Universidade do Estado do Rio de Janeiro
Centro de Ciências Sociais
Faculdade de Administração e Finanças

Tainá Ayres Sá

Métodos de classificação supervisionados aplicados à identificação de fraudes de fornecedores do Estado do Rio de Janeiro

Rio de Janeiro
2022

Tainá Ayres Sá

Métodos de classificação supervisionados aplicados à identificação de fraudes de fornecedores do Estado do Rio de Janeiro

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Contábeis, da Universidade do Estado do Rio de Janeiro. Área de concentração: Controle de Gestão.

Orientador: Prof. Dr. Francisco José dos Santos Alves

Coorientador: Prof. Dr. José Francisco Moreira Pessanha

Rio de Janeiro

2022

CATALOGAÇÃO NA FONTE
UERJ/REDE SIRIUS/BIBLIOTECA CCS/B

S111 Sá, Tainá Ayres.
Métodos de classificação supervisionados aplicados à identificação de fraudes de fornecedores do Estado do Rio de Janeiro / Tainá Ayres Sá. – 2022.
69 f.

Orientador: Prof. Dr. Francisco José dos Santos Alves.
Coorientador: Prof. Dr. José Francisco Moreira Pessanha.
Dissertação (mestrado) – Universidade do Estado do Rio de Janeiro, Faculdade de Administração e Finanças.
Bibliografia: f. 66-69.

1. Finanças públicas – Auditoria – Legislação – Teses. 2. Licitação pública – Teses. 3. Fraude – Prevenção – Teses. I. Alves, Francisco José dos Santos. II. Pessanha, José Francisco Moreira. III. Universidade do Estado do Rio de Janeiro. Faculdade de Administração e Finanças. IV. Título.

CDU 336.5(815.3):343.537

Bibliotecário: Fabiano Salgueiro CRB7/6974

Autorizo, apenas para fins acadêmicos e científicos, a reprodução total ou parcial desta dissertação, desde que citada a fonte.

_____ _____

Assinatura Data

Tainá Ayres Sá

Métodos de classificação supervisionados aplicados à identificação de fraudes de fornecedores do Estado do Rio de Janeiro

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Contábeis da Faculdade de Administração e Finanças, da Universidade do Estado do Rio de Janeiro. Área de concentração: Controle de Gestão.

Aprovada em 07 de julho de 2022.

Banca Examinadora:

_____

Prof. Dr. Francisco José dos Santos Alves (Orientador)

Faculdade de Administração e Finanças – UERJ

_____

Prof. Dr. José Francisco Moreira Pessanha (Coorientador)

Faculdade de Administração e Finanças – UERJ

_____

Prof. Dr. Frederico Antonio Azevedo de Carvalho

Universidade Federal do Rio de Janeiro – UFRJ

Rio de Janeiro

2022

DEDICATÓRIA

À minha avó, Carmen Luiza, por me fazer quem eu sou,
e à minha filha, Luiza, a quem eu me dedico ser.

AGRADECIMENTOS

Ao meu orientador, Francisco José dos Santos Alves, pela disponibilidade durante o processo de construção da dissertação e conclusão do curso de mestrado. Ao meu coorientador, José Francisco Moreira Pessanha, pelo apoio técnico e suas valiosas contribuições. E ao professor Frederico Antonio Azevedo de Carvalho pela parceria nos grupos de estudo de fraude e contribuições no processo de qualificação.

Ao meu marido, Marco Miranda, por todo o seu amor, por não me deixar desistir e ser meu suporte nas horas de estudo. À minha mãe Lilian e minha tia Eliane, pelo incentivo. Ao meu pai Paulo e meu irmão Guilherme, pelo exemplo.

À minha amiga Maria Lucia, por ter me despertado o entusiasmo pela pesquisa acadêmica. E a Ursula, Viviane e Renata, por serem as pessoas que me fizeram escolher o caminho do mestrado.

Ao Marcos Kovacs, pela recomendação e por ter estimulado em mim o interesse pelo tema, e ao Fernando Silva, por ter incentivado a me inscrever no curso.

Aos meus colegas de turma, Pauleniza e Leonardo, por sempre me ajudarem quando precisei.

A Deus, pela saúde e por fazer os astros conspirarem.

Não sabendo que era impossível, foi lá e fez.

Jean Cocteau

RESUMO

SÁ, Tainá Ayres. *Métodos de classificação supervisionados aplicados à identificação de fraudes de fornecedores do Estado do Rio de Janeiro*. 2022. 68 f. Dissertação (Mestrado em Ciências Contábeis) – Faculdade de Administração e Finanças, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2022.

A fraude nas licitações ocorre pela tentativa de frustrar o seu caráter competitivo, causando o superfaturamento e o favorecimento de um grupo específico. Esta pesquisa tem como objetivo verificar qual conjunto de indicadores e algoritmos possui melhores propriedades para identificar a propensão de um fornecedor ter praticado uma fraude contra a administração pública nos processos de contratação do Governo do Estado do Rio de Janeiro. A delimitação da pesquisa foram os fornecedores de pessoa jurídica do Governo do Estado do Rio de Janeiro que tiveram contratos ou participaram de licitações entre 2017 e 2021. Foram utilizados os dados de licitações e fornecedores publicados no Portal da Transparência do Governo do Estado do Rio de Janeiro, e os dados dos CNPJs publicados pela Receita Federal. O teste de hipótese baseado na Lei de Benford demonstrou que a base de contratos não estava em conformidade com a lei, podendo estar propensa a ter sofrido manipulação dos números.A pesquisa proposta demonstrou resultados satisfatórios para a classificação de fornecedores sancionados nesse estado. O algoritmo de classificação com melhor resultado nas três métricas de avaliação foi o *Support Vector Machine* (SVM), com precisão de 99,30%, e apenas um falso positivo na verificação da base-teste. Os atributos levantados por meio das bases de dados também se mostraram relevantes no resultado do modelo.

Palavras-chave: Fraude em licitações. Indicadores. Algoritmos de classificação.

ABSTRACT

SÁ, Tainá Ayres. *Supervised classification methods applied to identify supplier fraud in the State of Rio de Janeiro*. 2022. 68 f. Dissertação (Mestrado em Ciências Contábeis) – Faculdade de Administração e Finanças, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2022.

Bidding fraud occurs by the attempt to frustrate its competitive character, causing overpricing and favoring a specific group. This research aims to verify which set of indicators and algorithms have best properties to identify the propensity of a supplier to have committed fraud against the public administration in the contracting process of the Government of the State of Rio de Janeiro. The delimitation of the research was the suppliers of legal entities of the Government of the State of Rio de Janeiro that had contracts or participated in bids between 2017 and 2021. Data from bids and suppliers published on the Transparency Portal of the Government of the State of Rio de Janeiro, and data from CNPJs published by the Brazilian Internal Revenue Service were used. The hypothesis test based on Benford's Law demonstrated that the contracts database was not in compliance with the law, being prone to have had manipulation of the numbers. The proposed research showed satisfactory results for the classification of suppliers sanctioned by this state. The classification algorithm with the best result in the three metrics of evaluation was the Support Vector Machine (SVM), with an accuracy of 99.30%, and only one false positive in the verification of the test base. The attributes collected through the databases also proved to be relevant in the model result.

Keywords: Bidding fraud. Indicators. Classification algorithms.

LISTA DE FIGURAS

Figura 1 – Tipos de aprendizagem de máquina 31
Figura 2 – Modelo *K-Nearest Neighbor* 34
Figura 3 – Rede neural 35
Figura 4 – Estrutura do neurônio 36
Figura 5 – Árvore de decisão 38
Figura 6 – Máquina de vetor suporte (SVM) 39
Figura 7 – Matriz de confusão 40
Figura 8 – Esquema das bases de dados 45
Figura 9 – Diagrama de validação cruzada 50
Figura 10 – Matriz de correlação 56

LISTA DE GRÁFICOS

Gráfico 1 – Representação gráfica da frequência do primeiro dígito 29
Gráfico 2 – Reta de regressão linear *versus* curva logística 33
Gráfico 3 – Balanceamento das bases de dados 48
Gráfico 4 – Distribuição de fornecedores por região 51
Gráfico 5 – Distribuição de fornecedores por ano de cadastro 52
Gráfico 6 – Distribuição de fornecedores por porte da empresa 52
Gráfico 7 – Percentual dos motivos das sanções 53
Gráfico 8 – Tipos de sanções aplicadas aos fornecedores 54
Gráfico 9 – Valor dos contratos por ano 54
Gráfico 10 – Histograma de fornecedores por valor em contratos 55
Gráfico 11 – Valor das licitações por modalidade 55
Gráfico 12 – Lei de Benford do primeiro dígito 57
Gráfico 13 – Lei de Benford dos dois primeiros dígitos 58
Gráfico 14 – Curva ROC do modelo KNN 60
Gráfico 15 – Curva ROC do modelo *random forest* 60
Gráfico 16 – Curva ROC do modelo rede neural 61
Gráfico 17 – Curva ROC do modelo de regressão logística 61
Gráfico 18 – Curva ROC do modelo SVM 62

LISTA DE QUADROS

Quadro 1 – Bases de dados utilizadas 43
Quadro 2 – Atributos considerados no modelo final 45
Quadro 3 – Hiperparâmetros executados 50

LISTA DE TABELAS

Tabela 1 – Resumo dos limites por modalidade de licitação 23
Tabela 2 – Probabilidade conforme posição dos dígitos 29
Tabela 3 – População e amostra do sistema Siga 44
Tabela 4 – Motivos das sanções 53
Tabela 5 – Precisão dos algoritmos 59
Tabela 6 – Resultado da matriz de confusão 59
Tabela 7 – Coeficientes da regressão logística 63
Tabela 8 – *Features_importances* do *random forest* 64

LISTA DE ABREVIATURAS E SIGLAS

ACFE	*Association of Certified Fraud Examiners*
ACL	*Audit Command Language*
CAAT	*Computer Assisted Audit Techniques* (Técnicas de Auditoria Auxiliadas por Computador)
CNAE	Classificação Nacional de Atividade Econômica
CNPJ	Cadastro Nacional da Pessoa Jurídica
CRISP-DM	*Cross Industry Standard Process for Data Mining*
EPP	Empresa de Pequeno Porte
ICMS	Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação
KNN	*K-Nearest Neighbor*
LASSO	*Least Absolute Shrinkage and Selection Operator*
ME	Microempresa
MPF	Ministério Público Federal
ROC	*Receiver Operating Characteristic*
SEFAZ-BA	Secretaria da Fazenda do Estado da Bahia
SICAF	Sistema de Cadastramento de Fornecedores do Governo Federal
SIGA	Sistema Integrado de Gestão de Aquisições
SVM	*Support Vector Machine* (Máquina de Vetor Suporte)

SUMÁRIO

	INTRODUÇÃO	17
1	**REFERENCIAL TEÓRICO**	20
1.1	**Evolução do processo de auditoria**	20
1.2	**O processo licitatório**	21
1.3	**Fraude e definição de indicadores**	25
1.4	**Lei de Benford**	29
1.5	**Aprendizado de máquina**	30
1.6	**Algoritmos de classificação**	32
1.6.1	Regressão logística	32
1.6.2	KNN	34
1.6.3	Redes neurais artificiais	35
1.6.4	*Random forest*	37
1.6.5	Máquina de vetor suporte (SVM)	38
1.7	**Métricas de avaliação**	39
1.7.1	Precisão	39
1.7.2	Matriz de confusão	40
1.7.3	Curva ROC	41
1.8	**Metodologia CRISP-DM**	41
2	**METODOLOGIA**	43
2.1	**População e amostra**	43
2.2	**Coleta de dados**	44
2.3	**Tratamento dos dados**	47
2.3.1	Verificação de dados ausentes	47
2.3.2	Balanceamento das bases	47
2.3.3	Divisão da base entre treino e teste	49
2.3.4	Normalização da base de dados	49
2.3.5	Aplicação dos algoritmos	49
3	**ANÁLISE E DISCUSSÃO DOS RESULTADOS**	51
3.1	**Análise descritiva da base de dados**	51
3.2	**Teste de hipótese – Lei de Benford**	56
3.3	**Resultados**	58

3.4 **Melhoria da *performance* dos algoritmos** 64
CONSIDERAÇÕES FINAIS 65
REFERÊNCIAS 66

INTRODUÇÃO

De acordo com pesquisa publicada pela *Forbes*, em 2015, o volume de dados criado nos últimos dois anos era superior à quantidade de informações produzida em toda a história da humanidade (MARR, 2015). A partir da pandemia de covid-19, que fez com que o trabalho em casa se tornasse cada vez mais comum, essa produção se acentuou ainda mais.

Em pesquisa realizada pela Domo (c2022), empresa do ramo de tecnologia, a cada minuto de 2020 foram gerados aproximadamente 347 mil *stories* no Instagram, 42 milhões de mensagens no WhatsApp e 500 horas de conteúdo foram publicadas no Youtube. Também, nessa nova era digital, cada curtida na internet é um dado que pode revelar mais sobre o usuário.

Além da quantidade de informações ter aumentado, o poder de processamento dos computadores foi otimizado de forma exponencial. Em reportagem publicada pelo *Estadão* em outubro de 2019, o Google anunciou a realização de uma operação em 200 segundos por um computador quântico, atividade que levaria 10 mil anos para ser feita por um computador clássico (ROMANI, 2019).

Em se tratando de auditoria, esses fatos demonstram que o trabalho que era realizado de uma certa maneira há poucos anos não pode se manter igual. A atividade do auditor como era feita no passado, de selecionar uma amostra limitada de casos e revisá-la manualmente, não será mais eficaz, uma vez que esse processo deve ser atualizado de acordo com o surgimento das novas tecnologias.

Quando o objetivo é identificar fraudes praticadas por fornecedores contra a administração pública no processo de contratação, a utilização da tecnologia é essencial para descobrir os rastros deixados pelos fraudadores. Com o crescimento exponencial do volume de dados, o desafio se torna cada vez maior para saber quais são as situações que merecem mais atenção dos auditores, de modo que o tempo gasto nas investigações seja melhor utilizado e os esquemas de fraude revelados.

Diversos indicadores já foram propostos pela literatura, os quais poderiam sinalizar que uma fraude possa ter ocorrido. Dentre esses indicadores, uma opção muito utilizada é a Lei de Benford, de 1938, que analisa a regularidade dos primeiros dígitos de uma base de números naturais com a frequência esperada (BENFORD, 1938). Cella e Zanolla (2018) propuseram, por exemplo, a análise da Lei de Benford sobre o *ranking* de transparência dos municípios do estado de Goiás e a base de despesas empenhadas. Os autores verificaram que os municípios com mais aderência a essa teoria também são os considerados mais transparentes.

Apesar do grande volume de pesquisas já desenvolvidas, muitas vezes a avaliação de um indicador ainda pode não ser determinante, gerando falsos positivos, e no meio de uma grande quantidade de dados pode ficar difícil para o auditor conseguir identificar uma transação correta de uma fraudulenta. A partir daí, entram outras técnicas de inteligência artificial que podem ajudar a detectar mais padrões das atividades fraudulentas em que o profissional está interessado em comprovar.

Morais (2016) utilizou técnicas de mineração de dados e reconhecimento de padrões estatísticos para identificar cartéis em licitações de obras de engenharia realizadas pelo Estado do Ceará. A pesquisa aplicou um algoritmo de agrupamento (*K-means*) e como resultado obteve um grupo de empresas que poderiam estar atuando em licitações fraudulentas. Este é um tipo de análise que pode ser utilizada como ponto de partida para um processo de auditoria.

Os algoritmos de classificação já foram muito empregados na caracterização de diversos tipos de fraude. Oliveira (2016), por exemplo, aplicou o método de regressão logística para identificar fraudes nos cartões de crédito. Já Olokodana e Fernandes (2020) utilizaram o *K-nearest Neighbors* (KNN), entre outros algoritmos, para identificar fraudes nas demonstrações contábeis de empresas listadas na Bovespa. Oliveira e Santos (2020), por sua vez, propuseram um modelo baseado em redes neurais para a classificação de contribuintes mais propensos a assumir a condição de sonegadores do ICMS. Em outro estudo, Santos (2020) utilizou o *random forest* para a constatação de fraude no ramo da saúde. Além destes, Severino e Yaohao (2019) realizaram uma predição com para sinistros patrimoniais com fraudes mediante a aplicação de modelos da *Support Vector Machine* (SVM), entre outros algoritmos.

Esta pesquisa pretende responder ao seguinte problema de pesquisa: qual conjunto de indicadores e algoritmos possui melhores propriedades para identificar a propensão de um fornecedor ter praticado uma fraude contra a administração pública nos processos de contratação do Governo do Estado do Rio de Janeiro.

Assim, tem-se como objetivos específicos:

a) desenvolver uma lista de indicadores relevantes na identificação de fraudes;
b) investigar quais algoritmos de classificação possuem melhores propriedades na identificação dos padrões dos fornecedores fraudadores para definição de um modelo.

A seleção do estado do Rio de Janeiro deveu-se aos diversos esquemas de corrupção já divulgados pela mídia envolvendo os governadores Anthony Garotinho, Sérgio Cabral, Luiz

Fernando Pezão e Wilson Witzel. Além disso, os dados de licitações estão disponíveis no Portal da Transparência do Estado do Rio de Janeiro.[1]

Diante dos diversos relatos de fraudes veiculados diariamente nos noticiários brasileiros, a pesquisa tem relevância uma vez que pode ser utilizada pelos órgãos de controle do governo para identificar esses atos no processo de contratação, e ainda por empresas privadas, fazendo as adaptações de acordo com as características do processo de contratação utilizado. Cada organização terá o seu próprio processo e estrutura de controles internos. Entretanto, a pesquisa e seus resultados poderão ser um passo inicial para a construção de um modelo que irá atender ao objetivo de identificar fraudes em processos licitatórios.

A pesquisa foi realizada a partir de licitações e contratos executados pelo Governo do Estado do Rio de Janeiro entre 2017 e 2021. Os dados foram extraídos do Portal da Transparência do Estado do Rio de Janeiro e dos dados dos CNPJs registrados na Receita Federal e publicados no Portal da Transparência da União. Foi feito então um levantamento dessas informações com base na fundamentação do referencial teórico, que resultou na compilação em um único arquivo base. Em seguida, realizou-se uma análise descritiva desses elementos e, por fim, os algoritmos foram aplicados para verificar quais atendiam ao questionamento do estudo.

A pesquisa está estruturada da seguinte forma: esta introdução, que apresenta o problema de pesquisa, os objetivos e a justificativa do estudo. Em seguida, na primeira seção, encontra-se o referencial teórico, em que se descreve o processo licitatório conforme a legislação brasileira, a evolução do processo de auditoria, o levantamento de trabalhos com a proposta de indicadores para a identificação de fraudes e os diferentes algoritmos que serão utilizados na metodologia. Na segunda seção estão detalhados os passos realizados na metodologia. Os resultados são apresentados na terceira seção. Por fim, a quarta seção traz as considerações finais e reflexões sobre a possibilidade de trabalhos futuros, seguida das referências consultadas e utilizadas na elaboração deste trabalho.

[1] O endereço eletrônico do Portal da Transparência do Estado é: http://www.transparencia.rj.gov.br.

1 REFERENCIAL TEÓRICO

Nesta seção, será revisada a evolução do processo de auditoria, o processo licitatório conforme legislação brasileira, as principais teorias relacionadas às fraudes e o levantamento de indicadores e a Lei de Benford. Além disso, serão debatidos os conceitos sobre o aprendizado de máquina, os algoritmos e as métricas de avaliação que serão utilizados na metodologia para o desenvolvimento do trabalho. Por fim, será apresentada a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM) como ferramenta para o desenvolvimento de modelos para análises de dados.

1.1 Evolução do processo de auditoria

A tecnologia trouxe transformações nos processos de auditoria, assim como nas demais áreas das organizações. Se antes esse processo era realizado por meio de amostras, com a verificação manual de documentações baseadas em acontecimentos passados, hoje já é possível analisar grandes bases de dados, cruzando com informações externas à organização e fazendo estudos preditivos. Ademais, é possível que robôs façam levantamentos documentais sem intervenção humana.

A auditoria contínua surgiu para dar mais segurança ao processo e garantir que as exceções identificadas nas auditorias fossem tratadas tempestivamente. Segundo Bumgarner e Vasarhelyi (2015), esse processo de auditoria contínua pode ser definido como a metodologia que permite aos auditores proverem uma avaliação sobre uma exceção identificada simultaneamente ou pouco depois da sua ocorrência, e pode envolver modelos preditivos e completar controles organizacionais. O ambiente deve ser cada vez mais automatizado e os profissionais devem ser capazes de assumir papéis mais julgamentais. Assim, o procedimento utilizará métodos analíticos, por exceção, adaptáveis e cobrindo cada vez mais funções não financeiras.

Lombardi, Bloch e Vasarhelyi (2014) verificaram em sua pesquisa que está ocorrendo uma evolução das auditorias tradicionais (ou seja, dados históricos em uma data de corte) para uma versão mais contínua de todo o processo de negócios e riscos associados. Isso permite que os atuais problemas de negócios sejam abordados de maneira mais adequada.

Desse modo, as auditorias passaram de periódicas para contínuas, abrangendo um percentual muito maior de dados da empresa e monitoramento de processos de negócios. A

economia em tempo real exige garantia também em tempo real, e essa metodologia fornece a garantia por meio de verificação e controle de qualidade. O monitoramento constante dos processos de negócios pode identificar problemas emergentes e possivelmente levar a oportunidades de não auditoria dentro das empresas. Por outro lado, a amostragem usada na atualidade fornece garantia em dados históricos e não avalia processos de negócios.

Um método que já é utilizado há algum tempo são as técnicas de auditoria auxiliadas por computador (em inglês, *computer-assisted audit tools* – CAATs), que consistem na utilização de ferramentas computacionais para fazer as análises de dados. Segundo Porto (2011), o uso dos CAATs é muitas vezes associado com mineração de dados (*Data Mining)* e análise de dados, que pode ser relacionada tanto com a coleta de evidências quanto com os testes de certos objetivos de auditoria (como testes de anomalias, encontrar pagamentos duplicados, números de cheque faltando em uma sequência numérica, realizar o recálculo de uma folha de pagamento).

Essas técnicas são utilizadas para identificar exceções em um grande volume de dados, o que pode ser considerado eficaz, porque um auditor sozinho demoraria muito tempo revisando toda a informação. Esses testes são realizados durante as auditorias tradicionais, que são programadas em uma data e têm um período de escopo determinado.

De acordo com Lombardi, Bloch e Vasarhelyi (2014), os auditores devem estar preparados e devidamente treinados para lidar com os desafios associados à coleta, processamento e incorporação de novos formulários e grandes volumes de dados, que provavelmente exigirão a aplicação e a compreensão de tecnologias sofisticadas.

1.2 **O processo licitatório**

Licitação é o processo pelo qual a administração pública obtém a proposta mais vantajosa para a celebração de um contrato administrativo. As normas para a sua realização foram instituídas pela Lei n. 8.666/1993, que abrange as atividades realizadas no âmbito dos poderes da União, de estados, do Distrito Federal e de municípios, e podem ser realizadas nas modalidades: concorrência, tomada de preços, convite, concurso e leilão (BRASIL, 1993).

A Lei n. 10.520/2002 instituiu ainda a modalidade de pregão para a aquisição de bens e serviços comuns, estendendo a sua utilização para estados e municípios. A principal diferença para as outras modalidades é a inversão das fases, uma vez que a análise das propostas é feita antes da habilitação dos participantes (BRASIL, 2002). A disputa pode ser de modo presencial

ou eletrônico e se caracteriza pela realização de lances pelos concorrentes, além da presença de um pregoeiro.

A Lei n. 14.133/2021 trouxe mudanças nas modalidades de licitação, extinguindo o convite e a tomada de preço, e trazendo o diálogo competitivo como opção (BRASIL, 2021). A extinção total dessas categorias deverá ocorrer até 2023.

A nova lei também trouxe uma revisão das características e dos critérios de julgamento de cada modalidade. Antes, o valor estimado da contratação era levado em consideração; agora, apenas a natureza do objeto deve ser avaliada.

As seguintes características podem ser destacadas para cada modalidade:

a) Tomada de preços: realizada com concorrentes cadastrados e qualificados até o terceiro dia anterior à data de recebimento das propostas. A estimativa da compra para propostas de engenharia deve ser de até R$ 3,3 milhões, e para compra de outros serviços até R$ 1,43 milhão;
b) Convite: a unidade administrativa deve convidar pelo menos três concorrentes, que podem estar cadastrados ou não. Os valores estipulados são de até R$ 330 mil para obras e serviços de engenharia, e até R$ 176 mil para outras contratações. As micro e pequenas empresas têm prioridade nesses casos;
c) Concorrência: utilizada para a contratação de bens e serviços especiais e de obras e serviços comuns e especiais de engenharia. Os critérios para escolha podem ser o menor preço, melhor técnica, melhor preço e técnica, melhor retorno ou maior desconto;
d) Concurso: empregado para escolha de trabalho técnico-científico ou artístico em que haverá uma avaliação para o recebimento de prêmio ou remuneração pelo vencedor;
e) Leilão: válido para qualquer interessado na alienação de bens imóveis, de bens móveis inservíveis ou legalmente apreendidos. Ganha quem oferecer o maior lance. O cadastro dos interessados poderá ser feito após a conclusão da etapa de lances;
f) Pregão: modalidade aplicada para a aquisição de bens e serviços comuns, cujo critério poderá ser o menor preço ou o maior desconto;
g) Diálogo competitivo: utilizado para a contratação de obras, serviços e compras. A administração pública, a partir de critérios objetivos, realiza

> diálogos com licitantes previamente selecionados para desenvolver alternativas para o atendimento de suas necessidades. Ao final, uma proposta deve ser apresentada pelos participantes da conversa.

A Lei n. 8.666/1993 também institui os casos em que há a possibilidade de dispensa e inexigibilidade de licitação, ambos tipos de contratação direta. Na dispensa, apesar de haver a possibilidade de competição, a lei enumera os casos em que o procedimento licitatório é dispensado. Há ainda algumas hipóteses em que é possível realizar a dispensa, como quando o procedimento operacional for superior ao valor do contrato ou quando houver a urgência decorrente de calamidade pública por exemplo. Na inexigibilidade, a licitação é inviável visto que apenas um fornecedor seria capaz de atender às necessidades da administração.

Desde a primeira lei de licitações (Lei n. 8.666/1993), os valores por modalidade de compras foram modificados. Para esta pesquisa, o limite que estava em vigor na maior parte do tempo pesquisado (2017-2021) foi o instituído pelo Decreto n. 9.412/2018 (BRASIL, 2018). Os limites por modalidade de licitação podem ser resumidos de acordo com a Tabela 1.

Tabela 1 – Resumo dos limites por modalidade de licitação

Modalidade	**Compras – R$**	**Obras e Serviços – R$**
Dispensa	Até 17.000	Até 33.000
Convite	Até 176.000	Até 330.000
Tomada de preço	Até 1.430.000	Até 3.300.000
Concorrência	Acima de 1.430.000	Acima de 3.300.000
Pregão	Aplicável a qualquer valor	Aplicável a qualquer valor

Fonte: Adaptada de BRASIL, 2018.

A licitação é realizada por meio de um procedimento administrativo. Seu rito processual é dividido em duas fases: a fase interna, que ocorre antes da publicação do edital, e a fase externa, que ocorre depois. Em geral, de acordo com a Lei n. 8.666/1993, a licitação é composta pelas seguintes etapas:

a) Realização de pesquisa de mercado para determinação do preço estimado do processo licitatório;
b) Divulgação do edital com as principais características da licitação;
c) Habilitação dos licitantes – quando a documentação pertinente é entregue;
d) Recebimento das propostas dos licitantes;
e) Julgamento e classificação das propostas, que pode seguir diferentes critérios, como menor preço, prazo de entrega, qualidade;

f) Homologação – quando a licitação é aprovada;
g) Adjudicação – ato final, quando o objeto é atribuído ao vencedor. (BRASIL, 1993).

A licitação pode ser anulada se for verificado vício de ilegalidade no procedimento. Nesse caso, a anulação deve ser justificada e publicada, não gerando indenização ao licitante. Já a revogação se dá em razão de interesse público, como consequência de fato superveniente devidamente comprovado, e pode gerar indenização ao licitante vencedor.

Apesar dos diversos controles estipulados pela legislação ao longo dos anos, não é raro ouvir sobre casos de corrupção descobertos em licitações públicas.

As leis de 1993 e 2002 estabelecem sanções administrativas que devem ser aplicadas quando alguma infração for identificada. A finalidade dessas sanções é reprovar a conduta dos licitantes ou contratados e evitar a sua reincidência.

Segundo publicação do Ministério do Planejamento, as sanções administrativas em licitações e contratos

> […] são consequências de um ato ou um conjunto de atos, praticados por licitantes e contratados da Administração Pública que causem prejuízo à Administração ou violem normas de observância obrigatória. (BRASIL, 2014, p. 5).

E as penas que podem ser aplicadas são:

a) Advertência: comunicação formal sobre o descumprimento da obrigação legal. Notificando do agravamento da sanção no caso de reincidência;
b) Multa: de natureza pecuniária, empregada no caso de atraso no cumprimento contratual injustificado ou inexecução total ou parcial do contrato. Pode ser aplicada em conjunto da advertência, suspensão e inidoneidade;
c) Suspensão temporária: interrupção na possibilidade de participar em licitações por até dois anos. É executada no caso de inexecução total ou parcial do contrato;
d) Impedimento de licitar: barra o fornecedor de licitar e formalizar contrato com o ente que aplicar a sanção;
e) Descredenciamento: alteração do fornecedor no Sistema de Cadastramento de Fornecedores do Governo Federal (SICAF) para o status de inativo;

f) Declaração de inidoneidade: proíbe o fornecedor de participar de licitações em todos os órgãos da administração pública, União, estados, Distrito Federal e municípios.

A lei de 2021 enumera várias condutas reprováveis e é mais específica sobre quais penalidades devem ser aplicadas em cada situação.

Para Guedim (2021), os processos sancionatórios em licitações trazem inúmeros benefícios. Num primeiro momento, o fornecedor infrator, ou em vias de infringir, resolve cumprir o seu papel contratual ao enxergar uma administração capaz de aplicar-lhe uma punição.

Schramm (2018) destaca que, ao longo do tempo, a legislação brasileira que trata das aquisições públicas não concentrou esforços em conceder benefícios àqueles que observarem os ditames legais e optou por consolidar a cultura de que a certeza de punição é suficiente para mitigar a adoção de posturas desonestas nos processos. Entretanto, essas mesmas medidas foram incapazes de reprimir as práticas de corrupção nas contratações públicas.

1.3 **Fraude e definição de indicadores**

A fraude nos processos de contratação pública está essencialmente relacionada a uma tentativa de frustrar o caráter competitivo. Em outras palavras, seu objetivo direto é distorcer a livre disputa entre os participantes, natural em um processo competitivo, para que alguém seja indevidamente favorecido e contratado (TRANSPARÊNCIA BRASIL, 2019).

A fraude em licitações acontece em todo o mundo, tanto em organizações governamentais quanto em instituições privadas. Em um estudo internacional sobre a fraude ocupacional, aquela cometida por indivíduos contra a organização que os emprega, a *Association of Certified Fraud Examiners* (ACFE, 2020) verificou que, para todos os tipos de organização, o esquema mais observado é o de corrupção. Nas organizações governamentais, 48% dos casos reportados foram de corrupção. Essa categoria abrange os esquemas de fraudes em licitações e outros envolvendo conflitos de interesse, propinas, gorjetas ilegais e extorsão.

Jensen e Meckling (1976) desenvolveram a teoria da agência, que determina que a relação de agência é regida por termos contratuais. O principal seria aquele que está no centro da relação, no caso, a instituição governamental, e o agente é o profissional contratado para agir no melhor interesse do principal. Entretanto, nem sempre o agente fará isso, o que ocasionará

custos com problemas de agência. A fraude, nesse sentido, ocorre quando o agente atua em seu próprio interesse.

Para uma fraude ocorrer é preciso que alguns elementos estejam presentes. Em 1953, Donald Cressey formulou a hipótese que viria a ser conhecida como o modelo do Triângulo de Fraudes, segundo a qual pessoas que ocupam cargos de confiança em corporações se tornam violadoras dessa confiança quando se aproveitam da posição que ocupam para obter vantagens financeiras, de forma ilícita, para si (CRESSEY, 1953).

O modelo proposto por Cressey (1953) pressupõe três dimensões: *pressão*, *oportunidade* e *racionalização*. A *pressão* decorre da existência de problemas financeiros não compartilhados; a *oportunidade* se torna presente quando os fraudadores têm o conhecimento e a chance para realizar a fraude; e a *racionalização* ocorre quando o fraudador considera aceitável o ato fraudulento, justificando-o.

Essa teoria foi modificada ao longo do tempo. Por exemplo, Wolfe e Hermanson (2004) adicionaram o elemento *capacidade*, formando o Diamante da Fraude. A *capacidade* diz respeito à habilidade que o profissional possui para cometer esse ato, dado o seu conhecimento sobre o processo e suas fragilidades. Santos (2016) propôs ainda o Pentágono da Fraude, incorporando a *disposição ao risco*, que é quando o fraudador avalia o risco de ser pego em flagrante. Esse último elemento, *disposição ao risco*, está em linha com a ideia de Becker (1968) sobre a economia do crime:

> […] a decisão de cometer ou não o crime resultaria de um processo de maximização de utilidade esperada, em que o indivíduo confrontaria, de um lado, os potenciais ganhos resultantes da ação criminosa, o valor da punição e as probabilidades de detenção e aprisionamento associadas e, de outro, o custo de oportunidade de cometer crime, traduzido pelo salário alternativo no mercado de trabalho. (BECKER, 1968 *apud* CERQUEIRA; LOBÃO, 2004, p. 247).

Bonfim e Silva (2019) realizaram um experimento para verificar a honestidade dos indivíduos. Eles participaram de um jogo em que deveriam informar o número que aparecia nos dados para ganhar o prêmio. Porém, o pagamento foi atrelado ao número informado e não ao visto, tornando-os desonestos. Quando submetidos aos mecanismos de controle de auditoria e à leitura do código de ética, o índice de desonestidade foi de 9%. Quando não submetidos, o percentual subiu para 27%.

Esse experimento demonstra que, apesar de as pessoas terem uma tendência desonesta, controles podem ser implementados para inibi-las. Alguns desses controles podem ser preventivos, como o código de ética, e outros detectivos, pois quando há a possibilidade de ser

identificado e punido por uma fraude, o participante terá a percepção de risco maior e poderá não realizar o ato fraudulento.

A fraude ocorre por meio da facilitação realizada pelo agente público responsável pelo processo no órgão licitador (ou contratante) que, direcionando a licitação com a imposição de exigências referentes à diversas especificações técnicas, exclui a maioria das empresas que poderiam participar. Assim, beneficiando determinado fornecedor que ele deseja que ganhe, ambos são favorecidos.

A fraude ainda pode ocorrer apenas entre os participantes mediante a combinação de preços, muitas vezes majorados. Dessa forma, os concorrentes podem estabelecer um rodízio para o vencedor, tornando difícil a sua detecção pela análise de um único processo licitatório (MORAIS, 2016).

Na prática, quando o objetivo é identificar fraudes em um processo licitatório, é possível implementar indicadores específicos. Como exemplo, os indicadores sugeridos pela Transparência Brasil (2019), que foram divididos entre as fases interna e externa das licitações:

a) Fase interna:

— Contratações sem licitação ou tipicamente com baixa competitividade (dispensas, inexigibilidades e convites): quanto maior é o poder da administração em restringir os participantes, maior é o risco de que ocorra um favorecimento;

— Prazo de publicidade exíguo e publicação restrita: o prazo pode ser reduzido para restringir o tempo hábil dos potenciais licitantes prepararem os documentos para a proposta;

— Licitações exclusivas para microempresas (ME) ou empresas de pequeno porte (EPP): nesse caso, empresas de médio e grande porte podem participar como fachada;

— Valor estimado muito próximo ao limite da modalidade adotada: indicador para o fracionamento de despesas;

— Edital com indícios de direcionamento ou cláusulas restritivas: podem ser incluídas para direcionar fornecedores que atendam aos critérios estabelecidos.

b) Fase externa:

— Inabilitações em desacordo com o edital ou por excessiva formalidade: reprovação intencional de um participante com o objetivo de afastá-lo do certame;

— Escassez de licitantes e fraca disputa: a baixa competitividade pode ser um indício de direcionamento;

— Desconto atípico (excessivamente baixo ou alto) em relação ao valor estimado para a licitação: quando há um conluio entre os participantes, as propostas podem apresentar valores muito próximos, com variação atípica;

— Licitações vencidas por ME ou EPP: empresas de médio ou grande porte podem criar MEs ou EPPs para aproveitar o tratamento preferencial.

A análise de casos divulgados na imprensa também pode trazer indicativos para serem considerados. Por exemplo, a Operação Desmascarados, que deflagrou um esquema de fraude na contratação de uma empresa especializada em manutenção predial pelo Hospital Universitário Gaffrée e Guinle, no Rio de Janeiro. Segundo notícia veiculada pelo *G1*, em março de 2022, a investigação apontou que a empresa vencedora do certame havia sido constituída há apenas um mês do procedimento licitatório, com o quadro societário composto por laranjas (PF APURA…, 2022). Essa análise pode ser realizada de forma massiva na base de dados, por meio da utilização da base de CNPJs divulgada pela Receita Federal.

Outro caso trouxe a suspeita com relação à Classificação Nacional de Atividade Econômica (CNAE). O jornal *Mais Oeste* divulgou, em setembro de 2021, um caso de fraude na contratação do serviço de transporte escolar para a prefeitura de Wanderley/BA (PREFEITA…, 2021). Durante a investigação realizada pelo Ministério Público Federal (MPF), foi verificado que a atividade “transporte escolar” foi incluída no CNAE do vencedor por alteração societária em agosto de 2019, mais de cinco meses depois da licitação ter ocorrido. Além de inclusões próximas à data da disputa, é possível verificar o número de CNAEs que o CNPJ possui no seu cadastro. Quanto maior for a quantidade de registros, é possível que a empresa não seja especialista em nenhum negócio, e realize inclusões apenas para participar de licitações conforme requerido.

1.4 **Lei de Benford**

A Lei de Benford, também conhecida como Newcomb-Benford, foi criada por Frank Benford, em 1938, inspirada em Simon Newcomb. Por volta de 1881, Newcomb verificou que as primeiras páginas das tabelas de logaritmos se desgastavam mais do que as últimas, e que, portanto, qualquer lista de números tirados de um conjunto aleatório terá uma frequência maior para os números começados por "1", diminuindo progressivamente até o dígito "9".

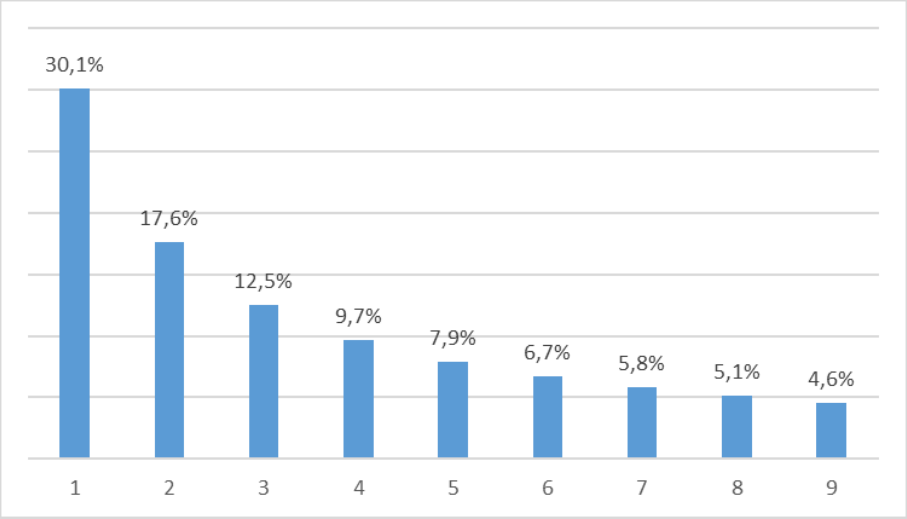
Benford coletou diversos conjuntos numéricos, como populações, taxas de mortalidade, áreas de rios, e verificou que todos seguiam a mesma distribuição. Assim, sempre que houver um conjunto numérico em que a ocorrência se dá de forma natural e não manipulada, a frequência do primeiro dígito deve ser conforme o padrão esperado. A análise pode ser realizada tanto a partir do primeiro dígito quanto do segundo, e ainda do terceiro e do quarto. As frequências devem seguir o padrão apresentado na Tabela 2 e no Gráfico 1.

Tabela 2 – Probabilidade conforme posição dos dígitos

Dígito	**Posição**			
	Primeiro	**Segundo**	**Terceiro**	**Quarto**
0		.11968	.10178	.10018
1	.30103	.11389	.10138	.10014
2	.17609	.10882	.10097	.10010
3	.12494	.10433	.10057	.10006
4	.09691	.10031	.10018	.10002
5	.07918	.09668	.09979	.09998
6	.06695	.09337	.09940	.09994
7	.05799	.09035	.09902	.09990
8	.05115	.08757	.09864	.09986
9	.04576	.08500	.09827	.09982

Fonte: NIGRINI, 2012, p. 6.

Gráfico 1 – Representação gráfica da frequência do primeiro dígito



Fonte: Adaptado de NIGRINI, 2012, p. 6.

A Lei de Benford pode ser utilizada na investigação de fraudes. Quando uma base de dados não está aderente a essa teoria, ou seja, se a frequência dos registros estiver fora do esperado, os seus desvios devem ser analisados. Nigrini (2012) sugere que a Lei de Benford deve ser empregada como ponto de partida e o teste preferido seria o dos dois primeiros dígitos, a menos que o conjunto de dados seja pequeno. A partir dos desvios, pode-se fazer a combinação com outras avaliações, como a de registros duplicados, associados a conhecimentos específicos sobre o processo ou negócio.

O teste de aderência do qui-quadrado pode ser realizado para aferir a hipótese de que a base de dados está aderente à Lei de Benford. Para isso, a seguinte fórmula poderá ser utilizada:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

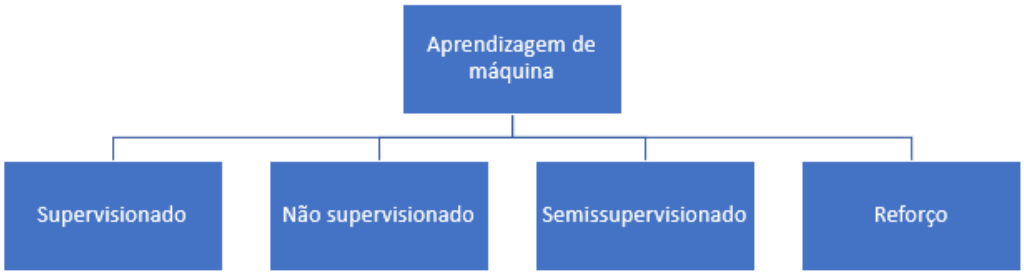
Nesse caso, o termo O_i representa a frequência observada para o dígito, e E_i, a frequência esperada. Quando o resultado do teste de qui-quadrado for superior ao valor crítico para o nível de significância estipulado, a hipótese de aderência à Lei de Benford deverá ser rejeitada.

Como exemplo da sua aplicação em estudos brasileiros, Cella e Zanolla (2018) propuseram a análise da Lei de Benford sobre o *ranking* de transparência dos municípios do estado de Goiás e a base de despesas empenhadas. A pesquisa verificou que os municípios com mais aderência à lei também são os considerados mais transparentes, de acordo com o *ranking* divulgado pelo MPF.

1.5 **Aprendizado de máquina**

O aprendizado de máquina aborda a questão de como construir programas de computador que melhoram o seu desempenho em alguma tarefa, por meio da experiência (MITCHELL, 1997). A partir dessa declaração, é possível perceber que o aprendizado de máquina não é muito diferente daquele visto em humanos, uma vez que é necessário obter exemplos para poder reproduzi-los e gerar conhecimento. Esse método pode ser desenvolvido mediante técnicas de aprendizagem supervisionada, não supervisionada, semissupervisionada e por reforço (Figura 1).

Figura 1 – Tipos de aprendizagem de máquina



Fonte: A autora, 2022.

O aprendizado supervisionado ocorre quando existe o objetivo de prever uma variável dependente a partir de variáveis independentes. Para realizá-lo, a base de dados precisa conter exemplos de casos com a resposta desejada. Assim, os algoritmos aprenderão conforme as respostas já conhecidas e farão uma previsão. Nessa técnica, temos os algoritmos de classificação que irão rotular os dados da base. Como exemplo, é possível citar as análises de crédito de clientes em que, a partir do histórico de pagamentos, eles podem ser rotulados entre possíveis adimplentes e inadimplentes.

No aprendizado não supervisionado, o rótulo dos dados não é conhecido. Nesse caso, são utilizados algoritmos de agrupamento que irão gerar grupos por semelhança. Um exemplo que se pode citar é a Netflix, que agrupa os filmes sugeridos pelo histórico das produções já visualizadas.

O aprendizado semissupervisionado é utilizado quando há o desbalanceamento da base de dados, ou seja, existem poucos exemplos de dados rotulados e muitos não rotulados. Em um primeiro momento, pode-se realizar técnicas para agrupamento das informações e fazer a rotulação levando em conta os padrões verificados. Em seguida, as previsões poderão ser feitas através de técnicas de aprendizado supervisionado.

No aprendizado por reforço, a máquina irá aprender por intermédio das tentativas realizadas, das recompensas quando o objetivo for atingido e das punições quando ocorre a falha. Essa abordagem é utilizada nos jogos, em que se estabelece o objetivo de vencer uma partida e, com base nas jogadas corretas e erradas, a máquina irá aprender a jogar.

Nesta pesquisa será utilizada uma abordagem de aprendizado supervisionado mediante a utilização de algoritmos de classificação. A rotulação dos dados para identificar se um fornecedor foi sancionado ou não é disponibilizada pelo portal da Transparência do Governo do Estado do Rio de Janeiro.

1.6 **Algoritmos de classificação**

Nesta seção serão apresentados os algoritmos de classificação utilizados na pesquisa: regressão logística, KNN, redes neurais artificiais, *random forest* e máquina de vetor suporte (SVM). Além do conceito utilizado individualmente para os algoritmos, serão apresentados exemplos de aplicações na identificação de fraudes em cada um deles.

1.6.1 Regressão logística

A regressão logística é um algoritmo utilizado para resolver problemas de classificação. A variável dependente tem apenas duas categorias, que pode ser "1" no caso de o teste ser positivo e "0" no caso negativo. Segundo Fernandes *et al*. (2020), para melhor entender o funcionamento desse algoritmo, é necessário compreender a lógica da análise de regressão linear de forma geral:

$$Y = \alpha + \beta X + \varepsilon \tag{2}$$

Y – variável dependente (o que queremos predizer);
α – valor de Y quando X assume valor zero;
β – variação observada em Y, associada ao aumento de uma unidade em X;
X – variável independente;
ε – erro do modelo.

O modelo de regressão logística pode ser escrito da seguinte maneira:

$$P(Y = 1) = \frac{1}{1+e^{-g(x)}} \tag{3}$$

Em que,

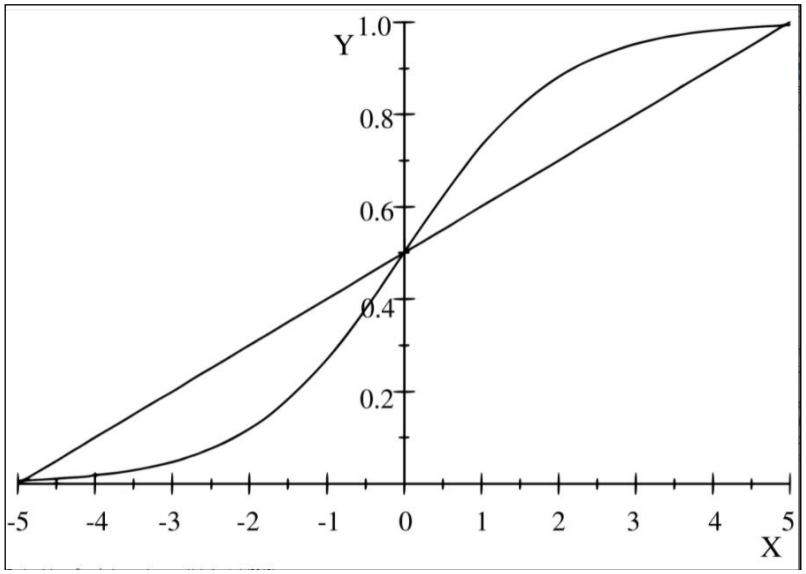
$$g(x) = B_0 + B_1X_1 + \cdots + B_pX_p \tag{4}$$

Os coeficientes B são estimados por meio do conjunto de dados pelo método da máxima verossimilhança, o qual encontra uma combinação de coeficientes que maximiza a probabilidade de a amostra ter sido observada. Observa-se que a curva logística tem um

comportamento probabilístico no formato da letra S, uma característica da regressão logística (HOSMER; LEMESHOW, 1989).

O Gráfico 2 ilustra a reta da regressão linear *versus* a curva da regressão logística.

Gráfico 2 – Reta de regressão linear *versus* curva logística



Fonte: FERNANDES *et al*., 2020, p. 3.

Como se pode observar no gráfico acima, quando a variável independente X assume valores menores, a probabilidade da variável dependente Y ser zero é maior, assim como o contrário também é válido. Como a variável dependente Y é aquela que pretendemos predizer, quando assume um valor mais próximo a 0 significa que a probabilidade de o fornecedor ser sancionado é menor. Em contrapartida, quando assume um valor mais próximo a 1, a probabilidade de sanção é maior.

> Uma das vantagens da regressão logística é que basta saber se um evento (por exemplo, comprar ou não, o sucesso ou falha de uma empresa) ocorreu e, então, usar um valor dicotômico como variável dependente. A partir desse valor, a regressão prediz a estimativa de probabilidade de ocorrência ou não do evento. Se a probabilidade prevista é maior do que 0,5, então a predição é correta; caso contrário, não. (ALVES, 2005, p. 143).

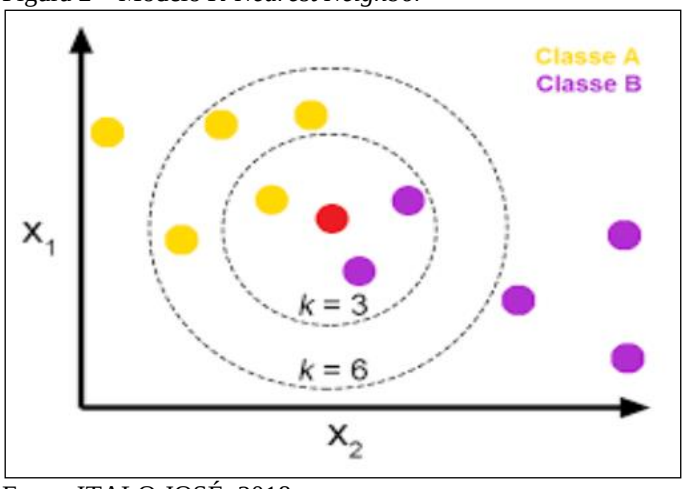
Oliveira (2016) aplicou o método de regressão logística para identificar fraudes nos cartões de crédito. O experimento propôs uma arquitetura baseada em regras de associação e regressão logística para produzir um conjunto de padrões de detecção de transações fraudulentas. Primeiro, o algoritmo *FP-Growth* foi aplicado para identificar os modelos de associação que levassem determinadas características das transações às fraudes. Em seguida, foram extraídos cinco conjuntos de regras, os quais foram submetidos à regressão logística. Por

fim, foi possível observar que o uso dessa técnica se mostrou efetivo, com resultados melhores do que as abordagens usuais de contagem de evidências de fraude.

1.6.2 KNN

O *K-Nearest Neighbor* (KNN) é um algoritmo de classificação que tem como objetivo determinar a qual grupo um registro pertencerá com base nos seus vizinhos. Cada registro é transformado em um vetor, que será classificado em grupos conforme a sua proximidade. A Figura 2 mostra vetores já classificados entre as classes A e B, o algoritmo deverá determinar em qual grupo classificar o novo registro (em vermelho).

Figura 2 – Modelo *K-Nearest Neighbor*



Fonte: ITALO JOSÉ, 2018.

Para realização do cálculo, utiliza-se a distância euclidiana, conforme fórmula a seguir:

$$E(x, y) = \sqrt{\sum_{i=0}^{n} (xi - yi)^2} \quad (5)$$

Nessa fórmula, x e y são os vetores que serão comparados. Quanto menor o resultado, menor será a distância entre os dois pontos. No caso da pesquisa, o algoritmo dividirá os dados em grupos de fornecedores, com e sem sanção, para depois identificar qual a menor distância entre os vizinhos e realizar a classificação.

Olokodana e Fernandes (2020) utilizaram o KNN, entre outros algoritmos de classificação, para identificar fraudes nas demonstrações contábeis de 110 empresas listadas na

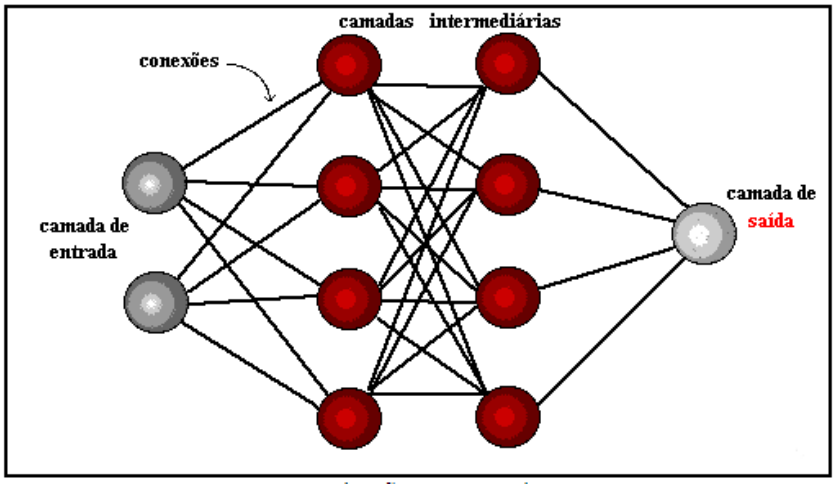
Bovespa. Como resultado, apesar da utilização do algoritmo, a pesquisa indicou que o modelo de regressão logística foi mais eficiente para a previsão das fraudes contábeis.

1.6.3 Redes neurais artificiais

Os estudos das redes neurais artificiais foram inspirados pela observação do sistema biológico de aprendizado, que é constituído de redes complexas de neurônios interconectados. Cada unidade de neurônio recebe *inputs* e produz *outputs*, que, por sua vez, servirão de *input* para o próximo neurônio (MITCHELL, 1997).

De uma maneira geral, podemos observar na Figura 3 um exemplo de esquema de rede neural artificial. Ela recebe os dados de entrada que são processados nos neurônios localizados nas camadas posteriores até trazerem o resultado final na camada de saída. As redes podem apresentar diversas combinações de camadas.

Figura 3 – Rede neural



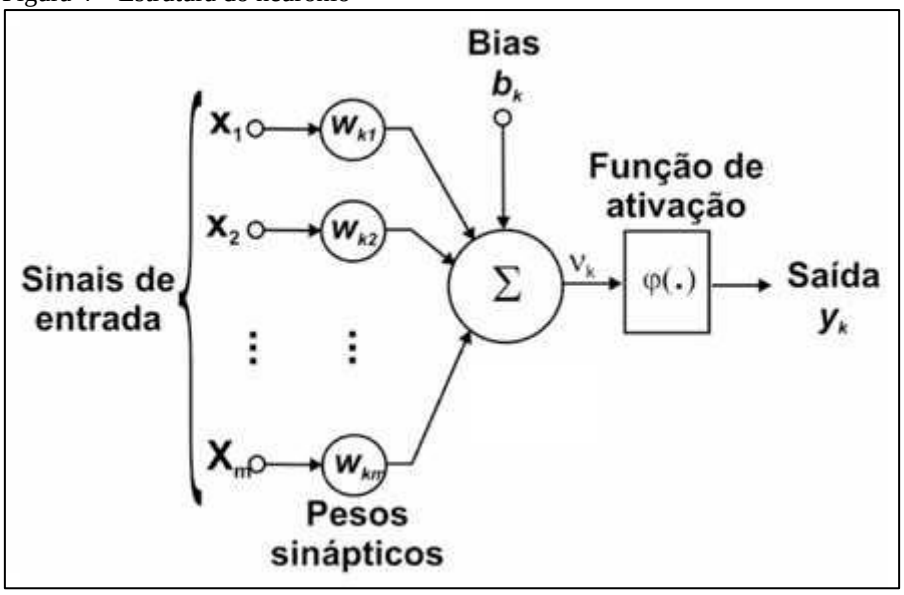
Fonte: CARVALHO, 2009.

Os principais elementos de uma rede neural artificial são:

a) Dados de entrada: informações que irão entrar na rede neural, passar pelos neurônios e produzir o *output* final;
b) Camadas escondidas: são as camadas intermediárias, com os neurônios responsáveis pelos cálculos que irão encontrar os valores finais;
c) Pesos: números reais aplicados aos neurônios, que expressam as respectivas importâncias das entradas para a saída;
d) Bias: medida utilizada para verificar a propensão do resultado da rede ser igual a 1;
e) Funções de ativação: pode ser entendida como uma camada a mais no processamento que irá decidir se o neurônio na rede deve ser ativado ou não para a solução do problema;
f) Dados de saída: resultado final da rede neural artificial.

A Figura 4 demonstra a estrutura de um neurônio e cada um dos componentes apresentados.

Figura 4 – Estrutura do neurônio



Fonte: OLIVEIRA, 2009, p. 15.

As arquiteturas são organizadas em camadas, com unidades conectadas à camada posterior. Os pesos de cada unidade são ajustados por treinamento que será dado a partir de exemplos.

As funções de ativação desempenham um papel importante na rede e podem ser de diferentes tipos. A seguir, são apresentados os principais exemplos dessas funções:

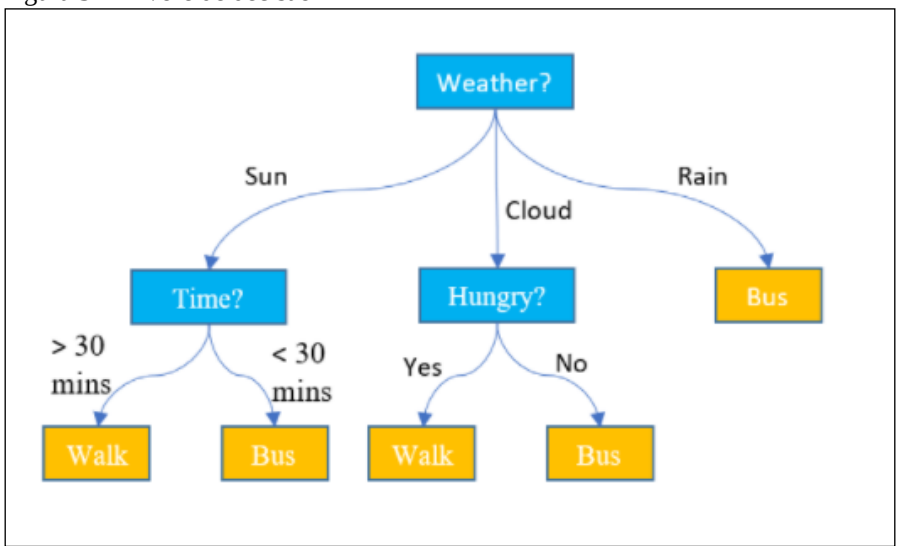
a) Linear: aplica um valor de multiplicação ao valor que recebe. É uma função mais limitada para entender relações mais complexas entre os dados;
b) Sigmoide: produz valores de intervalo entre [0, 1]. Sua utilização pode ser eficaz em problemas de classificação;
c) Softmax: semelhante à função sigmoide e pode ser usada em problemas de classificação com mais de duas classes. Transforma a saída de cada classe em valores entre 0 e 1 e a divide pela soma das saídas;
d) Tanh: semelhante à sigmoide, com variação de entre -1 e 1;
e) ReLU: função não linear que produz resultados no intervalo de [0, ∞]. Como transforma os resultados negativos em 0, faz com que os neurônios com esse valor não sejam ativados.

Oliveira e Santos (2020) propuseram um modelo baseado em redes neurais para a identificação de contribuintes mais propensos a assumir a condição de sonegadores do Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação (ICMS). Foram utilizadas informações extraídas dos bancos de dados da Secretaria da Fazenda do Estado da Bahia (SEFAZ-BA), que gerou uma amostra de 74.447 eventos. Na fase de treinamento, a rede apresentou um índice de acerto de 71% na classificação dos contribuintes passíveis de autuação (ou não). Em relação aos contribuintes que foram autuados, a performance foi de 94%.

1.6.4 *Random forest*

Traduzido para o português como floresta aleatória, o algoritmo *random forest* soluciona problemas de classificação criando diversas árvores de decisão de forma aleatória. A árvore de decisão faz previsões com base em perguntas para determinar qual ramo seguir e assim chegar a uma conclusão. As perguntas são consideradas os nós que serão ramificados nas respostas. Em uma base de dados, as perguntas serão os atributos e o resultado será o campo alvo, que no caso desta pesquisa é verificar se um fornecedor foi sancionado ou não. A Figura 5 mostra como uma árvore de decisão pode ser representada.

Figura 5 – Árvore de decisão



Fonte: HOARE, 2018.

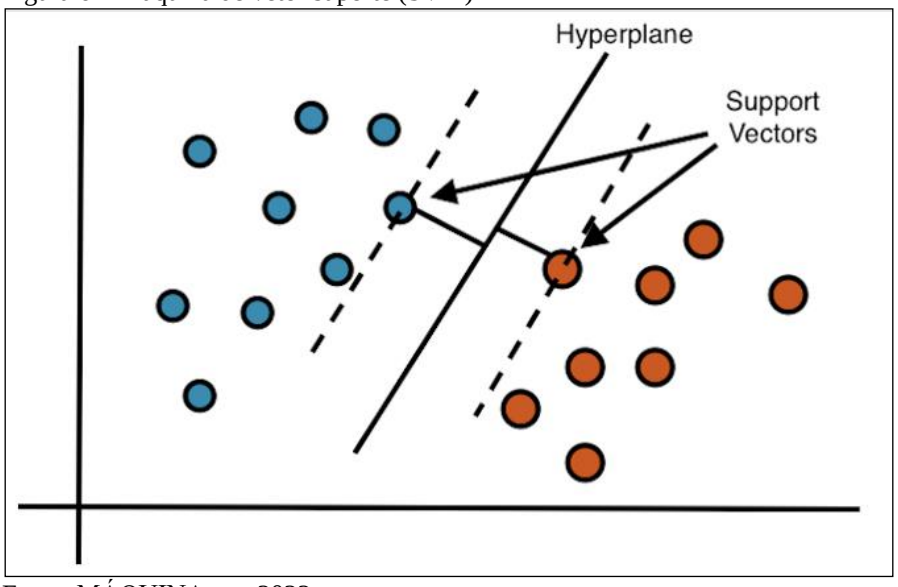
No caso da *random forest*, o algoritmo utilizará uma amostra de registros da base de dados de treino e selecionar dois ou três atributos de modo aleatório para definir qual será usado como o primeiro nó. O mesmo ocorrerá para as definições seguintes. O processo será repetido e diversas árvores de decisão serão criadas. No final, o resultado que mais vezes for apresentado será o escolhido para definição do modelo.

Santos (2020) utilizou o *random forest* em sua pesquisa para a identificação de fraude em um grupo internacional privado que faz a gestão dos seguros de saúde de algumas seguradoras. Foram utilizados os dados de pedidos de reembolso que haviam sido classificados como legítimos ou suspeitos por auditores. Os resultados demonstraram que o *random forest* foi um dos algoritmos que apresentou os melhores resultados, com 65% de precisão.

1.6.5 Máquina de vetor suporte (SVM)

A máquina de vetor suporte (em inglês, *support vector machine* — SVM) é um algoritmo que pretende encontrar uma linha de separação, também conhecida como hiperplano, entre diferentes classes. Os vetores suporte serão os pontos que se encontram mais próximos ao hiperplano. A Figura 6 demonstra graficamente como é a definição do hiperplano e dos vetores suporte.

Figura 6 – Máquina de vetor suporte (SVM)



Fonte: MÁQUINA…, c2022.

Esse algoritmo tem como objetivo encontrar o hiperplano ideal, em que as margens (distância entre os vetores suporte e o hiperplano) serão maiores. O *Kernel* no SVM consiste no conjunto de funções matemáticas utilizadas para determinar o hiperplano ideal, podendo ser do tipo linear, radial, polinomial ou sigmoide.

Severino e Yaohao (2019) realizaram uma predição para sinistros patrimoniais com fraudes, mediante a aplicação de modelos de aprendizado de máquina (SVM e *random forest*), com base em 19 variáveis explicativas. Os resultados preditivos foram comparados com a regressão logística e mostraram que ambos os modelos obtiveram desempenho médio superior à regressão logística.

1.7 **Métricas de avaliação**

Nesta seção serão apresentadas as métricas de avaliação utilizadas na pesquisa: precisão, matriz de confusão e curva ROC.

1.7.1 Precisão

Para avaliação do resultado dos algoritmos, verifica-se a precisão, que representa a média global de acertos na classificação. Essa métrica é calculada pela razão entre o total de acertos e o total de registros. Entretanto, o resultado não deve ser utilizado sozinho, pois ele

não é capaz de detalhar se os acertos ocorrem na classe majoritária (de fornecedores adequados) ou na minoritária (de fornecedores sancionados), que é o que realmente se pretende prever.

1.7.2 Matriz de confusão

Para complementar essa avaliação, utiliza-se a matriz de confusão, que é executada para entender a relação dos erros e acertos do modelo. Enquanto a acurácia mede o percentual total de acertos, a matriz de confusão irá detalhar esses acertos. No caso desta pesquisa, os resultados serão classificados conforme abaixo:

a) Verdadeiro positivo (TP): a classe prevista e a base observada correspondem a fornecedores não sancionados;
b) Falso negativo (FN): o fornecedor foi sancionado, mas a previsão era de que não seria;
c) Falso positivo (FP): o fornecedor não foi sancionado, mas a previsão era de que fosse;
d) Verdadeiro negativo (TN): tanto a previsão quanto a base observada indicam que o fornecedor deveria ser sancionado.

A Figura 7 ilustra o resultado da matriz de confusão, comparando o valor real da base com o valor obtido na predição.

Figura 7 – Matriz de confusão

		Valor Predito	
		1	**0**
Real	**1**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	0	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: A autora, 2022.

Na prática de uma auditoria, existem dois riscos: os falsos negativos podem deixar passar despercebido um fornecedor que esteja cometendo algum tipo de irregularidade, e os falsos positivos fazem com que o auditor perca tempo investigando um fornecedor que esteja

fazendo tudo adequado. Como o tempo do auditor é um recurso escasso, reduzir o número de falsos positivos inicialmente pode ser uma opção.

1.7.3 Curva ROC

A curva *Receiver Operating Characteristic* (ROC) é uma representação gráfica do desempenho do classificador. No gráfico, o eixo Y representa a sensibilidade, que é a capacidade de um evento ocorrer dentro da amostra (verdadeiro positivos / positivos totais), e o eixo X é a especificidade, que é a capacidade de um evento verdadeiro ocorrer dentro da amostra (verdadeiro negativo / negativos totais). Na leitura gráfica, a curva que tiver a parte superior esquerda maior será aquela que tem o melhor desempenho do classificador.

1.8 **Metodologia CRISP-DM**

A metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM) foi criada por Wirth e Hipp (2000) para orientar a execução de projetos de mineração de dados, na qual são propostas seis fases, conforme abaixo:

a) Compreensão do negócio: fase inicial, em que se entende quais são os objetivos do negócio, para que se possa entender como a mineração de dados poderá contribuir para o atingimento dos objetivos;
b) Compreensão dos dados: fase de coleta dos dados disponíveis, avaliação da qualidade e mapeamento das possíveis informações que podem ser obtidas;
c) Preparação dos dados: com os dados crus, realiza-se a geração das informações que serão utilizadas na modelagem;
d) Modelagem: fase em que são aplicadas as técnicas de modelagem. A calibragem é realizada para a obtenção dos melhores resultados;
e) Avaliação de desempenho: verificação dos resultados, conferência dos métodos utilizados e revisitação de etapas para garantir a efetividade do modelo;
f) Implantação: implementação do sistema em produção.

Esta pesquisa segue a metodologia CRISP-DM, pois segue as fases propostas por Wirth e Hipp (2000). Assim, num primeiro momento, deve-se buscar o entendimento do processo de

licitação e o objetivo de identificar fraudes cometidas pelos fornecedores; em seguida, devem ser realizadas a coleta de dados disponíveis, a preparação dos atributos, a aplicação dos algoritmos e a avaliação dos resultados. A última etapa seria a implementação propriamente dita, que poderia ser aplicada pelo Governo do Estado do Rio de Janeiro.

2 METODOLOGIA

Nesta seção será descrito o método e os passos utilizados para a elaboração da pesquisa, iniciando pela população e amostra, seguida da coleta e tratamento dos dados, que teve como produto final a construção da base de dados para o desenvolvimento do modelo (levantamento de características e definição de indicadores), na aplicação dos algoritmos. Por fim, tem-se a avaliação dos desempenhos de cada um deles.

2.1 População e amostra

O Sistema Integrado de Gestão (Siga), desenvolvido pela Secretaria de Estado de Planejamento e Gestão do Rio de Janeiro para atender toda a cadeia de suprimentos de bens e serviços do estado, possui, entre outras informações dos processos, os dados dos fornecedores, contratos e licitações realizadas. Apesar de disponibilizar informações desde 2010, optou-se por empregar na pesquisa apenas as mais atuais, referentes aos últimos cinco anos (2017-2021). Assim, pretende-se avaliar o comportamento mais recente dos fornecedores, eliminando o efeito de condutas que podem ter sido extintas ao longo do tempo por alterações na legislação ou nos processos. Utilizou-se também os dados da Receita Federal contendo o cadastro de todos os CNPJs registrados. Dessa forma, as bases de dados utilizadas nesta pesquisa estão listadas e descritas no Quadro 1.

Quadro 1 – Bases de dados utilizadas

Fonte	**Nome da tabela**	**Descrição**
Receita Federal	CNPJ	Dados do CNPJ principal da empresa, como capital social e porte.
	Estabelecimentos	Dados de todos os estabelecimentos (matriz e filial), contendo localização, data de abertura, situação cadastral, data da situação cadastral, CNAEs cadastrados e dados de contato.
	Sócios	Dados do quadro societário da empresa e representantes legais.
Portal Siga	Fornecedores	Fornecedores cadastrados para a participação em licitações e realização de compras.
	Contratos	Contratos firmados com os fornecedores, contendo *status*, valor contratado e valor empenhado.
	Licitações	Licitações realizadas, com *status* e valor total.
	Participantes licitações	Lista de participantes por licitação, contendo os lances dados e a informação dos vencedores.
	Compras diretas	Compras diretas aprovadas por fornecedor, com valor unitário e valor total da compra.
	Sanções	Sanções aplicadas por fornecedor, contendo a natureza e o *status* da sanção.

Fonte: A autora, 2022.

Ao final do processo de preparação da base de dados, foi observada a seguinte relação entre a população total das bases de dados originais e a amostra utilizada na pesquisa (Tabela 3).

Tabela 3 – População e amostra do sistema Siga

Nome da tabela	População total		Amostra (2017-2021)	
	Quantidade	Valor (R$ mil)	Quantidade	Valor (R$ mil)
Fornecedores	39.587	-	6.067	-
Contratos	82.019	68.100.627	24.758	31.063.203
Compras diretas	94.736	30.592.138	14.158	13.325.539
Fornecedores Sancionados	797	-	240	-

Fonte: A autora, 2022.

2.2 **Coleta de dados**

Para viabilizar esta análise, o primeiro passo consistiu na coleta dos dados de fornecedores, licitações e contratos divulgados no portal de compras do estado do Rio de Janeiro, e da base com todos os CNPJs cadastrados na Receita Federal.

As informações de fornecedores, licitações e contratos foram extraídas em 5 de janeiro de 2022, no Portal Siga.[2] O *site* contém todo o histórico de licitações realizadas desde 2010, data de implantação do sistema. A base de CNPJs, por sua vez, foi extraída do portal da Receita Federal,[3] em 20 de outubro de 2021. Tanto os arquivos do sistema Siga quanto os da Receita Federal são gerados no formato ".csv".

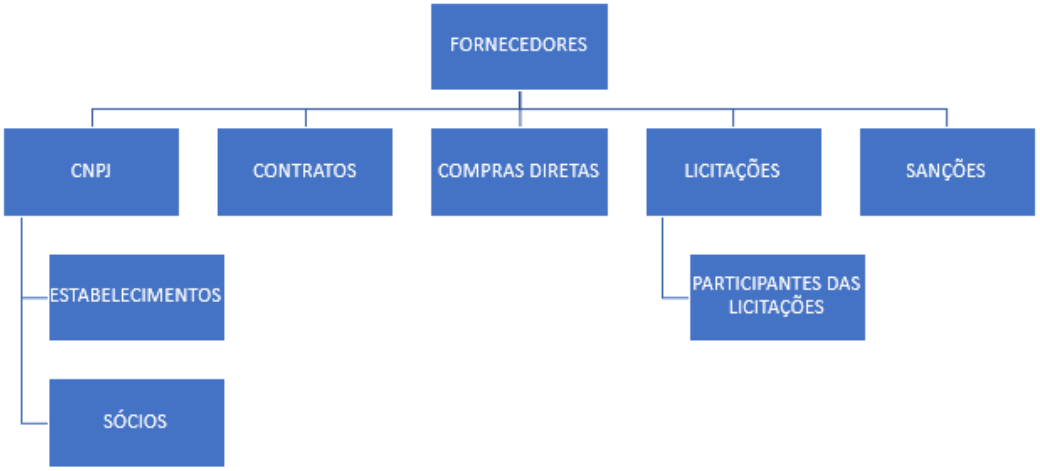
A primeira etapa do estudo consistiu na preparação da base de dados para utilização dos algoritmos de classificação. A ferramenta utilizada foi o *Audit Command Language* (ACL), em que se realizou a importação de todas as tabelas mencionadas no item anterior. Com isso, realizou-se o filtro dos anos de escopo, a combinação das informações entre as tabelas e, por fim, a geração dos atributos (indicadores) para serem utilizados no modelo de classificação.

Como já mencionado, apesar de o Siga ter informações desde 2010, foram utilizados nesta pesquisa apenas os dados dos últimos cinco anos (2017-2021), formando o primeiro filtro da base de dados.

O segundo passo consistiu na combinação entre as bases de dados, utilizando como tabela principal as informações de fornecedores, conforme esquema ilustrado na Figura 8.

[2] O endereço eletrônico do Portal Siga é: https://www.compras.rj.gov.br/Portal-Siga/Principal/extracaoTotal.action.
[3] O endereço eletrônico do portal da Receita Federal é: https://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoasjuridicas-cnpj/dados-publicos-cnpj.

Figura 8 – Esquema das bases de dados



Fonte: A autora, 2022.

Realizou-se um filtro na base de fornecedores e foram considerados somente aqueles que possuíam CNPJ (a base também continha pessoas físicas como fornecedores), que atuam no Brasil e que tiveram pelo menos um contrato firmado ou participaram de alguma licitação no período do recorte do estudo. Desse modo, foram criados os campos que serão utilizados como atributos (indicadores) no modelo de classificação. O *layout* final da base pode ser verificado no Quadro 2, com o total de 15 atributos e um campo alvo da classificação (SANÇÃO).

Quadro 2 – Atributos considerados no modelo final (continua)

Ordem	**Nome do campo**	**Descrição**	**Premissa utilizada para seleção do indicador**
1	NR_CNPJ	Número do CNPJ do fornecedor.	Campo de identificação do fornecedor.
2	ME_EPP	Identificação se a empresa é uma Microempresa (ME) ou Empresa de Pequeno Porte (EPP), sendo “1” caso positivo e “0” caso negativo.	Alguns processos licitatórios são destinados a microempresas ou empresas de pequeno porte. Empresas maiores podem querer se beneficiar desses autos e criar novos CNPJs para participar.
3	QTD_ CONTRATOS	Quantidade de contratos registrados para o fornecedor no período.	Campo informativo para identificar se a empresa tem relevância com relação à quantidade de contratos firmados.
4	VLR_ CONTRATOS	Valor dos contratos registrados para o fornecedor no período.	Campo informativo para identificar se os contratos possuem relevância de valor. Quanto menor o valor dos contratos, menor poderá ser a concorrência de outros fornecedores no processo licitatório.

Quadro 2 – Atributos considerados no modelo final (continuação)

Ordem	**Nome do campo**	**Descrição**	**Premissa utilizada para seleção do indicador**
5	TOTAL_LIC	Total de licitações em que o fornecedor participou no período.	Campo informativo para identificar a participação em licitações.
6	TOTAL_LIC_ VENC	Total de licitações que o fornecedor foi o vencedor no período.	Campo informativo para identificar a quantidade de licitações vencidas. A taxa de vitória baixa pode indicar licitantes que participam como fantasma, apenas para simular falsa concorrência. Por outro lado, uma taxa alta pode indicar favorecimento.
7	QTD_ COMPRAS_ DIRETAS	Quantidade de compras diretas registradas para o fornecedor no período.	Uma quantidade grande de compras diretas pode indicar processos criados com direcionamento a um determinado fornecedor.
8	VLR_ COMPRAS_ DIRETAS	Valor de compras diretas registradas para o fornecedor no período.	Identificação da relevância do valor das compras diretas.
9	CAPITAL_ SOCIAL	Capital social registrado para o CNPJ na Receita Federal.	Um capital social baixo, comparado ao serviço que está sendo contratado, pode indicar que o fornecedor não tem a estrutura necessária para prover o que se pede.
10	QTD_ CNAES	Total de CNAEs registrados para o CNPJ na Receita Federal.	Quanto maior a quantidade de CNAEs, menor é a especialização do fornecedor. As empresas podem incluir diferentes CNAEs para participarem de diferentes processos que não são a sua especialidade.
11	DISTÂNCIA	Distância entre a sede do fornecedor registrada no cartão no CNPJ e o local da realização do serviço (RJ), em que: “1” – fornecedores do RJ “2” – fornecedores da região Sudeste “3” – fornecedores da região Nordeste ou Sul “4” – fornecedores da região Norte	Quanto mais distante é o fornecedor do lugar onde o serviço será prestado, maior será a indicação de que ele não está participando de boa-fé do processo.
12	ANOS_ CRIAÇÃO	Diferença em anos entre a data de cadastro do fornecedor na base de dados do Governo do Estado do RJ e a criação da empresa na Receita Federal.	Empresas recém-criadas podem indicar que foram feitas exclusivamente para a participação no processo licitatório.
13	QTD_ EMPRESAS	Quantidade de empresas na base de fornecedores em que o fornecedor possui sócios em comum.	Fornecedores sancionados podem criar novos CNPJs para continuarem a participar de processos licitatórios, mantendo o mesmo quadro societário.
14	SITUAÇÃO_ RECEITA	Indica se a empresa se encontra ativa na Receita Federal “0” ou se tem outro *status*, como baixada, inapta ou suspensa “1”.	Empresas marcadas como baixadas, inaptas ou suspensas pela Receita Federal não deveriam estar prestando serviços para o estado.
15	BENFORD_ CONTRATOS	Quantidade de contratos firmados com o fornecedor cujos números iniciais falharam no teste de Benford (quantidade acima do limite superior).	Quanto maior a quantidade de contratos que falharam no teste de hipótese da Lei de Benford, maior será a indicação de que os valores dos contratos foram manipulados.

Quadro 2 – Atributos considerados no modelo final (conclusão)

Ordem	**Nome do campo**	**Descrição**	**Premissa utilizada para seleção do indicador**
16	MÉDIA_ IDADE	Média das idades dos sócios da empresa.	Média de idade baixa pode indicar a inclusão de sócios laranja na constituição da empresa.
17	SANÇÃO	Evento de interesse da pesquisa, em que "1" representa fornecedores sancionados e "0" não sancionados.	Campo alvo da classificação.

Fonte: A autora, 2022.

2.3 **Tratamento dos dados**

Depois da preparação da base de dados no ACL, a modelagem foi desenvolvida em *Python*, linguagem de programação de alto nível. Antes da execução dos algoritmos, foram executados os seguintes passos para tratamento das informações: verificação de dados ausentes, balanceamento das bases, divisão da base entre treino e teste, normalização da base de dados e aplicação dos algoritmos, os quais serão detalhados nas subseções a seguir.

2.3.1 Verificação de dados ausentes

Segundo Oliveira (2016), um dado é denominado ausente se o valor de um atributo não for medido ou atribuído, o que pode ser prejudicial ao resultado final do modelo. Portanto, devem ser praticadas estratégias para o contorno dos dados faltantes.

Realizou-se, então, uma verificação na base de dados para verificar a existência de dados ausentes, contudo, nenhuma linha com valores não atribuídos foi identificada.

2.3.2 Balanceamento das bases

Segundo Gadi, Lago e Mehnen (2010), um conjunto de dados para modelagem é perfeitamente balanceado quando a porcentagem de ocorrência de cada classe é 100/n, em que n é o número de classes. Se uma ou mais classes diferem significativamente das outras, esse conjunto de dados é chamado de assimétrico ou desbalanceado.

No caso desta pesquisa, como existem duas classes, cada uma deveria representar 50% da base de dados. Esse desbalanceamento costuma ser observado em problemas de detecção de fraudes, pois, na prática, ocorrem muitas transações legítimas e apenas uma pequena quantidade de transações fraudulentas.

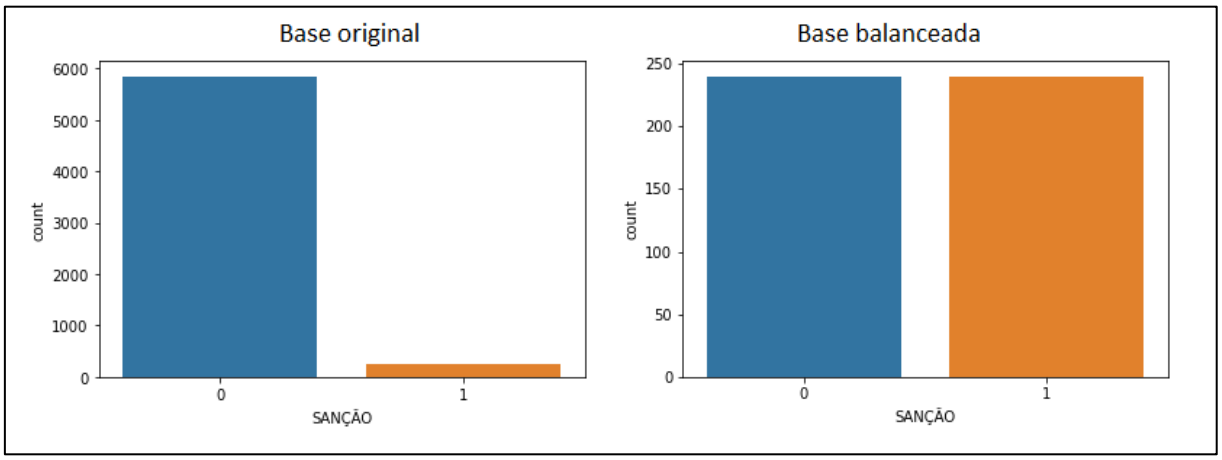
O problema de as classes estarem desbalanceadas é que o modelo pode apresentar uma acurácia muito grande, porque pode ser muito bom em acertar a classificação das classes majoritárias, mas pode não ser capaz de acertar a classificação das classes minoritárias, que é, na verdade, o evento que se tem interesse de prever.

A base de dados de fornecedores do estado do Rio de Janeiro é desbalanceada, uma vez que apenas 240 fornecedores foram sancionados, de um universo de 6.067, o que corresponde a apenas 4% do total.

Para resolver esse problema, utilizou-se a técnica de *undersampling* na base de dados, que consiste em selecionar uma amostra da classe majoritária para que se tenha a mesma quantidade de registros da base minoritária. O comando utilizado no *Python* para a realização do *undersampling* foi o *NearMiss*().

Conforme é possível observar nos gráficos abaixo (Gráfico 3), os fornecedores sancionados passaram de uma representatividade de 4% para um total de 50%.

Gráfico 3 – Balanceamento das bases de dados



Fonte: A autora, 2022.

Outra opção para o balanceamento de bases de dados seria o *oversampling* (utilizando o comando SMOTE() no *Python*), que consiste em gerar dados sintéticos da classe minoritária para que esta se iguale à classe majoritária. Nesse caso, a desvantagem observada seria que o *oversampling* pode transformar a base mais propícia ao *overfitting*. Durante a realização da pesquisa, o *oversampling* foi testado, mas os resultados obtidos foram inferiores ao do *undersampling*. Portanto, optou-se por seguir com os resultados utilizando este último.

2.3.3 Divisão da base entre treino e teste

Depois do rebalanceamento da base, foi realizada a divisão dos dados entre base de treino (70%) e teste (30%), com o objetivo de que o algoritmo fosse capaz de aprender com 70% dos registros, para depois aplicar esse conhecimento nos 30% dos registros faltantes. Nesse caso, é importante que os dados de testes sejam totalmente novos para que não haja influência no resultado final.

2.3.4 Normalização da base de dados

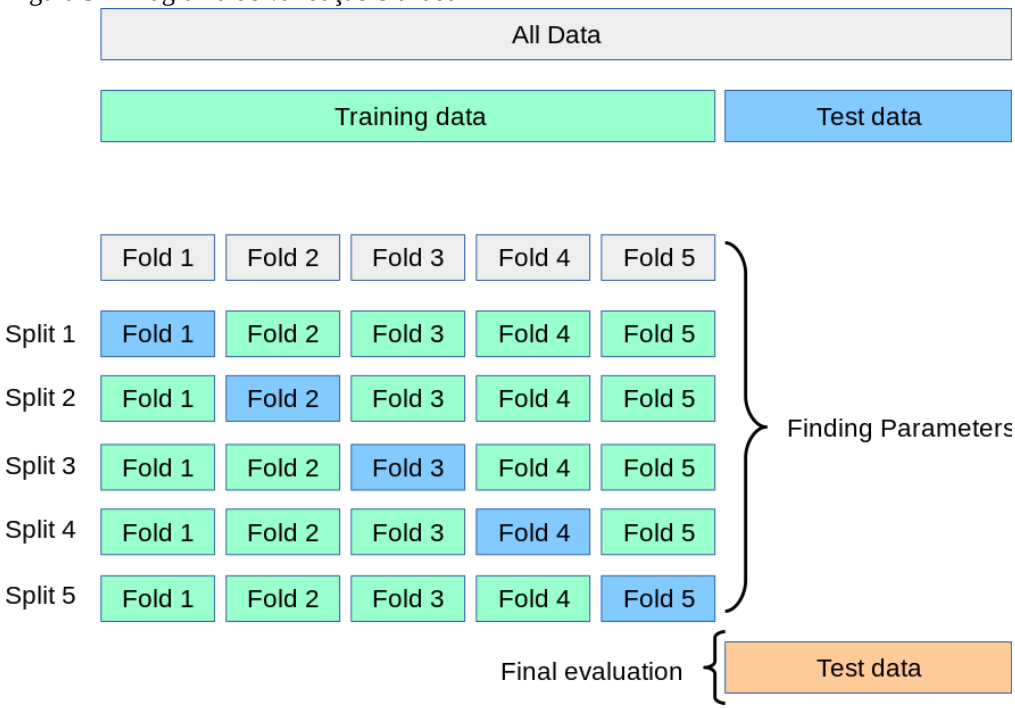
A normalização da base de dados é uma técnica utilizada para evitar que os algoritmos fiquem enviesados para variáveis com maior ordem de grandeza. Nesse sentido, algumas técnicas são utilizadas fazendo com que todos os atributos tenham o mesmo peso na modelagem dos dados. Para melhorar o desempenho dos algoritmos SVM e da rede neural, foi aplicado, para normalização dos dados, o comando *MinMaxScaler*, que dimensiona todos os atributos da base de dados no intervalo [0, 1]. Como o desempenho dos algoritmos KNN, *random forest* e regressão logística já tinham sido acima de 95% de precisão, a normalização não foi adotada

2.3.5 Aplicação dos algoritmos

Para identificar os hiperparâmetros que apresentariam o melhor resultado para cada algoritmo aplicado, utilizou-se o módulo GridSearchCV, cujo objetivo é fazer uma varredura entre uma possibilidade de hiperparâmetros e indicar qual combinação apresenta o melhor resultado para a solução do problema.

O objetivo do GridSearchCV é criar uma combinação de parâmetros e posteriormente avaliá-los. Neste processo é realiza a validação cruzada para garantir que não haja o *overfitting* do modelo, que ocorre quando o modelo tem um desempenho muito bom para classificar os casos na base de treino, mas não consegue classificar corretamente a base de teste. Seria como se o algoritmo decorasse os dados e não fosse capaz de generalizar esse aprendizado. A validação cruzada consiste em dividir o conjunto de treinamento em conjunto menores, e utilizar diferentes combinações para teste. No final, a medida de desempenho será calculada como a média dos valores calculados no processo. Este processo pode ser visualizado no diagrama a seguir (Figura 9):

Figura 9 – Diagrama de validação cruzada



Fonte: CROSS-VALIDATION…, c2022.

O Quadro 3 apresenta os hiperparâmetros utilizados em cada algoritmo.

Quadro 3 – Hiperparâmetros executados

Algoritmo	**Hiperparâmetro**	**Descrição**
KNN	KneighborsClassifier(n_neighbors=2)	Quantidade de vizinhos determinados igual a 2.
Random forest	RandomForestClassifier(max_depth=8,random_state=50)	Profundidade máxima da árvore igual a 8.
Rede neural	MLPClassifier(alpha=0.1, hidden_layer_sizes=12, max_iter=1000, random_state=7, solver='lbfgs')	Rede neural composta de 12 neurônios na camada escondida. Função de ativação padrão utilizada ReLU.
Regressão logística	LogisticRegression(C=0.01, penalty='l2', solver='newton-cg')	Algoritmo utilizado na otimização do problema 'newton-cg'.
SVM	SVC(C=0.1, gamma=0.0001, probability=True, random_state=0)	*Kernel* padrão utilizado no algoritmo rbf (radial).

Fonte: A autora, 2022.

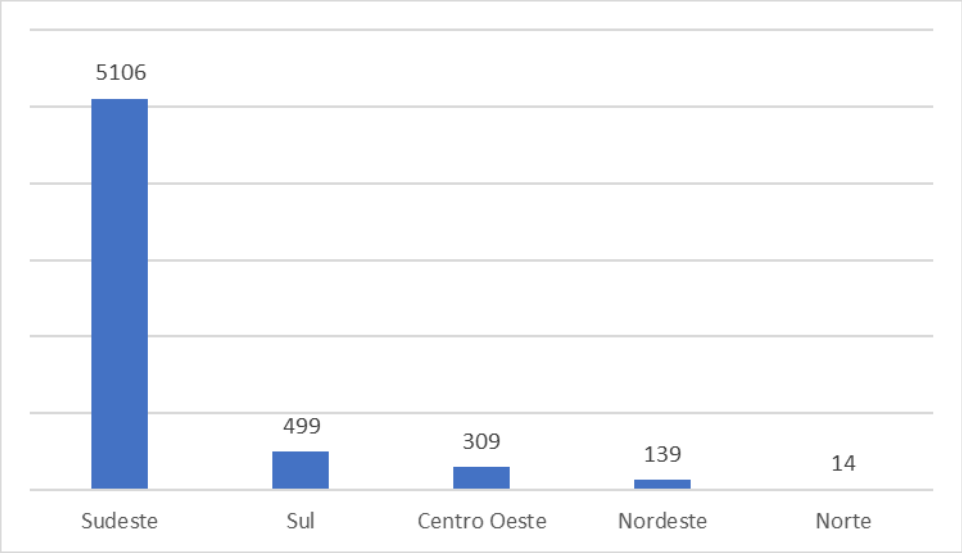
3 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Nesta seção serão detalhados os resultados obtidos na pesquisa, conforme aplicação da metodologia. Primeiro, será realizada uma análise descritiva dos dados. Em seguida, será aplicado o teste de hipótese segundo a Lei de Benford. Depois serão apresentados os resultados do processo, e, por fim, serão propostas sugestões de como melhorar a *performance* dos algoritmos.

3.1 Análise descritiva da base de dados

Durante a realização da análise da base de dados, como esperado, foi identificado que a maioria dos fornecedores está localizada na região Sudeste, próximo ao local de execução dos serviços (RJ). O Gráfico 4 representa a distribuição dos fornecedores avaliados, por região do Brasil em que o CNPJ está cadastrado.

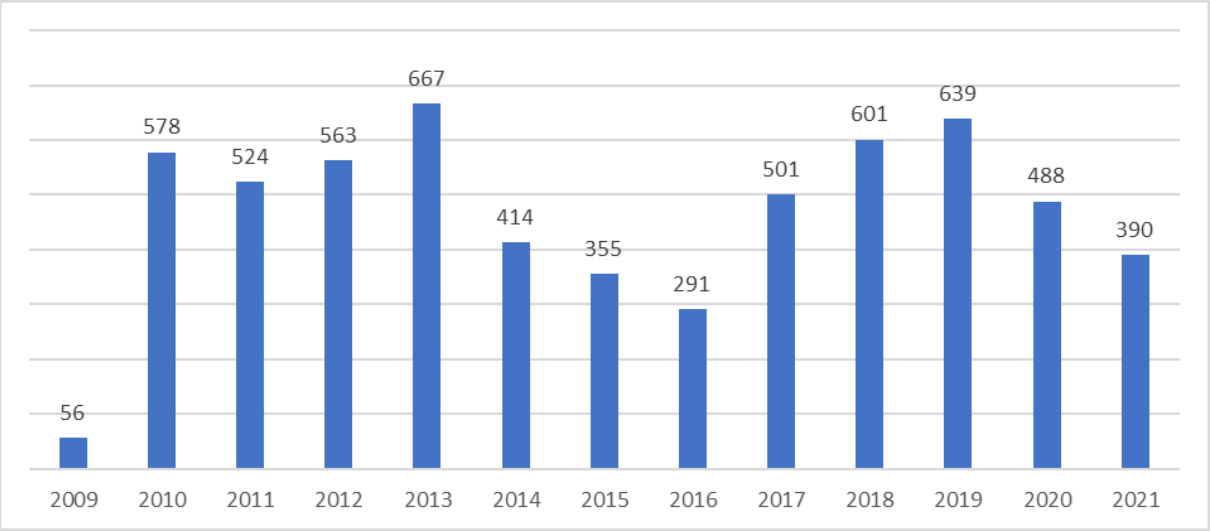
Gráfico 4 – Distribuição de fornecedores por região



Fonte: A autora, 2022.

Com relação ao ano de cadastro de fornecedores, é possível observar no Gráfico 5 que a cada ano uma quantidade relevante de novos fornecedores são catalogados na base de dados do Governo do Estado do Rio de Janeiro. Isso reforça a importância da pesquisa em identificar comportamentos associados à fraude, pelo constante nível de exposição ao risco.

Gráfico 5 – Distribuição de fornecedores por ano de cadastro

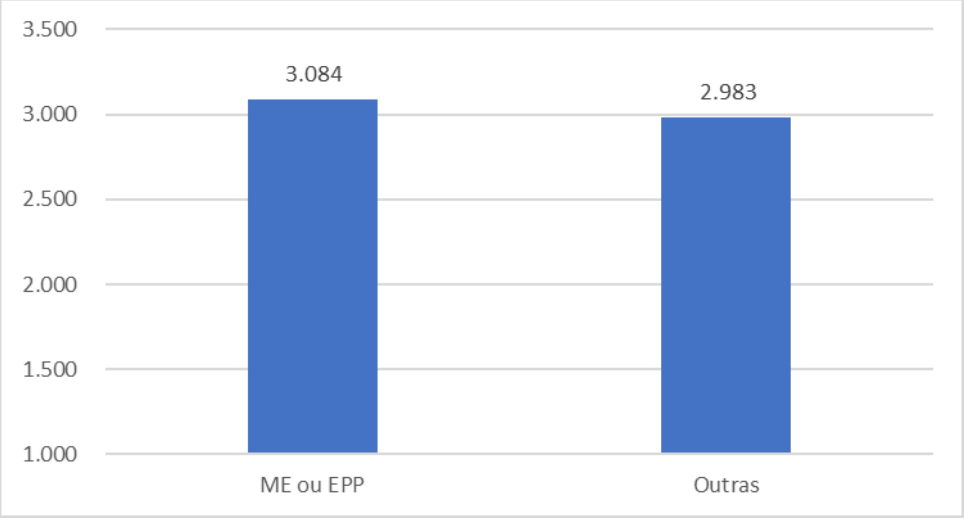


Fonte: A autora, 2022.

O porte das empresas observadas é balanceado entre microempresas (ME) e empresas de pequeno porte (EPP) e os demais portes (superiores) (Gráfico 6). Quando o serviço a ser prestado ou produto a ser vendido é de um valor relevante, este é um fator de destaque a ser observado para averiguar se a empresa será capaz de honrar com os seus compromissos.

O porte da empresa é um atributo importante de ser verificado, pois há processos de licitação exclusivos para participantes ME e EPP. Nesses casos, empresas de médio e grande porte também podem se aproveitar, criando novos CNPJs para participarem.

Gráfico 6 – Distribuição de fornecedores por porte da empresa

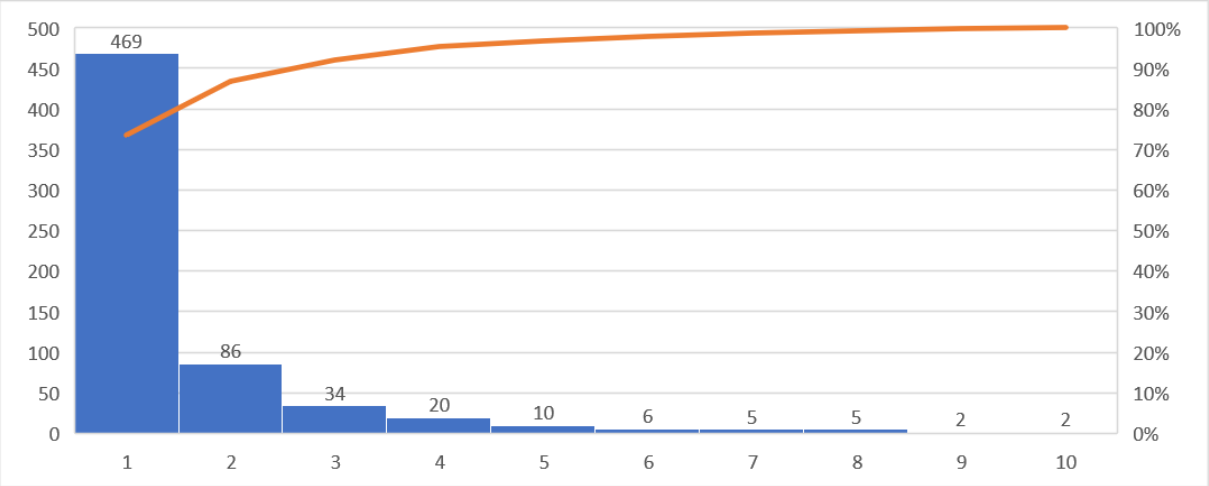


Fonte: A autora, 2022.

Os motivos pelos quais os fornecedores foram sancionados são, em sua maioria (73%), por inexecução total ou parcial do contrato (Gráfico 7). Ressalta-se que a base possui um total

de 240 fornecedores sancionados, de 639 sanções, o que indica que um fornecedor pode ter sido sancionado mais de uma vez.

Gráfico 7 – Percentual dos motivos das sanções



Fonte: A autora, 2022.

A descrição do motivo para cada sanção pode ser verificada na Tabela 4.

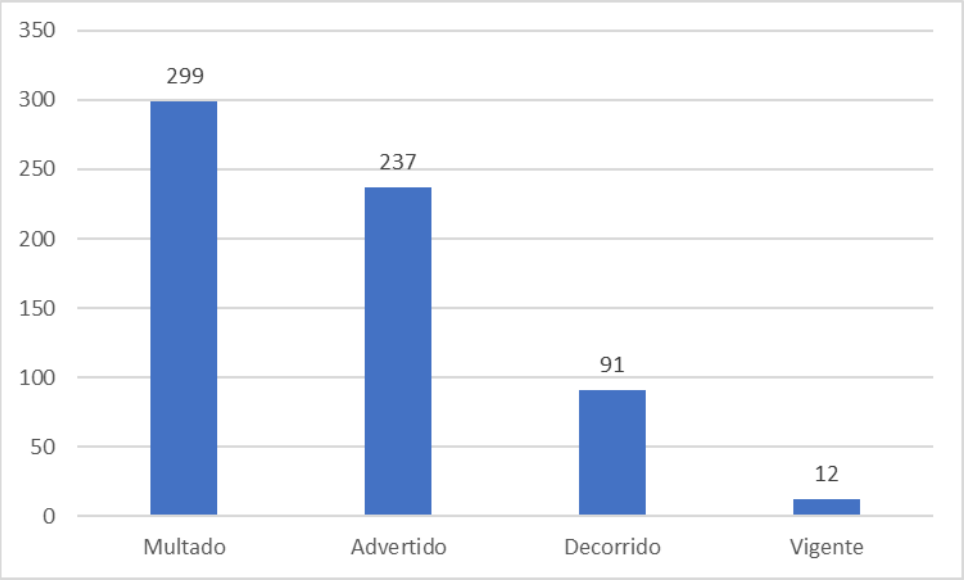
Tabela 4 – Motivos das sanções

Legenda	Motivo	Quantidade
1	Inexecução total ou parcial do contrato	469
2	Outros	86
3	Não apresentação de documentação exigida no certame ou apresentação de documentação falsa	34
4	Atraso injustificado na execução do contrato	20
5	Falha ou fraude na execução do contrato	10
6	Recusa em celebrar contrato	6
7	Prática de atos ilícitos visando a frustrar os objetivos da licitação ou contratação, tais como conluio, fraude, adulteração de documentos, documentação ou declaração falsa, entre outros	5
8	Inabilitação ou desclassificação por irregularidade ou inexequibilidade da proposta	5
9	Retardamento na execução do objeto ou não manutenção da proposta	2
10	Improbidade administrativa	2

Fonte: A autora, 2022.

Os tipos de sanções aplicadas aos fornecedores foram, na maioria dos casos, multa (47%), seguidos de advertência (37%). Os demais tipos encontrados indicam que a sanção já terminou o seu prazo (decorrido), ou que não possui um prazo determinado (vigente), conforme ilustra o Gráfico 8.

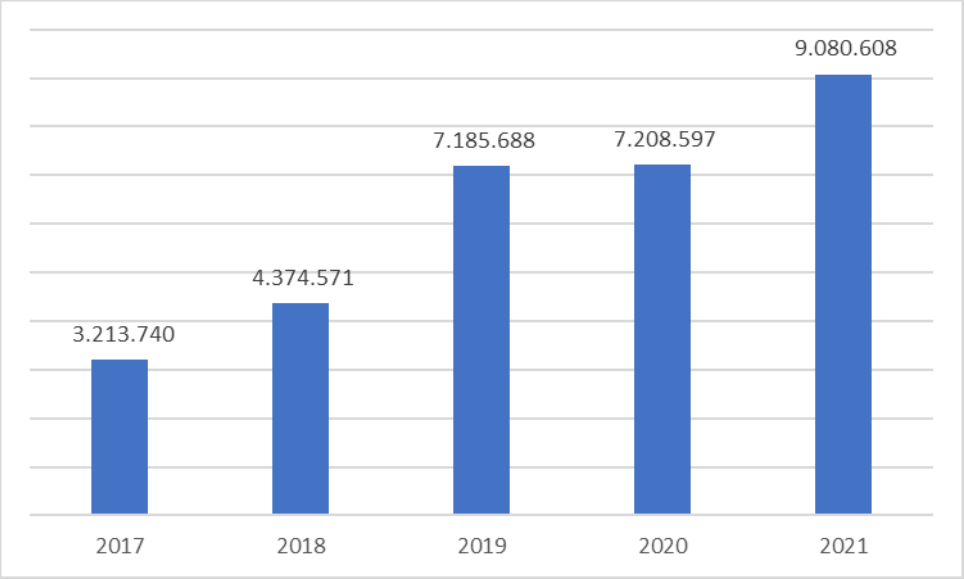
Gráfico 8 – Tipos de sanções aplicadas aos fornecedores



Fonte: A autora, 2022.

O valor dos contratos firmados no período vem crescendo a cada ano. Desde 2017, o montante das contratações quase triplicou (Gráfico 9).

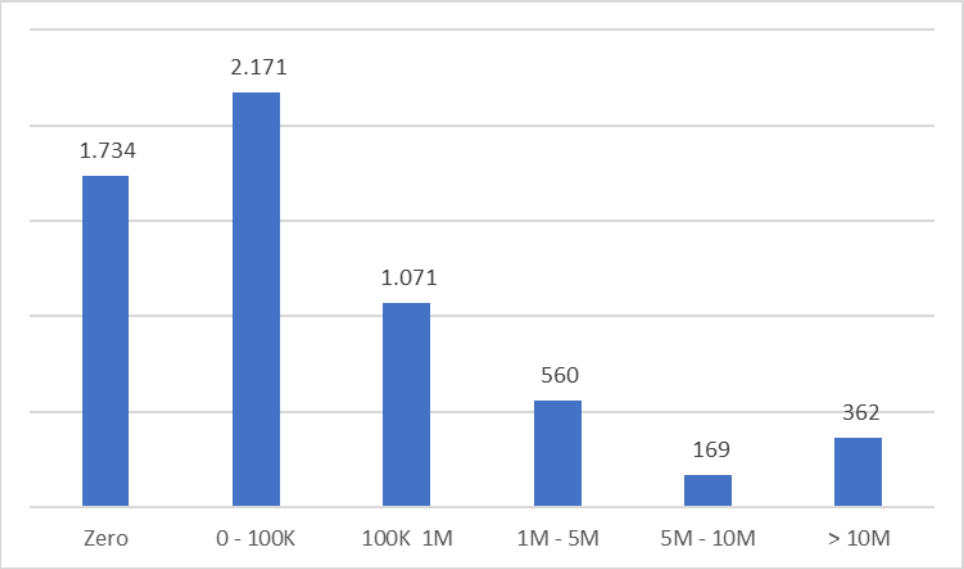
Gráfico 9 – Valor dos contratos por ano (em R$ mil)



Fonte: A autora, 2022.

Com relação ao valor dos contratos por fornecedor (Gráfico 10), verificou-se que a maioria foi de até R$ 100 mil. Nota-se ainda uma quantidade de fornecedores que foram cadastrados na base de dados, mas que não conseguiram vencer nenhuma licitação até o momento. Mesmo sem ter nenhum contrato firmado no período, 49 foram sancionados.

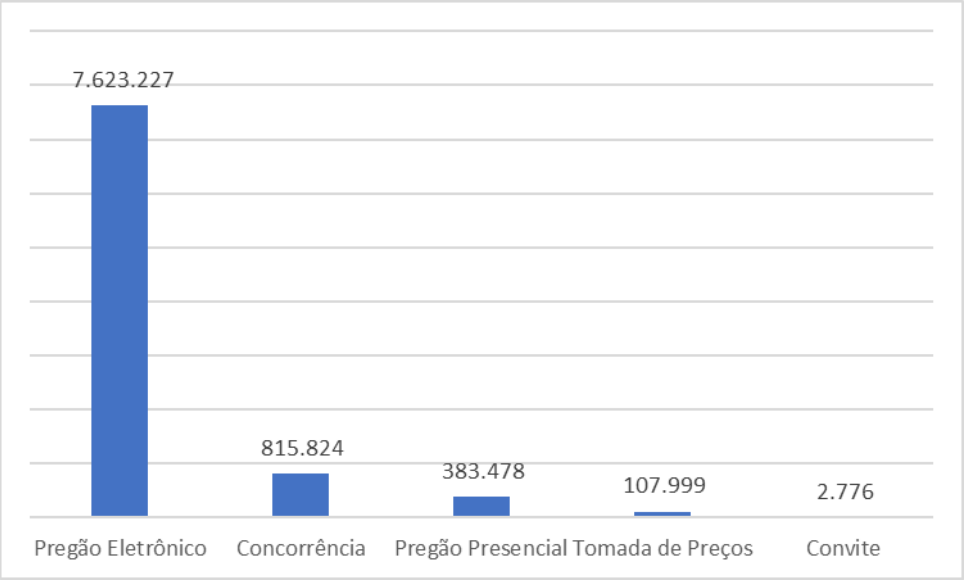
Gráfico 10 – Histograma de fornecedores por valor em contratos



Fonte: A autora, 2022.

Foi possível observar ainda que a modalidade de licitação mais frequente foi o pregão eletrônico, representando 85% do total (Gráfico 11).

Gráfico 11 – Valor das licitações por modalidade (em R$ mil)



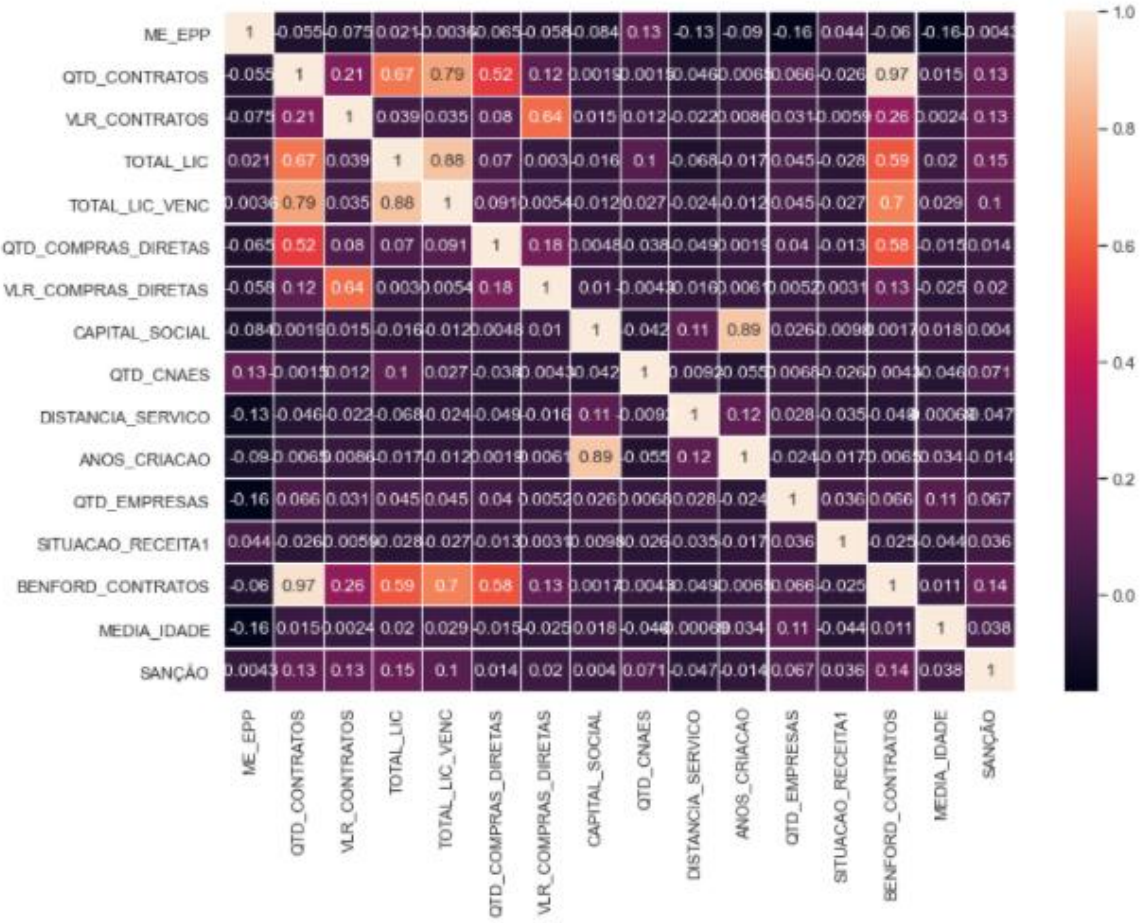
Fonte: A autora, 2022.

O coeficiente de correlação de Pearson é utilizado para verificar a relação entre duas variáveis. Quando negativa, indica que elas caminham em direções opostas, e quando positiva, caminham na mesma direção.

A matriz de correlação (Figura 10) foi executada para apurar quais variáveis têm maior correlação com a variável sanção. Foi possível verificar que as variáveis com maior correlação

foram: o total de licitações em que o fornecedor participou, a quantidade e o valor total de contratos e o total de contratos que falharam no teste da Lei de Benford.

Figura 10 – Matriz de correlação

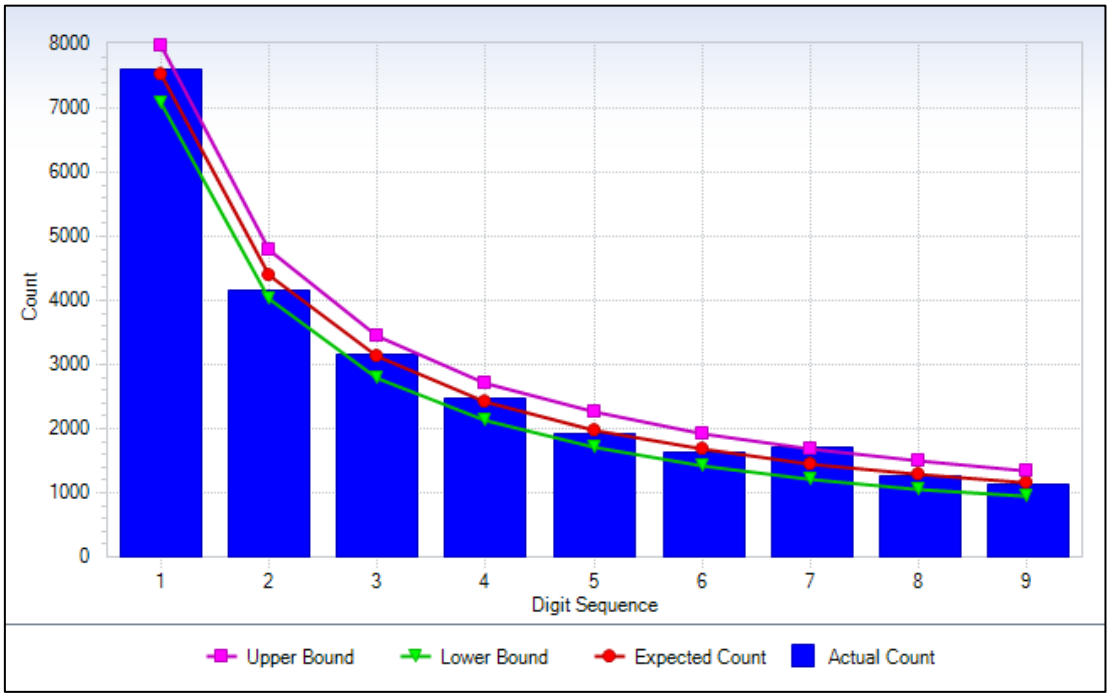


Fonte: A autora, 2022.

3.2 **Teste de hipótese – Lei de Benford**

A análise da Lei de Benford foi executada na base de contratos, através da ferramenta ACL, para verificar a hipótese de que os dados estariam adequados conforme essa teoria, ou se seriam suspeitos de alguma manipulação. O primeiro teste considerou o primeiro dígito do valor total do contrato. O Gráfico 12 representa a análise da Lei de Benford para o primeiro dígito.

Gráfico 12 – Lei de Benford do primeiro dígito

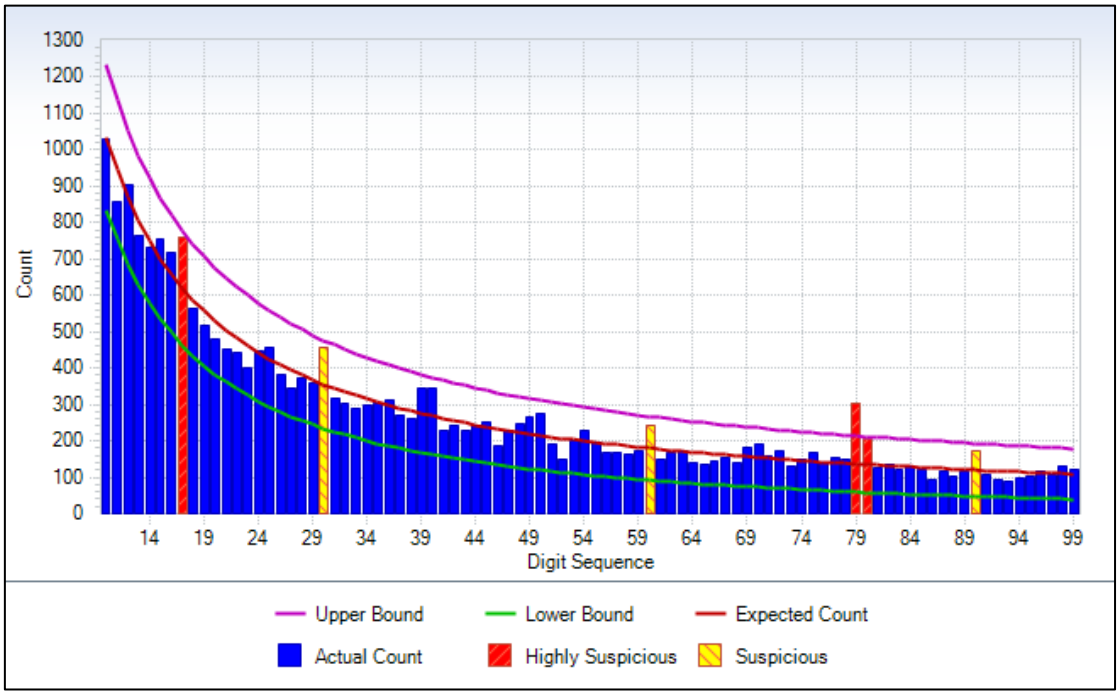


Fonte: A autora, 2022.

Ao executar o teste do qui-quadrado, foi obtido o valor de 67,94, que está acima do valor crítico de 15,51. Portanto, pode-se presumir que a base de contratos firmados pelo Governo do Estado do Rio de Janeiro não passou pelo teste de hipótese de aderência à Lei de Benford.

A segunda verificação foi realizada na mesma base, considerando os dois primeiros dígitos do contrato. Vários números apresentaram uma frequência maior do que a esperada. Entretanto, o que mais se destacou foi o 79. Para esse dígito, foi calculado o *z-score*, que representa a quantidade de vezes que a frequência da base está acima do desvio padrão observado. O valor obtido foi de 14,211. O Gráfico 13 ilustra o quanto a repetição do dígito 79 está acima do esperado.

Gráfico 13 – Lei de Benford dos dois primeiros dígitos



Fonte: A autora, 2022.

A quantidade esperada para os registros iniciados por 79 era de 137, mas foram constatados 303. Ao analisar os registros, foi verificado que 131 estavam em uma faixa entre R$ 7.900,00 e R$ 7.999,82, classificados como o tipo de aquisição de pequenas compras. Antes de 2018, o limite para realização da dispensa de licitação para contratos era R$ 8 mil. Esse fato pode explicar a quantidade de registros acima do esperado na faixa encontrada, pois os usuários podem estar fazendo as compras no limite para evitar um processo mais complexo de contratação.

Como resultado dessa análise, o auditor poderia se aprofundar nos 131 casos observados e verificar se existe algum comprador frequente, órgão ou fornecedor. As duplicidades também podem revelar se alguma despesa maior está sendo quebrada em diversos certames para não passar por todo o processo de seleção da licitação.

3.3 **Resultados**

Para avaliação do resultado dos algoritmos, foi verificada a precisão. Os resultados da pesquisa estão descritos na Tabela 5.

Tabela 5 – Precisão dos algoritmos

Algoritmo	Precisão	Erro na predição
KNN	95,83%	6
Random forest	95,83%	6
Rede neural	85,41%	21
Regressão logística	95,13%	7
SVM	99,30%	1

Fonte: A autora, 2022.

Considerando a métrica de precisão, o algoritmo que apresentou o melhor resultado foi o SVM, uma vez que exibiu apenas um erro na previsão. O pior modelo foi o de rede neural, que errou 21 previsões no total.

O segundo método para avaliação utilizado foi a matriz de confusão. Os resultados consolidados podem ser observados na Tabela 6.

Tabela 6 – Resultado da matriz de confusão

Algoritmo	TP	TN	FP	FN
KNN	76	62	6	0
Random forest	74	64	4	2
Rede neural	68	55	13	8
Regressão logística	76	61	7	0
SVM	76	67	1	0

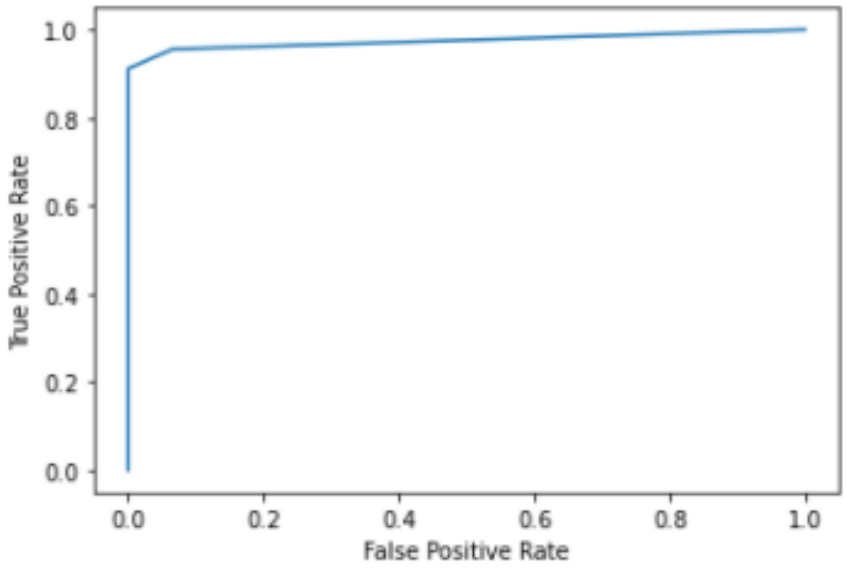
Fonte: A autora, 2022.

Considerando que uma área de auditoria deseja fazer investigações mais assertivas e não gastar tempo na investigação de fornecedores para saber quais são adequados, o melhor modelo a seguir seria o que apresenta o menor número de falsos positivos. Mais uma vez, o modelo SVM também pode proporcionar o melhor resultado, segundo a métrica da matriz de confusão, e a rede neural seria o algoritmo de pior resultado.

Verificando os resultados obtidos, o modelo SVM também pode proporcionar o melhor resultado, segundo a métrica da matriz de confusão, e a rede neural seria o algoritmo de pior resultado.

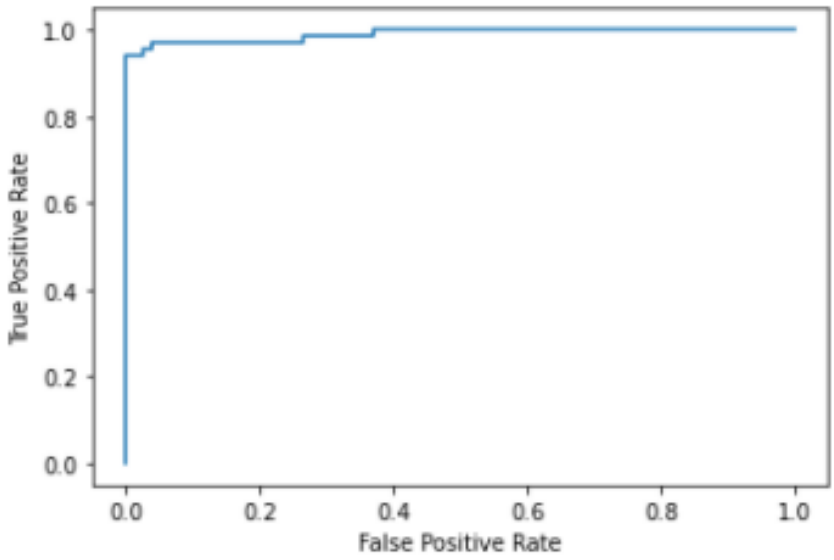
Por fim, apurou-se a curva ROC para cada modelo. O exame dos gráficos confirmou os resultados anteriores da precisão e da matriz de confusão (Gráficos 14 a 18).

Gráfico 14 – Curva ROC do modelo KNN



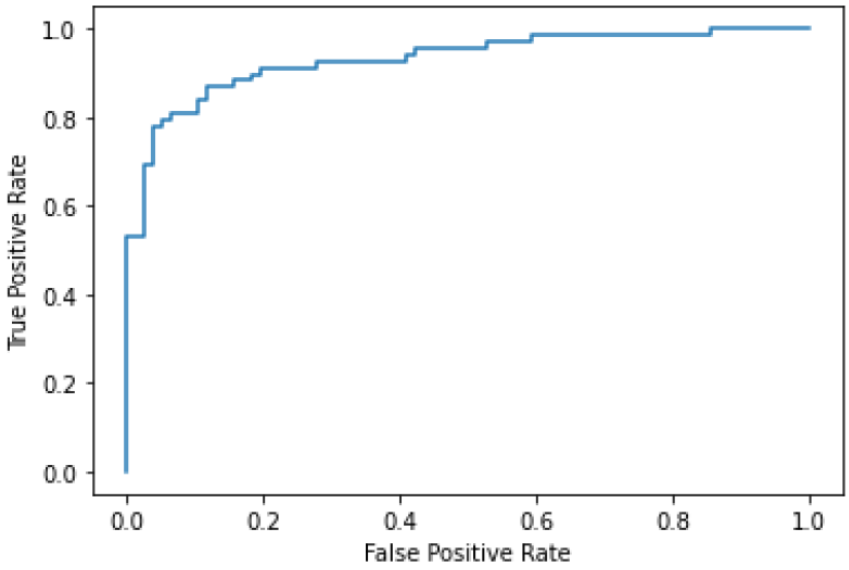
Fonte: A autora, 2022.

Gráfico 15 – Curva ROC do modelo *random forest*



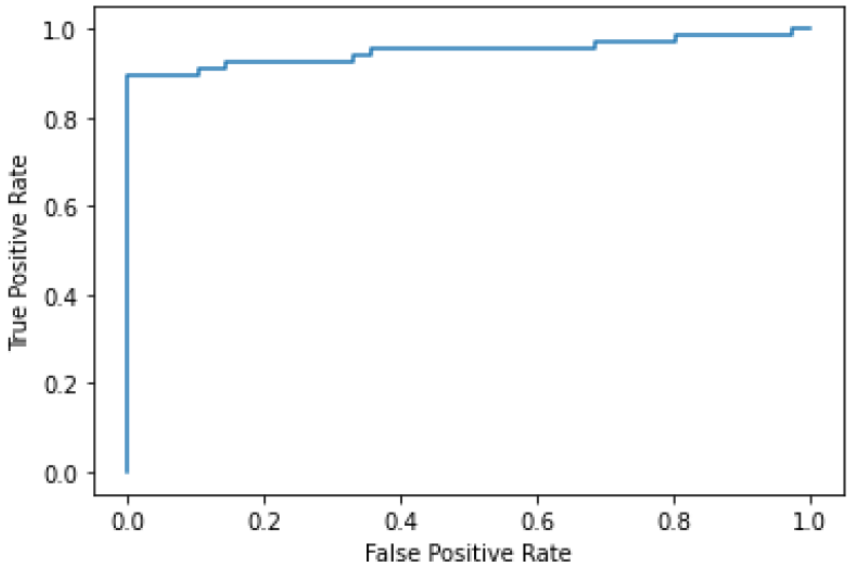
Fonte: A autora, 2022.

Gráfico 16 – Curva ROC do modelo rede neural



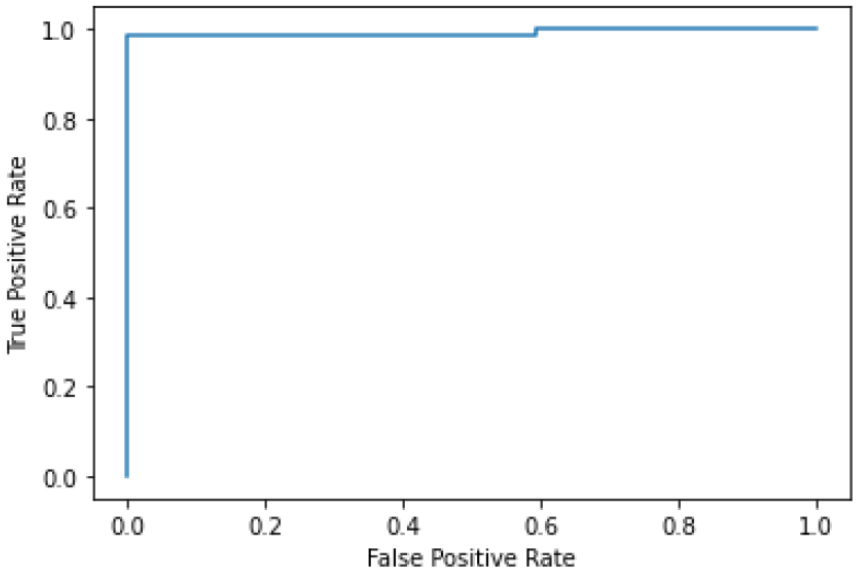
Fonte: A autora, 2022.

Gráfico 17 – Curva ROC do modelo de regressão logística



Fonte: A autora, 2022.

Gráfico 18 – Curva ROC do modelo SVM



Fonte: A autora, 2022.

Para a interpretação dos resultados, foram analisados os coeficientes da regressão logística. O comando executado no *Python* (coef_) informa o efeito da variável independente sobre a variação do logaritmo natural da chance da variável dependente ser igual a 1 (fornecedor sancionado), o que traz um resultado não intuitivo e não colabora para o entendimento da relação entre as variáveis.

Para transformar os resultados e facilitar o seu entendimento, foram obtidos os exponenciais dos coeficientes. Os resultados maiores que 1 indicam que, quanto maior o valor do atributo, maior a possibilidade de uma fraude ocorrer. Ao passo que os resultados menores que 1 indicam que quanto menor o valor, menor a probabilidade de uma fraude ocorrer. Os resultados mais próximos a 1 são os que possuem menor impacto no resultado.

A Tabela 7 mostra que quanto maior o número de licitações que o fornecedor participar, maior a probabilidade de ele ser sancionado. Por outro lado, quanto menor a média de idade dos sócios, a probabilidade de ele ser sancionado também é maior. Além disso, a variável que menos impacta no resultado é a quantidade de contratos que o fornecedor possui em que os números iniciados falharam no teste da Lei de Benford.

Tabela 7 – Coeficientes da regressão logística

Atributo	P-Valor	Coeficiente	Exp(Coeficiente)
TOTAL_LIC	0.00000000e+000	1.96250765e-02	1,01981891
SITUACAO_RECEITA	1.59973632e-001	1.44723512e-02	1,01457758
TOTAL_LIC_VENC	2.63905358e-201	1.41588459e-02	1,01425956
QTD_CNAES	1.14683516e-031	8.39435909e-03	1,00842969
VLR_CONTRATOS	0.00000000e+000	2.85894866e-03	1,00286304
VLR_COMPRAS_DIRETAS	0.00000000e+000	1.61989227e-03	1,00162121
QTD_CONTRATOS	0.00000000e+000	1.59662366e-06	1,00000160
CAPITAL_SOCIAL	0.00000000e+000	1.04976435e-06	1,00000105
QTD_COMPRAS_DIRETAS	9.64645466e-113	4.67301931e-07	1,00000047
BENFORD_CONTRATOS	1.93630257e-234	-2.82737286e-07	0,99999972
ME_EPP	1.21848565e-002	-1.57573079e-02	0,98436619
ANOS_CRIAÇÃO	4.72129702e-004	-2.19751507e-02	0,97826454
QTD_EMPRESAS	7.84282552e-003	-2.19798595e-02	0,97825994
DISTÂNCIA	2.12315989e-001	-2.67669292e-02	0,97358813
MÉDIA_IDADE	1.23799885e-006	-4.46694087e-02	0,95631358

Fonte: A autora, 2022.

Também é importante comentar as avaliações iniciais sobre os indicadores que foram tomadas como premissa. Empresas marcadas como baixadas, inaptas ou suspensas estão mais propensas a serem sancionadas, assim como as que possuem a maior quantidade de CNAES cadastrados, confirmando as premissas iniciais. A distância da sede do fornecedor para o Estado do Rio de Janeiro, entretanto, demonstrou que as empresas mais próximas do Rio de Janeiro são as mais propensas à sanção, contrariando a premissa de que fornecedores mais distantes seriam os com maior risco.

Por fim, foi executado o comando *features_importances*, do *random forest*, para verificar quais atributos têm uma importância maior para resolver o problema desta pesquisa. Esse comando identifica quais os atributos tiveram maior peso na definição da classificação final e só está disponível para o *random forest*.

Os atributos da Tabela 8 foram ordenados por maior importância. Pode-se observar que os mais relevantes no modelo são o valor e a quantidade total dos contratos firmados e o capital social registrado no cartão do fornecedor na Receita Federal.

Tabela 8 – *Features_importances* do *random forest*

Atributo	**Importância**
VLR_CONTRATOS	0,254642
QTD_CONTRATOS	0,236141
CAPITAL_SOCIAL	0,134538
VLR_COMPRAS_DIRETAS	0,067872
BENFORD_CONTRATOS	0,064277
QTD_COMPRAS_DIRETAS	0,040073
TOTAL_LIC	0,035075
TOTAL_LIC_VENC	0,033542
QTD_CNAES	0,030138
MÉDIA_IDADE	0,028579
ANOS_CRIAÇÃO	0,025063
SITUAÇÃO_RECEITA	0,020881
DISTÂNCIA_SERVIÇO	0,013249
ME_EPP	0,011066
QTD_EMPRESAS	0,004864

Fonte: A autora, 2022.

Nota-se que os resultados do *features importance* apresenta os indicadores de maior importância para a construção da árvore de decisão, mas não detalha sobre a relação entre valores maiores ou menores. Além disso, pode-se observar que os resultados são diferentes dos coeficientes da regressão logística. Portanto, seria importante avaliar os dois resultados em conjunto.

3.4 **Melhoria da *performance* dos algoritmos**

Para a rede neural, que apresentou o pior resultado dentre os algoritmos avaliados, uma opção para melhorar a sua *performance* seria utilizar a matriz de correlação, os coeficientes de regressão logística ou os resultados do *feature_importances*, e assim fazer uma seleção de atributos. Ao retirar aqueles com menor relevância, seria possível que o resultado melhorasse. Como os demais resultados obtidos na pesquisa foram satisfatórios, com uma precisão maior que 95%, essa avaliação não foi executada.

4 CONSIDERAÇÕES FINAIS

O teste de hipótese baseado na Lei de Benford foi realizado previamente na base de contratos firmados pelo Governo do Estado do Rio de Janeiro, e o resultado demonstrou que a base não estava em conformidade com essa teoria, estando propensa a ter tido manipulação dos números. Posteriormente, a pesquisa proposta apresentou resultados satisfatórios para a classificação de fornecedores sancionados pelo Governo. O algoritmo de classificação com melhor resultado nas três métricas de avaliação foi o SVM, com precisão de 99,30% e apenas um falso positivo na verificação da base teste. Os atributos levantados por meio das bases de dados também foram relevantes no resultado do modelo.

O resultado do trabalho pode contribuir não apenas para empresas públicas, mas também para as privadas, a partir de adaptações conforme os processos de licitação praticados por cada instituição.

A metodologia utilizada na pesquisa poderá ser replicada em auditorias para o levantamento de fornecedores suspeitos, além de auxiliar na análise preditiva para a identificação de fraudes. Como sugestão, pode-se realizar um comutê de máquinas para seleção de um fornecedor para análise. Ao invés de utilizar apenas o modelo com melhor desempenho, é possível se fazer uma espécie de votação para cada modelo, e o fornecedor selecionado será aquiele que receber o maior número de votos.

À medida que novos fornecedores sejam sancionados, o modelo deve ser retroalimentado para que o aprendizado seja adquirido e replicado em novos casos. Assim como, com a disponibilização de dados atualizados, deve-se revisitar os atributos a fim de verificar novas variáveis para esse modelo.

Para pesquisas futuras, sugere-se uma análise transacional das licitações, não consolidada por comportamento do fornecedor. Para isso, é preciso que as licitações sejam classificadas individualmente como fraudulentas ou não. Essa metodologia poderá contribuir para a avaliação do processo, uma vez que o auditor seria mais específico no momento de realizar uma investigação. Indica-se também o aprofundamento dos indicadores para a definição de novos atributos, e ainda a verificação de outros algoritmos de classificação.

Sugere-se ainda a aplicação do método LASSO (*Least Absolute Shrinkage and Selection Operator*) no modelo de regressão logística para a seleção de variáveis que melhor irão resolver o problema de identificação de fornecedores sancionados.

REFERÊNCIAS

ASSOCIATION OF CERTIFIED FRAUD EXAMINERS (ACFE). *Report to the Nations*: 2020 Global Study on Occupational Fraud and Abuse. Austin, USA: ACFE, 2020. Disponível em: <https://legacy.acfe.com/report-to-the-nations/2020/>. Acesso em: 22 mar. 2022.

ALVES, F. J. S. *Adesão do contabilista ao código de ética da sua profissão*: um estudo empírico sobre percepções. 2005. 273 f. Tese (Doutorado em Ciências Contábeis) – Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2005.

BENFORD, F. The law of anomalous numbers. *American Philosophical Society*, v. 78, n. 4, p. 551-572, 1938.

BONFIM, M. P.; SILVA, C. A. T. Mecanismos inibidores da desonestidade individual e em grupo. *Contabilidade Gestão e Governança*, v. 22, n. 2, p. 205-226, 2019.

BRASIL. Decreto nº 9.412, de 18 de junho de 2018. Atualiza os valores das modalidades de licitação de que trata o art. 23 da Lei nº 8.666, de 21 de junho de 1993. *Diário Oficial da União*, Brasília, DF, 19 jun. 2018. Disponível em: <https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/26370034/do1-2018-06-19-decreto-n-9-412-de-18-de-junho-de-2018-26369935>. Acesso em: 24 mar. 2022.

BRASIL. Lei nº 8.666, de 21 de junho de 1993. Regulamenta o art. 37, inciso XXI, da Constituição Federal, institui normas para licitações e contratos da Administração Pública e dá outras providências. *Diário Oficial da União*, Brasília, DF, 6 jul. 1993. Disponível em: <http://www.planalto.gov.br/ccivil_03/Leis/L8666compilado.htm>. Acesso em: 1 ago. 2021.

BRASIL. Lei nº 10.520, de 17 de julho de 2002. Institui, no âmbito da União, Estados, Distrito Federal e Municípios, nos termos do art. 37, inciso XXI, da Constituição Federal, modalidade de licitação denominada pregão, para aquisição de bens e serviços comuns, e dá outras providências. *Diário Oficial da União*, Brasília, DF, 18 jul. 2002. Disponível em: <http://www.planalto.gov.br/ccivil_03/Leis/2002/L10520.htm>. Acesso em: 1 ago. 2021.

BRASIL. Lei nº 14.133, de 1º de abril de 2021. Lei de Licitações e Contratos Administrativos. *Diário Oficial da União*, Brasília, DF, 1 abr. 2021. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14133.htm>. Acesso em: 21 mar. 2022.

BRASIL. Ministério do Planejamento, Orçamento e Gestão. *Caderno de Logística*: Sanções Administrativas em Licitações e Contratos. Brasília, DF: Ministério do Planejamento, Orçamento e Gestão, 2014. Disponível em: <https://www.gov.br/compras/pt-br/agente-publico/cadernos-de-logistica/midia/manual-sancoes-22-09.pdf>. Acesso em: 27 maio 2022.

BUMGARNER, N.; VASARHELYI, M. *Continuous auditing*: a new view, audit analytics and continuous audit: looking toward the future. New York: AICPA, 2015.

CARVALHO, A. P. L. F. *Redes neurais artificiais*. São Paulo: Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP), última atualização: 4 mar. 2009. Disponível em: <https://sites.icmc.usp.br/andre/research/neural/>. Acesso em: 21 mar. 2022.

CELLA, R. S.; ZANOLLA, E. *A Lei de Benford e a transparência*: uma análise das despesas públicas municipais. Goiânia: BBR, 2018.

CERQUEIRA, D.; LOBÃO, W. Determinantes da criminalidade: arcabouços teóricos e resultados empíricos. *Dados: Revista de Ciências Sociais*, Rio de Janeiro, v. 47, n. 2, 2004, p. 233-269.

CRESSEY, D. R. *Other people's money*: a study in the social psychology of embezzlement. Glencoe, IL: The free press, 1953.

CROSS-VALIDATION: evaluating estimator performance. *Scikit-learn developers (BSD License)*, c2022. Disponível em: < https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation>. Acesso em: 11 jul. 2022.

DOMO. *Data never sleeps 8.0.* c2022. Disponível em: <https://www.domo.com/learn/infographic/data-never-sleeps-8>. Acesso em: 23 ago. 2021.

FERNANDES, A. A. T. *et al*. Leia este artigo se você quiser aprender regressão logística. *Revista de Sociologia e Política*, Recife, v. 28, n. 74, e006, 2020.

GADI, M. F. A; LAGO, A. P.; MEHNEN, J. Data mining with skewed data. *In*: ZHANG, Y. (Ed.). *New advances in machine learning*. Norderstedt, Germany: BoD – Books on Demand, 2010. p. 173-187.

GUEDIM, T., Jr. Sanções administrativas e seus avanços na nova Lei de Licitações: aspectos comparativos das Leis de nº 8.666/1993, nº 10.520/2002 e nº 14.133/2021. *Zênite Fácil*, categoria Doutrina, 20 nov. 2021. Disponível em: <http://www.zenitefacil.com.br>. Acesso em: 21 mar. 2022.

HOARE, J. What is a Decision Tree? *Displayr Blog*, 1 Aug. 2018. Disponível em: <https://www.displayr.com/what-is-a-decision-tree/>. Acesso em: 21 mar. 2022.

HOSMER, D.; LEMESHOW, S. *Applied logistic regression*: Wiley series in probability and mathematical statistics. New York: John Wiley & Sons, 1989.

ITALO JOSÉ. KNN (K-Nearest Network) #1: como funciona? *AI Brasil*, 1 jul. 2018. Disponível em: <https://medium.com/brasil-ai/knn-k-nearest-neighbors-1-e140c82e9c4e>. Acesso em: 20 mar. 2022.

JENSEN, C.; MECKLING, H. Theory of the firm: managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, v. 3, n. 4, p. 305-360, 1976.

LOMBARDI, D.; BLOCH, R.; VASARHELYI, M. The future of audit. *JISTEM: Journal of Information Systems and Technology Management*, [*S. l.*], v. 11, n. 1, p. 21-32, 2014.

Disponível em: <https://www.revistas.usp.br/jistem/article/view/80174>. Acesso em: 20 mar. 2022.

MARR, B. 20 fatos sobre a internet que você (provavelmente) não sabe. *Forbes*, 1 out. 2015. Disponível em: <https://forbes.com.br/fotos/2015/10/20-fatos-sobre-a-internet-que-voce-provavelmente-nao-sabe/#foto1>. Acesso em: 26 out. 2019.

MÁQUINA de vetores de suporte (SVM) explicada. *ICHI.PRO*, c2022. Disponível em: <https://ichi.pro/pt/maquina-de-vetores-de-suporte-svm-explicada-97743104690915>. Acesso em: 21 mar. 2022.

MORAIS, C. M. M. *Proposição de indicadores para investigação de licitações por meio de técnicas de reconhecimento de padrões estatísticos e mineração de dados*. 2016. 126 f. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Tecnologia, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 2016.

MITCHELL, T M. *Machine learning*. New York: McGraw-Hill, 1997.

NIGRINI, M. J. *Benford's law*: applications for forensic accounting, auditing, and fraud detection. New Jersey: John Wiley & Sons, Inc., 2012.

OLIVEIRA, F. N.; SANTOS, L. P. G. Estratégias para combater a sonegação fiscal: um modelo para o ICMS baseado em redes neurais artificiais. *Revista de Gestão, Finanças e Contabilidade*, Salvador, v. 10, n. 1, p. 42-64, jan./abr. 2020.

OLIVEIRA, P. H. M. A. *Detecção de fraudes em cartões*: um classificador baseado em regras de associação e regressão logística. 2016. 117 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Pauo, 2016.

OLIVEIRA, T. E. A. *Redes neurais aplicadas à identificação de indivíduos através do som dos passos*. 2009. 86 f. Monografia (Bacharelado em Ciências da Computação) – Universidade Federal do Tocantins, Palmas, 2009. Disponível em: <https://www.researchgate.net/figure/Figura-23-Exemplo-de-neuronio-artificial-Adaptado-de-Haykin-2004_fig3_201549563>. Acesso em: 23 ago. 2021.

OLOKODANA, N.; FERNANDES, R. Previsión de fraude en informes contables a la luz del aprendizaje de máquinas. *Innovation and Technological Development*, v. 1, n. 1, 29 mar. 2020.

PF APURA fraude na contratação de manutenção predial do Hospital Universitário Gaffrée e Guinle. *G1*, Rio de Janeiro, 22 mar. 2022. Disponível em: <https://g1.globo.com/rj/rio-de-janeiro/noticia/2022/03/22/pf-deflagra-2a-fase-da-operacao-desmascarados-no-rio.ghtml>. Acesso em: 23 mar. 2022.

PORTO, J. CAATs – Técnicas de Auditoria Auxiliada por Computadores. *Audiit: Autitoria, Risco & Compliance*, 7 nov. 2011. Disponível em: <https://audiit.wordpress.com/2011/11/07/caats-tecnicas-de-auditoria-auxiliada-por-computadores/>. Acesso em: 26 out. 2019.

PREFEITA de Wanderley assume fraude no transporte escolar e deve cumprir TAC. *Mais Oeste*, 16 set. 2021. Disponível em: <http://www.maisoeste.com.br/2021/09/16/prefeita-de-wanderley-assume-fraude-no-transporte-escolar-e-deve-cumprir-tac/>. Acesso em: 23 mar. 2022.

ROMANI, B. Pesquisadores atingem feito histórico na computação quântica. *Estadão*, 23 out. 2019. Disponível em: <https://link.estadao.com.br/noticias/empresas,google-atinge-feito-historico-na-cmputacao-quantica,70003060160>. Acesso em: 26 out. 2019.

SANTOS, F. *Modelos supervisionados aplicados à detecção de fraude em seguros de saúde*. 2020. 74 f. Dissertação (Mestrado em Engenharia e Análise de Big Data) – Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisboa, 2020.

SANTOS, R. *Modelo preditivo de fraude ocupacional nas organizações privadas*. 2016. 205 f. Tese (Doutorado em Administração) – Programa de Estudos Pós-Graduados em Administração, Pontifícia Universidade Católica de São Paulo, São Paulo, 2016.

SEVERINO, M. K.; YAOHAO, P. Previsão de fraudes em seguros patrimoniais com algoritmos de aprendizado de máquina. ENCONTRO DA ANPAD, 43 – EnANPAD 2019. São Paulo, SP, 2019.

SCHRAMM, F. S. *O compliance como instrumento de combate à corrupção no âmbito das contratações públicas*. 2018. 412 f. Dissertação (Mestrado em Direito) – Programa de Pós-Graduação em Direito, Universidade Federal de Santa Catarina, Florianópolis, 2018.

TRANSPARÊNCIA BRASIL. *Métodos de detecção de fraude e corrupção em contratações públicas.* São Paulo: Transparência Brasil, 2019. Disponível em: <https://www.transparencia.org.br/downloads/publicacoes/Metodos%20Detec%C3%A7%C3%A3o%20de%20Fraude.pdf>. Acesso em: 29 ago. 2020.

WIRTH, R.; HIPP, J. Crisp-dm: towards a standard process model for data mining. INTERNATIONAL CONFERENCE ON THE PRACTICAL APPLIED OF KNOWLEDGE DISCOVERY AND DATA MINING, 4. *Proceedings*… Practical Application Company, 2000.

WOLFE, D.; HERMANSON, D. The fraud diamond: considering the four elements of fraud. *CPA Journal*, v. 74, n. 12, p. 38-42, 2004.