram a declaração ao SIOPE:  5 regiões
  ['CENTRO OESTE', 'NORDESTE', 'NORTE', 'SUDESTE', 'SUL']

 Estados cujos Municipios entregaram a declaração ao SIOPE:  4989 Municípios em  26  estados.
  ['Acre', 'Alagoas', 'Amapa', 'Amazonas', 'Bahia', 'Ceara', 'Espirito Santo', 'Goias', 'Maranhao',
 'Mato Grosso', 'Mato Grosso do Sul', 'Minas Gerais', 'Para', 'Paraiba', 'Parana', 'Pernambuco', 'Pia
 ui', 'Rio Grande do Norte', 'Rio Grande do Sul', 'Rio de Janeiro', 'Rondonia', 'Roraima', 'Santa Cat
 arina', 'Sao Paulo', 'Sergipe', 'Tocantins']

 Lista de Programas:  9 valores
  ['Não se aplica', 'PNAE', 'Vinculadas a Contribuição Social do Salário-Educação', 'PNATE', 'Outras
 Transferências de Recursos do FNDE', 'Transferências de Convênios - Educação', 'Outros Recursos Dest
 inados à Educação', 'PDDE', 'Ação Judicial FUNDEF - Precatórios']

 Lista de Grupos de Despesa:  3 valores
  ['Desp proprias - não EF', 'Desp FUNDEF', 'Desp com Recursos Vinculados']

 Lista de Pastas Pai:  2 valores
  ['Ensino Fundamental', 'Ensino Fundamental - Exceto FUNDEB']

 Lista de Pastas:  21 valores
  ['Administração Financeira', 'Administração Geral', 'Alimentação e Nutrição - Merenda Escolar', 'Co
 municação Social', 'Despesas Custeadas com Recursos de Royalties de Petróleo e de Indenizações', 'Ed
 ucação Especial', 'Educação de Jovens e Adultos', 'Ensino Fundamental', 'Ensino Fundamental - Exceto
 FUNDEB', 'Formação de Recursos Humanos', 'Normatização  e Fiscalização', 'Normatização e Fiscalizaçã
 o', 'Outros Encargos Especiais', 'Planejamento e Orçamento', 'Proteção e Benefícios ao Trabalhador',
 'Refinanciamento da Dívida Externa', 'Refinanciamento da Dívida Interna', 'Serviço da Dívida Extern
 a', 'Serviço da Dívida Interna', 'Tecnologia da Informação', 'Transporte Escolar']

 Lista de valores de Tipo de Gasto da despesa:  9 valores
  ['1. Remuneração', '2. Formação', '3. Didático', '4. Alimentação', '5. Transporte', '6. Manutençã
 o', '7. Investimentos', '8. Conveniadas', '9. Outros']

Variação da população dos Municípios:   781 a 12.252023  milhões
Variação do índice IDHM dos Municípios:   0.0 a  0.862
Variação da quantidade de escolas nos municípios:   1 a 1539
Variação da quantidade de professores nos municípios:   4 a 38406
Variação do número de matrículas no EF:   0 a 458634
Variação da taxa de evasão (EF) nos municípios:   0.0 a 21.5
Variação do IDEB - EF Anos Iniciais:  0.0 a 9.1
Variação do IDEB - EF Anos Finais:  0.0 a 7.2
Custo da educação por Aluno:  0.0 a 224567.0
Despesa da educação com Professor, por Aluno:  1401.95 a 186627.96
```

Fonte: Elaborada pelo autor (2020).

Tabela 6 – Descrição dos campos presentes no dataframe de despesas

CodUF, NomeUF, SigUF	Informações da UF.
CodMun, CodIBGE, CodIBGE_ Completo, NomeMunicipio	Informações do Município.
Regiao, MesoRegiao, NomeMeso-Regiao, MicroRegiao, NomeMi-croRegiao	Informações da Região.
NomePrograma	Nome do Programa do FNDE (somente para despesas vinculadas).

GrupoDespesa	Próprias, FUNDEB ou Vinculadas.
Tipo de Gasto	Classificação da conta contábil em um tipo de gasto mais genérico.
CodPasta_Pai e NomePasta_Pai	Representa a modalidade de ensino.
CodPasta e NomePasta	Representa a subfunção da educação.
CodCC	Número da conta contábil, sem pontos. Ex. 34490523400.
CodCC_f	Número da conta contábil com pontos que separam grupos de dois dígitos. Ex. 3.44.90.52.34.00
NomeCC	Nome da conta contábil. Ex. Máquinas, Utensílios e Equip. Diversos.
Cod_NomeCC	Campo que junta o Nome ao Código da Conta Contábil - necessário porque há contas contábeis diferentes que apresentam o mesmo nome.
Cod_NomeCC_f	Campo que junta o Nome ao Código da Conta Contábil formatado.
DA, DE, DL, DP	Tipo da Despesa (Dotação Atualizada, Despesa Empenhada, Despesa Liquidada e Despesa Paga).
Vlr_FUNDEB_STN	Campo adicionado ao dataframe por solicitação da CGEBC (fora do escopo do presente trabalho).
Pop_estimada	População estimada do município (fonte: IBGE).
IDHM (*)	IDHM (Fonte: PNUD).
IDHM_E (*)	IDHM Educação (Fonte: PNUD).
IDHM_L (*)	IDHM Longevidade (Fonte: PNUD).
IDHM_R (*)	IDHM Renda (Fonte: PNUD).
QtdEscolas	Quantidade de escolas (Fonte: INEP).
QtdDocentes	Quantidade de professores (Fonte: INEP).
NUM_MATR_361 (*)	Número de alunos matriculados no Ensino Fundamental (Fonte: INEP).
NUM_MATR_362	Número de alunos matriculados no Ensino Medio (Fonte: INEP).
NUM_MATR_363	Número de alunos matriculados no Ensino Profissional (Fonte: INEP).
NUM_MATR_365	Número de alunos matriculados no Ensino Superior (Fonte: INEP).
NUM_MATR_365_1	Número de alunos matriculados na Educação Infantil (creche) (Fonte: INEP).
NUM_MATR_365_2	Número de alunos matriculados na Educação Infantil (pré-escola) (Fonte: INEP).
NUM_MATR_366	Número de alunos matriculados na Educação de Jovens e Adultos (Fonte: INEP).
NUM_MATR_367	Número de alunos matriculados na Educação Especial (Fonte: INEP).
IDEB_AI (*)	Nota IDEB no Ensino Fundamental - Anos Iniciais (Fonte: INEP).
IDEB_AF (*)	Nota IDEB no Ensino Fundamental - Anos Finais (Fonte: INEP).
IDEB_EM (*)	Nota IDEB no Ensino Médio (Fonte: INEP).
TxEvasao_EF	Taxa de Evasão no Ensino Fundamental (Fonte: INEP). Valores zero significam que não houve evasão de alunos.
CustoAluno (*)	Valor do Custo por Aluno (cálculo realizado pelo FNDE). Não será utilizado no presente trabalho.
DespesaProf	Valor das Despesas por Professor (cálculo realizado pelo FNDE). Não será utilizado no presente trabalho.

(*) Valores zero significam valores ausentes, não informados.

Fonte: Elaborada pelo autor (2020).

4.3. PREPARAÇÃO DOS DADOS – DATAFRAME DE MUNICÍPIOS

A execução de determinados algoritmos de mineração de dados pressupõe um formato de dados de entrada no qual cada objeto seja representado por um vetor de atributos (ou vetor de características). Em virtude disso, outras preparações foram realizadas nos dados.

4.3.1. Pivoteamento dos dados

Para a geração de vetores de características, foi realizado um pivoteamento no dataframe de despesas – o resultado é um novo dataframe de municípios, no qual cada linha é um município e as colunas são todos os seus atributos.

Tabela 7 – Criação de novas colunas após o pivoteamento dos dados

Grupo Despesa	Despesas Próprias, Despesas FUNDEB, Despesas Vinculadas.
Tipo de Gasto	tgRemun (Remuneração), tgFormacao (Formação), tgDidatico (Material Didático), tgAlim (Alimentação), tgTransp (Transporte), tgManut (Manutenção), tgInvest (Investimentos), tgConv (Conveniadas), tgOutros (Outros)
Nome do Programa	Ação Judicial FUNDEF – Precatórios, Outros Recursos Destinados à Educação, Outras Transf Recursos do FNDE, Transf Convênios – Educação, PDDE, PNAE, PNATE, Vincul a Contrib Social do Salário-Educação.
Nome Pasta (subfunção)	Administração Financeira, Administração Geral, Planejamento e Orçamento, Alimentação e Nutrição - Merenda Escolar, Transporte Escolar, Comunicação Social, Tecnologia da Informação, Despesas Custeadas com Recursos de Royalties de Petróleo e de Indenizações, Educação Especial, Educação de Jovens e Adultos, Ensino Fundamental, Ensino Fundamental - Exceto FUNDEB, Formação Recursos Humanos, Normatização e Fiscalização, Outros Encargos Especiais, Proteção e Benefícios ao Trabalhador, Refinanciamento da Dívida Externa, Refinanciamento da Dívida Interna, Serviço da Dívida Externa, Serviço da Dívida Interna
Conta Contábil	de 3.31.90.01.00.00 a 3.46.00.00.00.00 (mais de 300 contas contábeis analíticas)

Fonte: Elaborada pelo autor (2020).

4.3.2. Consolidação de Contas Contábeis

O dataframe de municípios resultante do pivoteamento apresentou um grande número de colunas (387), principalmente de contas contábeis. As análises exploratórias de dados, realizadas inicialmente, comprovou um alto número de valores zerados para muitas dessas contas, o que motivou a substituição de algumas contas analíticas pela conta sintética equivalente, pois não se pretendeu analisar as contas contábeis em um nível muito detalhado. Em outras palavras, basta identificar os municípios que possuem anomalias em gastos relacionados com, por exemplo, Obrigações Patronais (composta de 14 contas analíticas, conforme listado na Figura 11), e não em cada uma dessas 14 contas.

Figura 11 – Exemplo de uma conta contábil sintética a ser considerada no dataframe

Conta Contabil	Descricao_Conta_Contabil	Nivel	Analit/ Sint
3.31.90.13.00.00	**Obrigações Patronais**	**3**	**Sintético**
3.31.90.13.01.00	FGTS	4	Analítico
3.31.90.13.02.00	Contribuições Previdenciárias - INSS	4	Analítico
3.31.90.13.03.00	Contribuições Previdenciárias no Exterior	4	Analítico
3.31.90.13.04.00	Contribuição de Salário-Educação	4	Analítico
3.31.90.13.08.00	Plano de Seguridade Social do Servidor - Pessoal Ativo	4	Analítico
3.31.90.13.09.00	Seguros de Acidentes do Trabalho	4	Analítico
3.31.90.13.11.00	FGTS - PDV	4	Analítico
3.31.90.13.13.00	Sesi/Sesc Ativo Civil	4	Analítico
3.31.90.13.14.00	Multas Indedutíveis	4	Analítico
3.31.90.13.15.00	Multas Dedutíveis	4	Analítico
3.31.90.13.17.00	Juros	4	Analítico
3.31.90.13.18.00	Contribuição para o PIS/PASEP S/Folha Pagto	4	Analítico
3.31.90.13.40.00	Encargos de Pessoal Requisitado de Outros Entes	4	Analítico
3.31.90.13.99.00	Outras Obrigações Patronais	4	Analítico

Fonte: Elaborada pelo autor (2020).

A lista abaixo apresenta as contas contábeis nas quais se fez a substituição das contas analíticas pela sintética equivalente, gerando um novo dataframe com menos de 200 colunas.

- Contas (3.31.90.04.**) em Contratação por Tempo Determ. (3.31.90.04.00.00)
- Contas (3.31.90.11.**) em Vencimentos e Vantagens Fixas - Pessoal Civil (3.31.90.11.00.00)
- Contas (3.31.90.13.**) em Obrigações Patronais (3.31.90.13.00.00)
- Contas (3.33.50.43.**) em Subvenções Sociais (3.33.50.43.00.00)
- Contas (3.33.90.30.**) em Material de Consumo (3.33.90.30.00.00)
- Contas (3.33.90.36.**) em Outros Serviços de Terceiros - PF (3.33.90.36.00.00)
- Contas (3.33.90.39.**) em Serviços de Terceiros - PJ (3.33.90.39.00.00)
- Contas (3.33.90.47.**) em Obrigações Tribut. e Contributivas (3.33.90.47.00.00)
- Contas (3.44.90.51.**) em Obras e Instalações (3.44.90.51.00.00)
- Contas (3.44.90.52.**) em Equipam. e Material Permanente (3.44.90.52.00.00)

4.3.3. Resumo do dataframe de municípios

Após a realização do pivoteamento e da consolidação de algumas contas contábeis, tem-se um dataframe de municípios com suas características: informações do IBGE (UF, região, população total); informações do INEP (quantidade de matrículas, escolas e professores; notas IDEB; taxa de evasão); informações do PNUD (IDHM); e informações de despesas, tipos de gastos, subfunções e contas contábeis.

Figura 12 – Resumo dos atributos do dataframe de municípios

```
Total de registros do dataframe df_consol:  4989 registros
Total de Municípios:  4989 Municípios

Total de colunas do dataframe df_consol:  175 colunas

Colunas do dataframe consolidado  :
 ['CodUF' 'NomeUF' 'SigUF' 'CodMun' 'CodIBGE' 'CodIBGE_Completo'
 'NomeMunicipio' 'Regiao' 'MesoRegiao' 'NomeMesoRegiao' 'MicroRegiao'
 'NomeMicroRegiao' 'Pop_estimada' 'IDHM' 'IDHM_E' 'IDHM_L' 'IDHM_R'
 'QtdEscolas' 'QtdDocentes' 'NUM_MATR_361' 'IDEB_AI' 'IDEB_AF'
 'TxEvasao_EF' 'CustoAluno' 'DespesaProf' 'DespFUNDEB' 'DespVinc'
 'DespProp' 'tgRemun' 'tgFormacao' 'tgDidatico' 'tgAlim' 'tgTransp'
 'tgManut' 'tgInvest' 'tgConv' 'tgOutros'
 'Ação Judicial FUNDEF - Precatórios' 'Outras Transf Recursos do FNDE'
 'Outros Recursos Destinados à Educação' 'PDDE' 'PNAE' 'PNATE'
 'Transf Convênios - Educação'
 'Vincul a Contrib Social do Salário-Educação' 'AdmFinanc' 'AdmGeral'
 'MerEscolar' 'ComunSocial' 'DespCusteadasRecRoyPetrIndeniz'
 'EducEspecial' 'EducJA' 'EnsFund' 'EnsFund_exc' 'FormRH' 'NormatFisc1'
 'NormatFisc2' 'OutrosEE' 'PlanOrc' 'ProtBenefTrab' 'RefinDivExt'
 'RefinDivInt' 'ServDivExt' 'ServDivInt' 'TI' 'TranspEsc'
 '3.31.90.01.00.00 - Aposentadorias' '3.31.90.03.00.00 - Pensões'
 '3.31.90.05.00.00 - Outros Benefícios Previdenciários'
 '3.31.90.07.00.00 - Contribuição a Entidades Fechadas de Previdência'] ... e demais contas contábeis.
```

Fonte: Elaborada pelo autor (2020).

5. FASE DE MODELAGEM

Essa fase compreende a seleção e aplicação de técnicas para criar modelos e descobrir conhecimentos. No presente trabalho, Análises Exploratórias de Dados (AED) foram realizadas para a descoberta de fatos relevantes. Em seguida, foram utilizadas as técnicas de Clusterização (para criar agrupamentos de municípios semelhantes) e Detecção de Anomalias (para identificar, em um determinado cluster, as despesas anômalas).

5.1. ANÁLISE EXPLORATÓRIA DE DADOS

A AED objetiva resumir e visualizar os dados antes de se criar modelos, permitindo entender as suas propriedades, inspecionar as suas características qualitativas e descobrir novos padrões (LEEK, 2015). No presente trabalho, procedeu-se à AED com estatísticas descritivas e plotagem de gráficos, a fim de conhecer os dados e sintetizar suas características mais relevantes; de detectar padrões ocultos; e de identificar correlações entre as variáveis.

No dataframe de despesas, a intenção foi explorar extensivamente as despesas pagas dos municípios com o Ensino Fundamental em 2018, de forma a:

- ter uma noção de ordem de grandeza dos montantes das despesas;
- detectar comportamentos dos totais de despesas pagas por grupo de despesa, subfunção, tipo de gasto e por contas contábeis; e
- preliminarmente, verificar a existência de despesas de valores anormais.

No dataframe de municípios, o objetivo foi o estudo mais específico dos comportamentos das outras variáveis. Como exemplos destes estudos, pode-se citar:

- Como é a distribuição da população, das quantidades de escolas e professores? E dos indicadores do IDHM e do IDEB?
- Como se comporta cada grupo de despesa?

• Quais as correlações existentes entre as variáveis?

Toda a AED realizada em ambos os dataframes se encontra em cadernos jupyter, que poderão ser consultados, caso necessário.

5.2. CLUSTERIZAÇÃO DE MUNICÍPIOS SEMELHANTES

5.2.1. Objetivos da clusterização de municípios

Nas análises exploratórias, foi possível perceber algumas correlações entre os dados – como a população com os quantitativos do INEP (acima de 0,86) e os índices IDHM com IDEB_AI (acima de 0,50) - motivaram a clusterização de municípios semelhantes, sem considerar os dados de despesas ou de contas contábeis. Parte-se do pressuposto que municípios semelhantes (com populações similares, indicadores próximos e quantidades parecidas de escolas, professores e alunos matriculados) devem ter despesas educacionais também semelhantes, ao menos em mesma ordem de grandeza.

Foram testados quatro algoritmos de clusterização – este capítulo detalha os dois algoritmos com os melhores resultados: k-Means e Agglomerative Clustering. Alguns resultados obtidos com o algoritmo DBSCAN são também demonstrados.

Após a criação dos agrupamentos dos municípios, alguns gráficos foram criados para facilitar a visualização da coerência destes conjuntos - ou seja, como forma de validar se determinado algoritmo foi capaz de separar bem os dados.

5.2.2. Decisão sobre o escalonamento dos dados

O escalonamento de variáveis é um método utilizado para padronizar um intervalo de variáveis independentes, sendo também denominado de normalização de dados (SRIVASTAVA, 2019). Deve-se, portanto, uniformizar os dados de características dos municípios (IBGE, INEP e PNUD) em uma mesma escala, antes de se utilizar quaisquer algoritmos de clusterização - principalmente quando a similaridade se baseia no cálculo de distâncias entre os pontos.

Várias são as técnicas de escalonamento de variáveis – no presente trabalho, foram testadas as técnicas de escalonamento definidas na tabela abaixo para cada algoritmo de clusterização utilizado. Gráficos de população com quantidade de alunos matriculados (todos em valores escalonados) foram criados para fundamentar a escolha pelo scaler mais apropriado.

Tabela 8 – Técnicas utilizadas para a transformação de variáveis

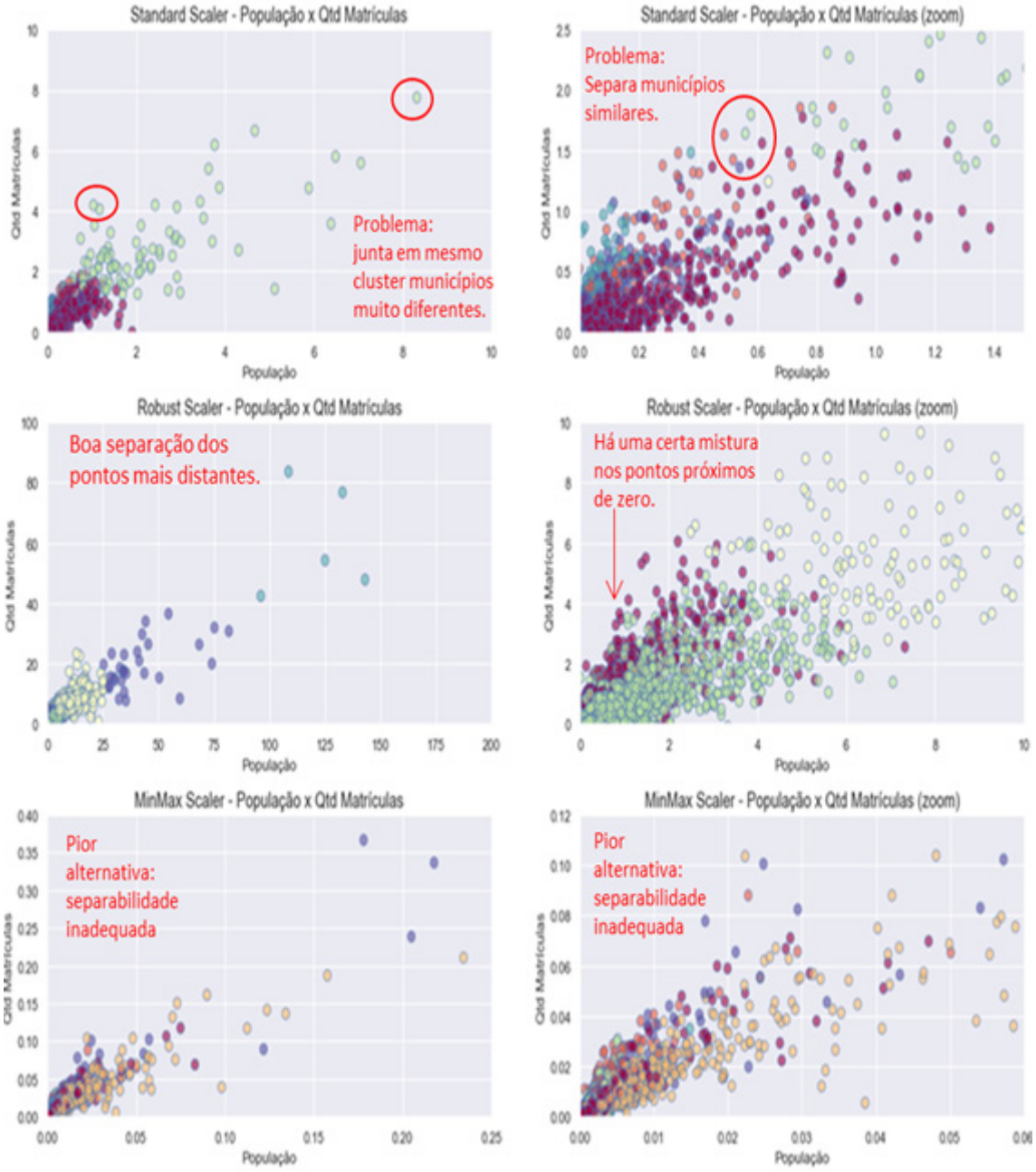
Standard Scaler	Assume que os dados são normalmente distribuídos em cada variável e, por isso, dimensiona para que a distribuição tenha média de valor zero com desvio padrão de valor 1. Se dados não são normalizados, não é uma boa alternativa de escalonamento.
Robust Scaler	Utiliza abordagem semelhante ao escalonamento MinMax, mas usa o intervalo inter-quartil ao invés do intervalo entre 0 a 1. É adequada para dados com presença de outliers.
MinMax Scaler	Reduz os dados para que assumam o intervalo de valores entre 0 e 1, ou -1 a 1 se houver valores negativos. Essa técnica é adequada para distribuições que não sejam gaussianas ou quando o desvio padrão das variáveis é de valor reduzido. É sensível a outliers.

Fonte: Elaborada pelo autor (2020), adaptado de (SRIVASTAVA, 2019).

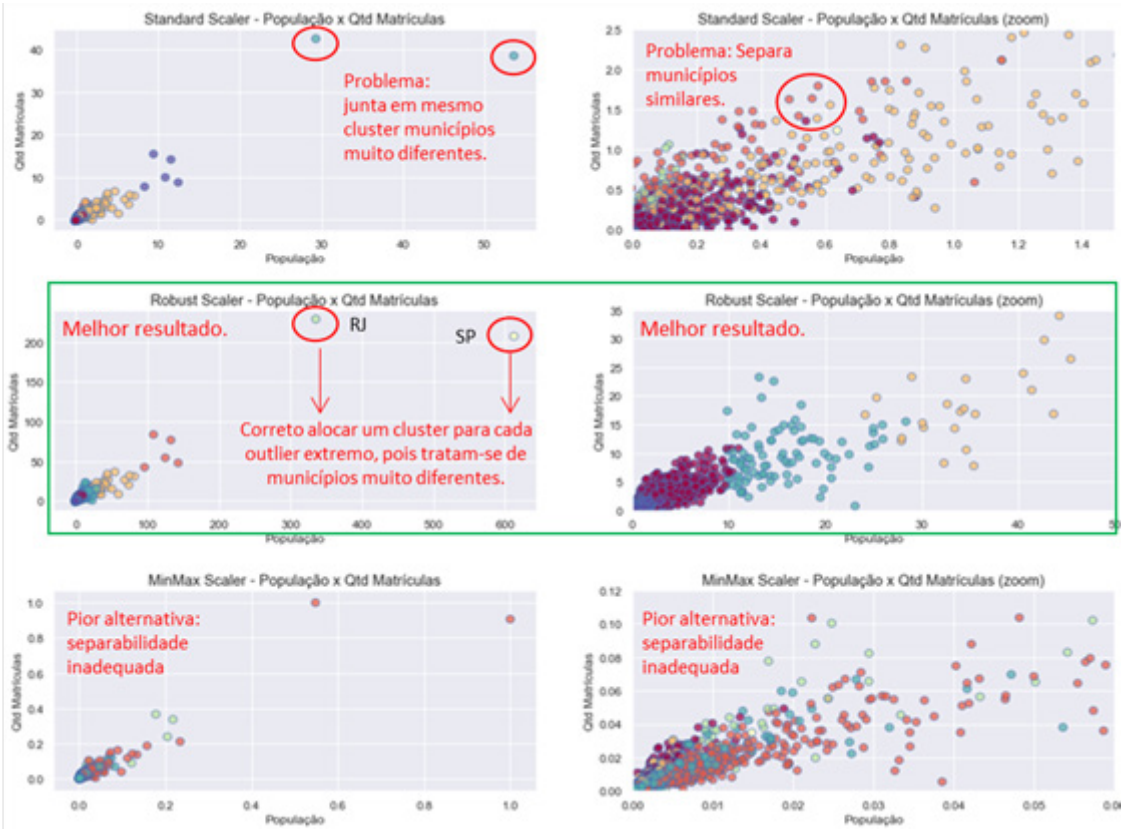
Nos gráficos da Figura 13, o lado esquerdo exibe todos os pontos (para visualizar os grupos com os pontos mais distantes), enquanto o lado direito realiza uma espécie de zoom, de forma a apresentar os grupos que são formados com os pontos mais concentrados.

Figura 13 – Escalonamento de dados com K-Means, DBSCAN e Aggl. Clustering

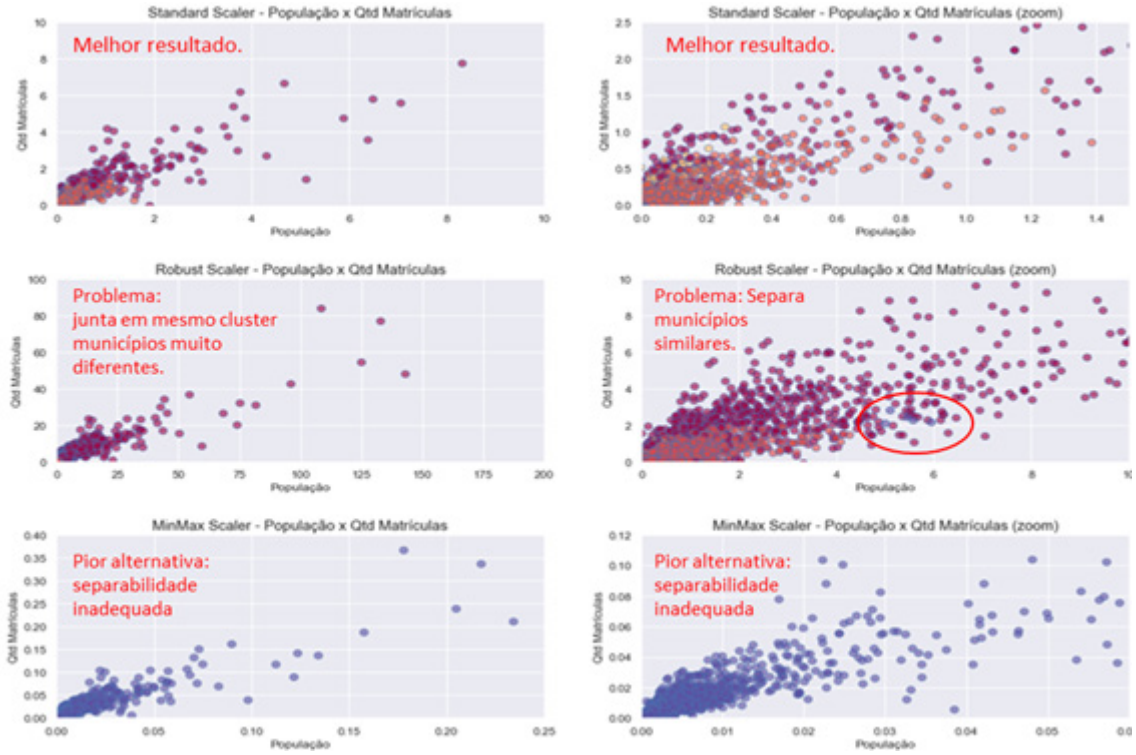
K-Means



Agglomerative Clustering



DBSCAN



Fonte: Elaborada pelo autor (2020)

Os resultados apresentados comprovam:

• K-Means: o RobustScaler apresentou os melhores resultados – os dados são bem separados em ambos os lados do gráfico, embora haja uma pequena mistura dos pontos mais próximos

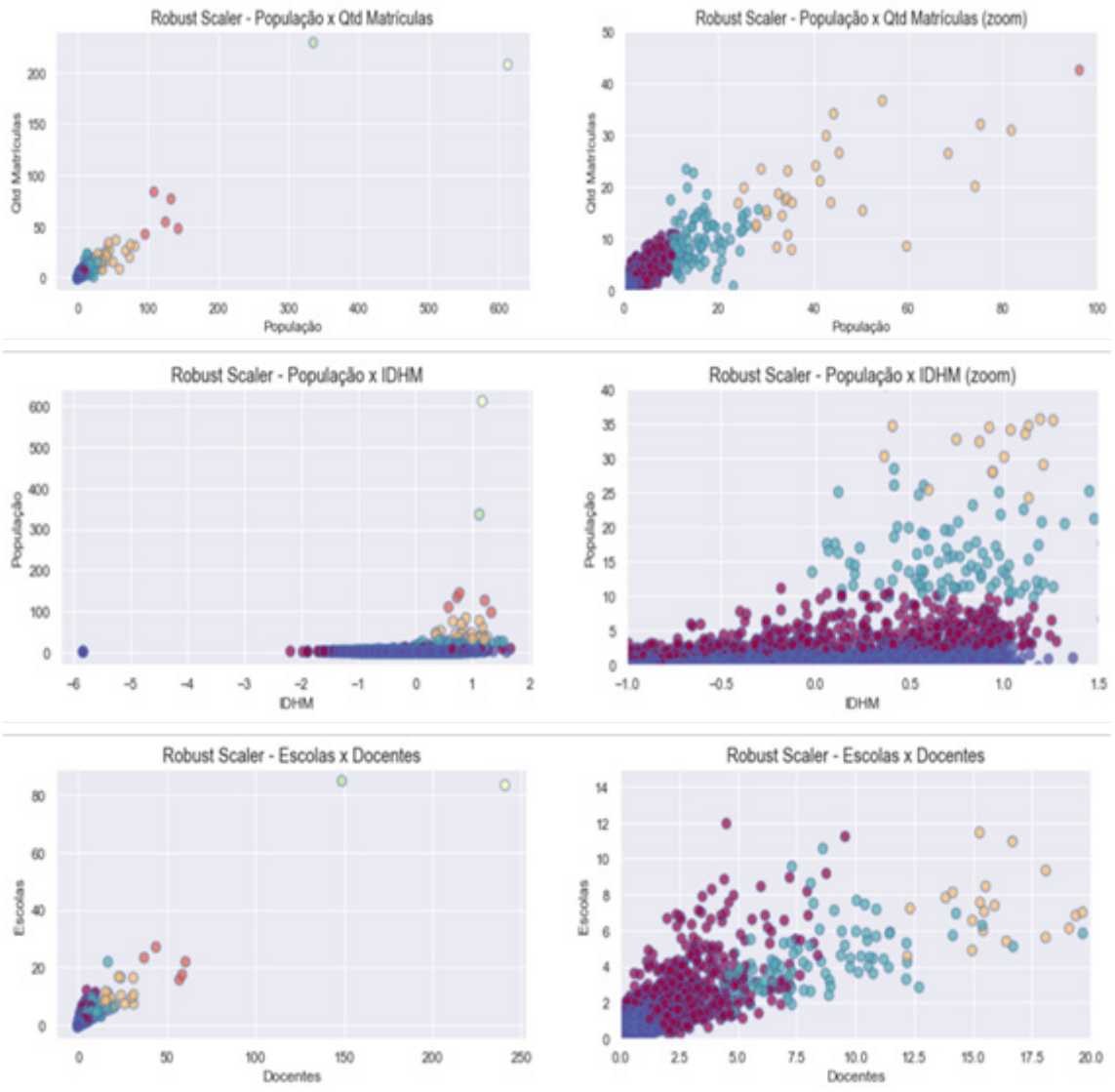
de zero;

• Agglomerative Clustering: nota-se que o RobustScaler é a melhor alternativa entre todos os gráficos apresentados (conforme delineado na cor verde);

• DBSCAN: embora haja poucos grupos, o StandardScaler se mostrou o mais adequado para separar os dados desses grupos.

A figura abaixo mostra outros gráficos criados, com outras variáveis, com o uso do algoritmo Agglomerative Clustering e técnica de escalonamento Robust Scaler.

Figura 14 – Escalonamento de dados com Aggl. Clustering e RobustScaler



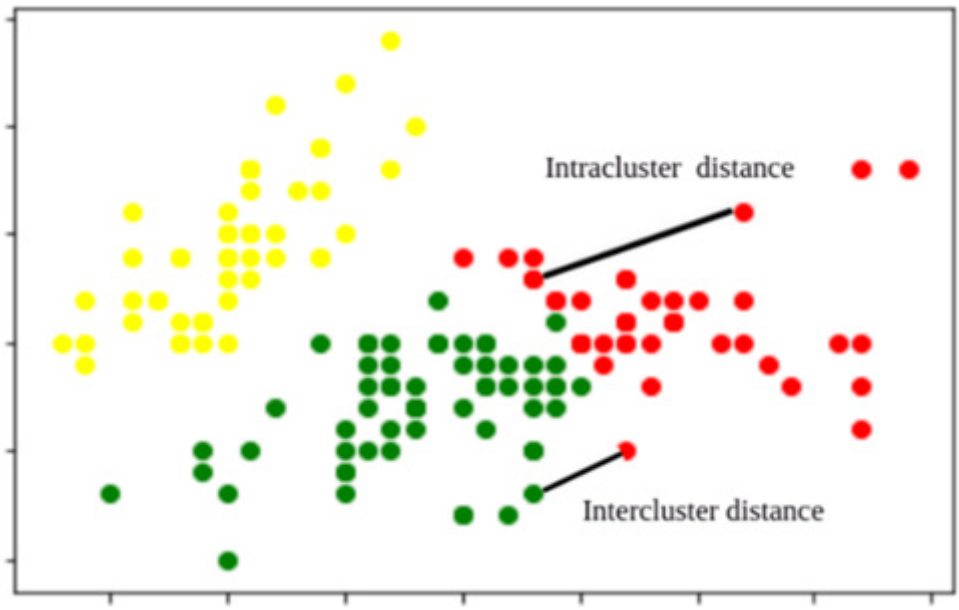
Fonte: Elaborada pelo autor (2020)

5.2.3. Clusterização k-Means

Dado o número k de clusters, K-Means particiona um conjunto de pontos em K grupos (ZAKI e MEIRA JR., 2014), de forma que cada ponto seja alocado ao grupo que lhe esteja mais próximo. Desta forma, os agrupamentos são criados com base nas distâncias mínimas entre os pontos pertencentes a cada cluster e seu centroide (ponto central do cluster). Um dos desafios em se utilizar o K-Means para agrupamentos é encontrar o número ideal de clusters – aquele que maximiza as diferenças entre

clusters (inter-cluster) e minimiza as variações dentro de um clusters (intra-cluster), conforme se pode visualizar na figura abaixo.

Figura 15 – Distância inter-cluster e intra-cluster.

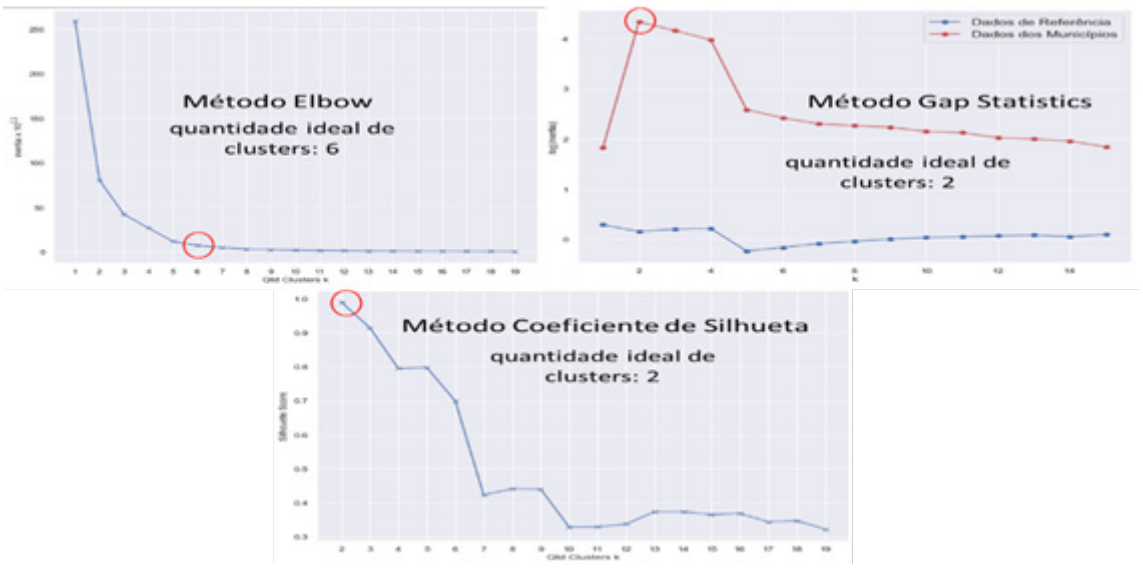


Fonte: SEN (2019)

Quando poucos clusters são formados, tem-se a maximização inter-cluster, mas as variações intra-cluster são prejudicadas, pois pontos muito distantes podem estar em um mesmo cluster. Por outro lado, ao se aumentar o número de clusters, as diferenças entre clusters se torna pequena, embora tem-se a vantagem de diminuir as variações em um cluster (intra-cluster).

É preciso, portanto, achar o ponto ótimo, no qual os pontos de cada cluster sejam os mais homogêneos possíveis e que clusters formados sejam suficientemente diferentes um dos outros. Foram utilizados alguns métodos de seleção desse ponto ótimo com os dados dos municípios, a saber: Elbow (cálculo de inertias), Gap Statistics e Coeficiente de Silhueta. A figura abaixo exibe os resultados encontrados em cada método.

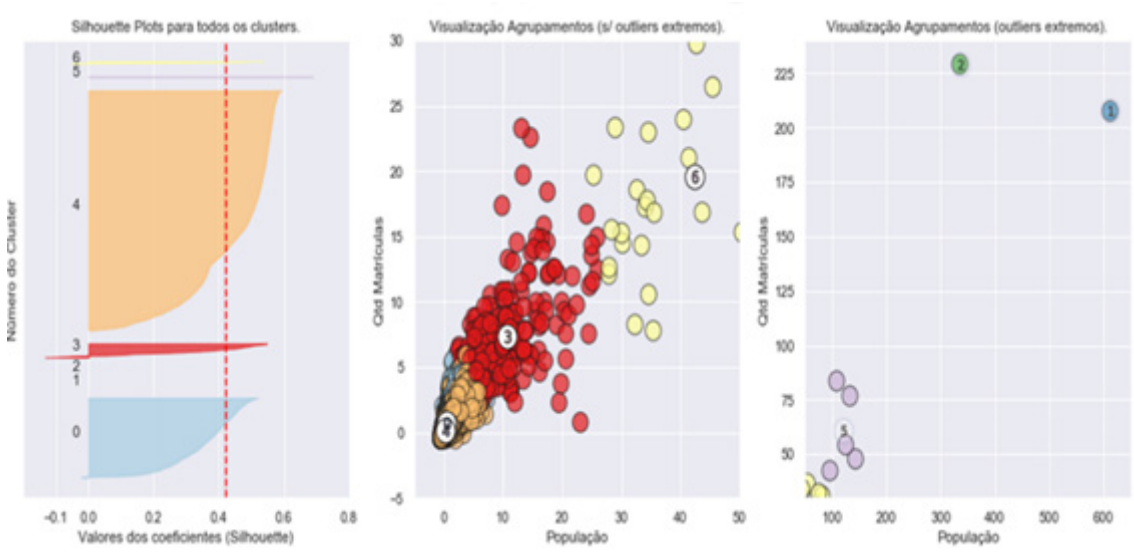
Figura 16 – Métodos para seleção do número ideal de clusters para k-Means.



Fonte: Elaborada pelo autor (2020)

Embora as sugestões variem, foi escolhido o valor de 7 clusters, pois 3 clusters já são alocados a poucos municípios anômalos. A Figura 17 mostra o gráfico de Análise de Silhueta (ALVES, 2019) para o valor de 7 clusters, utilizando-se população e número de alunos. Percebe-se que cada ponto extremo (lado direito) ficou em clusters diferentes (trata-se dos municípios de SP e RJ que são, de fato, distintos), enquanto pontos concentrados (lado esquerdo) foram adequadamente separados conforme faixas da população e número de alunos.

Figura 17 – Análise de Silhueta para clusterização k-Means com 7 clusters.

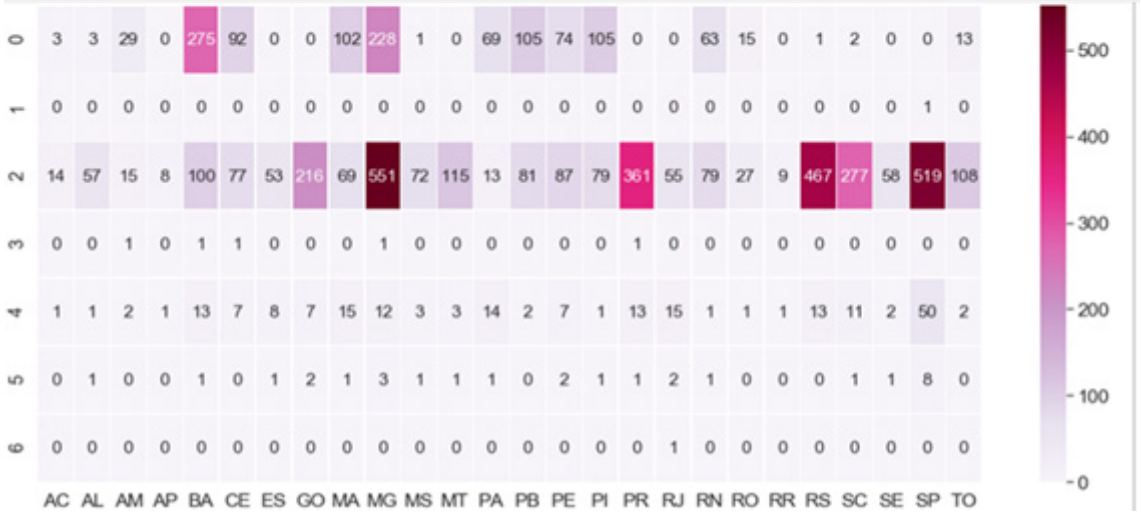


Fonte: Elaborada pelo autor (2020)

Baseado na escolha do valor de 7 clusters, a Figura 18 apresenta os resultados da clusterização com o algoritmo k-Means e escalonamento RobustScaler, exibindo a quantidade de municípios em cada cluster, região e UF.

Figura 18 – Detalhes dos clusters com k-Means e escalonamento RobustScaler.

Cluster	Total	Total em %
4	3567	71.50
0	1180	23.65
3	206	4.13
6	29	0.58
5	5	0.10
2	1	0.02
1	1	0.02



Fonte: Elaborada pelo autor (2020)

Percebe-se algumas características da clusterização k-Means:

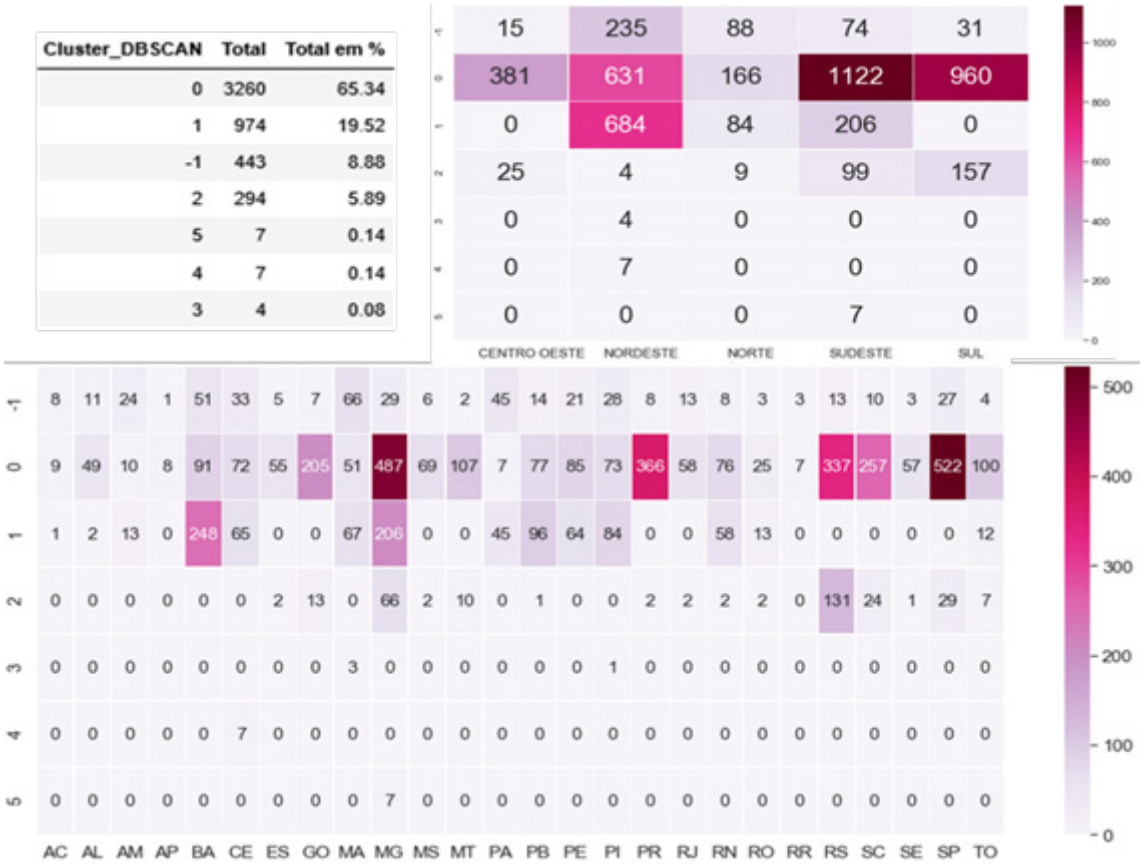
- são criados clusters específicos com municípios de valores extremos (municípios de população máxima, de maior quantidade de alunos ou de docentes);
- SP, de maior população, ficou em um cluster; enquanto RJ (metade da população de SP, mas maior quantidade de matrículas), ficou em outro cluster;
- ao se utilizar um valor K maior do que 10, são criados clusters específicos com municípios de valores mínimos (municípios com poucas escolas, por exemplo).
- geralmente ocorre a criação de clusters de poucos municípios (menos de 1% do total), que podem ser considerados anomalias de clusters.

5.2.4. Clusterização DBSCAN

Diferente de k-Means, que busca proximidade pela distância entre os pontos, DBSCAN usa a densidade local dos pontos – ou seja, utiliza o cálculo de áreas de maior e menor densidade de pontos – como critério de formação dos agrupamentos (ZAKI e MEIRA JR., 2014). Não requer o número de clusters e permite a definição de cluster irregulares.

No presente trabalho, convencionou-se por utilizar determinados parâmetros de forma que não ocorresse muitos ruídos (pontos que não são alocados em nenhum cluster), mas que houvesse ao menos 6 clusters (exceto o cluster -1, composto de ruídos). Baseado nos parâmetros eps=0.9 e min_samples=5, a Figura 19 apresenta os resultados da clusterização com o algoritmo DBSCAN e escalonamento StandardScaler, exibindo a quantidade de municípios em cada cluster, região e UF.

Figura 19 – Detalhes dos clusters com DBSCAN e escalonamento StandardScaler.



Fonte: Elaborada pelo autor (2020)

Percebe-se algumas características da clusterização DBSCAN:

- Os clusters não são bem separados como ocorre no k-Means, pois a separação de clusters considera faixas nas quais há grande densidade;
- os pontos extremos (municípios com população elevada, com poucos alunos ou poucos professores) não são alocados em algum cluster, mas inseridos no cluster noise. DBSCAN identifica os pontos outliers, ao contrário do k-Means (que os aloca em um cluster de tamanho pequeno);
- não é o algoritmo mais apropriado para a identificação de anomalias nas despesas, pois muitos pontos são alocados em poucos clusters (e o estudo requer alguns grupos semelhantes, e não grandes grupos de densidade similar).

5.2.5. Clusterização Hierárquica – Agglomerative Clustering

Algoritmos hierárquicos criam uma estrutura hierárquica de pontos aninhados conforme uma estratégia de agrupamento e um critério de ligação (métrica de dissimilaridade). As estratégias de agrupamento podem ser (ZAKI e MEIRA JR., 2014):

- aglomerativa (abordagem bottow-up): cada ponto é um cluster, e pares de clusters são mesclados sucessivamente à medida que se sobe na hierarquia;
- divisiva (abordagem top-down): todos os pontos estão em um cluster, e as divisões são executadas recursivamente à medida que se desce a hierarquia.

A dissimilaridade entre clusters é calculada conforme método escolhido no algoritmo:

- Single: a distância entre 2 clusters é dada pela menor distância entre dois pontos (distância

mínima, vizinho mais próximo). É sensível a outliers;

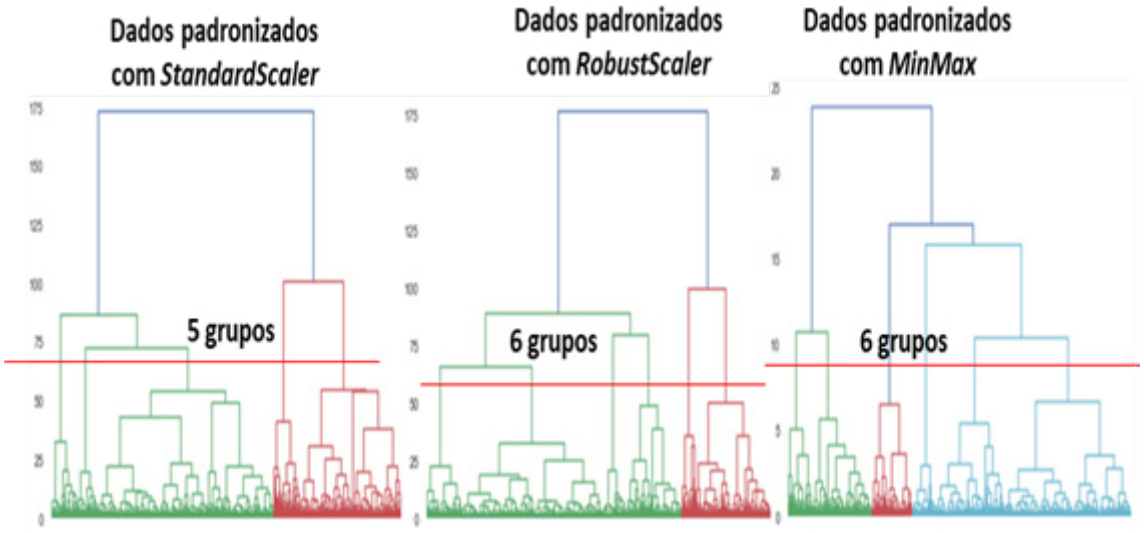
• Complete: a distância entre 2 clusters é dada pela maior distância entre dois pontos (distância máxima, vizinho mais distante). Tende a gerar clusters muito grandes;

• Average: a distância entre 2 clusters é dada pela média das distâncias entre cada dois pontos (distância entre os centroides). É menos sensível a outliers, mas tende a gerar clusters grandes, globulares;

• Ward: Minimiza a soma das diferenças entre pontos nos clusters. Mais efetivo quando se tem outliers; tende a gerar clusters de tamanhos mais regulares.

Os resultados da clusterização hierárquica são normalmente mostrados como uma árvore de grupos ou dendogramas, a partir dos quais é possível obter o número de clusters. No presente trabalho, foi utilizado o algoritmo hierárquico aglomerativo. Conforme a Figura 20, apenas para que fosse possível obter os dendogramas abaixo, filtrou-se o dataframe para conter os registros de municípios com população de até 11 mil habitantes (50% dos dados) e com a utilização do método Ward.

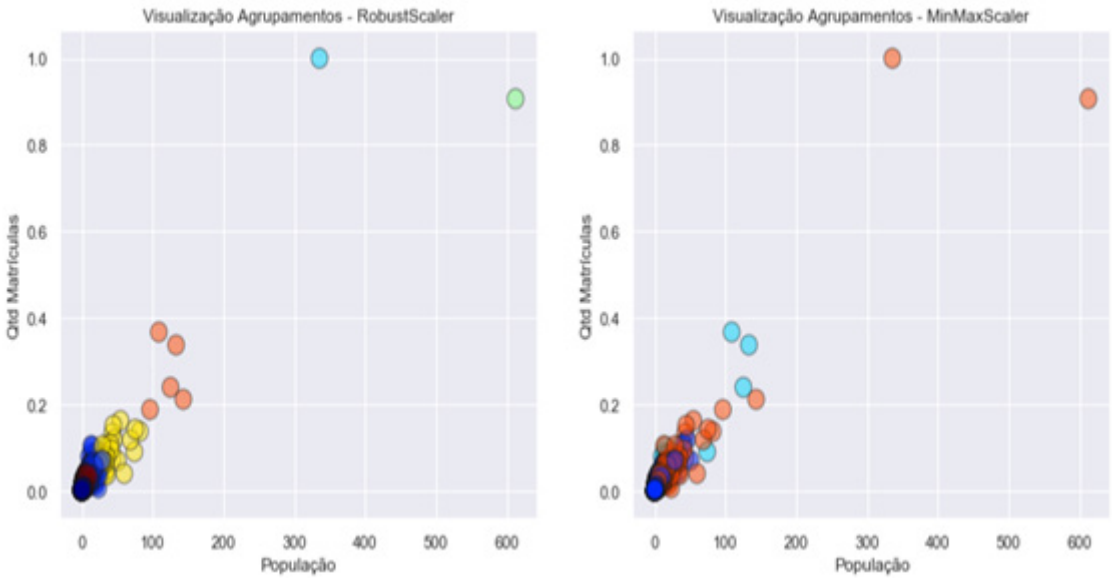
Figura 20 – Geração de dendogramas com método Ward.



Fonte: Elaborada pelo autor (2020)

Apesar das sugestões indicadas pelos dendogramas, seguiu-se a escolha de 7 clusters. A figura seguinte mostra o gráfico resultante da aplicação do Agglomerative Clustering com 7 clusters, com as variáveis população e número de alunos. Claramente, a utilização do RobustScaler traz uma boa separação dos clusters se comparado com MinMaxScaler.

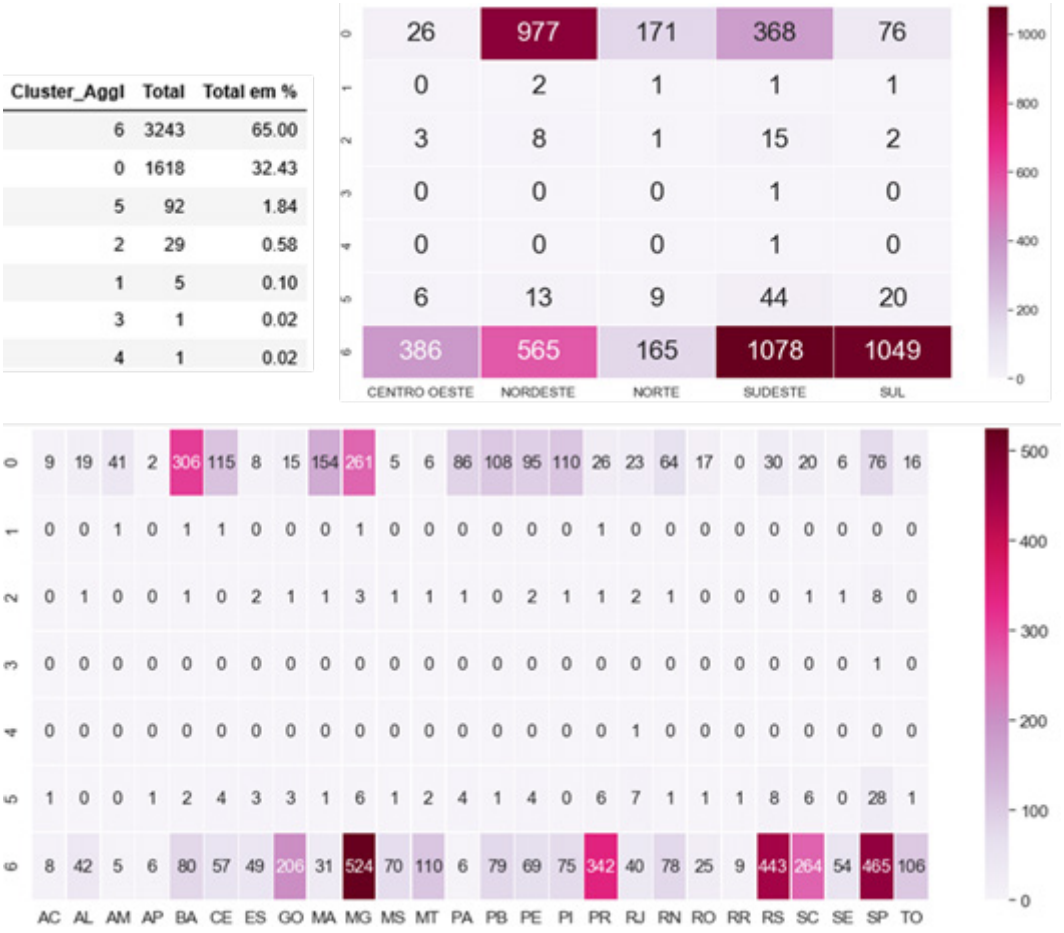
Figura 21 – Geração de gráfico com clusters gerados pelo Agglomerative Clustering



Fonte: Elaborada pelo autor (2020)

A Figura 22 apresenta os resultados da aplicação do Agglomerative Clustering com RobustScaler, exibindo a quantidade de municípios em cada cluster, região e UF.

Figura 22 – Detalhes dos clusters com Aggl. Clustering e escalonamento RobustScaler.



Cluster_Aggl	Total	Total em %
6	3243	65.00
0	1618	32.43
5	92	1.84
2	29	0.58
1	5	0.10
3	1	0.02
4	1	0.02

Fonte: Elaborada pelo autor (2020)

Embora algoritmos hierárquicos sejam simples e eficazes, com facilidade na visualização dos clusters em dendogramas e flexibilidade na escolha do corte para determinar o número de clusters,

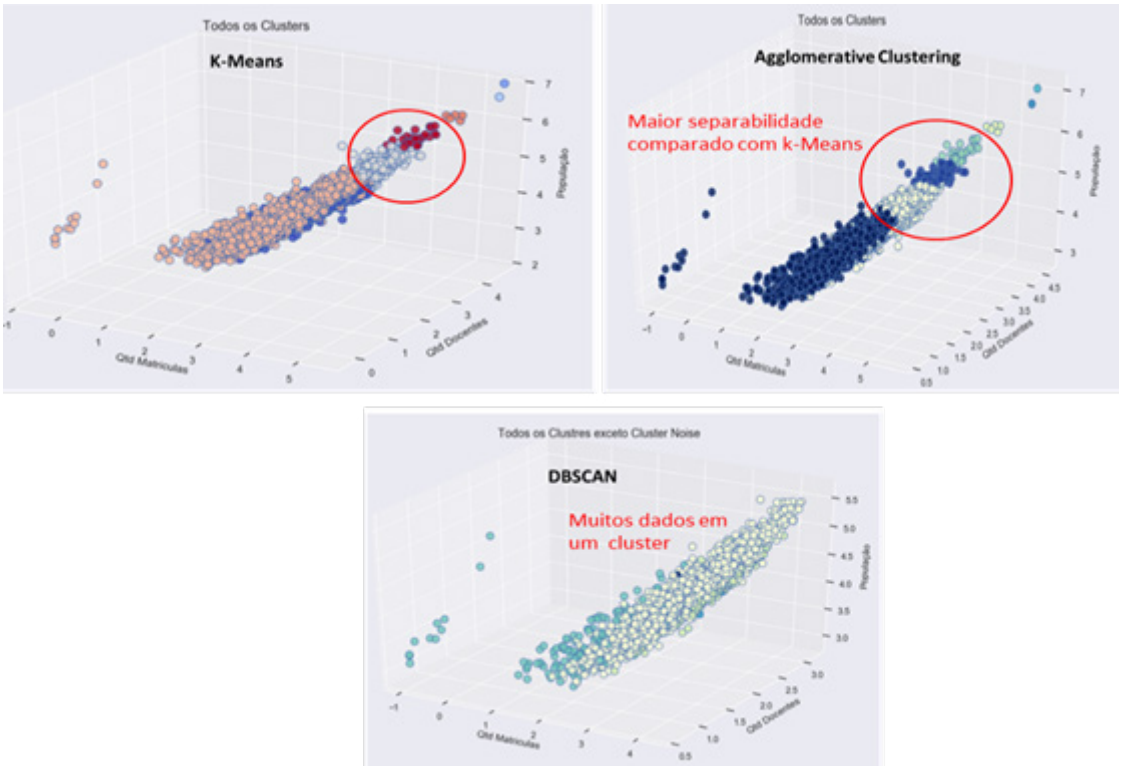
existem alguns dificultadores. Similar ao que ocorre no k-Means, a escolha do número de clusters não é trivial – deve-se decidir onde passar a linha no dendograma para se obter o número de clusters; e, além disso, métodos de dissimilaridade podem produzir resultados bem diferentes, e não há critério objetivo para avaliar qual o melhor método.

5.2.6. Validação dos algoritmos de clusterização

Alguns gráficos produzidos no caderno jupyter (com variáveis normalizadas em log) foram escolhidos para serem apresentados no presente trabalho, com o objetivo de comparar os resultados de alguns algoritmos utilizados. Essa comparação não só procura demonstrar a coerência da separabilidade dos clusters, mas orientou a escolha do algoritmo para a tarefa de detecção de anomalias.

A Figura 23 mostra os clusters formados para os algoritmos k-Means, DBSCAN e Agglomerative Clustering (AgrC), considerando os dados de população, quantidade de alunos matriculados e quantidade de docentes. Os grupos formados pelo k-Means e AgrC são bem parecidos, mas o AgrC apresentou uma melhor separabilidade em áreas densas; o DBSCAN já não é muito adequado, pois agrupou muitos pontos em um mesmo grupo e inseriu os pontos extremos no cluster noise (que não aparecem no gráfico).

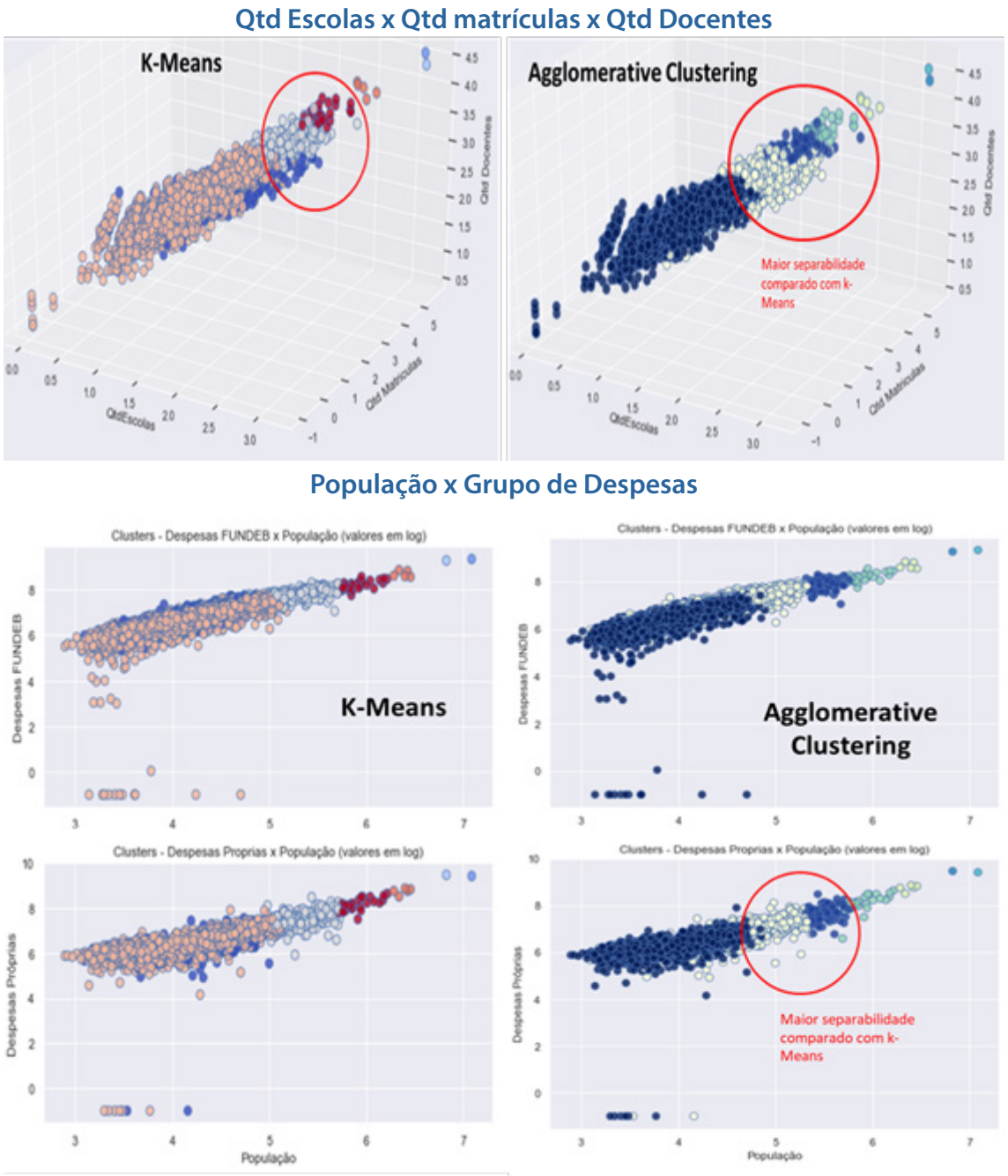
Figura 23 – Comparação dos resultados de alguns algoritmos de clusterização



Fonte: Elaborada pelo autor (2020)

As figuras seguintes comparam apenas os resultados dos algoritmos k-Means e Agglomerative Clustering (AgrC).

Figura 24 – Comparação dos resultados de alguns algoritmos de clusterização



Fonte: Elaborada pelo autor (2020)

De fato, o algoritmo com os melhores resultados, considerando os objetivos do presente trabalho e a geração dos gráficos com diversas variáveis (disponíveis nos cadernos jupyter), é o Agglomerative Clustering com dados escalonados com RobustScaler.

Em virtude da escolha pelo Agglomerative Clustering, foi utilizada, no caderno jupyter, uma medida da avaliação da qualidade da separação entre os clusters – o índice Davies-Bouldin, que compara a distância entre os clusters com o tamanho dos próprios clusters (DAVIES e BOULDIN, 1979). É utilizada principalmente quando os dados não são rotulados. Um valor mais próximo de zero significa um modelo com melhor separação entre os clusters.

Os melhores índices foram alcançados com 2 e 5 clusters (abaixo de 0,4); e o índice não foi satisfatório para 7 clusters (0,69). Entretanto, manteve-se este valor, pois é de interesse um maior número de grupos para a detecção de anomalias locais. Se há poucos grupos, poucos pontos anômalos serão identificados.

5.3. DETECÇÃO DE ANOMALIAS

A detecção de outliers ou anomalias tem como objetivo a identificação de itens inesperados no conjunto de dados, que diferem da norma (GOLDSTEIN e UCHIDA, 2016). Em geral, tais valores discrepantes são removidos antes do uso de algoritmos de mineração de dados. Para o presente trabalho, porém, o objetivo é, de fato, a detecção de despesas discrepantes em um dado grupo de municípios, que podem indicar falhas de preenchimento de registros, possíveis irregularidades ou eventos atípicos que devem ser justificados (uma eventual obra em uma escola, por exemplo).

Neste sentido, o foco é o uso de algoritmos de detecção de anomalias não supervisionados, que usam apenas as informações intrínsecas dos dados para detectar os pontos que se desviam dos demais. Procurou-se utilizar algoritmos baseados em diferentes critérios de detecção - distância, similaridade e densidade.

5.3.1. Delimitação das estratégias para detecção de outliers

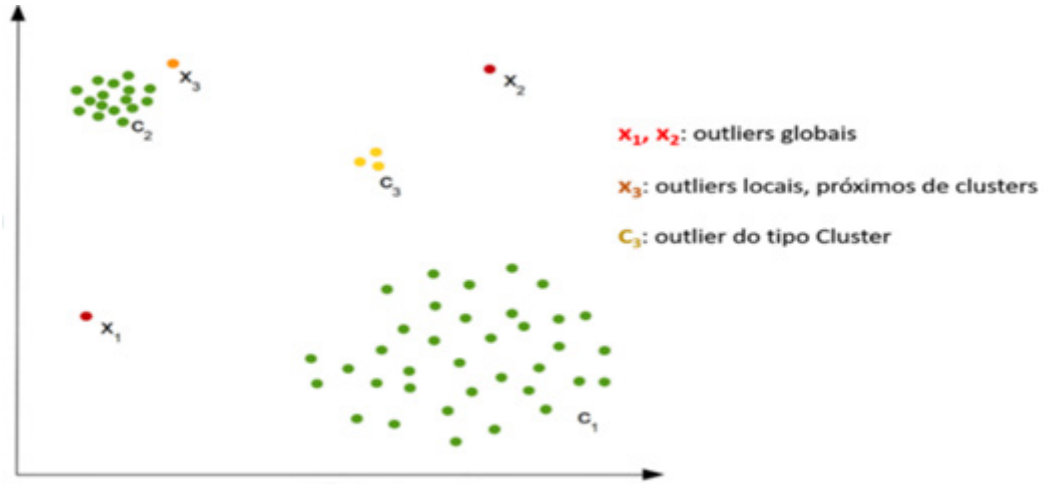
A análise exploratória permitiu a identificação de outliers globais (valores atípicos quando comparados com todo o conjunto de dados). Entretanto, há o interesse de se detectar outros tipos de outliers, listados e representados abaixo.

Tabela 9 – Os tipos de anomalias

Globais	Valores extremos (de baixo ou alto valor) muito diferentes dos demais pontos, facilmente detectáveis por técnicas estatísticas, histogramas e gráficos de dispersão.
Clusters	Poucos pontos juntos em um cluster, bem distantes de outros clusters mais densos, detectáveis pela aplicação de algoritmos de clusterização, como k-Means.
Locais	Poucos pontos próximos aos outros conjuntos densos, detectáveis por algoritmos específicos de detecção de outliers, como LOF. Esses pontos parecem normais, e são percebidos quando se analisa, isoladamente, um determinado cluster.

Fonte: Elaborada pelo autor (2020), adaptado de GOLDSTEIN e UCHIDA (2016).

Figura 25 – Representação gráfica dos tipos de anomalias



Fonte: Elaborada pelo autor (2020), adaptado de GOLDSTEIN e UCHIDA (2016).

Algoritmos de detecção de anomalias geram dois resultados: um rótulo indicando se uma instância é anômala ou não; e uma pontuação (score) que indica o grau de anormalidade (GOLDSTEIN e UCHIDA, 2016). Em algoritmos não supervisionados, as pontuações são mais comuns e permitem

a classificação (ranking) dos pontos mais anômalos. Por meio de um limite (threshold), a classificação pode ser convertida em um rótulo.

5.3.2. A biblioteca Python Outlier Detection (PyOD)

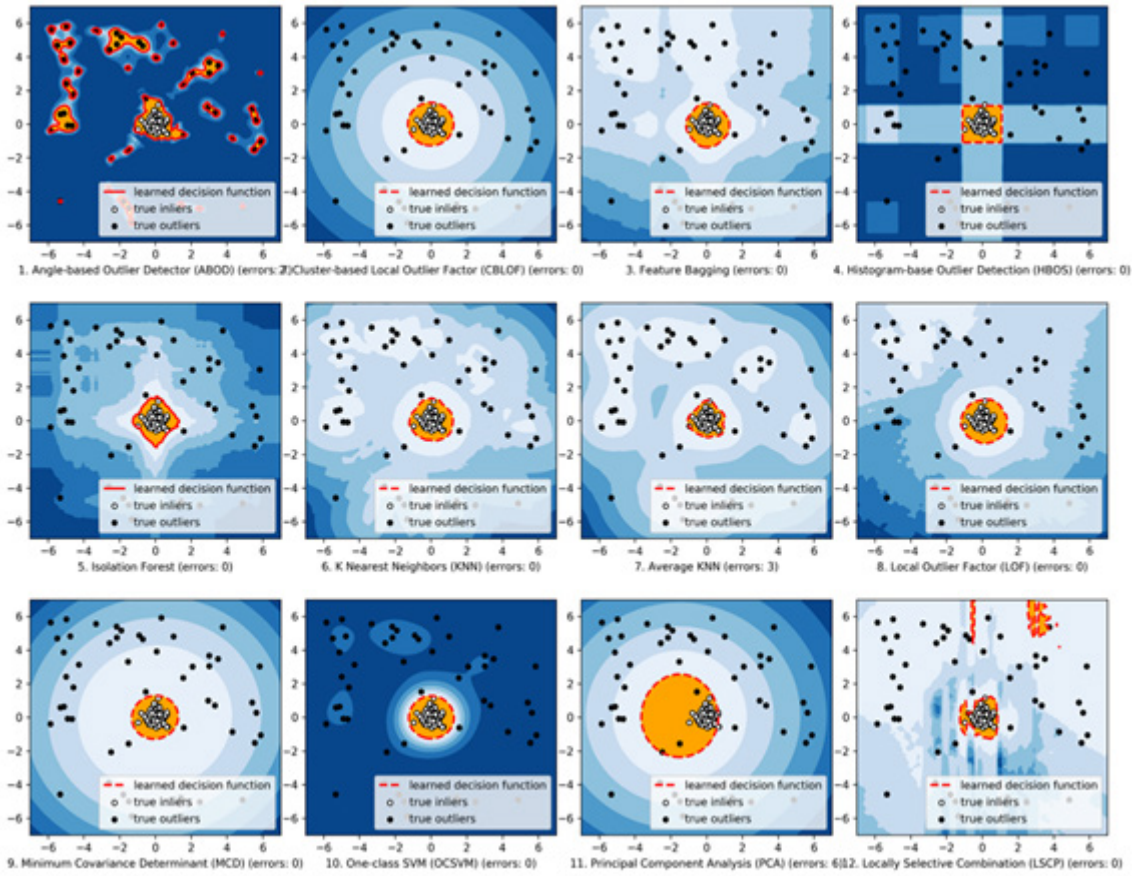
Para identificar as despesas discrepantes dos municípios, foi utilizada a biblioteca Python Outlier Detection (PyOD), que conta com uma variedade de modelos para a detecção de anomalias em dados multivariados (ZHAO, NASRULLAH e LI, 2019). A figura seguinte apresenta gráficos de alguns algoritmos da PyOD (os pontos pretos representam outliers); e a tabela abaixo lista os algoritmos aplicados no presente trabalho.

Tabela 10 – Algoritmos de detecção de anomalias utilizados nos dados

Angle-based Outlier Detection (ABOD)	Probabilístico
Local Outlier Factor (LOF)	Baseado em distância
Cluster-based Local Outlier Factor (CBLOF)	Baseado em distância/ densidade
Histogram-based Outlier Detection (HBOS)	Baseado em estatística
K Nearest Neighbors (KNN)	Baseado em distância
Average KNN (A_KNN)	Baseado em distância
Isolation Forest (IF)	Ensemble
Feature Bagging (FB)	Ensemble

Fonte: Elaborada pelo autor (2020), adaptado de ZHAO, NASRULLAH e LI (2019)

Figura 26 – Listagem de alguns modelos disponíveis na biblioteca PyOD

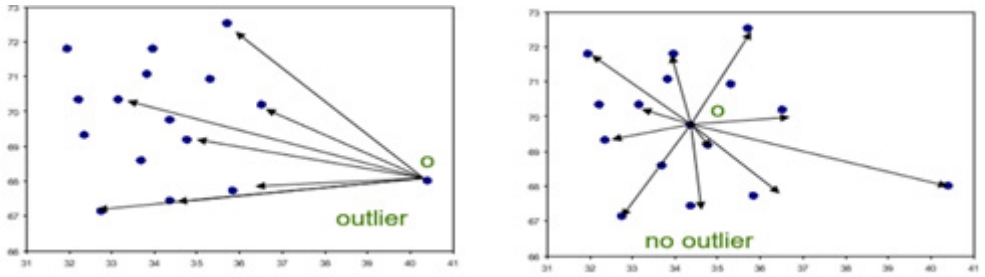


Fonte: ZHAO, NASRULLAH e LI (2019)

O algoritmo ABOD tem como diferencial o uso de medidas baseadas em ângulos. Em espaços

multidimensionais, os ângulos são medidas mais estáveis do que as medidas de distância, e um bom exemplo são as medidas de similaridade baseadas em cosseno para análise de textos (KRIEGEL et al, 2008). Conforme a figura seguinte, um ponto no espaço é considerado um outlier se a maioria dos outros pontos estiver localizada em direções semelhantes.

Figura 27 – Detecção de anomalias no algoritmo ABOD



Fonte: KRIEGEL et al (2008)

O algoritmo LOF foi o primeiro a introduzir a ideia de identificar as anomalias locais (GOLDSTEIN e UCHIDA, 2016). A pontuação da anomalia é baseada numa medida de densidade local e o desvio dessa medida entre um ponto e a dos seus vizinhos. As instâncias normais, com densidades similares de seus vizinhos, contêm uma pontuação próxima de 1; instâncias anômalas, com densidade local substancialmente inferior que a dos seus vizinhos, possuem pontuações maiores. Requer um número para k (quantidade de vizinhos).

O algoritmo CBLOF usa clusterização para determinar as áreas densas nos dados, criar clusters e calcular a sua estimativa de densidade (GOLDSTEIN e UCHIDA, 2016). A pontuação da anomalia de um ponto se baseia no tamanho do cluster ao qual ele pertence (parâmetro desativado por padrão) e na distância do maior cluster mais próximo. Consequentemente, todos os pontos em clusters pequenos, distantes dos clusters maiores, são anômalos (HE et al, 2003) – assim, localiza anomalias globais e de cluster, mas não as locais.

O algoritmo HBOS assume que variáveis são independentes, e calcula o grau de anomalia de uma instância de dado através de histogramas (GOLDSTEIN e UCHIDA, 2016). É considerado um algoritmo simples e de rápida execução, mas com menor precisão.

O algoritmo KNN calcula a pontuação de anomalia com base na distância de um ponto ao k-ésimo vizinho mais próximo, sendo três métodos possíveis de cálculo – a maior distância, a média (A-KNN) ou mediana das distâncias de todos os k-ésimos vizinhos próximos (ZHAO, NASRULLAH e LI, 2019). Detecta, portanto, as anomalias globais, não as locais.

O algoritmo Isolation Forest realiza o particionamento de dados usando uma estrutura de árvores. A pontuação de anomalia considera o quanto isolado um dado ponto está na estrutura, quando poucas partições são necessárias para seu isolamento.

O algoritmo Feature Bagging utiliza vários algoritmos de detecção (podendo ser LOF, kNN e ABOD) em diferentes amostras de um conjunto de dados multivariados, e utiliza medidas de média, ou outras métricas combinadas, para o cálculo da pontuação da anomalia (ZHAO, NASRULLAH e LI, 2019). É uma técnica que procura reduzir a variação entre os algoritmos, a fim de melhorar a eficácia e evitar sobreajustes (overfitting).

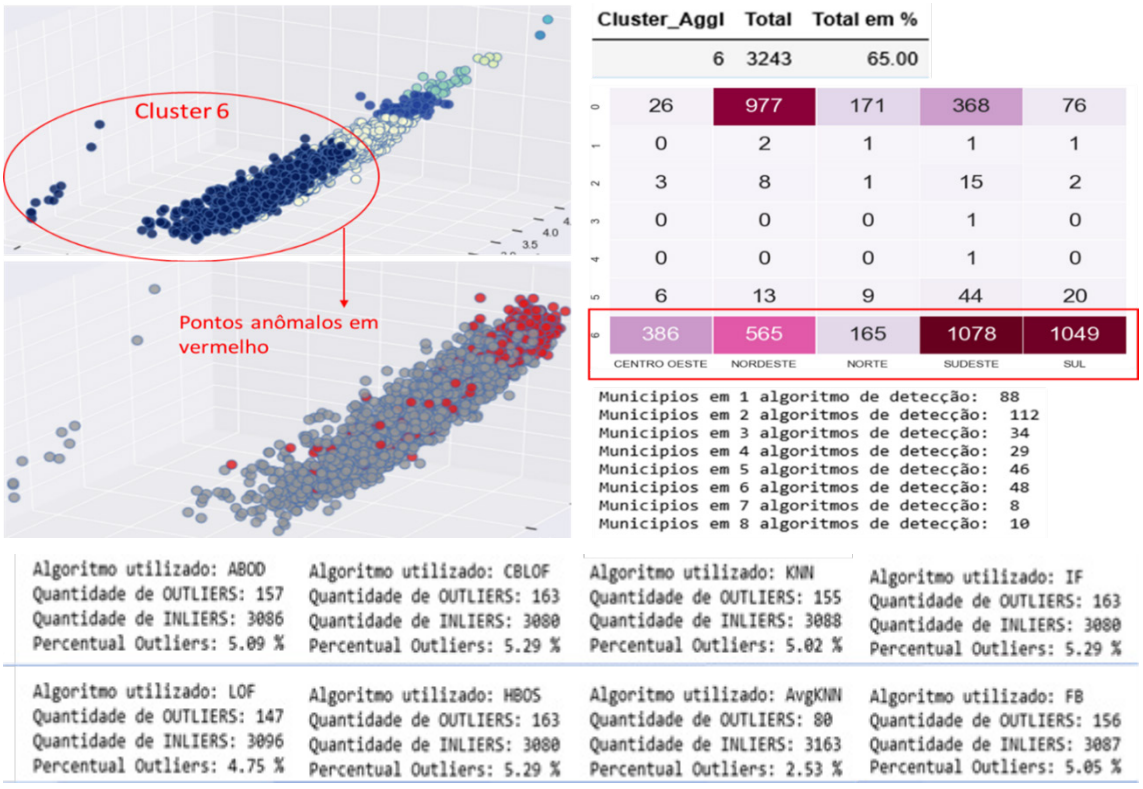
5.3.3. Escolha de Cluster para ser submetido aos algoritmos

Com base nos resultados do Agglomerative Clustering, escolheu-se o cluster 6, que foi submetido aos oito algoritmos de detecção escolhidos, em diferentes escopos de dados: todos os atributos (grupos de despesas, tipos de gasto, programas, subfunções e contas contábeis); despesas Próprias e FUNDEB; e tipos de gasto de remuneração e manutenção. As pontuações de anormalidade e os rótulos de cada município, produzidos por cada algoritmo de detecção, em cada escopo de dados, foram armazenadas no dataframe de municípios.

5.3.4. Resultados da detecção de anomalias

A figura abaixo exibe as características do cluster escolhido, o gráfico de dispersão com a indicação das anomalias e a quantidade de municípios anômalos em cada algoritmo, no escopo de todos os atributos.

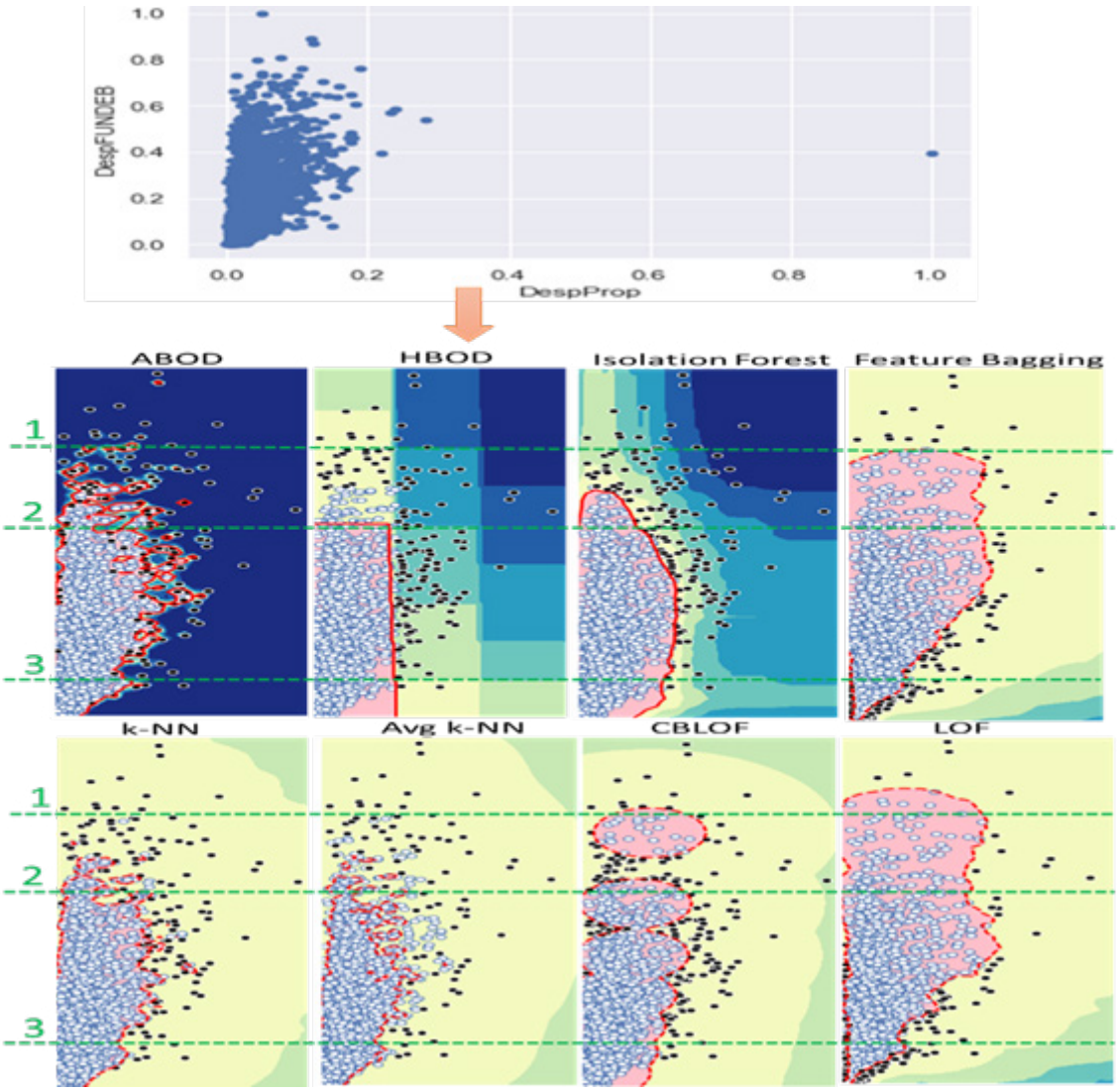
Figura 28 – Indicação e quantidades dos municípios anômalos (escopo: todos atributos)



Fonte: Elaborada pelo autor (2020)

O gráfico abaixo apresenta o gráfico de dispersão das Despesas Próprias com as Despesas FUNDEB, e logo abaixo exibe os resultados dos algoritmos no escopo dos grupos de despesa (ou seja, a entrada para os algoritmos são apenas os dados das despesas Próprias e das despesas FUNDEB). Todos os algoritmos (exceto LOF) consideraram os pontos extremos (alto valor das despesas), acima da faixa pontilhada 1, como anômalos. Entre as faixas pontilhadas 1 e 2, poucos algoritmos (ABOD, CBLOF, kNN) detectaram como anômalos alguns pontos mais internos, localizados entre regiões normais. Abaixo da faixa pontilhada 3, os algoritmos LOF e FB detectaram como anômalos uma grande concentração de pontos bem próximos aos pontos normais.

Figura 29 – Indicação dos municípios anômalos no escopo do grupo de despesas



Fonte: Elaborada pelo autor (2020), adaptado de ZHAO, NASRULLAH e LI (2019)

Algumas considerações se fazem necessárias:

• a linha vermelha representa o threshold (limite) da pontuação do grau de anomalia, estabelecido estatisticamente (score no percentil de 5%) - todos os pontos acima da linha são classificados como anomalias;

• as seis camadas de cores representam faixas de valores da pontuação de anomalia, sendo que a primeira camada se inicia após o valor limite.

5.3.5. Validação dos modelos de detecção de anomalias

É necessário aferir a confiabilidade dos modelos gerados. Há diversas métricas conhecidas para a validação de algoritmos supervisionados, como acurácia, precisão, matriz de confusão e validação cruzada. No caso dos algoritmos de detecção de anomalias, os métodos tradicionais para avaliar a qualidade das pontuações de anormalidade (scoring) são a curva ROC (Receiver Operating Characteristic) e a curva PR (Precision-Recall) – mas apenas quando os rótulos (classes) estão disponíveis (GOIX, 2016).

No presente trabalho, porém, não há rótulos – ou seja, não há dados reais sobre municípios que apresentaram despesas discrepantes em educação. Desta forma, a validação dos modelos gerados,

de forma não supervisionada, não é uma tarefa trivial. Uma alternativa viável foi a comparação das estatísticas e gráficos entre o conjunto de dados normais e os anômalos – para averiguar se tais anormalidades são consistentes.

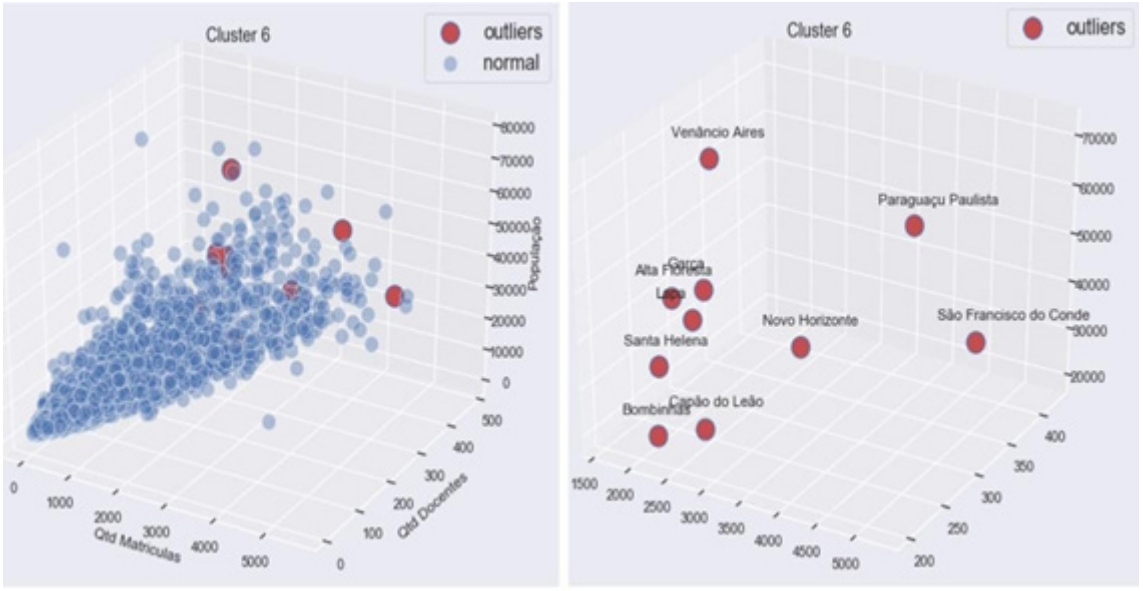
A execução dos 8 algoritmos, em todo o conjunto de dados, apontou 375 municípios anômalos (citados ao menos em um algoritmo), ou 11,5% do total de municípios, sendo 10 municípios anômalos em comum (citados por todos os algoritmos), listados na figura 30.

Realizou-se uma análise simplificada nesses dez municípios anômalos. Conforme estatísticas e histogramas de população e quantitativos de escolas, alunos e professores na figura 31 – são municípios que se encontram acima da faixa do percentil de 75% quando comparados com o conjunto todo. Além disso, os demais histogramas de despesas, de algumas funções e de algumas contas contábeis, nas figuras 32 e 33, apresentaram curvas mais deslocadas à direita – que podem ter influenciado na pontuação da anormalidade para certos municípios.

Em seguida, escolheu-se o município de Bombinhas para uma análise ainda mais detalhada (por ter a menor população), conforme a Figura 34. Comparando-se alguns histogramas de Bombinhas com todos os demais municípios, nota-se claramente que há despesas de valores atípicos quando comparadas com as mesmas despesas de seus semelhantes (os valores para Bombinhas estão representados pela linha vermelha vertical, localizadas ao final das curvas que representam os intervalos dos demais municípios). Tais valores devem ser apresentados aos órgãos de controle para as devidas investigações ou providências.

Pode-se dizer, desta forma, que os modelos gerados são válidos e permitem a identificação de fatos relevantes.

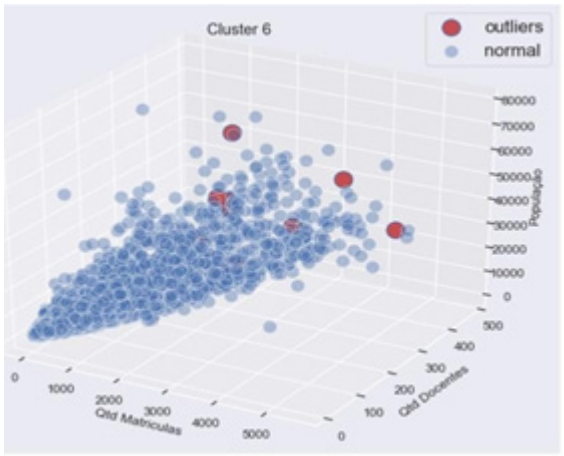
Figura 30 – Os 10 municípios anômalos em todos os algoritmos de detecção

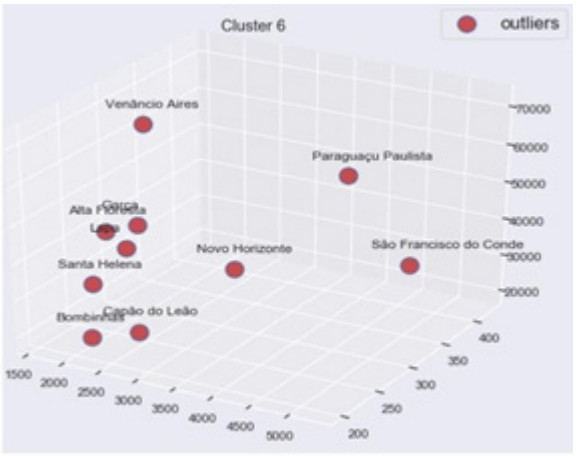


SigUF	NomeMunicipio	Pop_estimada	QtdEscolas	QtdDocentes	NUM_MATR_361	DespFUNDEB	DespVinc	DespProp	tgRemun	tgManut
SC	Bombinhas	19769	11	213	2102	6964060.50	4684067.90	8958981.20	12854264.40	1857706.80
RS	Capão do Leão	25354	13	206	2825	17223486.90	2453994.70	3126054.70	17824240.80	2542848.40
PR	Santa Helena	26490	17	261	1628	8179828.60	10247410.10	4392123.20	13582136.70	7906225.20
BA	São Francisco do Conde	39802	50	323	5288	11535926.70	881183.50	79260447.90	80184416.20	9554227.50
SP	Novo Horizonte	41052	21	248	3691	15192903.70	5539053.60	6177325.00	16685934.40	3971600.70
SP	Garça	44390	25	267	2211	11789378.10	3930066.50	7062476.10	16775069.50	2577109.40
SP	Paraguaçu Paulista	45703	24	423	3587	15898405.40	3986312.90	7336933.90	17539570.30	5233980.20
PR	Lapa	48163	31	201	2738	12753182.10	3036825.20	9656951.70	17016054.40	5073990.60
MT	Alta Floresta	51782	19	199	2491	10673361.60	4575425.50	4919555.80	13954297.20	3654192.30
RS	Venâncio Aires	71554	37	266	2334	16587104.50	4433957.50	4664847.80	19133679.30	1055143.60

Fonte: Elaborada pelo autor (2020)

Figura 31 – Os 10 municípios anômalos – Comparação dos dados IBGE e INEP



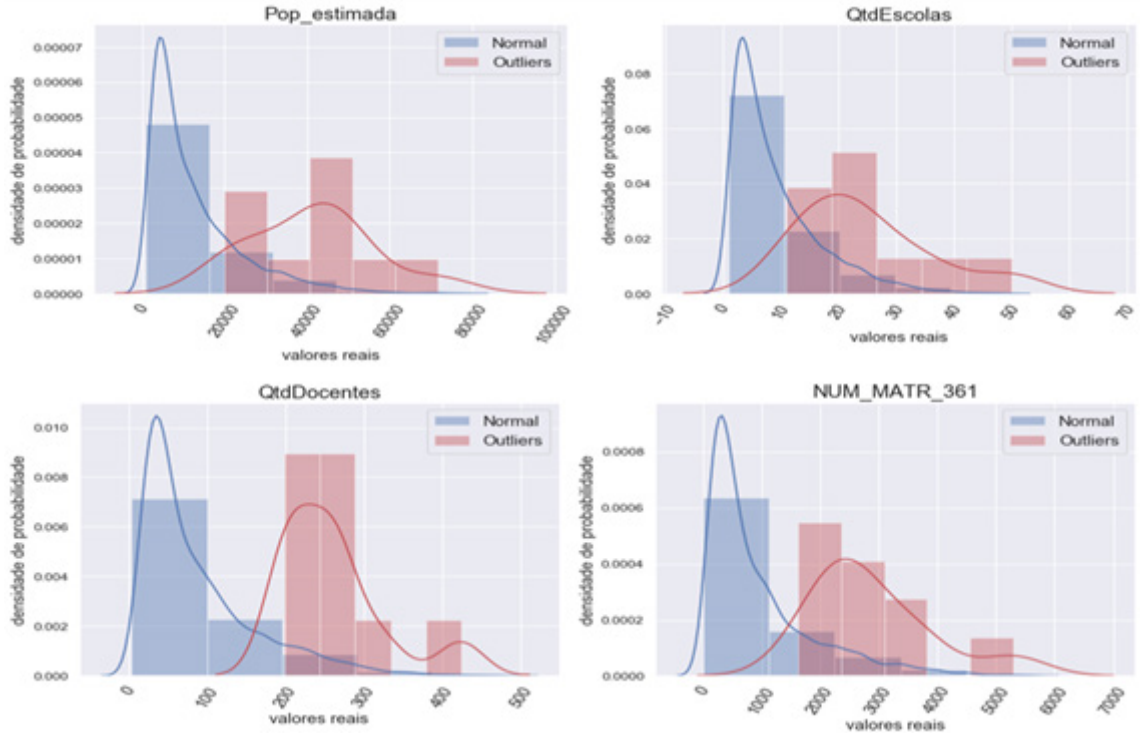


Estatísticas dos municípios não anômalos

	Pop_estimada	QtdEscolas	QtdDocentes	NUM_MATR_361
count	3233.00	3233.00	3233.00	3233.00
mean	12363.98	9.01	87.68	936.48
std	11164.00	7.53	72.89	908.83
min	781.00	1.00	4.00	0.00
25%	4377.00	3.00	35.00	298.00
50%	8347.00	6.00	61.00	587.00
75%	16684.00	12.00	118.00	1242.00
max	78013.00	49.00	482.00	5622.00

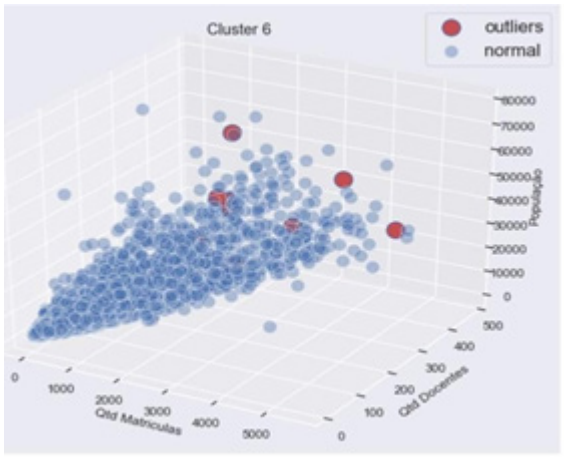
Estatísticas dos municípios anômalos

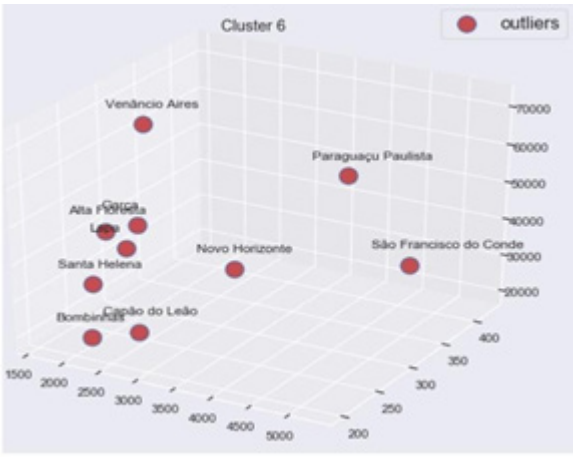
Pop_estimada	QtdEscolas	QtdDocentes	NUM_MATR_361
10.00	10.00	10.00	10.00
41405.90	24.80	260.70	2889.50
15057.67	11.84	69.19	1056.62
19769.00	11.00	199.00	1628.00
29818.00	17.50	207.75	2241.75
42721.00	22.50	254.50	2614.50
47548.00	29.50	266.75	3396.50
71554.00	50.00	423.00	5288.00



Fonte: Elaborada pelo autor (2020)

Figura 32 – Os 10 municípios anômalos – Comparação dos dados de despesas



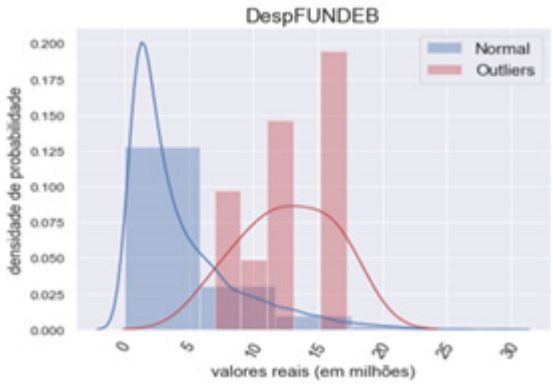


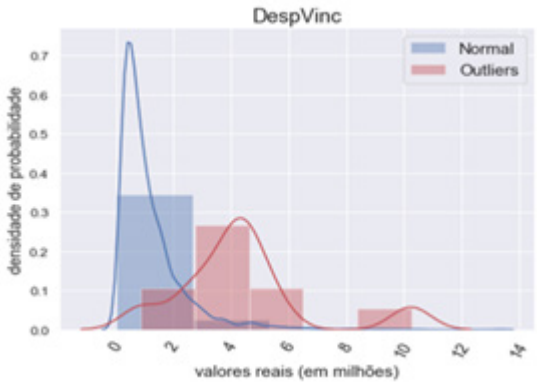
Estatísticas dos municípios não anômalos

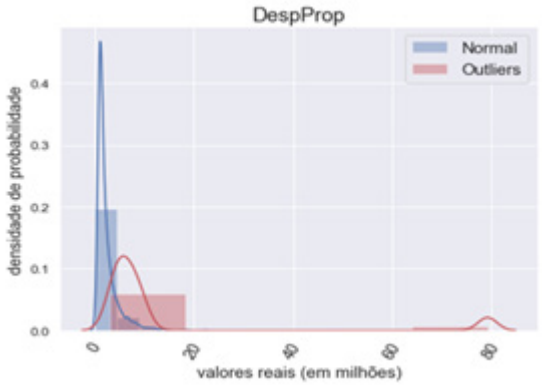
	DespFUNDEB	DespVinc	DespProp	DespVinc	tgRemun
count	3233.00	3233.00	3233.00	3233.00	3233.00
mean	4289156.39	1175863.79	2354977.16	1175863.79	5457233.62
std	4193131.01	1150610.20	2180574.85	1150610.20	4881259.90
min	0.00	2000.00	0.00	2000.00	0.00
25%	1334999.00	453094.20	1068547.60	453094.20	2069091.90
50%	2726222.20	822421.30	1594175.40	822421.30	3660412.80
75%	5878849.90	1486703.00	2788916.00	1486703.00	7243038.10
max	29267721.00	13313226.00	22283660.40	13313226.00	33520827.10

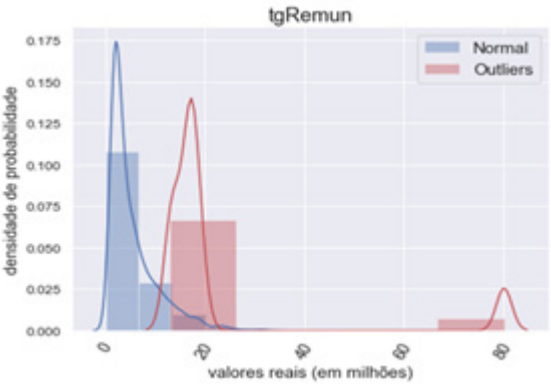
Estatísticas dos municípios anômalos

DespFUNDEB	DespVinc	DespProp	DespVinc	tgRemun
10.00	10.00	10.00	10.00	10.00
12679763.81	4376829.74	13555569.73	4376829.74	22554966.32
3521980.63	2449855.49	23178290.29	2449855.49	20351117.21
6964060.50	881183.50	3126054.70	881183.50	12854264.40
10889002.88	3260135.53	4728524.80	3260135.53	14637206.50
12271280.10	4210135.20	6619900.55	4210135.20	16895561.95
15722029.98	4656907.30	8553469.38	4656907.30	17753073.18
17223486.90	10247410.10	79260447.90	10247410.10	80184416.20



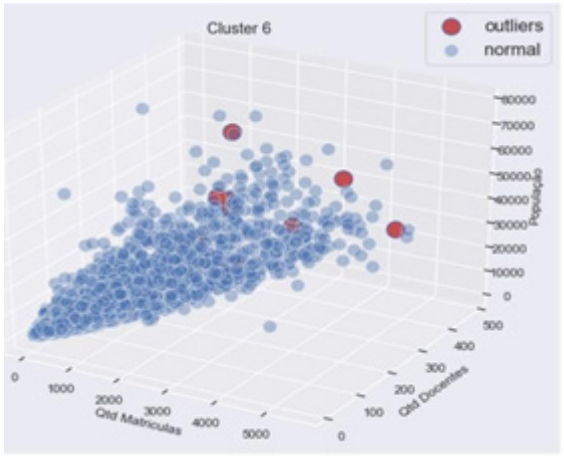


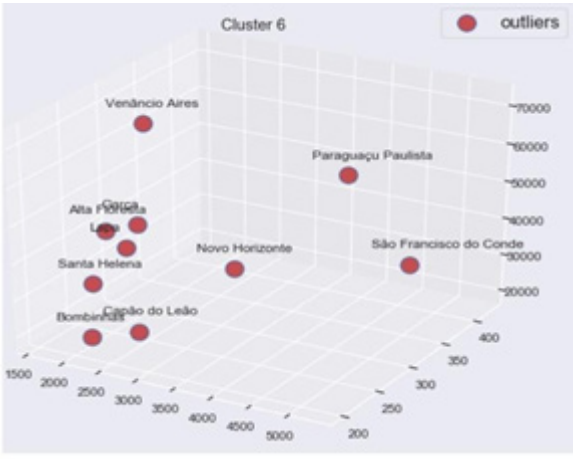




Fonte: Elaborada pelo autor (2020)

Figura 33 – Os 10 municípios anômalos – Comparação das subfunções e contas contábeis



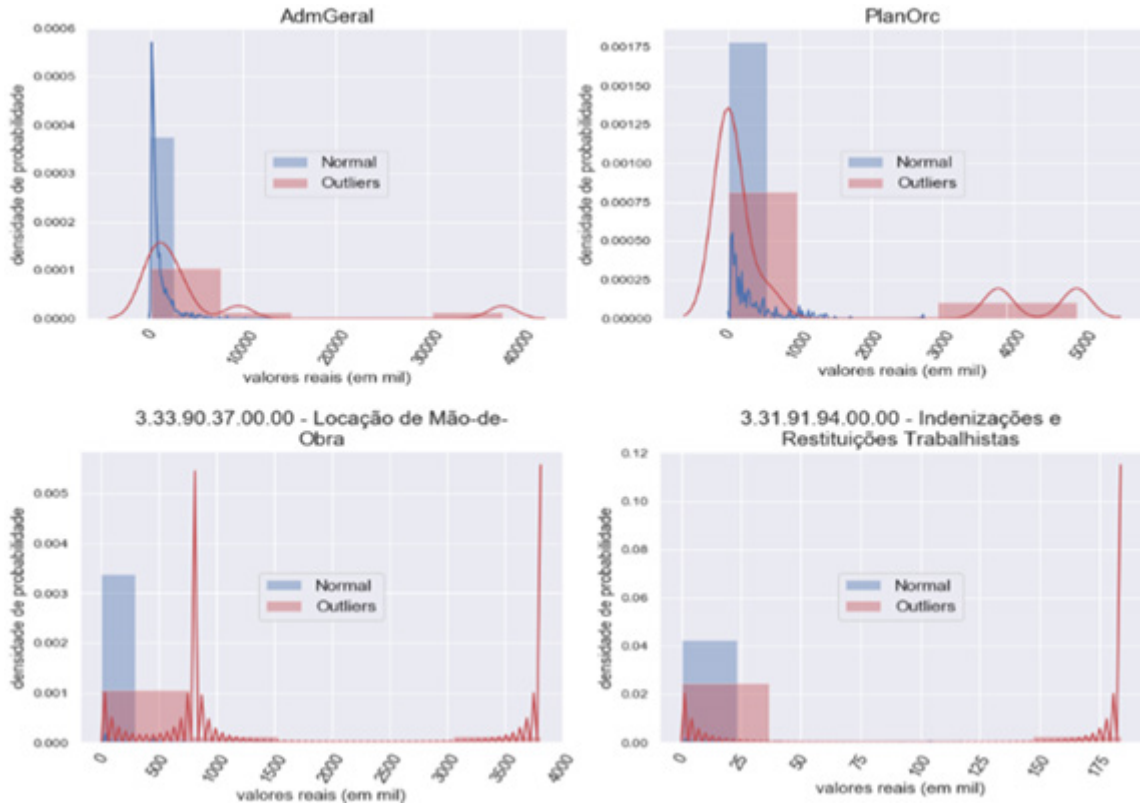


Estatísticas dos municípios não anômalos

	AdmGeral	PlanOrc	Loc Mão de Obra	Inden e Rest Trab
count	3233.00	3233.00	3233.00	3233.00
mean	385891.06	35789.62	2891.97	204.33
std	916691.88	159118.00	46714.57	3661.90
min	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00
50%	0.00	0.00	0.00	0.00
75%	357076.20	0.00	0.00	0.00
max	12942553.90	2756751.70	1475296.60	117179.20

Estatísticas dos municípios anômalos

	AdmGeral	PlanOrc	Loc Mão de Obra	Inden e Rest Trab
10.00	10.00	10.00	10.00	
5825102.61	922495.94	462361.03	18390.74	
11675656.98	1826477.83	1204416.11	58156.63	
0.00	0.00	0.00	0.00	
311712.32	0.00	0.00	0.00	
1485411.80	0.00	0.00	0.00	
3361079.18	412500.00	0.00	0.00	
38060507.70	4887124.40	3812547.40	183907.40	



Fonte: Elaborada pelo autor (2020)

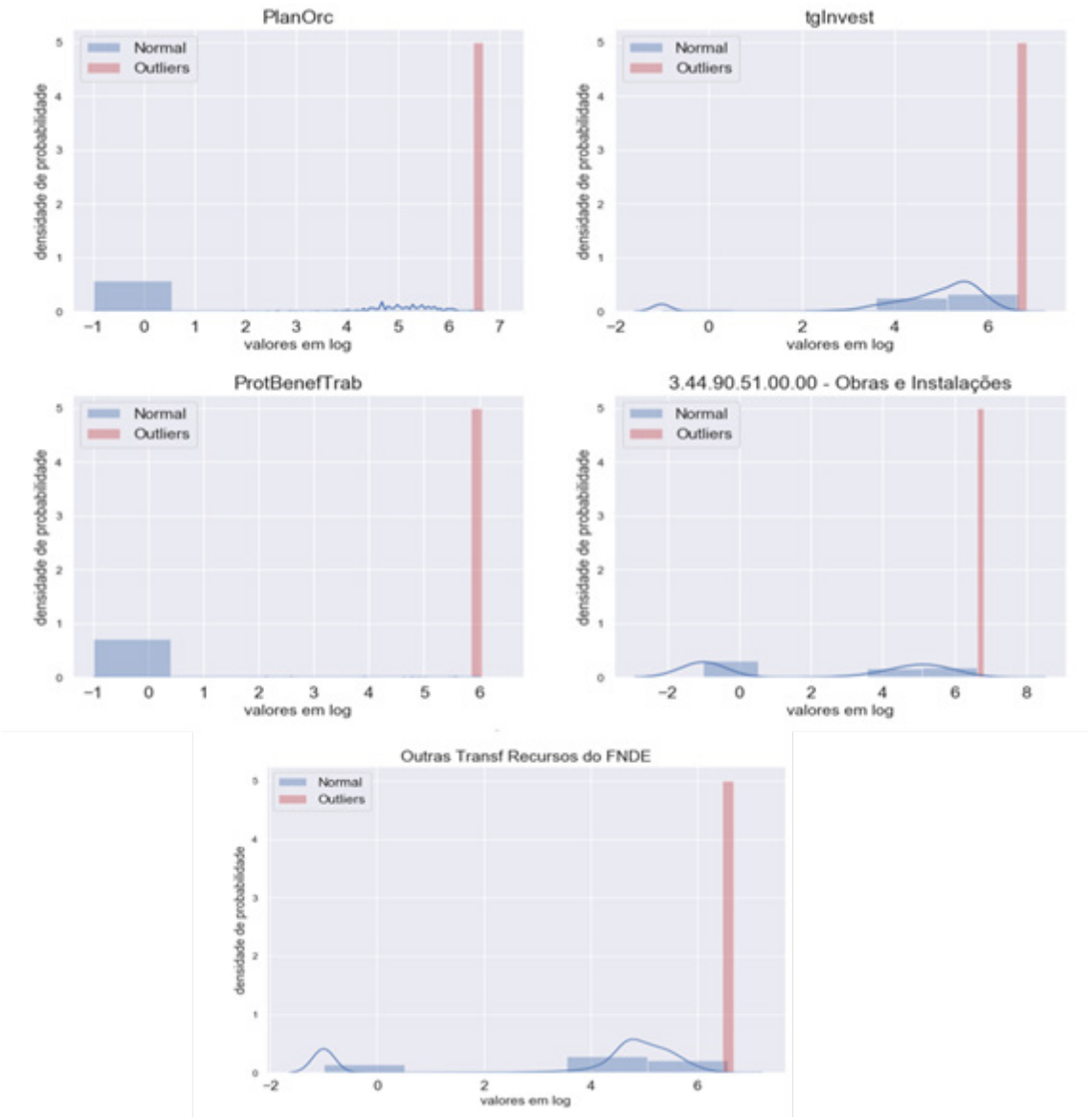
Figura 34 – Um exemplo de um município anômalo

Estatísticas do Cluster 6 (exceto Bombinhas)

	Pop_estimada	QtdEscolas	QtdDocentes	NUM_MATR_361	PlanOrc	tgInvest	ProtBenefTrab	Obras e Instalacoes	Outras Transf Rec FNDE
count	3242.00	3242.00	3242.00	3242.00	3242.00	3242.00	3242.00	3242.00	3242.00
mean	12451.27	9.06	88.18	942.15	37367.36	284108.24	1305.80	133039.89	148096.19
std	11291.43	7.59	73.48	915.48	180532.20	423471.48	30665.23	281902.37	253380.70
min	781.00	1.00	4.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	4384.25	3.00	35.00	298.25	0.00	22180.70	0.00	0.00	7602.82
50%	8399.00	7.00	61.00	591.50	0.00	130872.25	0.00	10029.80	54170.90
75%	16814.25	12.00	119.00	1247.00	0.00	378673.40	0.00	138883.98	171412.65
max	78013.00	50.00	482.00	5622.00	4887124.40	4440136.20	1047799.40	4062826.00	3796703.70

Bombinhas (SC)

Pop_estimada	QtdEscolas	QtdDocentes	NUM_MATR_361	PlanOrc	tgInvest	ProtBenefTrab	Obras e Instalacoes	Outras Transf Rec FNDE
19769	11	213	2102	3787835.00	5273938.50	871000.00	5167917.60	3787835.00



Fonte: Elaborada pelo autor (2020)

6. FASE DE AVALIAÇÃO E IMPLANTAÇÃO

A fase de Avaliação examina se os resultados do modelo atendem aos objetivos de negócio e seus critérios de sucesso; e determina a decisão de o modelo ser implantado ou submetido às novas iterações de fases. A fase de Implantação resume-se em colocar o modelo obtido em produção, in-

cluindo, quando aplicável, a confecção de relatório final.

Detectar anomalias em uma base de dados com centenas de atributos, de forma não supervisionada, mostrou-se uma tarefa complexa, pois há muitas questões desafiadoras, como:

- decidir qual o melhor algoritmo de escalonamento dos dados, bem como o algoritmo de clusterização mais adequado, dada a dificuldade de visualizar dados multidimensionais em apenas duas ou três dimensões;
- definir os atributos mais relevantes como entrada para os algoritmos (nas palavras da CGE-BC, todas as despesas e contas contábeis são imprescindíveis e não devem ser descartadas durante a análise de dados);
- estabelecer o limite (threshold) da pontuação de anormalidade de forma objetiva, visto que a fronteira entre o normal e o anômalo não é precisa;
- identificar os pontos anômalos mais internos, localizados em regiões com dois ou mais agrupamentos; e
- validar a acurácia dos resultados em uma situação sem dados históricos.

Além disso, frente à grande diversidade de algoritmos de detecção (de diferentes critérios – estatístico, distância, densidade, similaridade), é necessário escolher quais aplicar e otimizar os parâmetros de cada um. Alguns algoritmos ainda requerem a definição de critérios específicos, como o número de clusters (CBLOF) e o número de vizinhos próximos (kNN e Average kNN). Devido aos poucos recursos disponíveis (pessoal e tempo), todas estas questões não puderam ser trabalhadas em profundidade adequada.

Por outro lado, pode-se afirmar que o presente trabalho alcançou a finalidade principal de detectar anomalias em despesas dos municípios com o Ensino Fundamental (o dataframe de municípios é atualizado com os rótulos e as pontuações de cada algoritmo), tendo em vista que não havia nenhum trabalho prévio de análise estatística ou de dados no SIOPE pela CGEBC – e foi possível também atingir os objetivos definidos no item 3.5:

- obter ao menos 1% de entes federativos com discrepâncias nos seus gastos educacionais declarados (quando se consideram todos os atributos) – conforme a Figura 28, o percentual de cada algoritmo variou entre 2 a 5%; e
- de fato, alguns entes foram identificados como anômalos em todos os algoritmos de detecção escolhidos (vide Figura 28, foram dez municípios identificados como anômalos pelos oito algoritmos aplicados).

Entretanto, percebeu-se que, para uma maior eficácia desta atividade de detecção, é necessário, ainda, estabelecer algumas estratégias mais individualizadas. Um exemplo seria determinar as despesas mais relevantes – caso fosse alimentação, pode-se cruzar os dados do SIOPE com bases de dados do SigPC (base de prestação de contas do FNDE que contém informações sobre o PNAE), a fim de detectar anomalias em gastos com merenda escolar. Desta forma, um caminho recomendado para o presente trabalho seria voltar à fase de entendimento do negócio para a reformulação de objetivos mais específicos, e apresentar os resultados consolidados em um painel gerencial.

Não obstante, o presente trabalho apresentou uma proposta, um caminho viável e produtivo para se chegar às anomalias nas despesas dos municípios com o Ensino Fundamental – através do uso concomitante de exploração de dados, de algoritmos de clusterização e de algoritmos focados na

detecção de anomalias. É necessário reaplicar tais caminhos para a detecção de anomalias nas demais modalidades de ensino (Educação Infantil, Ensino Médio, etc.). O presente trabalho, portanto, não se encontra finalizado – na verdade, precisa ser submetido às novas iterações de fases, com reformulações de objetivos de negócio e de mineração de dados, e ser apresentado à área de auditoria da Educação Básica. Uma possibilidade seria a formulação de um painel consolidado, com a indicação dos municípios anômalos para cada modalidade de ensino, incluindo, ainda, para cada caso, os atributos envolvidos que influenciaram em uma pontuação maior das anomalias.

7. CONCLUSÃO

Entende-se que explorar os dados de despesas públicas, a fim de obter conhecimento estratégico e de valor agregado, representa um importante objetivo para o controle interno e também para a Administração Pública – pois possibilita, no caso do SIOPE, a identificação de despesas atípicas (anomalias) que podem indicar possíveis falhas ou irregularidades nos investimentos públicos em educação. Consequentemente, vislumbra-se a oportunidade de oferecer os resultados da detecção dessas despesas anômalas como insumos para uma série de atividades, a saber: planejamento das ações de controle (no caso, complementando a Matriz de Vulnerabilidade dos municípios); acompanhamento e adoção de providências cabíveis por parte das instâncias de controle; e criação de trilhas de auditoria voltadas para o monitoramento dos gastos públicos.

Desta forma, o presente trabalho teve como objetivo inicial uma extensa análise exploratória dos dados do SIOPE, cujos resultados delimitaram o escopo das despesas a serem investigadas (ou seja, as despesas municipais pagas no Ensino Fundamental, no ano de 2018) e determinaram as estratégias seguintes – o uso das técnicas de clusterização e detecção de anomalias.

Com base na ideia de que municípios semelhantes devem apresentar despesas educacionais também semelhantes, ao menos em ordem de grandeza – a clusterização buscou agrupar municípios semelhantes, ou seja, similares quanto aos dados de população; quantidades de alunos, docentes e escolas; e indicadores IDEB e IDHM. Várias execuções foram realizadas com diferentes algoritmos (k-Means, DBSCAN, MeanShift e Agglomerative Clustering), diferentes formas de escalonamento dos dados (StandardScaler, RobustScaler e MinMaxScaler) e diferentes parâmetros (número de clusters, epsilon, número de amostras, critério de ligação, etc.). Alguns métodos de cálculo do número ideal de clusters foram também utilizados (métodos Elbow, Gap Statistics e Coeficiente de Silhueta). Como medida da avaliação da qualidade da separação entre os clusters, foi utilizado o índice Davies-Bouldin, o qual foi interpretado em conjunto com gráficos que permitiram a verificação da coerência da separabilidade dos clusters. Apesar de sugestões (2 ou 6 clusters) pelos métodos mencionados, definiu-se como sete a quantidade ideal de clusters – pois já era esperada a criação de dois clusters com apenas um município (SP e RJ), e dois outros cluster com poucos municípios (menos de 1%). Concluiu-se que o Agglomerative Clustering, com dados escalonados com RobustScaler, apresentou os melhores resultados.

Finalmente, deu-se preferência ao cluster de maior número de municípios para submetê-lo a oito algoritmos de detecção de anomalias da biblioteca PyOD. Um parâmetro adicional foi a definição de quais atributos submeter aos algoritmos (escopo de despesa) – ou seja: todos os atributos de

despesas; apenas grupos das despesas próprias e despesas FUNDEB; e tipos de gasto de remuneração e manutenção. Os resultados obtidos (classificação do município, se anômalo ou não; e pontuação da anormalidade, calculada por cada algoritmo em cada escopo de despesa) foram consolidados ao dataframe de municípios. Outras métricas também foram adicionadas, como a quantidade de vezes em que um município foi marcado como anômalo. No escopo de todos os atributos, foi possível identificar as despesas que influenciaram a pontuação da anormalidade (através de histogramas comparativos).

Pode-se afirmar que foi possível, por todo o trabalho, identificar as anomalias globais, locais e de cluster – ou seja, as despesas atípicas de cada município com relação aos seus municípios semelhantes.

O conjunto dos modelos gerados atendeu aos objetivos de negócio (indicar os municípios com discrepâncias nos seus gastos educacionais) e aos critérios de sucesso (obter ao menos 1% de municípios com discrepâncias nos seus gastos educacionais, sendo alguns apontados como anômalos em todos os algoritmos) definidos no item 3.6. Da mesma forma, pode-se dizer que todos os objetivos da mineração de dados (realizar tarefas de AED, de clusterização e de detecção de anomalias) foram também alcançados.

Como recomendações e sugestões para trabalhos futuros, há uma infinidade de possibilidades:

- detectar as anomalias nas despesas do Ensino Fundamental para os demais clusters de municípios, bem como reaplicar as técnicas utilizadas (AED, clusterização e detecção de anomalias em despesas) nas demais modalidades de ensino (Educação Infantil, Ensino Médio, Ensino Superior);
- realizar os mesmos estudos para os dados do SIOPE Estadual;
- realizar análises de dados considerando-se os valores de despesas per capita (dividindo-se as despesas pela população estimada ou pela quantidade de alunos matriculados);
- criar perfil normal de gastos com educação, para cada município ou grupo de municípios semelhantes (envolve analisar os dados do SIOPE dos anos anteriores a 2018) – para possibilitar a detecção mais imediata de anomalias a partir desse perfil de gastos, a cada bimestre (por ser a periodicidade da transmissão de dados do SIOPE);
- propor ou buscar medidas de acurácia dos resultados dos algoritmos não supervisionados;
- consolidar as classificações e pontuações dos algoritmos de detecção de anomalia em uma plataforma de business intelligence, de forma a apresentar painéis gerenciais, com múltiplas visões dos dados, às equipes de auditoria;
- estudar a viabilidade de cruzamento do SIOPE com outras bases de dados dos entes federativos, como a base SIAFEM (trata-se de um sistema equivalente ao SIAFI, porém, no âmbito dos estados e municípios);
- verificar possibilidades de aplicar técnicas de deep learning para a detecção de anomalias, tais como autoencoders e redes generativas adversariais.

REFERÊNCIAS

AGUIAR, Gilson. 2012. **Pior que a corrupção é a má gestão. 2012.** Disponível em: <https://www.cbnmaringa.com.br/noticia/gilson-aguiar-pior-que-a-corrupcao-e-a-ma-gestao>. Acesso em: 20 fev 2020.

ALVES, Gisely. **Aprendizado não supervisionado com K-means.** Disponível em: <https://medium.com/neuronio-br/aprendizado-n%C3%A3o-supervisionado-com-k-means-f4272dee98a0>. Acesso em 20 dez 2019.

ANGÉLICO, Fabiano. **Má gestão + corrupção = nota baixa.** 2012. Disponível em <https://apublica.org/2012/07/ma-gestao-corrupcao-nota-baixa>. Acesso em: 20 fev 2020.

ARCOVERDE, Léo, TOLEDO, Luiz Fernando. CGU aponta uso irregular de quase R$ 51 milhões do Fundeb em todo o país. **G1 Portal de Notícias**, 2019. Disponível em: <https://g1.globo.com/educacao/noticia/2019/08/15/cgu-aponta-uso-irregular-de-quase-r-51-milhoes-do-fundeb-em-todo-o-pais.ghtml>. Acesso em 20 nov 2019.

Associação de Jornalistas de Educação (JEDUCA). **Financiamento da Educação Básica - Guia de Cobertura. São Paulo: Editora Moderna**, 2019. Disponível em: <http://jeduca.org.br/arquivos/Financiamento-da-Educacao-basica-121822.pdf>. Acesso em 02 fev 2020.

Controladoria lança ferramenta para avaliação preventiva e automatizada de editais de licitação. **Governo Federal**, 2015. Disponível em: <https://www.gov.br/cgu/pt-br/assuntos/noticias/2015/06/controladoria-lanca-ferramenta-para-avaliacao-preventiva-e-automatizada-de-editais-de-licitacao>. Acesso em 18 fev 2020

DAVIES, David L.; BOULDIN, Donald W. (1979). "A Cluster Separation Measure" IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224-227. doi:10.1109/TPAMI.1979.4766909.

PROVOST, Foster; FAWCETT, Tom. **Data Science para Negócios. O que você precisa saber sobre mineração de dados e pensamento analítico de dados.** 1.ed. Rio de Janeiro: Alta Books, 2016. ISBN: 9788576089728.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. Brasília, 1988.
Disponível em:
<http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm>. Acesso em: 19 nov 2019.

BRASIL. **Lei nº 9.394, de 20 de dezembro de 1996.** Estabelece as diretrizes e bases da educação nacio-

nal. Brasília, 1996. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L9394compilado.htm>. Acesso em: 19 nov 2019.

BRASIL. **Lei nº 11.494, de 20 de junho de 2007**. Regulamenta o Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação - FUNDEB. Brasília, 2007. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/lei/l11494.htm>. Acesso em: 19 nov 2019.

BRASIL, Controladoria-Geral da União. **Relatório de Fiscalização nº 201602219 - Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (Fundeb)**. Disponível em <https://auditoria.cgu.gov.br/download/12682.pdf>. Acesso em: 19 nov 2019a.

BRASIL, Controladoria-Geral da União. **Relatório de Fiscalização nº 201602218 - Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (Fundeb)**. Disponível em <https://auditoria.cgu.gov.br/download/12685.pdf>. Acesso em: 19 nov 2019b.

BRASIL, Controladoria-Geral da União. **Relatório de Auditoria Anual de Contas nº 201900673 - Fundo Nacional de Desenvolvimento da Educação - Exercício 2018**. Disponível em: <https://auditoria.cgu.gov.br/download/13670.pdf>. Acesso em 20 dez 2019c.

BRASIL, Controladoria-Geral da União. **Orientações para o acompanhamento das ações do Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (FUNDEB)**. Disponível em <https://www.cgu.gov.br/Publicacoes/controle-social/arquivos/fundeb2012.pdf>. Acesso em: 20 dez 2019d.

BRASIL, Controladoria-Geral da União. **Entenda os indicadores** (Matriz de Vulnerabilidade). Disponível em: <https://www.cgu.gov.br/assuntos/auditoria-e-fiscalizacao/programa-de-fiscalizacao-em-entes-federativos/1-ciclo/1o-ciclo/entenda-os-indicadores>. Acesso em 21 dez 2019e.

BRASIL, Controladoria-Geral da União. **Relatório de Fiscalização nº 201902570 - Mata Roma (MA) - Educação e Saúde.** Disponível em <https://auditoria.cgu.gov.br/download/13842.pdf>. Acesso em: 20 jan 2020a.

BRASIL, Controladoria-Geral da União. **Portaria nº 3.553, de 12 de novembro de 2019.** Aprova o Regimento Interno e o Quadro Demonstrativo de Cargos em Comissão e das Funções de Confiança da Controladoria-Geral da União - CGU e dá outras providências. Disponível em <http://www.in.gov.br/web/dou/-/portaria-n-3.553-de-12-de-novembro-de-2019-227654932>. Acesso em 02 fev 2020b.

BRASIL, Fundo Nacional de Desenvolvimento da Educação. **Portal SIOPE: Sistema de Informações sobre Orçamentos Públicos em Educação.** Disponível em: <https://www.fnde.gov.br/fnde_sistemas/siope>. Acesso em 06 ago 2019.

BRASIL, Instituto Brasileiro de Geografia e Estatística. **Estimativas da População.** Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?=&t=resultados>. Acesso em 08 out 2019.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Resultados do IDEB.** Disponível em: <http://portal.inep.gov.br/web/guest/educacao-basica/ideb/resultados>. Acesso em 08 out 2019a.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Resultados do IDEB.** Municípios – Ensino Fundamental Regular – Anos Iniciais.
Disponível em: <http://download.inep.gov.br/educacao_basica/portal_ideb/planilhas_para_download/2017/ divulgacao_anos_iniciais_municipios2017-atualizado-Jun_2019.xlsx >. Acesso em 08 out 2019b.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Resultados do IDEB.** Municípios – Ensino Fundamental Regular – Anos Finais.
Disponível em: <http://download.inep.gov.br/educacao_basica/portal_ideb/planilhas_para_download/2017/divulgacao_anos_finais_municipios2017-atualizado-Jun_2019.xlsx >. Acesso em 08 out 2019c.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Resultados do IDEB.** Municípios – Ensino Médio. Disponível em: <http://download.inep.gov.br/educacao_basica/portal_ideb/planilhas_para_download/2017/divulgacao_ensino_medio_municipios2017-atualizado-Jun_2019.xlsx >. Acesso em 08 out 2019d.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Instruções para utilização dos Microdados do Censo da Educação Básica 2018.** Disponível em: <http://download.inep.gov.br/microdados/microdados_educacao_basica_2018.zip>. Acesso em 12 out 2019e.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Taxas de Transição 2014/2015.** Disponível em: <http://portal.inep.gov.br/web/guest/indicadores-educacionais>. Acesso em: 12 out 2019f.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Notas Estatísticas - Censo da Educação Básica 2019.** Disponível em: <http://portal.inep.gov.br/censo-escolar>. Acesso em 02 fev 2020.

BRASIL, Ministério da Educação. **SIOPE Municipal 2018: Manual de Orientações para o Usuário.** Disponível em: https://www.fnde.gov.br/index.php/centrais-de-conteudos/publicacoes/category/139-siope?download=11869:manual-siope-municipal. Acesso em 06 ago 2019.

BRASIL, Ministério do Planejamento, Desenvolvimento e Gestão. **Portaria no 42, de 14 de abril de 1999.** Atualiza a discriminação da despesa por funções e dá outras providências. Disponível em <http://www.orcamentofederal.gov.br/orcamentos-anuais/orcamento-1999/Portaria_Ministerial_42_de_140499.pdf>. Acesso em 02 fev 2020.

BRASIL. Ministério do Planejamento, Desenvolvimento e Gestão. Secretaria de Orçamento Federal. **Manual técnico de orçamento - MTO 2017**. Brasília, 2016. Disponível em: <http://www.orcamentofederal.gov.br/informacoes-orcamentarias/manual-tecnico/mto_2017-1a-edicao-versao-de-06-07-16.pdf >. Acesso em 02 fev 2020.

BRASIL. Programa das Nações Unidas para o Desenvolvimento. **O que é o IDHM.** Disponível em: <https://www.br.undp.org/content/brazil/pt/home/idh0/conceitos/o-que-e-o-idhm.html>. Acesso em 13 out 2019.

BRASIL. Tribunal de Contas da União. **Acórdão nº 618/2014.** Plenário. Relator: Ministro Valmir Campelo. Sessão de 19/3/2014. Disponível em: <https://pesquisa.apps.tcu.gov.br/#/redireciona/acordao-completo/%22ACORDAO-COMPLETO-1300927%22>. Acesso em: 02 fev 2020.

CHAPMAN, Pete et al. **CRISP-DM 1.0**: Step-by-step data mining guide. Tecnical report.
The CRISP-DM consortium, 2000. Disponível em: <https://pdfs.semanticscholar.org/5406/1a4aa0cb241a726f54d0569efae1c13aab3a.pdf>. Acesso em: 29 jan 2020.

GOLDSTEIN, Markus e UCHIDA, Selichi. **A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data.** PLoS ONE, 11(4): e0152173, April, 2016. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152173>
Acesso em: 23 dez 2019.

GOIX, Nicolas. **How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?** 2016. Disponível em: <https://arxiv.org/abs/1607.01152> Acesso em: 25/02/2020.

HE, Zengyou; XIAOFEI, Xu e SHENGCHUN, Deng. **Discovering cluster-based local outliers.** Pattern Recognit. Lett., vol. 24, p. 1641-1650, 2003.

IBM. **IBM SPSS Modeler CRISP-DM Guide.** Disponível em: <https://www.ibm.com/support/knowledgecenter/SS3RA7_18.2.1/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html>, 2019. Acesso em 29 jan 2020.

KRIEGEL, Hans-Peter., SCHUBERT, Matthias, and ZIMEK, Arthur. **Angle-based outlier detection in high-dimensional data.** In Proceedings of the 14th ACMKDD International Conference on Knowledge Discovery and Data Mining (pp. 444-452). Association for Computing Machinery. Las Vegas, NV, 2008.

LEEK, Jeff. **The Elements of Data Analytic Style**. Leanpub, Victoria British Columbia, 2015.

PEDREGOSA, Fabian et all. **Scikit-learn: Machine Learning in Python.** Journal of Machine Learning Research, 2011. volume 12, p. 2825–2830.

PROJECT JUPYTER. **Project Jupyter®**. Disponível em: <https://jupyter.org>. Acesso em: 30 ago 2019.

PYTHON, Software Foundation. **Python.** Disponível em: <https://www.python.org/>. Acesso em: 30 ago 2019.

QUEIROZ, Christina. **Pesquisa FAPESP: Engrenagem Complexa. Alimentados por arrecadação tributária, regimes de financiamento à educação como o Fundeb, que expira em 2020, constituem desafio ao governo federal.** Disponível em: https://revistapesquisa.fapesp.br/2019/03/12/engrenagem-complexa>. Acesso em 02 fev 2020.

SEABORN. **SEABORN statistical data visualization.** Disponível em <https://seaborn.pydata.org/index.html>. Acesso em: 30 ago 2019.

SEN, Soumya. **Intercluster and Intracluster Distance.** Disponível em: <https://www.geeksforgeeks.org/ml-intercluster-and-intracluster-distance>. Acesso em: 20 dez 2019.

SRIVASTAVA, Shobhit. **Feature Scaling in Scikit-learn.** Disponível em: <https://medium.com/analytics-vidhya/feature-scaling-in-scikit-learn-b11209d949e7>. Acesso em: 20 dez 2019.

THE PANDAS PROJECT. **Pandas - Python Data Analysis Library.** Disponível em: <https://pandas.pydata.org>. Acesso em: 30 ago 2019.

THESING , Ana Paula. **Analytics e Big Data são poderosas armas contra a corrupção.** 2019. Disponível em: <https://www.itforum365.com.br/analytics-e-big-data-sao-poderosas-armas-contra-a-corrupcao>. Acesso em 20 fev 2020.