An Intelligent Model for Generating Indications of Tax Gap in Service Companies

Wellington Franco[1], Elioenai Alves[2], Fábio Sousa[2], Zairo Bastos[1], Vládia Pinheiro[2]

[1]Campus Crateús – Universidade Federal do Ceará (UFC)
Crateús – CE – Brasil

[2]Universidade de Fortaleza (UNIFOR) – Mestrado em Informática Aplicada
Foratelza – CE – Brasil.

wellington@crateus.ufc.br, l.oenaialves@edu.unifor.br

fabiosantos@edu.unifor.br , zairo.vianahd@alu.ufc.br, vladiacelia@unifor.br

Abstract. *Tax gap is still one of the main problems of the Brazilian Tax Administration. In the case of the Service Tax (ISS), recognizing and estimating the tax loss becomes more difficult, as the ISS is a self-assessing tax and the services provided are volatile and cannot be verified after delivery. Aiming to promote tax self-regulation by service companies, this paper proposes a model for generating indications of tax gap, based on forecasting costs and arbitrating the billing of such companies. The model is composed by a committee of Artificial Intelligence (AI) and Data Science (CD) algorithms that infer a probability of a given company presenting outlier behavior. The differential of the model is the possibility of inferring such indications even in the absence of data on the costs of companies. The evaluation of the proposed model was carried out in a case study in the city of Fortaleza. As a result of the experiment, 1,839 service companies, contained in a universe of 22,071 companies, were recognized with strong indication of tax gap, resulting in loss of ISS tax revenue calculated at approximately R$ 10 million.*

Keywords— Tax Gap, Outliers, Classifiers Committee

Resumo. *Evasão fiscal é ainda um dos principais problemas da Administração Tributária Brasileira. No caso do Imposto Sobre Serviço (ISS), reconhecer e estimar a perda tributária se torna mais difícil, pois o ISS é um imposto auto-lançável e os serviços prestados são voláteis e não podem ser verificados após sua prestação. Visando promover a autorregularização fiscal de empresas prestadoras de serviço, este trabalho propõe um modelo de geração de indícios de evasão fiscal, a partir da previsão dos custos e do arbitramento do faturamento de tais empresas. O modelo é operacionalizado por um comitê de algoritmos de Inteligência Artificial (IA) e Ciência de Dados (CD) que infere a probabilidade de determinada empresa apresentar comportamento atípico. O diferencial do modelo é a possibilidade de inferir tais indícios mesmo na ausência de dados sobre os custos das empresas. A avaliação do modelo proposto foi realizada em um estudo de caso no município de Fortaleza. Como resultado do experimento foram reconhecidas 1.839 empresas de serviço, contidas em um universo de 22.071 empresas, com fortes indícios de evasão fiscal, importando em uma perda receita tributária do ISS estimada em, aproximadamente, 10 milhões de reais.*

1. Introdução

Os diversos órgãos da Administração Tributária Brasileira, nos níveis federal, estadual, municipal e distrital, empreendem esforços e recursos para mitigar a evasão fiscal, visando, sobretudo, melhorar a captação de receitas tributárias e promover a justiça fiscal. Evasão fiscal (ou "*tax gap*"), como definida em [Franzoni 1998], é uma deficiência específica da arrecadação aferida pela diferença entre os pagamentos efetivos e a obrigação legalmente prevista.

Especificamente, no âmbito dos municípios brasileiros e do Distrito Federal, destaca-se o Imposto sobre Serviços de Qualquer Natureza (ISSQN ou ISS) como o principal tributo gerido por estes entes federativos. Por exemplo, em Fortaleza, a cidade com o maior PIB do Nordeste, a arrecadação do ISS é responsável por 45% de suas receitas tributárias próprias[1]. A relevância do ISS advém do fato que o setor de serviços representa 70% do PIB do Brasil[2]. No entanto, o ISS é um imposto auto-lançável, ou seja, o próprio contribuinte declara o valor de sua receita bruta, calcula o valor do imposto e realiza o pagamento. Ademais, os fatos geradores do ISS - os serviços prestados a empresas e consumidores finais - são voláteis e, muitas vezes, desaparecem ou não se podem verificar ao fim de sua prestação. Uma possível consequência é que a arrecadação tributária oriunda da prestação de serviços é mais susceptível à evasão fiscal e de difícil auditagem a posteriori. Como exemplo podemos citar empresas de serviço de cuidados pessoais, estética, de conserto mecânico, de educação, dentre tantas outras. É difícil rastrear, verificar, ou auditar o serviço que foi efetivamente realizado em um procedimento estético realizado em uma pessoa ou em um conserto mecânico realizado em um automóvel.

Visando dirimir a sonegação de receita tributária proveniente do ISS, através de ações de auditoria e de promoção da autorregularização fiscal, este trabalho propõe um modelo de geração de indícios de evasão fiscal em empresas prestadoras de serviço, a partir da previsão dos custos e do arbitramento do faturamento de tais empresas. O modelo é operacionalizado por um comitê de algoritmos da área de Inteligência Artificial (IA) e Ciência de Dados (CD) que, a partir de dados fiscais e financeiros, infere a probabilidade de determinada empresa apresentar comportamento anômalo. O diferencial do modelo é a possibilidade de inferir indícios de evasão fiscal mesmo na ausência de dados sobre todos os custos operacionais e todas as operações de compra e venda das empresas. Para isso, usa-se uma regra heurística segundo a qual duas empresas que prestam serviços de mesmo tipo (por exemplo, duas oficinas mecânicas), de mesmo porte (p.ex., considerando a área edificada do imóvel ou consumo de energia elétrica), e localizadas na mesma região (p.ex. mesmo bairro ou distrito) devem apresentar um padrão de receitas e/ou despesas similar.

O modelo foi executado e avaliado em empresas de serviço do município de Fortaleza de seis segmentos econômicos: Academias, Escolas, Hotéis, Lavanderias, Oficinas, e Salões de beleza, considerando os dados coletados no período de Janeiro a Junho de 2022, e somente para os principais bairros da cidade. Como resultado, gerou-se indícios de evasão fiscal (anomalia do tipo 3) para 1839 empresas que totalizaram R$ 200 milhões de faturamento, possivelmente não declarado, importando em, aproximadamente, R$ 10 milhões de perda de receita tributária do ISS. Os dados gerados pelo modelo foram validados por auditores da Secretaria Municipal das Finanças do Município de Fortaleza (SEFIN).

Este artigo está estruturado como descrito a seguir. A seção 2 apresenta os trabalhos relacionados. O modelo de geração de indícios de evasão fiscal é detalhado na seção 3. A seção 4 apresenta os experimentos e os resultados do estudo realizado em empresas de serviço do município de Fortaleza. Por fim, na última seção, apresentam-se as considerações finais e os trabalhos

[1]Considerando a média da arrecadação tributária própria da cidade de Fortaleza nos últimos três anos - 2020, 2021 e 2022

[2]https://www.ibge.gov.br/explica/pib.php

futuros.

2. Trabalhos Relacionados

Os controles tributários envolvem tanto verificações puramente formais, baseadas em dados e elementos que podem ser deduzidos diretamente da declaração fiscal, quanto procedimentos mais complexos. Entre estes últimos, as administrações fiscais realizam (1) verificações dos dados da declaração contra outros dados que permitem presumir a veracidade dos dados declarados (tais como os apresentados por outros contribuintes ou recolhidos por outras autoridades fiscais) e (2) investigações por meio de inspeções, auditorias e solicitações de informações destinadas a detectar evasão ou fraude fiscal. Na literatura existem diversos trabalhos de prevenção e detecção de anomalias contábeis, focadas em tributos diferentes e com o uso de estratégias diferentes [Dias and Becker 2017, Oliveira 2019, de Vasconcelos Soares and Cunha 2020, Xavier et al. 2022, Oliveira et al. 2022].

No trabalho de [Oliveira 2019] buscou-se investigar o uso de redes neurais artificiais para identificar os riscos de inadimplência fiscal, utilizando a base de dados do cadastro fiscal do Distrito Federal focando no ICMS. O trabalho apresentou como resultado o uso de dois modelos de predição: regressão logistica e redes neurais do tipo perceptron multicamadas. O modelo de regressão logística se mostrou eficiente para identificar as variáveis mais importantes e ajudar a entender os resultados da rede neural. O resultado alcançado foi um modelo que obteve, na tarefa de predição, uma taxa de erro menor que 11%.

[Xavier et al. 2022] propõem uma solução inteligente capaz de identificar perfis de potenciais sonegadores utilizando apenas dados abertos e públicos fornecidos pela Receita Federal e pelo Conselho de Administração Tributária do Estado de Goiás, além de outros cadastros públicos. Como resultado, o trabalho obteve mais de 98% de precisão na previsão do perfil padrão. Por fim, foi criada uma solução de software de visualização para ser utilizada e validada pelos auditores fiscais do Estado de Goiás.

[Oliveira et al. 2022] apresenta uma abordagem para identificação de incongruências entre o tipo dos itens da licitação (produtos e serviços) e a atividade econômica dos licitantes que participam de processos de compras públicas, usando técnicas de Processamento de Linguagem Natural (PLN) (e.g., uma empresa do ramo de peças automotivas participando de licitações envolvendo apenas gêneros alimentícios). Como resultado, apresentou acurácia de 60,87%.

Em relação ao ISS, podemos citar os trabalhos de [Dias and Becker 2017] e [de Vasconcelos Soares and Cunha 2020]. O trabalho de [Dias and Becker 2017] busca identificar empresas, através de um sistema baseado em aprendizado não supervisionado, que apresentam mudança de endereço tributário para fins de evasão fiscal. Como estudo de caso, é mostrado um estudo feito pela Procempa, Empresa de Informática de Porto Alegre, em parceria com a Secretaria Municipal da Fazenda da cidade. O estudo de caso na cidade de Porto Alegre relatou precisão de 80% a 90%. O trabalho de [de Vasconcelos Soares and Cunha 2020] utiliza técnicas de aprendizado de maquina para prever o risco de empresas adimplentes ou inadimplentes se tornarem devedoras contumazes no próximo exercício fiscal. Os modelos foram desenvolvidos e testados em um conjunto de dados com cerca de 60 mil registros de declarações fiscais agrupadas trimestralmente ao longo de cinco anos. Os resultados preliminares mostram que o modelo desenvolvido identifica esse tipo de comportamento irregular com 85,21% de precisão.

3. Modelo de Geração de Indícios de Evasão Fiscal em Empresas de Serviço

Este trabalho propõe um modelo baseado em técnicas de Inteligência Artificial e Ciência de Dados para reconhecimento de indícios de evasão fiscal em empresas prestadoras de serviço. O princi-

pal diferencial do modelo é o uso de uma heurística de similaridade que permite inferir empresas com comportamento anômalo mesmo sem todos os dados referentes aos custos e faturamento de determinadas empresas. Por exemplo, oficinas de grande porte, localizadas em um mesmo bairro, devem apresentar uma relação similar entre as receitas e despesas realizadas em determinado período. Com base nesta heurística, é possível inferir o valor de custos e/ou faturamento que não foram recuperados das bases de dados oficiais. Dessa forma, esperamos que o modelo aprimore a fiscalização dos auditores fiscais ajudando no processo de autorregularização.

A Figura 1 apresenta as dimensões do modelo proposto e os conjuntos de dados previstos. As dimensões permitem direcionar a seleção das empresas e dos dados a serem analisados. São elas:

- **Segmento Econômico -** define quais segmentos econômicos serão analisados pelo modelo, por exemplo, escolas e oficinas. Importante salientar que as empresas são comparadas somente com outras empresas de mesmo segmento. Os segmentos econômicos são definidos pelos códigos dos itens lista de serviços;
- **Localização Geográfica -** define o nível de localização geográfica - bairro, distrito, quadra, setor censitário, etc. - que as empresas serão analisadas. Esta dimensão considera que as dinâmicas sociais e econômicas por regiões da cidade, pois, dependendo do bairro em que a empresa está localizada, há um aumento de custos com impostos (p.ex. IPTU), com logística de entrega e até concorrência de mercado;
- **Porte/Tamanho -** nesta dimensão é definido se o porte das empresas baseia-se na área edificada do imóvel onde a empresa está instalada ou em algum custo operacional, p.ex., energia elétrica;
- **Tempo -** define o corte temporal da análise, pois, devido a sazonalidade, alguns serviços podem apresentar importantes variações de custo e faturamento;
- **Receitas/Despesas -** esta dimensão define os conjuntos de dados a serem importados para o modelo - dados de faturamento (Escrituração Fiscal, Simples Nacional, de Operações Financeiras, Arbitrado) e dos Custos (pessoal, impostos, compras, despesas de custeio, etc.). Estes dados são coletados, normalmente, das seguintes bases de dados: CPBS – Cadastro de Produtores de Bens e Serviços; CIM – Cadastro Imobiliário Municipal; NFSe – Notas Fiscais de Serviço Eletrônicas; DAM – Documentos de Arrecadação Municipal; Energia Elétrica – gastos das contas de energia elétrica da concessionária; Operações Financeiras - dados de recebimento através de operações financeiras como cartão de crédito, de débito, por pix ou por boleto, etc.; PGDAS – dados do Simples Nacional; Folha de pagamento – dados de pagamento de pessoal, contribuições e impostos relacionados; Água / Comunicação – dados de custeio com água e comunicação (telefonia e Internet); Tributos Federais - dados dos tributos como Imposto de Renda e outros; Compras – dados de aquisição de mercadorias.

Na Figura 2, ilustra-se a arquitetura do modelo proposto, que é composta das seguintes etapas para geração dos indícios de evasão fiscal:

1. **Entrada de dados** - o modelo recebe conjuntos de dados a partir de fontes internas (secretarias de finanças municipais) e externas (secretarias da fazenda estaduais, Receita Federal do Brasil (RFB), instituições financeiras, concessionárias e empresas de água, esgoto, energia elétrica e comunicação. A complexidade desta etapa reside, principalmente, no

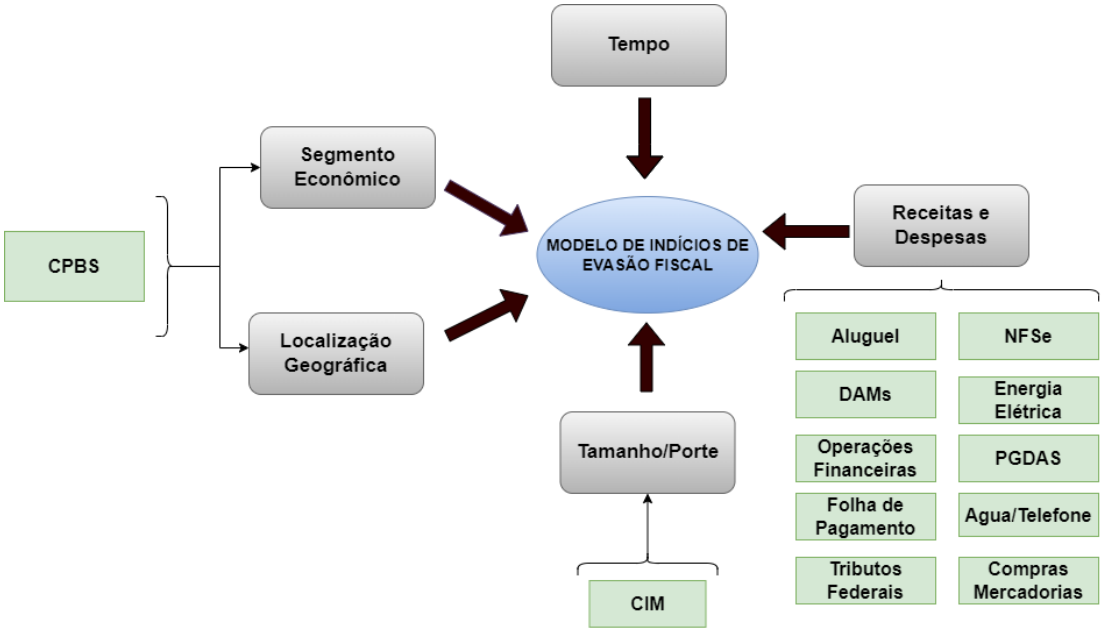

Figura 1. Dimensões para Seleção e Análise das Empresas Prestadoras de Serviço.

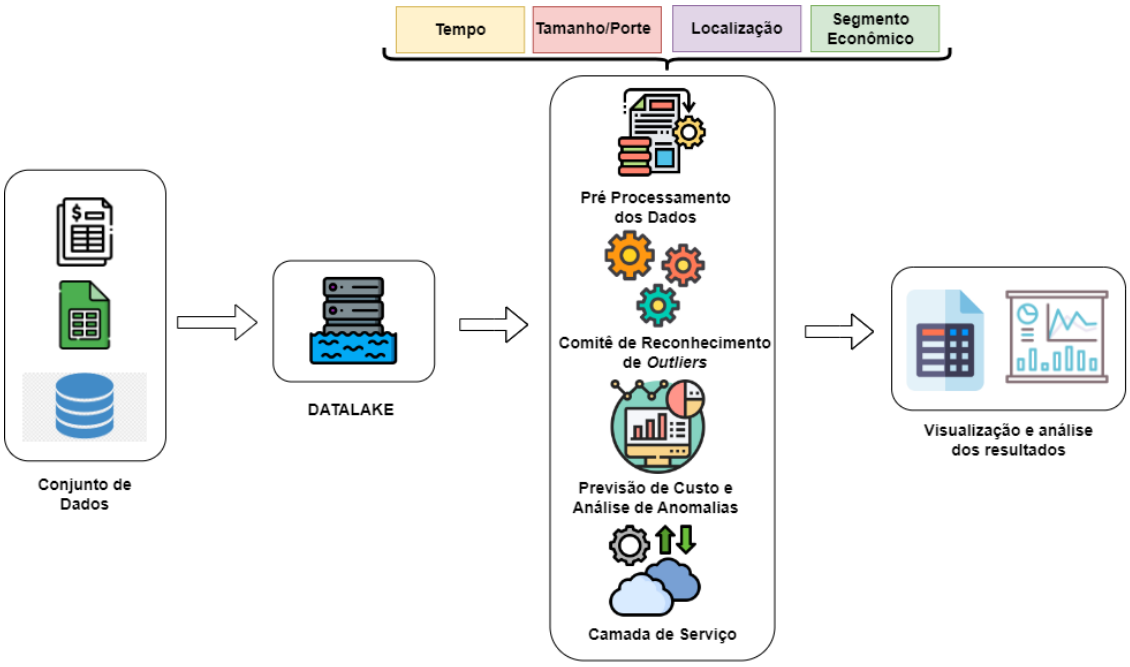

Figura 2. Arquitetura do Modelo Inteligente de geração de indícios de Evasão Fiscal.

processo de extração e transformação e carga (*ETL - Extract, Transform and Load*), visto que os dados e informações são coletados em diversos formatos (bases de dados, planilhas, csv, etc), em níveis de agregação distintos. Por fim, um módulo de ETL para cada conjunto de dados é desenvolvido para gerar um *datalake* com os dados de entrada do modelo.

2. **Pré-processamento** - a partir dos dados dispostos no *datalake* são desenvolvidas e aplicadas rotinas computacionais para cálculo, seleção, e integração dos dados. Alguns novos dados devem ser calculados e normalizados, por exemplo, com base na área edificada define-se a classificação do porte da empresa pequeno, médio ou grande, e com base no valor de mercado do imóvel estima-se o valor da despesa de aluguel. Em seguida, os dados são selecionados de acordo com as dimensões segmento econômico, localização geográfica, tamanho/porte e tempo. Por fim, para integração dos dados usa-se a chave de relacionamento por CNPJ. No entanto, é necessário aplicar algumas estratégias para melhorar a completude da integração. Por exemplo, nos dados de consumo de energia elétrica nem sempre a conta de energia está associada ao CNPJ da empresa, então busca-se integrar pelo CPF dos sócios ou representates legais da empresa em conjunto com o endereço. Em Fortaleza, apenas 9,72% das empresas foram associadas a seus dados de consumo de energia pela chave primária do CNPJ. Após aplicação da estratégia acima, este percentual subiu para 28,15%.

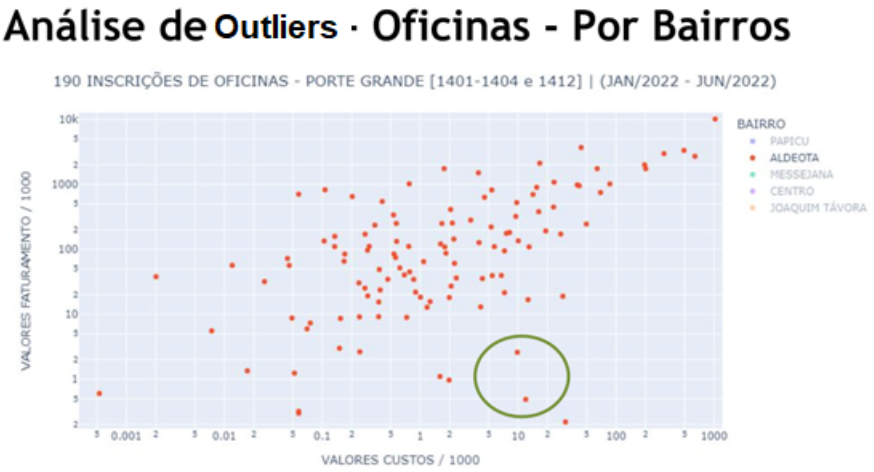

Figura 3. Gráfico de Dispersão de empresas do segmento econômico "*oficina*".

3. **Comitê de Reconhecimento de *Outliers*** - nesta etapa são inferidos padrões de faturamentos (receitas) e custos (despesas) das empresas de serviço, considerando as dimensões espacial, temporal, atividade econômica e porte das empresas, e, em consequência, são identificadas empresas que apresentam comportamento atípico em relação aos padrões. Por exemplo, na Figura 3 é apresentado um gráfico de dispersão de oficinas de grande porte, localizadas no Bairro Aldeota, na cidade de Fortaleza, considerando as variáveis faturamento e custos. No destaque, tem-se duas empresas (circuladas) que apresentam custos similares ao padrão, mas que apresentam baixo valor de faturamento em relação ao padrão das oficinas do bairro, no período de Janeiro a Junho/2022. Para descoberta de empresas com comportamento atípico (*outliers*), seja de custos ou de faturamento, aplica-se um um comitê de algoritmos e técnicas de IA e CD, detalhados a seguir:

    - **Boxplot -** de acordo com [Jiawei et al. 2012], o boxplot, também conhecido como gráfico de caixa, consegue resumir e exibir a distribuição de um conjunto de dados de maneira eficiente, além de identificar valores atípicos que divergem da média e não obedecem ao comportamento geral dos dados. Os boxplot mostram visualmente a distribuição de dados numéricos e assimetria por meio da exibição de quartis de dados (ou percentis) e médias. Ao usar um boxplot, um *outlier* é

definido como um ponto de dados localizado fora dos bigodes do boxplot. Por exemplo, fora de 1,5 vezes o intervalo interquartil acima do quartil superior e abaixo do quartil inferior (Q1 - 1,5 * IQR ou Q3 + 1,5 * IQR);

- **Regressão linear -** para [Pedregosa et al. 2011], a regressão linear é um modelo supervisionado que busca estabelecer uma relação linear entre uma variável dependente (Y) e uma ou mais variáveis independentes (X). A regressão linear pode ser expressa em termos de matrizes como $y = X\beta + \alpha$, onde y é o vetor $n \times 1$ de valores de resposta observados, X é np matriz de precursores (matriz de design), β contém os coeficientes de regressão $p \times 1$ e α é o vetor de termos de erro $n \times 1$. Para encontrar os coeficientes, foi usado o método dos mínimos quadrados ordinários que minimiza a soma das distâncias quadradas para todos os pontos desde a observação real até a superfície de regressão e calcula uma medida de confiança para estimar os *outliers*. A medida de confiança $Z - score$ calcula o intervalo em que se espera que os reais valores das receitas e despesas estejam, com uma certa probabilidade de confiança.

- **Nadaraya-Watson -** observando o comportamento dos dados, foi usada também a técnica de regressão baseada em kernel conhecida como Nadaraya-Watson. De acordo com [Ivezić et al. 2019], é um algoritmo não-paramétrico de suavização de kernel utilizado em análise de dados e aprendizado de máquina. Ele é usado principalmente em problemas de regressão, com o objetivo de estimar uma função contínua a partir de um conjunto de dados. O identificador de outliers de kernel Nadaraya-Watson funciona estimando a função contínua em cada ponto dos dados, a partir de uma média ponderada dos valores da variável dependente nos pontos próximos, com os pesos determinados pelo kernel (uma função de ponderação que determina a influência dos pontos próximos na estimativa). Quando encontramos o kernel é necessario definir um intervalo de confiança, para definir possíveis **outliers** com base no kernel. A medida de confiança utilizada foi o $Z - score$, a mesma utilizada no modelo de **regressão linear**;

- **Desvio Padrão -** de acordo com [Altman and Bland 2005], o desvio padrão é uma medida de variabilidade que permite estimar a dispersão da população a partir de uma amostra. Quanto maior o desvio padrão, mais os dados estão distantes da média. Para utilizar esta medida na identificação de *outliers*, o algoritmo realiza os seguintes passos: (i) cálculo da média (μ) e do desvio padrão (σ) do conjunto de dados; (ii) define-se um limite superior (*upper*) e inferior (*lower*) para identificar valores que estão mais distantes da média, através das fórmulas $upper = \mu + (k * \sigma)$ e $lower = \mu - (k * \sigma)$, onde "k" é o número de desvios padrão a serem usados (valores usuais de "k" são 2, 2.5 ou 3); (iii) identificam-se os valores que estão acima do limite superior ou abaixo do limite inferior, os quais são justamente os considerados *outliers*.

Por fim, nesta etapa, calcula-se o grau de confiança (*GC*) na atipicidade de uma dada empresa **X** com base no número de algoritmos que a identificaram como *outlier*. Em outras palavras, *GC* representa a probabilidade da empresa **X** apresentar um comportamento atípico em relação ao padrão de empresas similares. A equação 1 apresenta a fórmula de cálculo do *GC*, onde *p* é o número de algoritmos do comitê que indicou a empresa **X** como *outlier*, e *n* é o número de algoritmos em execução no comitê. Importante notar que o modelo prevê que o comitê pode ser alterado, com a inclusão e exclusão de algoritmos.

$$GC = 100 * \left(\frac{p}{n}\right) \% \quad (1)$$

4. **Previsão de Custo e Análise de Anomalias** - nesta etapa, são realizados dois processos importantes e interdependentes. O primeiro é a predição de valores de custos ausentes das empresas *outliers* com $GC \geq$ lc (limiar de confiança), com base nos valores médios das empresas similares (após desconsiderar da amostra os valores das empresas *outliers*). Este processo visa completar os dados de custos, considerando que a empresa deveria apresentar comportamento padrão. Por exemplo, seja uma empresa *outlier* **X** com *GC*=75% e para a qual não foi possível aferir o valor de custo de energia elétrica em um dado período de tempo. Então, o processo acima prediz o valor deste custo com base nos gastos médios de energia elétrica das empresas típicas, considerando a mesma atividade econômica, o mesmo porte, o mesmo período de tempo, e a mesma localização geográfica de **X**.

   Ainda nesta etapa, um segundo processo classifica os indícios de anomalia de cada empresa **X** em três tipos:

   - **Anomalia do Tipo 1**: quando a empresa **X** possui receitas e despesas iguais a zero;
   - **Anomalia do Tipo 2**: quando a empresa **X** apresenta receitas superior às despesas, considerando um limiar, por exemplo de 30%; e
   - **Anomalia do Tipo 3**: quando a empresa **X** tem um valor de receitas inferior às despesas, considerando um limiar, por exemplo de 30%.

   Na Figura 4, tem-se uma representação gráfica dos tipos de anomalias em relação aos limites inferiores e superiores. É importante destacar que, dentre os três tipos de indícios acima, os que decorrem da Anomalia do Tipo 3 são os que indiciam evasão fiscal, e, portanto, são os mais importantes para o contexto deste trabalho.

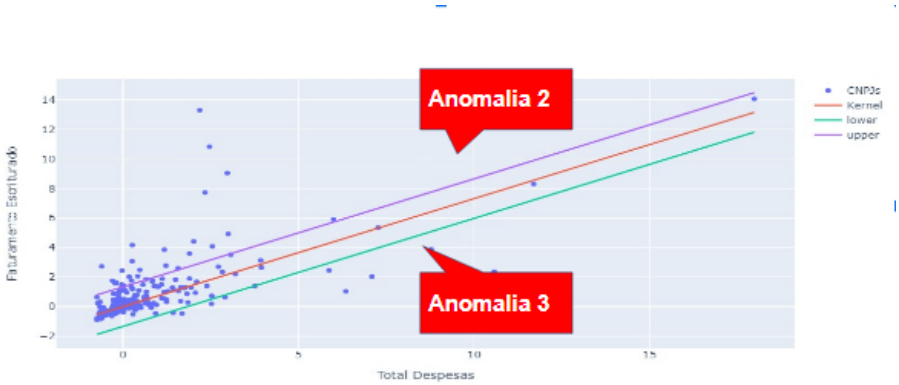

Figura 4. Exemplo dos tipos de anomalias detectadas pelo modelo.

5. **Camada de Serviço** - por fim, os dados agregados das receitas e despesas de cada empresa, aferidos ou inferidos, bem como a classificação dos indícios e o grau de confiança, são gerados e disponibilizados através de uma camada de serviço. Os dados são importados para ferramentas de visualização e análise dos resultados do modelo, para que auditores e gestores possam tomar a decisão de quais empresas serão auditadas e/ou notificadas.

4. Avaliação Experimental no Município de Fortaleza, Ceará

Para validação do modelo proposto, foi realizado um estudo de caso em empresas de serviço da cidade de Fortaleza, capital do Estado do Ceará, com os seguintes parâmetros, condições e restrições:

- Os dados utilizados para treinamento do modelo foram extraídos da Secretária Municipal das Finanças do município de Fortaleza (SEFIN) e foram anonimizados pela área de tecnologia da SEFIN. O acesso aos dados respeitou a Lei Geral de Proteção aos Dados (LGPD) e a mesma secretaria autorizou a divulgação das análises descritivas desde que fossem respeitados o sigilo fiscal;
- Os segmentos econômicos definidos para este estudo de caso foram Academias, Escolas, Lavanderias, Hotelaria, Oficinas, e Salões de Beleza;
- A dimensão temporal foi de janeiro a junho de 2022;
- Para agregar os dados por localização geográfica foram utilizados os bairros da cidade de Fortaleza;
- Para a definição do porte das empresas foi usada a área edificada dos imóveis onde as empresas estão localizadas. A categorização em pequeno, médio e grande porte foi realizada usando a técnica *boxplot*, onde 25% das empresas foram categorizadas como pequeno porte, 50% como médio porte e 25% como grande porte;
- Os dados de Receitas e Despesas foram coletados da SEFIN e entidades externas ou calculados conforme segue:
  - **Receitas** - Escrituração fiscal SEFIN, Simples Nacional, Vendas de Operações Financeiras (SEFAZ/CE) (p.ex. cartão de crédito, de débito, pix);
  - **Despesas** - Serviços Tomados (NFSe/SEFIN), Tributos Municipais como ISS, IPTU e ITBI, Taxas (DAMs/SEFIN), Gastos com Energia Elétrica (concessionária ENEL);
  - **Dados ausentes** - os seguintes conjuntos de dados não estavam disponíveis no momento deste trabalho: Tributos federais, Compras de Mercadorias, Folha de pagamento, gastos com Água.
  - **Dados preditos** - Custos de Aluguel foram calculados com base em 5% do valor de mercado do imóvel; Custos com Comunicação foram calculados a partir de uma média histórica dependendo do porte (R$ 250,00 - Pequena e R$ 500,00 - Médio e Grande por mês); Custos com Contabilidade foram calculados em um salário mínimo mensal, para o caso em que o contador da empresa não era funcionário da empresa; por fim, os demais custos (com valor = 0,00) de empresas atípicas foram inferidos com base na similaridade com empresas típicas, conforme explanado na etapa **Previsão de Custos e Análise de Anomalias**.

Com base nos parâmetros definidos para este estudo de caso, o recorte inicial foram 22.071 empresas ativas com faturamento ou custo, no período de janeiro a junho de 2022, e que não eram MEI (Microempreendedor Individual), assim distribuídas por segumento econômico: Oficinas (10.790), Escolas (6149), Salões de Beleza (2827), Hotelaria (759), Lavanderias (305) e Academias (1241).

Considerando a dimensão geográfica e visando otimizar a validação do modelo, foram selecionados os dois bairros com a maior quantidade de empresas em cada segmento. Assim, os bairros analisados por segmento econômico foram: Oficina (Aldeota e Centro), Escolas (Aldeota e Centro), Salão de Beleza (Aldeota e Centro), Hotelaria (Meireles e Centro), Lavanderia (Aldeota e Meireles), Academia (Aldeota e Meireles).

Na dimensão Porte/Tamanho, foram selecionadas apenas as empresas de médio e grande porte, simplesmente para que a validação do modelo, em tese, se concentrasse em empresas com maior faturamento.

Após os filtros e a execução do modelo, o conjunto de dados resultante continha 2.257 empresas, sendo que, destas, 2.246 empresas apresentaram algum indício de anomalia, como pode ser observado na Figura 5.

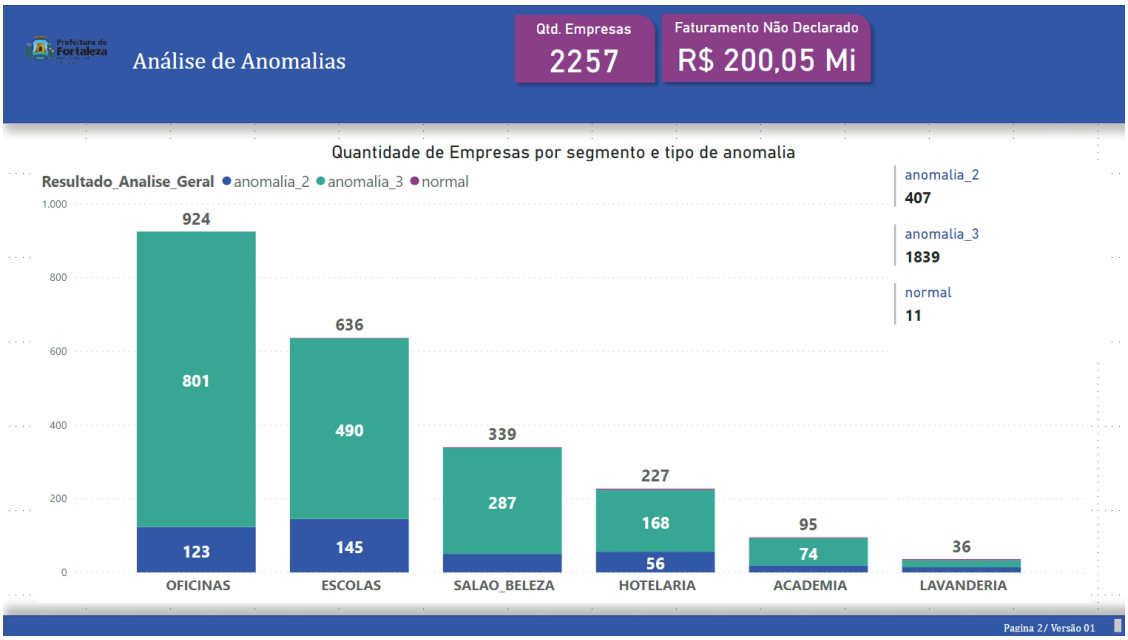

Figura 5. Distribuição de empresas por segmento e tipo de anomalia calculado pelo modelo.

Considerando apenas a Anomalia do Tipo 3 (Receitas $\leq$ Despesas), a Figura 6 apresenta dados de 1839 empresas com indício de evasão fiscal e $GC \geq 75\%$ (1526 empresas com $GC = 100\%$ e 313 empresas com $GC = 75\%$). Para cada empresa, foi calculado o valor de Faturamento Não-declarado com base na diferença entre o faturamento escriturado e os custos calculados pelo modelo, totalizando um motante de R$ 200 milhões de faturamento não-declarado. No gráfico à esquerda, tem-se a distribuição deste valor por segmento econômico. As escolas e oficinas apresentaram os maiores valores de faturamento não-declarado, sendo responsáveis por 72% do total não-declarado. O gráfico à direita apresenta a distribuição do valor não-declarado pelos bairros de Fortaleza (bairros com maior concentração das empresas analisadas - Centro, Aldeota e Meireles). Por exemplo, no bairro Meireles, os hotéis representam o segmento com maiores valores de evasão fiscal.

Como resultado do modelo, os dados analíticos de cada empresa são gerados na camada de serviço e podem ser importados para uma ferramenta de visualização e análise de dados. A Tabela 1 apresenta, de forma resumida, um extrato do arquivo gerado neste estudo. No caso, a Empresa 4, localizada no bairro CENTRO, do segmento OFICINAS, de porte GRANDE, apresentou fa-

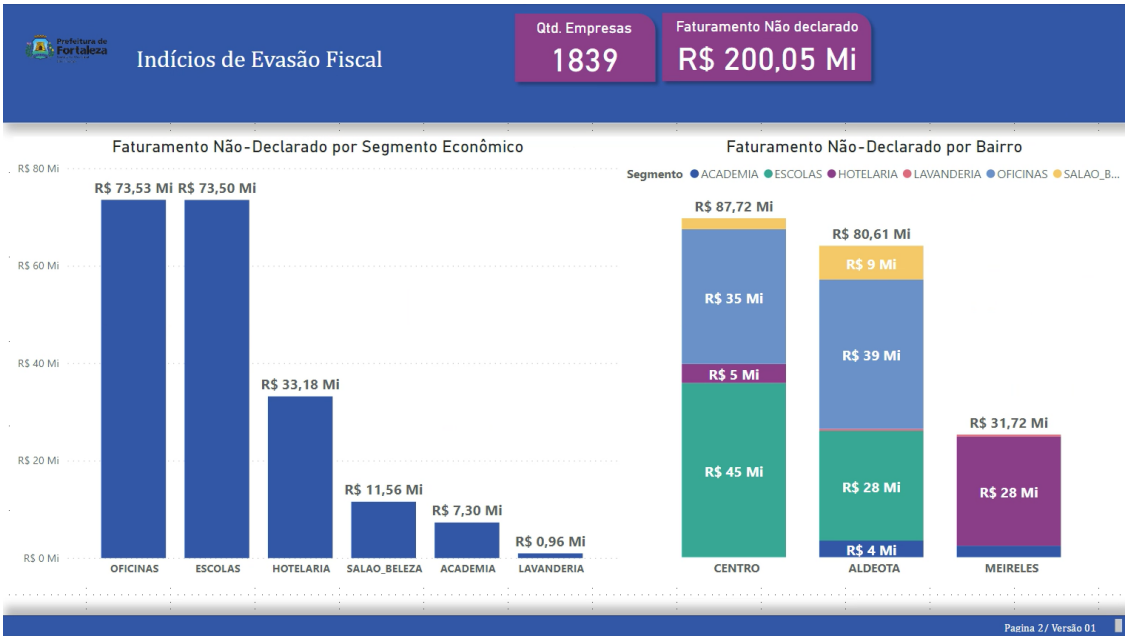

Figura 6. Análise da Evasão Fiscal em Fortaleza, por segmento econômico e bairro, no período de Janeiro a Junho de 2022 (considerando os recortes do conjunto de dados analisado)

turamento escriturado no valor de R$ 2.178,00 e total das despesas no valor de R$ 97.295,91, resultando em anomalia do tipo 3 com 100% de confiança. Um exemplo interessante é o caso das Empresas 0, 2, 3, 5 e 6, cujo faturamento registrado é igual a zero. Isso pode ocorrer devido a diferentes razões. Uma delas é a baixa declaração do ISS, um imposto autodeclarado. Além disso, é possível que essas empresas estejam ativas, mas sem operar efetivamente, e ainda assim enfrentem despesas como DAMs, energia, contador, entre outros. Outra possibilidade é que a empresa esteja ativa, mas sem atividade comercial real, o que pode resultar em custos estabelecidos arbitrariamente pelo modelo.

Tabela 1. Extrato dos resultados analíticos gerados neste estudo de caso, na cidade de Fortaleza.

Empresa	Bairro	Segmento	Porte	Faturamento	Despesas	Analise	GC[3]	Diferenca
Empresa 0	CENTRO	OFICINAS	medio	0.0	49091.54	anomalia_3	100	-49091.54
Empresa 1	CENTRO	OFICINAS	grande	86714.41	58411.09	anomalia_2	25	28303.32
Empresa 2	CENTRO	OFICINAS	indefinido	0.0	25609.61	anomalia_3	100	-25609.61
Empresa 3	ALDEOTA	OFICINAS	grande	0.0	38625.56	anomalia_3	100	-38625.56
Empresa 4	CENTRO	OFICINAS	grande	2178.0	97295.91	anomalia_3	100	-95117.91
Empresa 5	CENTRO	OFICINAS	indefinido	0.0	30270.6	anomalia_3	100	-30270.6
Empresa 6	ALDEOTA	OFICINAS	indefinido	0.0	35009.92	anomalia_3	100	-35009.92

Fonte: Elaborada pelos autores.

5. Conclusão

Este trabalho apresenta a proposta de um modelo de geração de indícios de evasão fiscal para empresas prestadoras de serviço. Os diferenciais do modelo são o reconhecimento de *outliers* por um comitê de algoritmos de Inteligência Artificial e Ciência de dados, e a possibilidade de inferir os valores de custos quando na ausência de dados sobre os custos operacionais e/ou as operações de compra e venda das empresas. A abordagem proposta possibilida uma melhoria no

processo de tomada de decisão dos auditores fiscais, pois reconhece e qualifica as empresas com comportamento atípico facilitando o processo de auditoria e notificação para autorregularização.

Foi realizado um estudo de caso com empresas de médio e grande porte da cidade de Fortaleza, capital do estado Ceará, de 6 (seis) segmentos econômicos (Oficinas, Escolas, Salões de Beleza, Hotelaria, Lavanderias e Academias), considerando apenas os bairros com maior número de empresas, e os dados de receitas e despesas de janeiro a junho de 2022. Ao final, a partir do recorte inicial de 22.071 empresas, foram reconhecidas 2.246 empresas com algum tipo de anomalia, das quais 1.839 apresentam fortes indícios de sonegação fiscal, da ordem de R$ 10 milhões em perda tributária do ISS. Os resultados foram validados pelos auditores da SEFIN que avaliaram o modelo como promissor no sentido de auxiliar o planejamento fiscal das secretarias de finanças dos municípios.

Como trabalhos futuros, pretende-se incluir outros algoritmos e técnicas no comitê de identificação de *outliers*, e definir outros tipos de anomalia com base do faturamento de Operações Financeiras e faturamento das empresas do Simples Nacional.

Referências

Altman, D. G. and Bland, J. M. (2005). Standard deviations and standard errors. *Bmj*, 331(7521):903.

de Vasconcelos Soares, G. and Cunha, R. (2020). Predição de irregularidade fiscal dos contribuintes do tributo iss. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, pages 223–228. SBC.

Dias, M. and Becker, K. (2017). Identificação de candidatos à fiscalização por evasão do tributo iss. In *5th Symposium on Knowledge Discovery, Mining and Learning, Uberlândia, MG*.

Franzoni, L. A. (1998). Tax evasion and tax compliance. *Available at SSRN 137430*.

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., and Gray, A. (2019). *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*, volume 8. Princeton University Press.

Jiawei, H., Micheline, K., and Jian, P. (2012). *Data Mining: Concepts and Techniques.-3rd*. Morgan kaufmann.

Oliveira, G. P., Reis, A. P., Freitas, F. A., Costa, L. L., Silva, M. O., Brum, P. P., Oliveira, S. E., Brandão, M. A., Lacerda, A., and Pappa, G. L. (2022). Detecting inconsistencies in public bids: An automated and data-based approach. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 182–190.

Oliveira, V. D. (2019). Redes neurais artificiais aplicadas à identificação de riscos de inadimplência fiscal de icms e iss no distrito federal.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Xavier, O. C., Pires, S. R., Marques, T. C., and Soares, A. d. S. (2022). Identificação de evasão fiscal utilizando dados abertos e inteligência artificial. *Revista de Administração Pública*, 56:426–440.