MÉTODOS DE CLASSIFICAÇÃO SUPERVISIONADA APLICADOS À IDENTIFICAÇÃO DE FRAUDES DE FORNECEDORES

SUPERVISED CLASSIFICATION METHODS APPLIED TO THE IDENTIFICATION OF SUPPLIER FRAUD

Tainá Ayres Sá
Mestre em Ciências Contábeis pela Universidade Estadual do Rio de Janeiro, 2022
Universidade Estadual do Rio de Janeiro
tainayres@hotmail.com

Prof. Dr. José Francisco Moreira Pessanha
Pontifícia Universidade Católica – PUC/RIO, 2006
Universidade Estadual do Rio de Janeiro
professorjfmp@hotmail.com

Prof. Dr. Francisco José dos Santos Alves
Universidade de São Paulo – USP, 2005
Universidade Estadual do Rio de Janeiro
fjalves@globo.com

Resumo

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 e maior assertividade na identificação de potenciais fraudadores entre os fornecedores da administração pública nos processos de contratação do Governo do Estado do Rio de Janeiro. A pesquisa tem relevância uma vez que pode ser utilizada pelos órgãos de controle do governo e empresas privadas para identificar esses atos no processo de contratação. Como parte da metodologia, a Lei de Newcomb-Benford foi aplicada inicialmente aos dados de contratos publicados no Portal da Transparência do Governo do Estado do Rio de Janeiro e demonstrou que a base não estava em conformidade com a lei, podendo estar propensa a ter sofrido manipulação dos números. Na sequência foram avaliados diferentes métodos de classificação supervisionada, nomeadamente, vizinho mais próximo (KNN), regressão logística, máquina de vetores de suporte (SVM) e árvore de classificação com o objetivo de identificar fraudadores. Os resultados obtidos foram satisfatórios para a classificação de fornecedores sancionados. O algoritmo de classificação com melhor resultado nas duas métricas de avaliação foi o SVM, com precisão de 99,30%, e 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 supervisionada.

Abstract

Fraud in bids occurs by trying to thwart their competitive nature, causing overbilling, and favoring a specific group. This research aims to verify which set of indicators and algorithms has better properties and greater assertiveness in identifying potential fraudsters among public

Resumo do processo editorial:
a) Submissão em: 01/06/2022.
b) Envio para avaliação em: 15/06/2022.
c) Término da avaliação em : 16/06/2022.
d) Correções solicitadas em: 16/06/2022.
e) Recebimento da versão ajustada em: 22/06/2022.
f) Aprovação final em: 26/06/2022.

administration suppliers in the hiring processes of the State Government of Rio de Janeiro. The research is relevant since it can be used by government control bodies and private companies to identify these acts in the hiring process. The results obtained were satisfactory for the classification of sanctioned suppliers. As part of the methodology, the Newcomb-Benford Law was applied initially to the contract data published on the Transparency Portal of the Government of the State of Rio de Janeiro and demonstrated that the database was not in compliance with the law, and could be prone to having suffered manipulation of the numbers. Next, different supervised classification methods were evaluated, namely, nearest neighbor (KNN), logistic regression, support vector machine (SVM) and classification tree in order to identify fraudsters. The classification algorithm with the best result in the two evaluation metrics was the SVM, with an accuracy of 99.30%, and a false positive in the base-test verification. The attributes surveyed through the databases were also relevant in the model's result.

Keywords: bidding fraud; indicators; supervised classification algorithms.

1 Introdução

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 métodos já foram propostos pela literatura, os quais poderiam sinalizar que uma fraude possa ter ocorrido. Dentre esses, uma opção muito utilizada é a Lei de Newcomb-Benford, de 1938, que analisa a regularidade dos primeiros dígitos de uma base de números naturais com a frequência esperada (BENFORD, 1938).

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 fraudulenta. A partir daí, entram outras técnicas estatísticas e 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 para sinistros patrimoniais com fraudes mediante a aplicação de modelos da *Support Vector Machine* (SVM), entre outros algoritmos.

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

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.

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.

Para além desta breve introdução sobre o problema de pesquisa, os objetivos e a justificativa do estudo, o presente trabalho encontra-se organizado em quatro seções. A seguir, na primeira seção, tem-se o referencial teórico, no qual se descreve o processo licitatório conforme a legislação brasileira, a evolução do processo de auditoria e dos diferentes algoritmos que serão utilizados na metodologia. As etapas da metodologia proposta encontram-se detalhadas n segunda seção . 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.

2 Referencial Teórico

Nesta seção, apresenta-se 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. Ademais, apresentam-se os principais conceitos sobre o aprendizado de máquina, os algoritmos de classificação (supervisionada) e as métricas de avaliação que serão utilizados na metodologia para o desenvolvimento do trabalho.

2.1 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 pequenas, 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. As exceções podem ser anomalias, pagamentos duplicados, números de cheque faltando em uma sequência numérica ou divergências no recálculo de uma folha de pagamento comparadas à legislação trabalhista.

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.

2.2 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 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 (BRASIL, 1993). 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.

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.

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. Com a tentativa de coibir as atividades fraudulentas, as leis de 1993 e 2002 estabelecem sanções administrativas que devem ser aplicadas quando alguma infração for identificada.

2.3 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).

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.

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).

A análise de casos divulgados na imprensa também pode trazer indicadores para serem considerados no modelo, como empresas criadas pouco tempo antes do início da licitação ou com uma quantidade grande de CNAEs (Classificação Nacional de Atividade Econômica) listados no cartão de CNPJ, indicando que a empresa não é especializada em um serviço.

2.4 Lei de Benford

A Lei de 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".

A Lei de Newcomb-Benford pode ser utilizada na investigação de fraudes. Quando uma base de dados não está aderente a Lei, ou seja, se a frequência dos registros estiver diferente do esperado, os seus desvios devem ser analisados. Nigrini (2012) sugere que a Lei de Newcomb-Benford deve ser empregada como ponto de partida e o teste recomendado deve abranger os 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.

2.5 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, semi-supervisionada e por reforço.

Nesta pesquisa será utilizada uma abordagem de aprendizado supervisionado mediante a utilização de algoritmos de classificação. 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 categoria encontram-se os algoritmos de classificação que irão rotular os dados da base. 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.

2.6 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). A variável dependente corresponde à sanção sofrida pelo fornecedor, sendo "1" quando positivo e "0" quando negativo.

2.6.1 Regressão logística

De forma distinta da regressão linear, na qual a variável resposta (Y) é contínua, na regressão logística a variável resposta é binária (YÎ{0, 1}) e denota as categorias de uma variável qualitativa, por exemplo, fornecedor sancionado (Y=1) e fornecedor não sancionado (Y=0). Ademais, a regressão logística usa a função logística para modelar a probabilidade de a variável resposta assumir determinada categoria em função das variáveis independentes (X), i.e., P(Y=1|X), conforme indicado pelo modelo não a seguir:

$$P(Y = 1|X) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X)]} \quad (1)$$

A partir de um conjunto de treinamento rotulado, i.e., uma amostra contendo fornecedores sancionados e não sancionados, os coeficientes de regressão (b_0, b_1) em (1) são estimados pelo método da máxima verossimilhança.

No caso em tela, a equação (1) fornece uma estimativa da probabilidade do fornecedor avaliado ser sancionado. Assim, para um dado limite de corte λ, em geral fixado 0,5, a regressão logística classifica uma empresa como sancionada (Y=1) se P(Y=1|X)>0,5, caso contrário, a empresa é classificada como não sancionada Y=0.

2.6.2 KNN

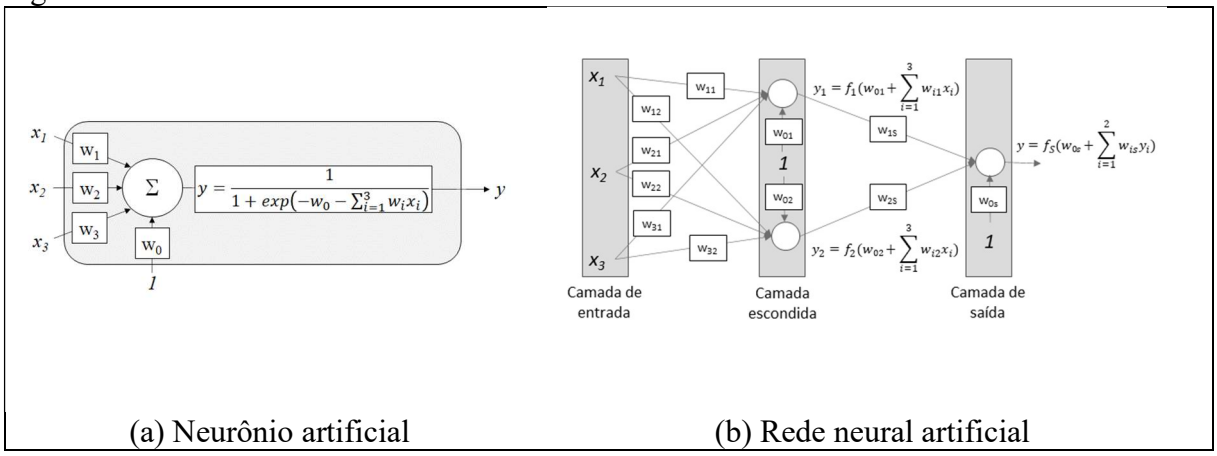
Dado um conjunto de treinamento rotulado, neste caso um conjunto formado por fornecedores identificados como sancionados e não sancionados, o algoritmo KNN (*K-Nearest Neighbor*) classifica novos fornecedores a partir da associação com os K vizinhos mais próximos (mais semelhantes) no conjunto de treinamento. A distância entre os casos é geralmente medida usando a distância Euclidiana. A classificação consiste em atribuir ao novo fornecedor o rótulo da classe mais frequente entre os K vizinhos.

A capacidade de generalização do KNN é afetada pelo valor de K, um hiperparâmetro que pode ser determinado por validação cruzada. Apesar de simples e rápido, o KNN não é robusto, pois pode ser afetado por ruído ou outliers nos dados.

2.6.3 Redes neurais artificiais

As redes neurais artificiais são inspiradas no sistema biológico de aprendizado, que é constituído de redes complexas de neurônios interconectados. Cada neurônio recebe *inputs* e produz *outputs*, que, por sua vez, servirão de *input* para o próximo neurônio (MITCHELL, 1997). Sem perda de generalidade, a estrutura e o funcionamento de um neurônio biológico podem ser modelados pelo neurônio artificial ilustrado na Figura 1(a), cuja resposta é uma função (função de ativação) da soma ponderada de três variáveis de entrada pelos respectivos pesos sinápticos (*w*). A organização de vários neurônios artificiais em uma estrutura forma uma rede neural artificial (RNA), cuja arquitetura mais usual é a rede *perceptron* com três camadas, conforme mostra a Figura 1(b).

Figura 1 – Neurônio artificial



(a) Neurônio artificial (b) Rede neural artificial

Fonte: autoria própria.

Note que as entradas do neurônio da camada de saída correspondem às saídas dos neurônios da camada escondida, determinados pelas variáveis de entrada. Assim, o diagrama na Figura 1(b) é a representação da seguinte equação de regressão não linear:

$$y = \frac{1}{1 + exp\left(-w_{0s} - \frac{w_{1s}}{1 + \exp\left(-w_{01} - \sum_{i=1}^{3} w_{i1}x_i\right)} - \frac{w_{2s}}{1 + \exp\left(-w_{02} - \sum_{i=1}^{3} w_{i2}x_i\right)}\right)} \quad (2)$$

A construção de uma RNA envolve a definição do número de camadas escondidas, a definição do número de neurônios em cada camada e a escolha da função de ativação dos neurônios. A identificação da configuração adequada exige a avaliação de diferentes configurações da rede. Entretanto, deve-se sempre empregar o princípio da parcimônia, e saber que uma rede com apenas uma camada escondida é capaz de aproximar qualquer tipo de função contínua (HAYKIN, 2009). Ademais, em problemas de classificação é comum que a função de ativação dos neurônios seja a função logística para que a saída da rede seja um número no intervalo [0,1] e possa ser interpretado como $P(y=1|x_1,x_2,x_3)$ (BAESENS; VAN VLASSELAER; VERBEKE, 2015).

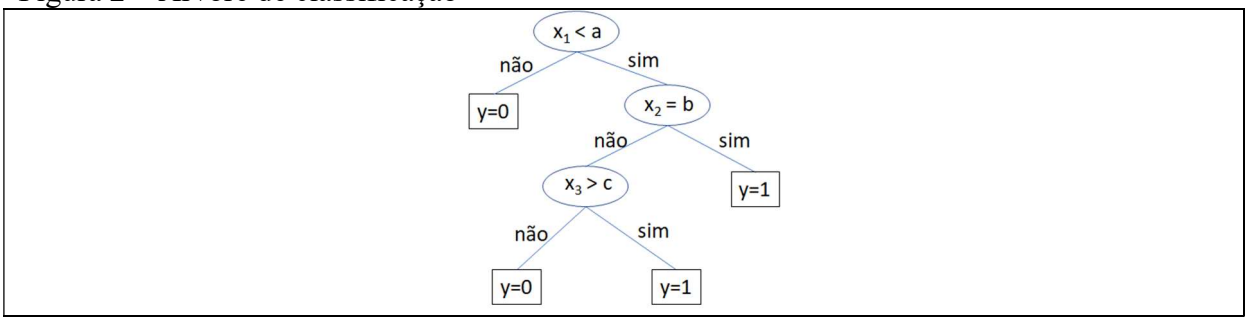
O ajuste dos pesos sinápticos é realizado por meio de um processo iterativo denominado aprendizagem. A classificação requer um aprendizado supervisionado, i.e., deve-se apresentar os padrões de entrada (x) e a respectivas saídas (rótulos binários em y) à RNA. O método de treinamento mais usado no aprendizado supervisionado é a retropropagação do erro (WERBOS, 1990), cujo objetivo consiste em ajustar os pesos sinápticos (w), de tal forma a minimizar a soma dos quadrados dos desvios (erros) entre a saída da RNA e os valores binários da variável resposta.

2.6.4 *Random forest*

Uma árvore de classificação representa um processo de decisão multi-estágio no qual em cada estágio o conjunto de dados rotulados é dividido em dois subconjuntos com base nos valores de uma das variáveis analisadas. A árvore é formada por nós e ramos, com nós designados por raiz, internos e terminais (TAN; STEINBACH; KUMAR, 2009). O nó raiz marca o início da árvore e divide o conjunto de dados em dois subconjuntos que são recursivamente divididos em subconjuntos menores pelos nós internos até que sejam alcançados os nós terminais, onde os subconjuntos são formados majoritariamente por observações da mesma classe, i.e., são mais puros.

Conforme ilustrado na Figura 2, a árvore de classificação faz previsões com base em perguntas (nós raiz e internos) para determinar qual ramo seguir e assim chegar a uma conclusão (nó terminal). A partir do nó raiz, a árvore é construída pela adição sucessiva dos nós internos, sendo que para cada nó deve ser escolhida uma variável e a respectiva condição para particionar o conjunto de dados. Outra decisão importante refere-se ao momento de interromper a adição de nós à árvore. Portanto, a construção da árvore de classificação requer a escolha da sequência de variáveis nos nós (x_1, x_2, x_3,...) e das respectivas condições (<a, =b, >c,...) usadas para particionar o conjunto de dados em cada estágio, uma tarefa de grande complexidade, dada a grande quantidade de possibilidades para construção de uma árvore. Felizmente, há alguns algoritmos para construção de árvores de classificação, entre os quais destacam-se o C4.5, CART e CHAID (BAESENS; VAN VLASSELAER; VERBEKE, 2015).

Figura 2 – Árvore de classificação



Fonte: autoria própria.

Traduzido para o português como floresta aleatória, o algoritmo *random forest* soluciona problemas de classificação criando diversas árvores de decisão de forma aleatória. No caso da *random forest*, o algoritmo utilizará uma amostra de registros da base de dados rotulados 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 (*ensemble*). No final, o resultado mais frequente será a classificação atribuída pelo *ensemble* de modelos.

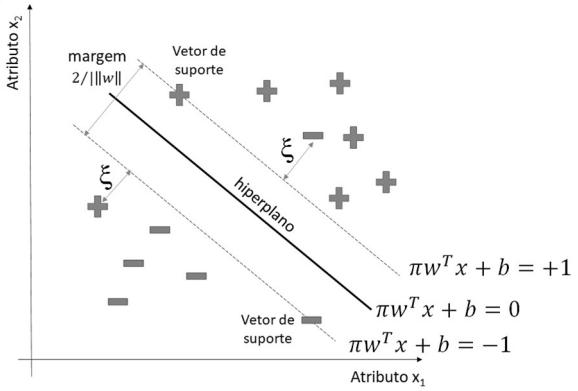
2.6.5 Máquina de vetor suporte (SVM)

O algoritmo máquina de vetor suporte (*support vector machine* – SVM) ajusta um hiperplano que separa as observações de duas classes (YÎ{-1,1}) presentes em um conjunto de dados rotulados. O hiperplano discriminante é ajustado pelo problema de otimização em (3) de forma que a margem de separação entre as duas classes seja máxima conforme ilustrado na Figura 3, sendo que as observações nas linhas tracejadas, na fronteira da margem, são os vetores de suporte, elas são as observações mais difíceis de serem classificadas e determinam o ajuste do hiperplano discriminante.

Sem perda de generalidade, em (2) N é o número de fornecedores, y é a variável resposta que indica se o fornecedor foi sancionado (Y=1) ou não sancionado (Y=-1), enquanto x denota as respectivas variáveis explicativas. A função *f*(x) mapeia o espaço das variáveis explicativas em um espaço de maior dimensão (*feature space*), onde as classes tornam-se linearmente separáveis. Adicionalmente, w e b denotam os parâmetros do hiperplano discriminante e x representa os erros de classificação ilustrados na Figura 3. O ajuste do hiperplano discriminante é determinado pela solução do problema em (2), controlada pelo hiperparâmetro C>0 com a finalidade de buscar uma solução de equilíbrio entre a largura da margem e a soma dos erros de classificação. A largura da margem e a constante C variam em sentidos opostos, um valor

grande para C implica na redução da margem de classificação, enquanto um valor pequeno implica em uma margem larga e com muitos erros de classificação em função dos ruídos nos dados.

Figura 3 – Máquina de vetor suporte



1. Hiperplano discriminante

$$\begin{aligned} \underset{w,b}{Min} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ s.a. \quad & y_i(wf(x_i)+b) + \xi_i \geq 1 \quad \forall i = 1, \\ & \xi_i \geq 0 \quad \forall i = 1, \end{aligned} \tag{3}$$

2. Problema de otimização

Fonte: autoria própria.

Para evitar a especificação explícita da função $f(x)$ deve-se trabalhar com o dual do problema em (3), na qual a função é definida implicitamente por uma função *kernel* $K(x_i,x_j)= f^T(x_i)f(x_j)$, por exemplo, a mais usada é função de base radial ($K(x_i,x_j)=\exp(-||x_i-x_j||^2/s^2)$ com hiperparâmetro adicional s^2 (BAESENS et al. 2015), determinado por validação cruzada juntamente com o hiperparâmetro C. Assim, a classificação de uma observação (x) é dada por:

$$classe\ da\ observação = sinal\left(\sum_{i=1}^{N}\lambda_i\, y_i K(x,x_i) + b\right) \tag{4}$$

Em que λ_i " i=1,N denotam os multiplicadores de Lagrange resultantes do dual do problema em (2), cujos valores são diferentes de zero apenas nos vetores de suporte. Portanto, apenas as observações classificadas como vetores de suporte contribuem para a soma em (4).

2.7 Métricas de avaliação

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

2.7.1 Precisão

Para avaliação dos resultados 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.

2.7.2 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 a Figura 4.

Figura 4 – Matriz de confusão

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

Fonte: autoria própria.

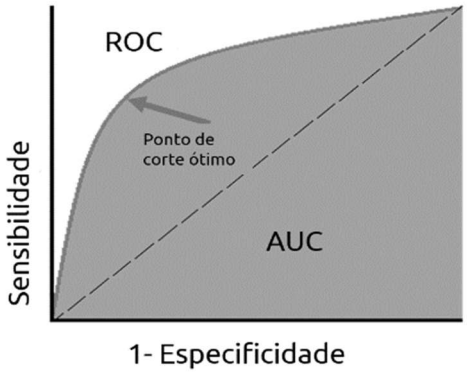
Na prática de uma auditoria, existem dois riscos: os falsos negativos podem deixar passar despercebido um fornecedor que esteja cometendo algum tipo de irregularidade, e os falsos positivos fazem com que o auditor perca tempo investigando um fornecedor que esteja fazendo tudo adequado. Como o tempo do auditor é um recurso escasso, reduzir o número de falsos positivos inicialmente pode ser uma opção.

Outras métricas possíveis para serem utilizadas na avaliação dos modelos são a sensibilidade e a especificidade. A sensibilidade avalia a capacidade do método em detectar resultados positivos, e pode ser calculado através da fórmula: VP / (VP+FN). Já a especificidade avalia a capacidade do método em detectar resultados negativos, sendo calculado pela fórmula: VN / (FP+VN).

2.7.3 Curva ROC

A curva *Receiver Operating Characteristic* (ROC) ilustrada na Figura 5 é uma representação gráfica do desempenho do classificador, medido pela sensibilidade e especificidade, para diferentes limiares de discriminação (*cutoff*). A forma da curva ROC decorre do *tradeoff* entre a sensibilidade e a especificidade. O classificador perfeito (sensibilidade e especificidade iguais a 1 ou 100%) corresponde ao vértice superior esquerdo do diagrama, enquanto a linha tracejada no sentido diagonal representa um classificador aleatório. Assim, um classificador é considerado efetivo se a sua curva ROC estiver acima da linha tracejada e o mais próximo do vértice superior esquerdo, i.e., a área sob a curva ROC ou *Area Under Curve* (AUC) deve ser maior que 0,5 e idealmente próximo de 1.

Figura 5 – Curva ROC



Fonte: Scudilio (2020).

3 Metodologia

Nesta seção será descrita a metodologia 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 dos algoritmos avaliados.

3.1 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. 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.
	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: autoria própria.

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 1).

Tabela 1 – 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: autoria própria.

3.2 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. 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, 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.

O segundo passo consistiu na combinação entre as bases de dados, utilizando como tabela principal as informações de fornecedores.

Realizou-se um filtro na base de fornecedores e foram considerados somente aqueles que possuíam CNPJ, atuam no Brasil e que tiveram pelo menos um contrato firmado ou participaram de alguma licitação no período de estudo. Desse modo, foram criados os campos que serão utilizados como atributos (indicadores) no modelo de classificação. O Quadro 2 apresenta o *layout* da base, com o total de 15 atributos e a variável dependente (SANÇÃO).

Quadro 2 – Atributos considerados no modelo final

Ordem	Nome do campo	Descrição
1	NR_CNPJ	Número do CNPJ 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.
3	QTD_CONTRATOS	Quantidade de contratos registrados para o fornecedor no período.
4	VLR_CONTRATOS	Valor dos contratos registrados para o fornecedor no período.
5	TOTAL_LIC	Nº licitações em que o fornecedor participou no período.
6	TOTAL_LIC_VENC	Total de licitações que o fornecedor foi o vencedor no período.
7	QTD_COMPRAS_DIRETAS	Quantidade de compras diretas registradas para o fornecedor no período.

8	VLR_COMPRAS_DIRETAS	Valor de compras diretas registradas para o fornecedor no período.
9	CAPITAL_SOCIAL	Capital social registrado para o CNPJ na Receita Federal.
10	QTD_CNAES	Total de CNAEs registrados para o CNPJ na Receita Federal.
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
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.
13	QTD_EMPRESAS	Quantidade de empresas na base de fornecedores em que o fornecedor possui sócios em comum.
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".
15	BENFORD_CONTRATOS	Quantidade de contratos firmados com o fornecedor cujos números iniciais falharam no teste de Benford (quantidade acima do limite superior).
16	MÉDIA_IDADE	Média das idades dos sócios da empresa.
17	SANÇÃO	Evento de interesse da pesquisa, em que "1" representa fornecedores sancionados e "0" não sancionados.

Fonte: autoria própria.

3.3 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 de classificação foram executados os seguintes passos para o tratamento de dados: 1) verificação de dados ausentes, 2) balanceamento da base de dados, 3) divisão da base entre treino e teste e 4) normalização da base de dados.

3.3.1 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. Realizou-se, então, uma avaliação da base de dados para verificar a existência de dados ausentes, contudo, nenhuma linha com valores ausentes foi identificada.

3.3.2 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. 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.

No caso desta pesquisa, como existem duas classes, cada uma deveria representar 50% da base de dados.

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*().

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.

3.3.3 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 demais 30% dos registros.

3.3.4 Normalização da base de dados

No caso das redes neurais artificiais e SVM os dados foram previamente normalizados para o intervalo [0,1] por meio do comando *MinMaxScaler*.

3.3.5 Aplicação dos algoritmos

Conforme ilustrado no Quadro 3, os algoritmos avaliados possuem hiperparâmetros, cujos valores devem ser previamente definidos antes do ajuste dos modelos. Neste trabalho os hiperparâmetros foram definidos com o apoio do módulo GridSearchCV que pesquisa uma grade com possíveis valores para os hiperparâmetros e realiza a validação cruzada [cabe uma referência] para minimizar a chance de *overfitting* do modelo, um problema que ocorre quando o modelo tem um desempenho muito bom para classificar os casos na base de treino, mas não consegue classificar corretamente os exemplos da base de teste. Seria como se o algoritmo decorasse os dados e não fosse capaz de generalizar esse aprendizado.

Quadro 3 – Hiperparâmetros dos modelos

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: autoria própria.

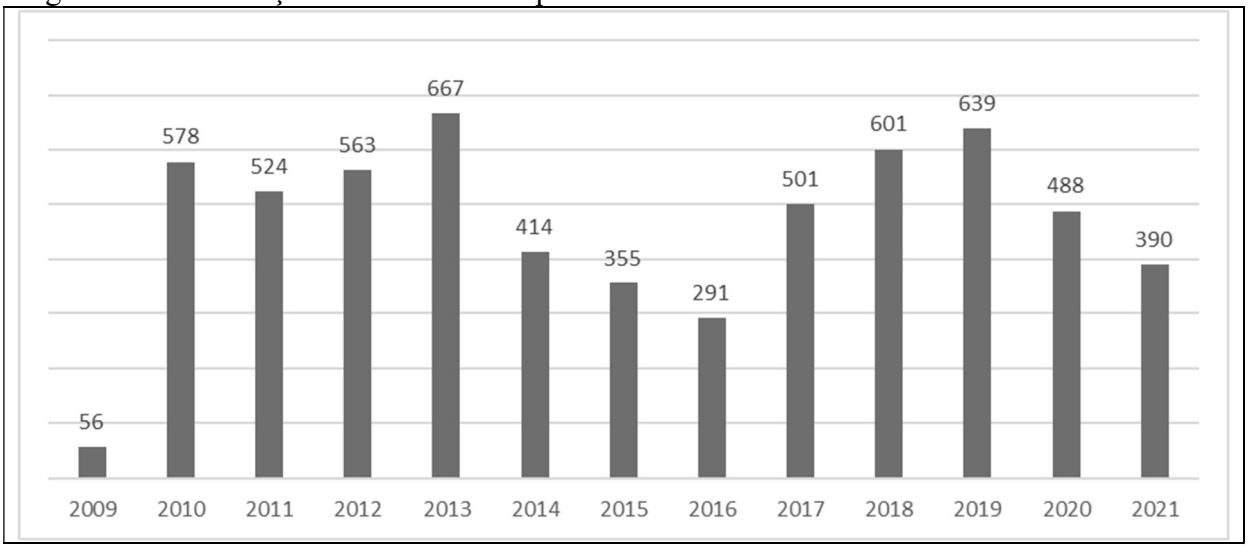
4 Análise e Discussão dos Resultados

Inicialmente apresenta-se uma análise descritiva dos dados, seguida pela avaliação da validade da Lei de Newcomb-Benford ao caso em tela. Na sequência apresentam-se os resultados dos classificadores avaliados, especialmente da regressão logística e da árvore de classificação, dois métodos que além de classificar fornecedores entre sancionados e não sancionados, também permitem identificar as principais variáveis responsáveis pelas sanções aos fornecedores.

4.1 Análise descritiva da base de dados

É possível observar na Figura 6 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.

Figura 6 – Distribuição de fornecedores por ano de cadastro

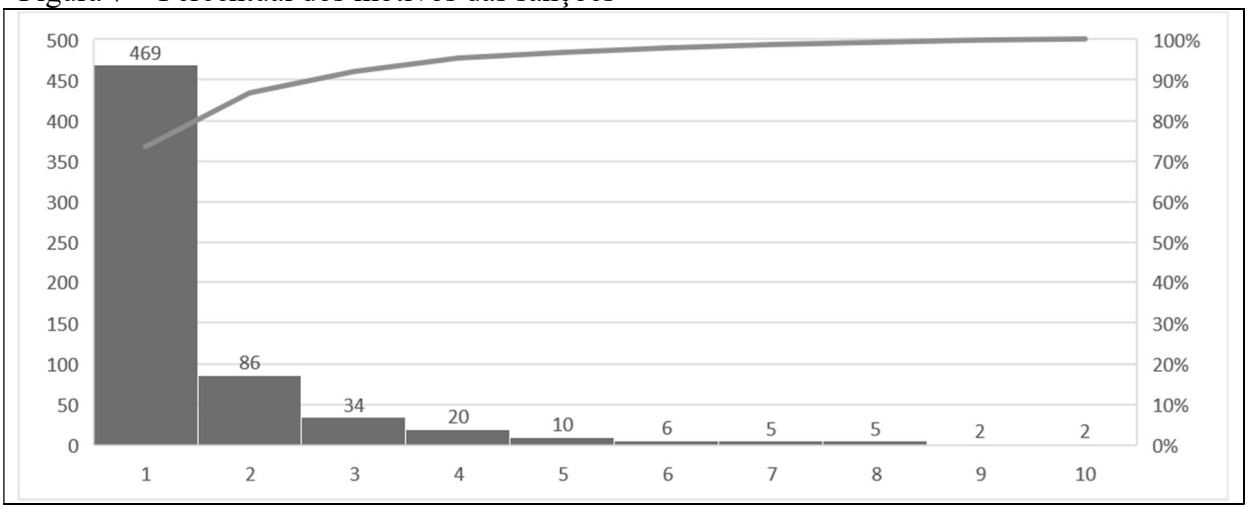


Fonte: autoria própria.

Os motivos pelos quais os fornecedores foram sancionados são, em sua maioria (73%), por inexecução total ou parcial do contrato (Figura 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.

Figura 7 – Percentual dos motivos das sanções



Fonte: autoria própria.

A descrição do motivo para cada sanção pode ser verificada na Tabela 2.

Tabela 2 – 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: autoria própria.

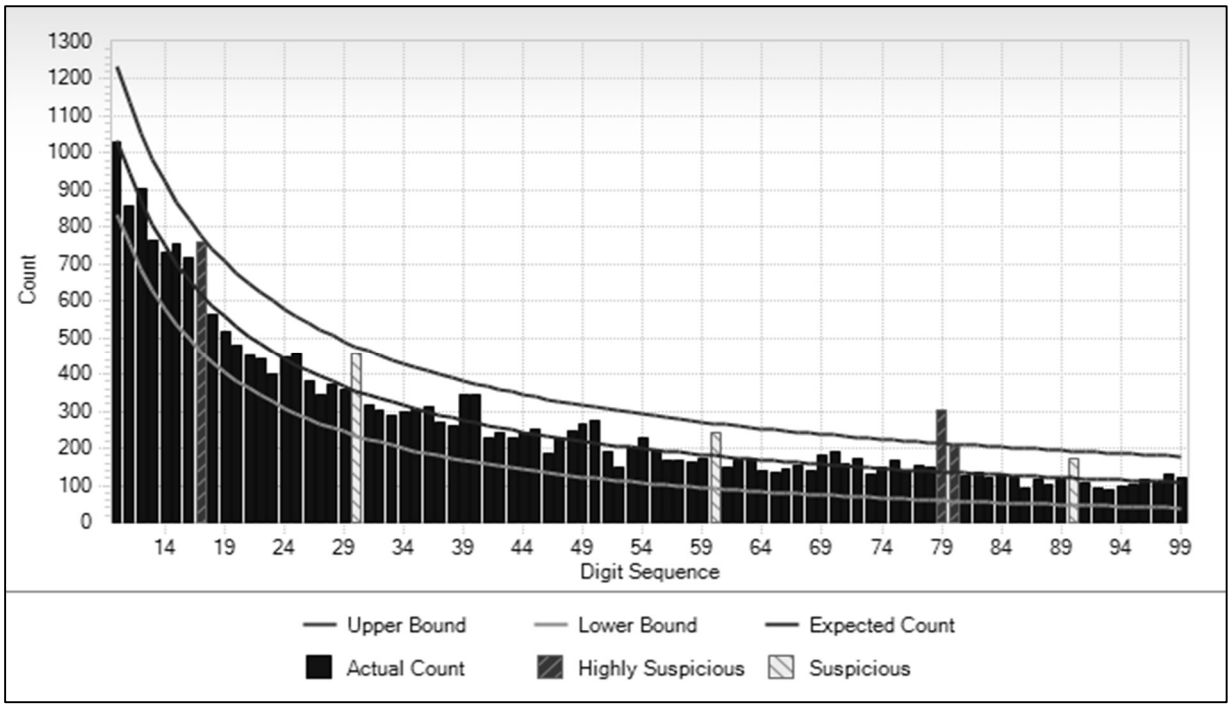
4.2 Lei de Newcomb-Benford

Nesta etapa buscou-se avaliar a hipótese de que os dados da base de contratos estão em conformidade com a Lei de Newcomb-Benford. Vale lembrar que a refutação desta hipótese implica na suspeita de algum tipo de manipulação dos dados.

A verificação foi realizada por meio da ferramenta IDEA e considerou os dois primeiros dígitos do contrato. O número que mais se destacou na análise foi o 79.

A Figura 8 ilustra o quanto a repetição do dígito 79 está acima do esperado. A quantidade esperada para os registros iniciados por 79 é 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.

Figura 8 – Lei de Benford dos dois primeiros dígitos
Fonte: autoria própria.



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.

4.3 Resultados dos modelos de classificação

Conforme indicado na Tabela 3, o algoritmo com a maior precisão (maior taxa de acerto) 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.

Tabela 3 – 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: autoria própria.

A seguir na Tabela 4, cada linha guarda os elementos da matriz de confusão resultante de cada método de classificação.

Tabela 4 – Resultado da matriz de confusão

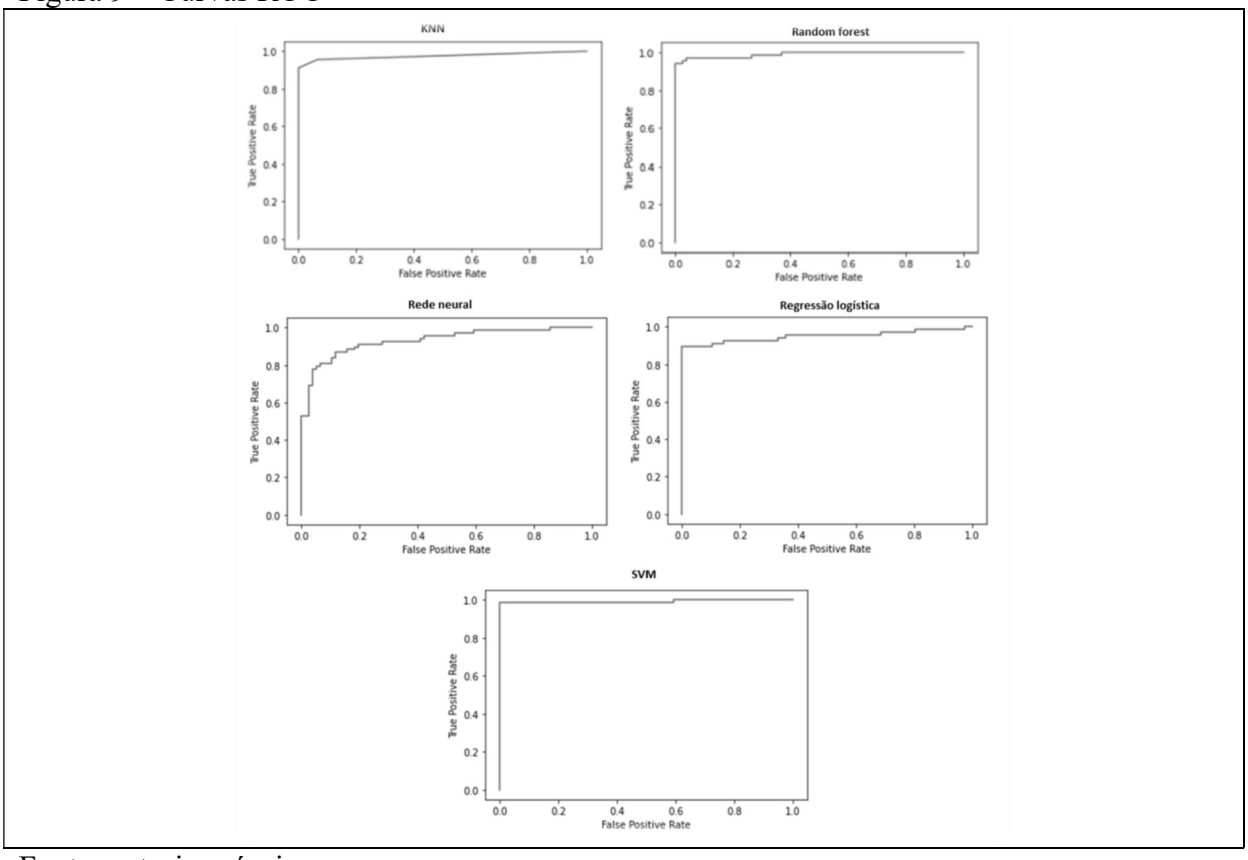
Algoritmo	**TP**	**TN**	**FP**	**FN**
KNN	76	62	6	0
Random forest	74	64	4	2
Rede neural	68	55	13	8
Regressão logística	76	61	7	0
SVM	76	67	1	0

Fonte: autoria própria.

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.

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, conforme demonstrado na Figura 9.

Figura 9 – Curvas ROC



Fonte: autoria própria.

Aprofundando a análise dos resultados, na Tabela 5 apresentam-se os coeficientes do modelo de regressão logística. As variáveis SITUAÇÃO_RECEITA e DISTÂNCIA não

apresentaram coeficientes de regressão estatisticamente significativos, conforme indicado pelos respectivos p-valores menores que o níveis usuais de significância. Por outro lado, as demais variáveis apresentaram coeficientes estatisticamente significativos. Ademais, o impacto de cada variável sobre a chance (*odds*) de sanção do fornecedor pode ser avaliado pela exponencial do respectivo coeficiente, por exemplo, cada licitação (variável TOTAL_LIC) aumenta a chance de ser sanção de um fornecedor em quase 2% ou mais precisamente (1,01981891-1) x 100%, já um ano adicional na média das idades dos sócios (variável MÉDIA_IDADE) reduz a chance de sanção do fornecedor em cerca de 4% ou mais precisamente (0,95631358 - 1) x 100%.

Tabela 5 – Coeficientes da regressão logística

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

Fonte: autoria própria.

Por fim, foi executado o comando *features_importances*, do *random forest*, para ordenar os atributos em ordem decrescente de importância, i.e., peso na definição da classificação final, conforme indicado na Tabela 6. Esse comando só está disponível para o *random forest*.

Tabela 6 – *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	0,013249
ME_EPP	0,011066
QTD_EMPRESAS	0,004864

Fonte: autoria própria.

Os coeficientes da regressão logística são uma medida da importância relativa das variáveis na previsão da variável dependente. Quanto maior o coeficiente de uma variável, maior é sua influência na previsão. Já na *random forest*, as árvores de decisão são construídas a partir de amostras aleatórias do conjunto de dados, e a importância das variáveis é determinada pela sua influência na redução da impureza ao longo das árvores de decisão. Assim, a importância das variáveis na *random forest* é determinada pela média da importância atribuída a cada variável em cada árvore da floresta. Portanto, não é possível estabelecer uma comparação direta entre os dois métodos de avaliação das variáveis pois eles trazem informações diferentes. Os coeficientes da regressão logística permitem avaliar o efeito de uma variável explicativa sobre a chance de um fornecedor ser sancionado e ainda informam a significância estatística deste efeito. Já o grau de importância informado pela regressão logística identificam quais as variáveis mais importantes para a classificação dos fornecedores.

4 Considerações Finais

O teste de hipótese baseado na Lei de Newcomb-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 comitê 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 aquele 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.

Referências

BAESENS, B.; VAN VLASSELAER, V., VERBEKE, W. **Fraud analytics using descriptive, predictive, and social network techniques**: a guide to data science for fraud detection. New Jersey: Wiley, 2015.

BENFORD, F. The Law of Anomalous Numbers. **Proceedings of the American Philosophical Society**, [*s. l.*], v. 78, n. 4, p. 551-572, 1938. Disponível em: http://www.jstor.org/stable/984802. 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, 11 jun. 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.

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

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

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.

HAYKIN, S. **Neural networks and learning machines**. New York: Pearson, 2009.

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

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.

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 Paulo, 2016.

OLOKODANA, N.; FERNANDES, R. Previsão de fraudes em relatórios contábeis à luz do aprendizado de máquina. **Innovation and Technological Development**, [*s. l.*], v. 1, n. 1, p. 85-98, 29 mar. 2020.

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.

SCUDILIO, J. Qual a melhor métrica para avaliar os modelos de Machine Learning? **Flai**, [*s. l.*], 26 jul. 2020. Disponível em: https://www.flai.com.br/juscudilio/qual-a-melhor-metrica-para-avaliar-os-modelos-de-machine-learning/. Acesso em: 6 fev. 2023.

SEVERINO, M. K.; YAOHAO, P. Previsão de fraudes em seguros patrimoniais com algoritmos de aprendizado de máquina. *In*: ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM ADMINISTRAÇÃO, 43 – EnANPAD 2019. São Paulo, 2 a 5 out. 2019. **Anais**… São Paulo: ANPAD, 2019. ISSN: 2177-2576. Disponível em: https://arquivo.anpad.org.br/eventos.php?cod_evento=&cod_evento_edicao=96&cod_edicao_subsecao=1665&cod_edicao_trabalho=26417. Acesso em: 6 fev. 2023.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao *DATAMINING* mineração de dados**. Rio de Janeiro: Ciência Moderna, 2009.

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.

WERBOS, P. J. Backpropagation through time: what it does and how to do it. *In*: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING – ICASSP, [*s. l.*], 1990. **Proceedings**…, [*s. l.*], v. 78, n. 10, p. 1550-1560, Oct. 1990. DOI: 10.1109/5.58337. Disponível em: https://ieeexplore.ieee.org/document/58337. Acesso em: 6 fev. 2023.