UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE GESTÃO E NEGÓCIOS
MESTRADO PROFISSIONAL EM GESTÃO ORGANIZACIONAL

VINÍCIUS SOUZA MORAIS

PREVISÃO DE FRAUDE EM LICITAÇÕES NO BRASIL

UBERLÂNDIA
2024

VINÍCIUS SOUZA MORAIS

PREVISÃO DE FRAUDE EM LICITAÇÕES NO BRASIL

Dissertação apresentada ao Programa de Mestrado Profissional em Gestão Organizacional da Faculdade de Gestão e Negócios da Universidade Federal de Uberlândia como requisito para conclusão do curso.

Orientador: Prof. Dr. Pablo Rogers

Linha de Pesquisa: Gestão Pública

UBERLÂNDIA

2024

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

M827p 2024

Morais, Vinícius Souza, 1991-
Previsão de fraude em licitações no Brasil [recurso eletrônico] / Vinícius Souza Morais. - 2024.

Orientador: Pablo Rogers.
Dissertação (Mestrado profissional) - Universidade Federal de Uberlândia, Programa de Pós-graduação em Gestão Organizacional.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2024.5149
Inclui bibliografia.
Inclui ilustrações.

1. Administração. I. Rogers, Pablo, 1980-, (Orient.). II. Universidade Federal de Uberlândia. Programa de Pós-graduação em Gestão Organizacional. III. Título.

CDU: 658

André Carlos Francisco
Bibliotecário Documentalista - CRB-6/3408

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Gestão Organizacional
Av. João Naves de Ávila, 2121, Bloco 5M, Sala 108 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3291-6333 - www.ppggo.fagen.ufu.br - ppggo@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Gestão Organizacional				
Defesa de:	Dissertação de Mestrado Profissional, 116, PPGGO				
Data:	Três de setembro de dois mil e vinte e quatro	Hora de início:	14:30	Hora de encerramento:	16:00
Matrícula do Discente:	12212GOM023				
Nome do Discente:	Vinícius Souza Morais				
Título do Trabalho:	Previsão de Fraude em Licitações no Brasil				
Área de concentração:	Gestão Organizacional				
Linha de pesquisa:	Gestão Pública				
Projeto de Pesquisa de vinculação:	Previsão de Condição Financeira dos Municípios Mineiros por Meio de Inteligência Artificial (Edital Universal FAPEMIG 01/2023 - APQ-00545-23)				

Reuniu-se, por meio de webconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Gestão Organizacional, assim composta: Professores Doutores: Flávio Luiz de Moraes Barboza (UFU), Marcelo Botelho da Costa Moraes (USP) e Pablo Rogers Silva, orientador do candidato.

Iniciando os trabalhos o presidente da mesa, Dr. Pablo Rogers Silva, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado.

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Pablo Rogers Silva**, **Professor(a) do Magistério Superior**, em 03/09/2024, às 16:10, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Flávio Luiz de Moraes Barboza**, **Professor(a) do Magistério Superior**, em 04/09/2024, às 11:36, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcelo Botelho da Costa Moraes**, **Usuário Externo**, em 04/09/2024, às 15:49, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5617691** e o código CRC **11B2B92E**.

Referência: Processo nº 23117.053027/2024-79 SEI nº 5617691

RESUMO

Os portais eletrônicos de licitação transformaram o processo de aquisição de bens e serviços, tornando-o mais eficiente, competitivo e transparente. Paralelemente, o processo de fiscalização também deve acompanhar essas mudanças e se adaptar a esse novo sistema. No contexto acadêmico, nota-se um avanço com a utilização de técnicas de Inteligência Artificial para a previsão de fraudes, tendo esse ferramental uma ampla gama de aplicações na gestão pública, em particular no acompanhamento de processos licitatórios. Assim, este estudo propõe a utilização da técnica de *Random Forest* para a previsão de indícios de fraudes em licitações na administração pública brasileira. O modelo *Random Forest* foi aplicado ao conjunto de dados relativos à 'Licitações e Contratos' do ano de 2022, obtidos no Portal Transparência da Controladoria-Geral da União. Onze variáveis independentes (incluindo orçamento, objetivo da licitação, tipo de processo, dia que o contrato foi assinado, período de execução, período de espera, intervalo entre a data de assinatura e a data de concessão, intervalo entre a data de início da execução e a data de assinatura, distância para a eleição mais próxima, categoria do produto, estado) foram utilizadas para prever se a empresa recebeu multas (variável dependente). O desempenho do modelo foi avaliado mensalmente e anualmente. O classificador foi eficiente na detecção de fraudes nas licitações com *F1 Score* mensal médio de 78,5% e anual de 80%. Destacou-se pela capacidade de detectar 90% (*recall*) de todos os casos reais de fraude para ambas análises. Na análise mensal, observou-se uma redução considerável nos parâmetros de qualidade dos casos de irregularidade com valores máximo em Janeiro (*recall*=1,00 e *F1 Score*=0,96) e mínimo em Novembro (*recall*=0,47 e *F1 Score*=0,54). As variáveis de maior importância flutuaram significativamente ao longo dos meses, mas no geral foram: Categoria do Produto, Objetivo da Licitação e Estado. Destaca-se o impacto potencial deste trabalho não apenas na academia, mas também para todos os cidadãos e gestores públicos, oferecendo uma ferramenta eficaz na detecção de potenciais desvios de conduta no setor público.

Palavras-chave: Licitação, Licitação Pública, e-Procurement, Fraudes, Serviço Público, Inteligência Artificial, *Random Forest*.

ABSTRACT

Electronic procurement portals have transformed the process of acquiring goods and services, making it more efficient, competitive, and transparent. Simultaneously, the audit process must also keep up with these changes and adapt to this new system. In the academic context, there has been progress with the use of Artificial Intelligence techniques for fraud prediction, having this tool a wide range of applications in public management, particularly in monitoring procurement processes. Thus, this study proposes the use of the Random Forest technique for predicting signs of fraud in public administration procurements in Brazil. The Random Forest model was applied to the dataset related to 'Bidding and Contracts' of the year 2022, obtained from the Transparency Portal of the General Comptroller of the Union. Eleven independent variables (including budget, purpose of the bid, type of process, day the contract was signed, execution period, waiting period, interval between the signing date and the grant date, interval between the start date of execution and the signing date, distance to the next election, product category, state) were used to predict whether the company received fines (dependent variable). The performance of the model was evaluated monthly and annually. The classifier was effective in detecting fraud in procurements with an average monthly F1 Score of 78,5% and annual of 80%. It stood out for its ability to detect 90% (recall) of all actual cases of fraud in both analyses. In the monthly analysis, a significant reduction in the quality parameters of the irregularity cases was observed, with maximum values in January (recall=1,00 and F1 Score=0.96) and minimum in November (recall=0,47 and F1 Score=0,54). The most important variables fluctuated significantly over the months, but overall were: Product Category, Purpose of the Bid, and State. The potential impact of this work is highlighted not only in academia but also for all citizens and public managers, offering an effective tool in detecting potential misconduct in the public sector.

Keywords: Bidding, Public Procurement, e-Procurement, Fraud, Public Service, Artificial Intelligence, Random Forest.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A atividade econômica do Estado desempenha um papel crucial para impulsionar a demanda efetiva, estimulando a produção e criação de empregos no país. Nos Estados Unidos, uma das maiores potências mundiais, as compras de bens e serviços do Governo Federal corresponderam a 17,3% do Produto Interno Bruto (PIB) (U.S. Bureau of Economic Analysis, 2024). Nos países pertencentes à Organização para a Cooperação e Desenvolvimento Econômico (OCDE), as aquisições realizadas pelo Estado aumentaram na última década, passando de 11,8% do PIB em 2007 para 12,9% do PIB em 2021 (OECD, 2024).

No Brasil, observa-se uma tendência similar (Costa; Terra, 2019). As contratações realizadas pelo setor público, considerando todos os entes federativos, foram em média de 12,5% do PIB para o período compreendido entre 2006 e 2017. No caso da União, as contratações públicas representaram, em média, cerca de 6,8% do PIB. Um estudo promovido pelo IPEA também indica duas características relevantes do mercado de aquisições públicas: em comparação com os países que compõem a OCDE, o mercado de compras públicas brasileiro possui maior participação do governo central, enquanto nos demais países observa-se maior participação dos municípios; além disso, a Petrobras representou, individualmente, 4,2% do mercado de aquisições entre 2006 e 2017, ocupando, por consequência, parcela relevante dos gastos da União (Ribeiro; Inácio, 2019).

As contratações do Estado devem ser realizadas por meio de licitação (Brasil, 1988). Nesse sentido, o sistema de leis que regulamenta as contratações públicas atualmente define o poder de compra estatal como um instrumento capaz de efetivar políticas públicas com finalidades distintas, tais como redução das desigualdades, promoção do desenvolvimento socioeconômico e temas relacionados à inovação e à sustentabilidade. Dessa forma, a função primária da contratação pública está relacionada à própria aquisição de determinado bem ou contratação de um serviço, enquanto a função derivada é a consecução de algum fim público adicional (Rejeb *et al.*, 2023; Zago, 2018).

A licitação é um procedimento formal da Administração Pública, no qual empresas são convocadas a apresentar propostas para a oferta de bens e serviços, conforme estabelecido em edital ou convite. O tema é regulamentado por diversas legislações, sendo a principal a Lei nº 14.133/2021, que revogou a Lei nº 8.666/1993, resultando na coexistência de dois modelos de contratações públicas no Brasil. As modalidades de licitação estabelecidas pela Lei nº 8.666/1993 foram atualizadas, destacando-se a criação da modalidade de diálogo competitivo e a substituição das modalidades convite e tomada de preço pelo pregão.

As aquisições públicas são um dos aspectos do Estado em que a probabilidade de ocorrer corrupção, fraude e conluio é maior (Gunasegaran; Basiruddin; Rizal, 2023; Mavidis; Folinas, 2022; Rejeb *et al.*, 2023). No entanto, a maioria dos estudos sobre a prevenção e detecção de fraudes está relacionada ao setor privado (Gunasegaran; Basiruddin; Rizal, 2023). A transparência nas contratações, a capacitação da equipe responsável e a atuação do controle interno e externo são aspectos centrais na prevenção de fraudes. Quanto à detecção das fraudes, a literatura acadêmica indica que o principal método é a utilização de *Red Flags*/Alertas para que os auditores aprofundem suas análises sobre cada caso. Outras formas de detecção de fraudes estão associadas à disponibilização de canais de denúncia e à utilização de equipes multidisciplinares para investigação (Gunasegaran; Basiruddin; Rizal, 2023).

As normas brasileiras que regulamentam as licitações e contratações públicas estabelecem uma série de sanções administrativas e penais para garantir a transparência, moralidade e eficiência nesses processos. A Lei nº 8.666/1993, conhecida como Lei de Licitações e Contratos, prevê punições como advertência, multa, suspensão temporária de participação em licitação e declaração de inidoneidade para licitar ou contratar com a Administração Pública. A Lei do Pregão (Lei nº 10.520/2002) e o Regime Diferenciado de Contratações Públicas (Lei nº 12.462/2011) também estipulam sanções semelhantes, enquanto a Nova Lei de Licitações e Contratos Administrativos (Lei nº 14.133/2021) introduz novas disposições, mantendo uma estrutura similar de sanções administrativas.

A legislação brasileira também aborda a responsabilidade administrativa e civil por atos de corrupção no contexto das licitações e contratações públicas. A Lei Anticorrupção (Lei nº 12.846/2013) estabelece punições para pessoas jurídicas envolvidas em atos lesivos contra a administração pública, enquanto o Código Penal tipifica crimes como fraude em licitação, peculato e corrupção. A implementação dessas leis visa promover a integridade no ambiente corporativo e governamental, além de combater a corrupção no país, alinhando-se aos esforços internacionais de governança e transparência.

No entanto, a aplicação eficaz dessas leis enfrenta desafios significativos. A complexidade dos processos de licitação e a grande quantidade de transações tornam a detecção de irregularidades uma tarefa árdua. A identificação precoce de possíveis fraudes e atos corruptos é essencial para a eficácia das medidas preventivas e punitivas. Nesse contexto, a tecnologia desempenha um papel crucial ao fornecer ferramentas avançadas para monitorar e analisar os dados das contratações públicas.

Contratações públicas eletrônicas são um tema multidisciplinar, que está relacionado a diversos campos do conhecimento, como o governo eletrônico, os sistemas de informação, a

administração pública e o direito (Mohungoo; Brown; Kabanda, 2020). A adoção de sistemas eletrônicos traz melhorias na eficiência e na eficácia do setor público, em especial quanto à transparência, à governança, à melhor utilização dos recursos públicos, à promoção do desenvolvimento regional e da competição entre os fornecedores (Adobor; Yawson, 2023; Gadour, 2024; Gunasegaran; Basiruddin; Rizal, 2023; Mavidis; Folinas, 2022). Além disso, os sistemas eletrônicos são ferramentas fundamentais para dificultar a ocorrência de fraudes nas contratações (Adobor; Yawson, 2023; Gadour, 2024; Gunasegaran; Basiruddin; Rizal, 2023).

Os sistemas eletrônicos de aquisições públicas estão em constante evolução, simplificando e automatizando processos. Um dos desafios atuais é realizar a transição tecnológica desses sistemas, com a utilização de tecnologias da indústria 4.0 (Gadour, 2024; Mavidis; Folinas, 2022; Siciliani *et al.*, 2023). Para que esse processo de transformação seja viabilizado, os países devem atentar-se para questões como a alteração da legislação, preocupação ética quanto à substituição de mão de obra por máquinas, capacitação de seus quadros funcionais nas tecnologias emergentes, mudanças na estrutura organizacional, bem como investimentos em infraestrutura tecnológica (Adobor; Yawson, 2023; Gadour, 2024; Mavidis; Folinas, 2022; Mohungoo; Brown; Kabanda, 2020).

A adoção de tecnologias como a inteligência artificial (IA) e o aprendizado de máquina tem se mostrado promissora para superar esses desafios. Ferramentas de análise de dados podem processar grandes volumes de informações rapidamente, identificando padrões e anomalias que podem indicar atividades ilícitas. Além disso, a integração de técnicas de inferência causal permite uma compreensão mais profunda das relações entre variáveis e a identificação de fatores de risco específicos.

A Administração Pública tem adotado com cada vez mais frequência ferramentas de inteligência artificial. O Supremo Tribunal Federal anunciou a incorporação ao seu acervo da ferramenta VitórIA, com a finalidade de realizar o agrupamento dos processos judiciais por tema (Supremo Tribunal Federal, 2023a). Além disso, já utiliza outras ferramentas de IA, como o Projeto Victor, destinado à identificação de processos com temas que já possuam julgados com repercussão geral, e a RAFA 2030 (Redes Artificiais Focadas na Agenda 2030), que auxilia na classificação de processos por associação com os Objetivos de Desenvolvimento Sustentável (ODS) (Supremo Tribunal Federal, 2021; Supremo Tribunal Federal, 2023a; Supremo Tribunal Federal, 2023b). Nesse mesmo sentido, o Tribunal de Contas da União, responsável por realizar o controle externo da Administração Pública Federal, também possui diversas iniciativas com a aplicação de ferramentas de IA. Este órgão já implementou o *chatbot* Zello, utilizado para realizar atendimentos por mensagem, além do ChatTCU e do Copilot

TCU, que são IAs generativas que utilizam os bancos de dados do próprio tribunal (Secom TCU, 2020; Tribunal de Contas da União, 2024).

No campo das compras públicas, a principal iniciativa é o *software* Alice (Analisador de Licitações, Contratos e Editais), desenvolvido pela Controladoria-Geral da União (2024). Esta ferramenta permite a análise automatizada dos documentos publicados nos portais de contratação do Governo Federal, do Banco do Brasil e da Caixa, sendo que recentemente foi disponibilizada para utilização pelos municípios. A atuação antecipada nas contratações públicas com suspeitas já possibilitou uma economia da ordem de 11 bilhões de reais (Biagini, 2024).

Os modelos de aprendizado de máquina oferecem uma abordagem inovadora para prever condutas ilegais na aquisição pública, utilizando análise de grandes conjuntos de dados para identificar padrões e variáveis associadas a comportamentos corruptos. Essas previsões podem ser utilizadas como pontuações de risco por agências anticorrupção e partes contratantes para orientar a seleção de operações de aquisição que precisam de monitoramento mais rigoroso ou auditorias.

Diversos trabalhos buscaram investigar a aplicação das ferramentas de inteligência artificial nas compras públicas. Marin *et al.* (2023) utilizaram algoritmos de Processamento de Linguagem Natural e Aprendizado de Máquina para realizar a classificação de uma lista de itens a serem adquiridos. Siciliani *et al.* (2023) desenvolveram um Sistema de Apoio à Tomada de Decisão utilizando informações do sistema de aquisições públicas italiano. Essas informações foram integradas em um banco de dados, pré-processadas e analisadas por meio de ferramentas de Processamento de Linguagem Natural e Aprendizado de Máquina. Torres-Berru, Lopez-Batista e Zhingre (2023) buscaram estudar, utilizando o Processamento de Linguagem Natural, a ocorrência de preconceito e favorecimentos nas contratações públicas do Equador, indicando que este é um forte alerta para corrupção. O modelo proposto obteve uma acurácia de 88% na detecção de preconceito e acurácia de 90% na detecção de favorecimento.

Nota-se estudos relevantes na literatura recente. Estudos como o de Gallego *et al.* (2021) destacam a eficácia dos modelos de aprendizado de máquina na previsão de condutas ilegais, fornecendo *insights* valiosos para políticas direcionadas e reformas regulatórias. Além disso, a combinação desses modelos com técnicas tradicionais de inferência causal representa uma abordagem promissora para fortalecer as políticas de combate à corrupção e melhorar a integridade dos processos de aquisição pública. Nai *et al.* (2023) aplicaram técnicas de aprendizado de máquina para prever reclamações nos tribunais administrativos em processos

de aquisição na Itália, alcançando uma acurácia de 80%. Oliveira *et al.* (2023) identificaram fraudes em licitações públicas por meio da análise das interações entre os participantes, reduzindo um grande volume de dados em indicadores e direcionando o trabalho de auditoria. Decarolis e Giorgiantonio (2020) destacam a importância de identificar indicadores de risco, promover a competição entre empresas e implementar práticas anticorrupção para fortalecer a integridade nas licitações públicas e reduzir a corrupção.

Diante disso, este estudo propõe um modelo de previsão de fraudes (alertas) em licitações no Brasil utilizando a técnica *Random Forest*. O objetivo principal é investigar a eficácia desse modelo, por meio da análise da acurácia, dos erros do tipo 1 (falsos positivos) e tipo 2 (falsos negativos), da identificação das variáveis mais importantes e pela comparação com estudos relacionados. A hipótese central da pesquisa é que o *Random Forest* pode ser uma ferramenta eficaz para a detecção de fraudes em licitações, contribuindo tanto para o avanço científico quanto para a aplicação prática na gestão pública.

Este estudo tem utilidade para os agentes públicos que atuam nas contratações públicas pois permite a identificação de casos com maiores chances de irregularidades de forma preditiva, facilitando a análise em tempo real (Siciliani *et al.*, 2023). Tendo em vista as competências desempenhadas pelo controle externo no âmbito da governança pública e a experiência no uso de ferramentas de inteligência artificial, este trabalho será especialmente útil no trabalho de auditoria.

Sob a perspectiva de contribuições acadêmicas, este trabalho preenche uma lacuna na literatura relacionada às compras públicas ao explorar o campo da aplicação de uma ferramenta de inteligência artificial, que é um campo com poucos trabalhos, conforme identificado por Rejeb *et al.* (2023) e Gunasegaran, Basiruddin e Rizal (2023).

Além desta seção introdutória, este estudo segue com o referencial teórico, que aborda: Licitação no Brasil e contratação pública eletrônica; Condutas ilícitas, fraudes e corrupção em contratações públicas; Dados públicos e modelos de aprendizagem de máquina para detecção prévia de fraudes, juntamente com estudos relacionados. A metodologia inclui a coleta de dados, *Random Forest*, o Algoritmo e a Análise dos resultados. Por fim, é apresentada a seção de Resultados e Discussões, seguida pela Conclusão com as considerações finais.

2 REFERENCIAL TEÓRICO

2.1 Licitação no Brasil e contratação pública eletrônica

A licitação é um procedimento administrativo formal em que a Administração Pública convoca, por meio de condições estabelecidas em ato próprio (edital ou convite), empresas interessadas na apresentação de propostas para o fornecimento de bens, serviços ou execução de obras (Brasil, 2010; Tribunal de Contas da União, 2023). Nesse sentido, o processo de licitação visa selecionar a proposta mais vantajosa, a partir de condições anteriormente estabelecidas e pautadas nos princípios da Administração Pública. Em outras palavras, a licitação pode ser definida como uma série de atos organizados e sistemáticos realizados por agentes públicos e particulares, com a finalidade de selecionar a proposta contratual mais vantajosa para a Administração Pública (Amorim, 2020).

Como desdobramento do dispositivo constitucional que estabelece a licitação como a forma pela qual o Estado realizará suas contratações, foram criadas diversas legislações para regulamentar o tema. As principais foram consignadas no Quadro 1.

Quadro 1 – Legislação relacionada às licitações

Lei/Decreto	Assunto
Lei Nº 14.133, de 1 de abril de 2021	Lei de Licitações e Contratos Administrativos
Decreto Nº 10.024, de 20 de setembro de 2019	Regulamenta a licitação, na modalidade pregão, na forma eletrônica
Decreto Nº 7.892, de 23 de janeiro de 2013	Regulamenta o Sistema de Registro de Preços
Lei Nº 10.520, de 17 de julho de 2002	Institui modalidade de licitação denominada pregão
Lei Nº 8.666, de 21 de julho de 1993	Institui normas para licitações e contratos

Fonte: Elaboração própria.

As aquisições públicas passaram por um processo de transformação recente, culminando na aprovação da Lei Nº 14.133, de 1º de abril de 2021, chamada de Lei de Licitações e Contratos Administrativos, a qual revogou, em 2023, a Lei Nº 8.666, de 21 de julho de 1993. Embora a Lei Nº 8.666/1993 não esteja mais vigente, todas as contratações iniciadas com base nessa legislação seguirão este regramento até o encerramento dos

respectivos contratos. Como efeito, coexistem dois modelos de contratações públicas no Brasil: aquele instituído pelos regramentos anteriores à Lei Nº 14.133/2021, válido até o término dos contratos originados sob este modelo, e aquele instituído pela Lei Nº 14.133/2021, que se tornará o único modelo a ser utilizado para novas contratações públicas (Brasil, 1993, 2021; Tribunal de Contas da União, 2023).

As modalidades de licitação estabelecidas na Lei Nº 8.666/1993 são concorrência, tomada de preços, convite, concurso e leilão. A Lei Nº 10.520, de 17 de julho de 2002, instituiu a modalidade de licitação denominada pregão, sendo esta a modalidade mais utilizada atualmente. Com a nova lei de licitações, as modalidades convite e tomada de preços deixaram de existir, enquanto foi criada a modalidade denominada diálogo competitivo. Também é importante destacar que as contratações de obras, antes realizadas de forma presencial nas modalidades de tomada de preços e concorrência, passaram a ser realizadas de forma eletrônica pela modalidade denominada concorrência eletrônica (Brasil, 1993, 2002; Tribunal de Contas da União, 2023).

Esses normativos também trazem os princípios que devem orientar a sua interpretação e aplicação. No Quadro 2, foram consignados os artigos contendo o rol de princípios que devem nortear as contratações públicas.

Para que a Administração Pública viabilize suas aquisições, os processos de contratação devem observar basicamente três etapas: a fase preparatória, a fase externa e a fase de execução contratual. A fase preparatória inicia-se com a formalização da demanda, ou seja, da necessidade que a Administração Pública precisa suprir. Em seguida, é nomeada uma equipe de planejamento, incumbida de construir os estudos técnicos preliminares e o termo de referência ou o projeto básico. Nessa fase, a necessidade é detalhada de forma a permitir ao mercado compreender o que a Administração Pública necessita, bem como definir o valor máximo que irá dispor para esse fim. Nessa etapa, também são definidas as regras de como funcionará a seleção da proposta e da empresa que irá atender à contratação (Amorim, 2020; Brasil, 2021; Tribunal de Contas da União, 2023).

A partir da publicação do edital de licitação no Diário Oficial do ente federativo, tem-se o marco do início da fase externa da licitação. No caso do pregão eletrônico, essa etapa ocorre a competição de fato, por meio do envio de lances no sistema ComprasGov, segundo critérios pré-estabelecidos. Em sequência, ocorre o julgamento da proposta mais bem classificada, bem como a avaliação dos critérios de habilitação das empresas participantes. Ao final, os itens adquiridos ou serviços contratados são adjudicados à empresa que oferecer a proposta de menor valor que cumpra os requisitos fixados no edital e no termo de referência,

bem como atenda às condições de habilitação. A homologação do resultado do processo licitatório pela autoridade máxima de cada órgão indica o fim da fase externa ou de seleção do fornecedor. Na fase de execução contratual, os materiais adquiridos são entregues ou os serviços contratados começam a ser executados (Amorim, 2020; Brasil, 2021).

Quadro 2 – Legislação versus Princípios

Lei/Decreto	**Artigo da norma contendo os Princípios**
Lei Nº 14.133, de 1 de abril de 2021	Art. 5º Na aplicação desta Lei, serão observados os princípios da legalidade, da impessoalidade, da moralidade, da publicidade, da eficiência, do interesse público, da probidade administrativa, da igualdade, do planejamento, da transparência, da eficácia, da segregação de funções, da motivação, da vinculação ao edital, do julgamento objetivo, da segurança jurídica, da razoabilidade, da competitividade, da proporcionalidade, da celeridade, da economicidade e do desenvolvimento nacional sustentável, assim como as disposições do Decreto-Lei nº 4.657, de 4 de setembro de 1942 (Lei de Introdução às Normas do Direito Brasileiro).
Lei Nº 8.666, de 21 de julho de 1993	Art. 3º A licitação destina-se a garantir a observância do princípio constitucional da isonomia, a seleção da proposta mais vantajosa para a administração e a promoção do desenvolvimento nacional sustentável e será processada e julgada em estrita conformidade com os princípios básicos da legalidade, da impessoalidade, da moralidade, da igualdade, da publicidade, da probidade administrativa, da vinculação ao instrumento convocatório, do julgamento objetivo e dos que lhes são correlatos.
Decreto Nº 10.024, de 20 de setembro de 2019	Art. 2º O pregão, na forma eletrônica, é condicionado aos princípios da legalidade, da impessoalidade, da moralidade, da igualdade, da publicidade, da eficiência, da probidade administrativa, do desenvolvimento sustentável, da vinculação ao instrumento convocatório, do julgamento objetivo, da razoabilidade, da competitividade, da proporcionalidade e aos que lhes são correlatos. § 1º O princípio do desenvolvimento sustentável será observado nas etapas do processo de contratação, em suas dimensões econômica, social, ambiental e cultural, no mínimo, com base nos planos de gestão de logística sustentável dos órgãos e das entidades. § 2º As normas disciplinadoras da licitação serão interpretadas em favor da ampliação da disputa entre os interessados, resguardados o interesse da administração, o princípio da isonomia, a finalidade e a segurança da contratação.

Fonte: Elaboração própria.

A intensificação do uso das Tecnologias da Informação e Comunicação (TIC) transformou a interação dos governos com a sociedade, proporcionando maior comodidade e conveniência no acesso a serviços, conceito conhecido como governo eletrônico (e-Gov). Com a evolução tecnológica, o governo eletrônico evoluiu para o governo digital, que busca modernizar a administração pública, utilizando a Tecnologia da Informação (TI) para reconstruir processos e otimizar serviços, reduzindo a burocracia e melhorando a experiência do cidadão (Tribunal de Contas da União, 2024).

Além disso, ao longo dos últimos anos, inovações no campo da Tecnologia da Informação foram sendo incorporadas e implementadas no campo das compras públicas pelo governo federal. Assim, municípios, estados e o Distrito Federal começaram a desenvolver seus próprios sistemas para a realização de compras públicas. Como consequência, com a utilização desses sistemas, como o Sistema Integrado de Administração de Serviços Gerais (SIASG), além do Comprasnet, a Administração Pública pôde contar com maior rapidez, aumento da concorrência, maior transparência e controle nos processos de compras. Além disso, os gastos públicos direcionados ao processo de compras foram reduzidos (Pinto, 2020).

Todas essas inovações têm sido importantes, uma vez que alguns dos desafios da Administração Pública são atender às demandas da população a partir de um serviço célere, de custo possível, mantendo a qualidade e eficiência, além da transparência (Pinto, 2020). A partir da análise de 70 trabalhos relacionados à temática de compras públicas, apresentados em congressos do Conselho Nacional de Secretários de Estado de Administração, entre 2008 e 2016, Fernandes (2019) identificou que, a respeito dos portais da *internet*, estes proporcionaram grande inovação no processo de licitação. No caso da Administração Federal, o governo desenvolveu seu próprio portal eletrônico, o Comprasnet, possibilitando o cadastro online de fornecedores, a participação nos pregões eletrônicos, além do acesso dos fornecedores às informações relacionadas às licitações.

De acordo com Fernandes (2019), quase todos os estados possuem portais eletrônicos de compras que possibilitam que as fases da licitação ocorram de forma eletrônica, principalmente a divulgação dos editais, o cadastro dos fornecedores e os pregões. A implementação dos portais se deu inicialmente a partir dos estados de São Paulo e Minas Gerais.

Assim, pode-se dizer que os sistemas eletrônicos de contratações públicas representam uma importante evolução no processo de licitações e aquisições governamentais, trazendo maior transparência, eficiência e garantias para o processo de concorrência. Esses sistemas são fundamentais na modernização da administração pública, uma vez que permitem a publicação

online de todos os processos licitatórios, o que facilita o acesso e a fiscalização por parte de fornecedores e cidadãos (controle social). Além disso, a digitalização reduz custos operacionais e acelera o processo de contratação, beneficiando tanto o governo quanto os fornecedores (Costa; Hollanagel; Bueno, 2019; Pinto, 2020).

Esses sistemas promovem uma competição mais ampla e justa, permitindo a participação de uma gama maior de fornecedores, incluindo empresas de diferentes regiões ou mesmo internacionais. A padronização dos processos assegura que todos os participantes sigam as mesmas regras, garantindo a legalidade e segurança jurídica nos processos. No entanto, esses sistemas enfrentam desafios, como a necessidade de treinamento dos usuários, a resistência à mudança e as questões de segurança cibernética (Cardoso, 2004).

No Brasil, há vários sistemas em que as contratações públicas podem ser realizadas. Um dos principais é o Portal de Compras do Governo Federal. Este sistema é obrigatório para a Administração Pública Federal direta, bem como para autarquias e fundações, além de ser aplicável a entes federativos que empregam recursos provenientes de transferências voluntárias da União. Outros sistemas relevantes são a Bolsa Eletrônica de Compras do Estado de São Paulo (BEC/SP), que oferece um ambiente específico para licitações no âmbito estadual, o Licitações-e, utilizado pelo Banco do Brasil em suas contratações, e o Portal de Compras do Governo de Minas Gerais (Contador; Cardoso, 2005).

O sistema Comprasnet possui diversos pontos fortes, incluindo a transparência, que permite a qualquer cidadão acompanhar as licitações públicas; a competitividade, que facilita a participação de fornecedores de todo o país; a economicidade, que permite aos órgãos públicos obter melhores preços desde a pesquisa até a negociação com fornecedores; e a agilidade, permitindo que as licitações sejam realizadas de forma eletrônica. No entanto, Comprasnet apresenta alguns pontos fracos, como seu modelo de negócio não ser específico para *marketplace*, o que dificulta a integração com outros sistemas governamentais e a aquisição de serviços. Além disso, seu processo de implementação antigo dificulta a atualização do sistema para atender às novas necessidades da administração pública (Yamaji; Vieira; Ferrer, 2022).

As iniciativas de compras eletrônicas (*e-procurement*) estão se tornando mais comuns em todas as esferas de governo. Embora a dimensão governo para negócios (G2B) tenha sido bem desenvolvida, as dimensões governo para o cidadão (G2C) e governo para governo (G2G) ainda necessitam de maior atenção e integração de dados entre diferentes níveis de governo. Exemplos como o portal Comprasnet, que obtém apenas 36% e 60% das pontuações nas dimensões G2C e G2G, respectivamente, ilustram essa necessidade. A avaliação preliminar

sugere que há um grande potencial de evolução na transparência e eficiência dos portais públicos de compras, apontando para um campo de estudo rico em perspectivas na administração pública brasileira (Chevitarese Alves; Dufloth, 2010).

2.2 Condutas ilícitas, fraudes e corrupção em contratações

De forma geral, corrupção refere-se a atos repreensíveis que prejudicam a população e estão diretamente relacionados à ética, ocorrendo em negociações com a Administração Pública ou entre particulares. A Transparência Internacional define corrupção como o "abuso de poder confiado para ganho privado", abrangendo diversos atos, como suborno, fraude, conflito de interesses, nepotismo e lavagem de dinheiro (Brasil, 2022).

Entre as principais condutas infratoras, destacam-se a fraude à licitação, que se refere a qualquer ato que comprometa a isonomia e a competitividade do procedimento licitatório, como combinação de preços ou simulação de propostas. Também é considerada infração a obstrução da fiscalização, que ocorre quando alguém impede ou dificulta a fiscalização dos órgãos competentes sobre os procedimentos licitatórios, contratos ou execução das obras e serviços (Brasil, 2021).

Outras condutas infratoras incluem a modificação ou pagamento irregular, que se refere a alterações indevidas no conteúdo ou na qualidade do objeto de um contrato, bem como a realização de pagamentos em desacordo com as normas contratuais ou legais. O afastamento indevido de licitante, por meio de fraude ou oferecimento de vantagem, e a frustração do caráter competitivo da licitação, ao comprometer ou impedir a realização de qualquer fase do procedimento licitatório, também são práticas sancionadas pela lei (Brasil, 2021).

Além disso, a prática de condutas anticompetitivas, o atraso injustificado na execução de obras ou serviços, a inexecução do contrato e a corrupção, que envolve oferecer ou receber vantagens indevidas relacionadas à licitação ou ao contrato, são igualmente consideradas infrações graves. Por fim, o conluio entre licitantes, com o objetivo de fraudar a licitação, é mais uma conduta passível de penalidades (Brasil, 2021).

A fraude nos processos de contratação pública tem como objetivo principal frustrar o caráter competitivo, distorcendo a livre disputa entre os participantes para favorecer indevidamente alguém. Esse tipo de fraude ocorre quando o agente público responsável pelo processo no órgão licitador direciona a licitação através da imposição de exigências técnicas específicas, excluindo a maioria das empresas que poderiam participar. Dessa forma, beneficia-se um fornecedor específico, favorecendo tanto o fornecedor quanto o agente público envolvido (Sá; Pessanha; Alves, 2024).

A Lei do Sistema Brasileiro de Defesa da Concorrência identifica duas formas principais de formação de cartel: concorrentes que combinam previamente suas ações em licitações para controlar preços e condições, e aqueles que combinam suas ações em outras situações de mercado. O Cade é responsável por punir esses cartéis, inclusive quando envolvem fraudes em licitações. Fraude, nesse contexto, é definida como qualquer ato destinado a obter vantagem ilícita em licitações públicas, incluindo superfaturamento, falta de ampla divulgação, dispensa indevida de licitação, direcionamento de contratos, apresentação de documentos falsos e outros. A Lei Anticorrupção confere à Controladoria-Geral da União (CGU) a competência para instaurar procedimentos contra pessoas jurídicas por essas práticas corruptas. Fraudes em licitações podem ou não envolver ajustes prévios entre licitantes ou com servidores públicos. Tais fraudes lesam o patrimônio público e comprometem a integridade do processo de escolha de fornecedores, protegida pela CGU, enquanto a formação de cartel prejudica a livre concorrência, justificando a intervenção do Cade (De Paula; Gobbes, 2024).

As normas brasileiras que regulamentam as licitações e contratações públicas estabelecem uma série de sanções administrativas e penais para garantir a transparência, moralidade e eficiência nesses processos. Entre as principais leis nesse contexto, destaca-se a Lei nº 8.666/1993, conhecida como Lei de Licitações e Contratos, que previu punições como advertência, multa, suspensão temporária de participação em licitação e declaração de inidoneidade para licitar ou contratar com a Administração Pública. A Lei do Pregão (Lei nº 10.520/2002) e o Regime Diferenciado de Contratações Públicas (Lei nº 12.462/2011) também estipulam sanções semelhantes para infrações em seus respectivos âmbitos. Recentemente, a Lei nº 14.133/2021, conhecida como Nova Lei de Licitações e Contratos Administrativos, foi instituída para atualizar e substituir as legislações anteriores, mantendo as sanções administrativas e introduzindo novas disposições.

Além dessas leis específicas, a Lei de Improbidade Administrativa (Lei nº 8.429/1992) aborda a punição de agentes públicos por atos de improbidade, que podem incluir irregularidades em licitações e contratações. Por fim, o Código Penal brasileiro tipifica crimes relacionados a esses processos, como fraude em licitação, peculato, corrupção passiva e ativa. Essas legislações formam um arcabouço legal que busca prevenir e punir irregularidades e ilegalidades nas licitações e contratações públicas, assegurando a integridade e eficácia desses procedimentos essenciais para a gestão dos recursos públicos.

Atos de corrupção podem levar a consequências legais nas esferas criminal, civil e administrativa. Atos criminosos contra a Administração Pública são definidos no Código Penal, enquanto ações civis podem ser intentadas sob a Lei de Improbidade Administrativa.

Responsabilidades administrativas por práticas relacionadas à corrupção são delineadas em leis que regem contratos públicos, como a Lei de Licitações e Contratos (Brasil, 2022).

A Lei Anticorrupção estabelece responsabilidade administrativa e civil para pessoas jurídicas envolvidas em atos lesivos contra a administração pública nacional ou estrangeira, realizados em seu interesse ou benefício. Investigações nessas diferentes esferas podem prosseguir de forma independente, com inquéritos administrativos não necessariamente aguardando processos civis ou criminais. Embora investigações administrativas, civis e criminais geralmente não influenciem os resultados umas das outras, existem exceções, como a disposição na Lei de Servidores Públicos que absolve a responsabilidade administrativa em caso de absolvição criminal negando a materialidade do delito ou sua autoria (Brasil, 2022).

A Lei Anticorrupção, Lei nº 12.846/2013, é um marco importante na luta contra a corrupção no Brasil, estabelecendo uma série de condutas infratoras que podem levar à responsabilização de pessoas jurídicas. Entre as principais condutas infratoras previstas pela lei estão a promessa, oferta ou entrega de vantagem indevida a agente público ou a terceiros relacionados, o financiamento ou patrocínio de atos ilícitos, a utilização de intermediários para ocultar interesses reais ou beneficiários de atos praticados, a fraude ou frustração do caráter competitivo de licitações públicas, a obtenção indevida de vantagens ou benefícios em detrimento da administração pública, a manipulação ou fraude do equilíbrio econômico-financeiro de contratos públicos e a interferência em investigações ou fiscalizações de órgãos públicos (Brasil, 2013).

As sanções para as empresas que infringirem a lei são severas, incluindo multas, publicação da decisão condenatória e a proibição de receber incentivos financeiros de entidades públicas. A Lei Anticorrupção é um passo importante para promover a integridade no ambiente corporativo e combater a corrupção no país (Brasil, 2013).

A Nova Lei de Licitações e Contratos Administrativos (Lei nº 14.133/2021) estabelece um conjunto de regras e diretrizes para as contratações públicas no Brasil, visando garantir a eficiência, a transparência e a integridade desses processos. Dentro desse contexto, a lei também define uma série de condutas consideradas infratoras, que podem levar à aplicação de sanções administrativas, civis e penais aos envolvidos (Brasil, 2021). É importante ressaltar que a Lei nº 14.133/2021 busca promover um ambiente de contratações públicas mais íntegro e eficiente, e a identificação e punição das condutas infratoras são fundamentais para alcançar esse objetivo. As sanções aplicáveis variam de acordo com a gravidade da infração e podem incluir multas, advertências, impedimento de contratar com a Administração Pública, e declaração de idoneidade.

O Brasil demonstra um compromisso contínuo com a governança global e os esforços internacionais de combate à corrupção por meio de sua participação ativa em várias convenções internacionais. A adesão à Convenção Interamericana Contra a Corrupção, da Organização dos Estados Americanos (OEA), em 1996, destaca a determinação do país em fortalecer as instituições democráticas, prevenir distorções econômicas, eliminar vícios na gestão pública e melhorar a moral social. Este compromisso é reforçado pela participação do Brasil em outras importantes convenções, como a da OCDE sobre o Combate da Corrupção de Funcionários Públicos Estrangeiros em Transações Comerciais Internacionais e a Convenção das Nações Unidas contra a Corrupção, que buscam erradicar a corrupção e promover a transparência nas transações internacionais e no setor público (Brasil, 2022).

Essas convenções formam um quadro robusto para a luta contra a corrupção, incentivando medidas preventivas, criminalização de atos corruptos, e fomentando a cooperação internacional nesta área. O envolvimento do Brasil nestes acordos internacionais não apenas reafirma seu papel como um ator global responsável, mas também contribui para os esforços internos de reforma para assegurar a integridade no setor público e privado. A implementação dessas convenções reflete a dedicação do Brasil em adotar as melhores práticas internacionais, visando fortalecer a confiança pública nas instituições e promover um ambiente de negócios justo e transparente (Brasil, 2021).

Um desempenho ruim nas contratações públicas pode acarretar uma série de consequências negativas tanto para as organizações envolvidas quanto para a sociedade em geral. Isso inclui o desperdício de recursos públicos, a falta de transparência e prestação de contas, bem como a seleção de fornecedores inadequados que resultam em serviços e produtos de baixa qualidade. Ademais, a corrupção e a ineficiência nas contratações públicas podem impactar negativamente o desenvolvimento econômico, afastar investidores e prejudicar a confiança nas instituições públicas, gerando riscos legais e danos à reputação das organizações envolvidas. Portanto, é crucial adotar práticas transparentes, éticas e eficazes nas contratações públicas para evitar tais implicações e promover a eficiência e a integridade nesse processo (Rodrigues; Reis, 2023).

O combate a fraudes em compras públicas é essencial para garantir a integridade e a confiança nas instituições governamentais, fortalecendo a transparência e a ética. Além disso, a luta contra as fraudes promove o uso eficiente dos recursos públicos, direcionando-os para áreas prioritárias como saúde, educação e infraestrutura, e assegurando que o dinheiro dos contribuintes seja empregado de forma responsável. Ao eliminar práticas fraudulentas, o governo fomenta a concorrência justa entre os fornecedores, promovendo igualdade de

oportunidades e um ambiente de negócios transparente e competitivo, ao mesmo tempo em que reforça a prestação de contas e a transparência perante a sociedade, fortalecendo a governança e a eficiência do setor público (Sampaio *et al.*, 2022).

Costa *et al.* (2023) apresentaram um conjunto de 12 trilhas de auditoria para identificar licitações públicas suspeitas de fraude, das quais sete são referentes aos licitantes, uma avalia os sócios e quatro relacionam-se ao tipo de vínculo. Essas trilhas foram modeladas como uma rede social, onde os nós representam empresas licitantes ou seus sócios, e as arestas representam vínculos entre eles, permitindo identificar características suspeitas. Os resultados mostram que a metodologia é eficaz para filtrar dados, reduzir o volume de informações a serem analisadas por especialistas e priorizar a análise das licitações. As descobertas fornecem subsídios para a criação de algoritmos de classificação de licitações como fraudulentas ou não, contribuindo no combate à corrupção.

Nesse sentido, existem diversas formas de promover o combate às fraudes em contratações públicas no Brasil. Sampaio *et al.* (2022) utilizaram a Lei de Newcomb-Benford para detectar fraudes em compras públicas no Brasil através da análise dos padrões de distribuição dos dígitos iniciais dos valores envolvidos nas transações. Ao utilizar essa lei, é possível identificar anomalias nos dados que podem indicar possíveis irregularidades, como manipulação de valores, duplicação de números ou arredondamentos não naturais. A detecção dessas discrepâncias pode sinalizar a presença de fraudes e auxiliar na investigação de práticas indevidas nas compras públicas, contribuindo para a promoção da transparência e integridade nos processos governamentais.

Outra iniciativa no combate à ocorrência de fraudes é a utilização de análise automatizada de artefatos textuais em compras públicas, como é o caso do *software* Alice, criado pela CGU. A implementação do robô Alice no processo de compras públicas resultou em diversos benefícios significativos. Entre os resultados obtidos, destacam-se a identificação tempestiva de indícios de irregularidades, fraudes, desvios e desperdícios de recursos públicos, a realização automatizada de auditoria preventiva para uma análise mais eficiente, e a contribuição para o aprimoramento da gestão de recursos públicos por meio da análise automatizada de grandes volumes de dados. Adicionalmente, a correção de falhas antes de sua ocorrência, a revogação de pregões irregulares, a redução do tempo necessário para auditorias e a promoção da transparência e do combate à corrupção demonstram o impacto positivo do robô Alice, fortalecendo a gestão pública no Brasil (Panis *et al.*, 2022).

Para combater fraudes de maneira eficaz, os resultados mostram que a prevenção e a mitigação de riscos em processos são essenciais. Isso envolve a implementação de controles

internos preventivos e de detecção, como capacitação, infraestrutura física, recursos materiais e tecnológicos, sistemas, estrutura organizacional e de governança, manuais, normatização, corregedoria e penalizações efetivas (Souza; Silva, 2021). Este trabalho terá foco nas soluções que utilizam a aprendizagem de máquinas para detecção prévia de fraudes em licitações.

2.3 Dados públicos e modelos de aprendizagem de máquinas para detecção prévia de fraudes

Os modelos de detecção de fraudes podem ser usados para prever condutas ilegais na aquisição pública, através da análise de grandes conjuntos de dados de contratos de aquisição pública, identificando padrões e variáveis associadas a condutas ilegais, como tamanho e duração dos contratos, atrasos na implementação e tendências específicas do setor. Com esses modelos, torna-se possível prever quais contratos têm maior probabilidade de resultar em investigações de corrupção, violação de contrato ou ineficiências na implementação. Essas previsões podem ser usadas como pontuações de risco por agências anticorrupção, organizações de monitoramento e partes contratantes para orientar a seleção de operações de aquisição que precisam ser monitoradas mais de perto ou auditadas.

Neste contexto, as técnicas de inteligência artificial (IA) são fundamentais para os órgãos de fiscalização. A inteligência artificial é uma tecnologia avançada que simula diversas capacidades cognitivas humanas, tais como raciocínio lógico, aprendizado adaptativo e tomada de decisões estratégicas (Xu *et al.*, 2021). Ela é desenvolvida por meio de algoritmos sofisticados e modelos computacionais que permitem que os sistemas executem tarefas complexas que normalmente exigiriam intervenção humana. Ao integrar esses processos cognitivos, a IA capacita computadores e outros dispositivos a analisar grandes volumes de dados, aprender com experiências passadas e tomar decisões autônomas em uma variedade de contextos, desde a identificação de padrões em dados financeiros até a detecção de fraudes em contratos de aquisições públicas.

Entre as principais técnicas de inteligência artificial destacam-se o aprendizado de máquinas (do inglês *machine learning* - ML), redes neurais e processamento de linguagem natural (do inglês *Natural Language Processing* - NLP). O aprendizado de máquina permite que o sistema aprenda a partir de dados históricos e identifique padrões que possam indicar a ocorrência de fraudes (Paes; Selmini, 2021). Por exemplo, a análise de rede é uma técnica que permite a identificação de conexões entre diferentes entidades envolvidas em aquisições públicas, como empresas e indivíduos, para detectar possíveis esquemas de corrupção. As redes neurais, por sua vez, permitem a criação de modelos preditivos complexos que podem

identificar padrões e anomalias em grandes conjuntos de dados (Nai; Sulis; Meo, 2022). O processamento de linguagem natural, outra técnica, permite a análise de grandes volumes de dados não estruturados, como documentos e e-mails, para identificar possíveis sinais de fraude. Os métodos automatizados extraem informações de fontes de aquisições públicas por meio da coleta de dados, NLP e análise de dados. Esses métodos buscam dados em registros de contratos, licitações e outras transações governamentais, e utilizam técnicas de NLP para analisar documentos não estruturados, como descrições de contratos e comunicações relacionadas a aquisições públicas. Em seguida, aplicam técnicas de análise de dados para identificar padrões, tendências e anomalias nos dados coletados, o que pode indicar possíveis casos de fraude (Lima *et al.*, 2020; Nai; Sulis; Meo, 2022).

Posteriormente, os modelos preditivos são desenvolvidos por meio de algoritmos de aprendizado de máquina, os quais utilizam os dados coletados para identificar padrões e prever possíveis casos de fraude com base em características específicas. Ao combinar essas técnicas, os métodos automatizados podem extrair informações relevantes das fontes de aquisições públicas e desenvolver modelos preditivos eficazes para detecção de fraudes (Nai; Sulis; Meo, 2022).

A detecção de fraudes é um desafio constante para organizações públicas e privadas, e a utilização de técnicas de inteligência artificial pode ajudar a identificar possíveis casos de corrupção de forma mais rápida e eficiente. É importante ressaltar que essas técnicas não substituem a necessidade de auditorias e investigações humanas, mas podem ser uma ferramenta valiosa para complementar esses processos e aumentar a eficácia na detecção de fraudes em aquisições públicas (Nai; Sulis; Meo, 2022).

Random Forest é uma técnica de aprendizado de máquina que combina múltiplas árvores de decisão durante o processo de treinamento para obter um modelo mais robusto e preciso (Paes; Selmini, 2021). O modelo *Random Forest* é de fácil implementação, amplamente utilizado em outros contextos de forma robusta e eficiente, com capacidade de lidar com grandes volumes de dados não lineares e muitas variáveis, sem a necessidade de redução de dimensionalidade extensiva. Destaca-se pela sua robustez a *overfitting* e alta capacidade de generalização. Considerando que os registros de licitação nem sempre são completos ou consistentemente detalhados, o modelo lida relativamente bem com dados incompletos ou faltantes sem a necessidade de imputação complexa. Possui, ainda, o diferencial de fornecer medidas quantitativas da importância de cada variável para a tomada de decisão. Essas vantagens tornam o modelo *Random Forest* ideal para sua aplicação na previsão de fraudes nas licitações.

2.4 Estudos Relacionados

Nai *et al.* (2023) aplicaram técnicas de aprendizado de máquina e sistemas de recomendação para melhorar a eficiência e transparência dos processos de aquisição pública, bem como para prever possíveis reclamações nos tribunais administrativos. Os resultados do estudo mostraram que é possível utilizar técnicas de aprendizado de máquina em conjuntos de dados jurídicos reais relacionados aos processos de aquisição na Itália para prever reclamações nos tribunais administrativos com base nas características dos processos de aquisição. Os autores desenvolveram um modelo de aprendizado de máquina capaz de reconhecer anomalias em processos de aquisição e identificar as características mais relevantes para a classificação dos processos com ou sem reclamações. O modelo apresentou uma acurácia de 80% na previsão de reclamações, o que demonstra a eficácia da abordagem proposta.

Além disso, os autores desenvolveram um sistema de recomendação para retornar processos de aquisição semelhantes e encontrar empresas para licitantes, com base nos requisitos do processo de aquisição. O sistema de recomendação foi desenvolvido utilizando técnicas de aprendizado de máquina e redes neurais profundas, e apresentou resultados promissores na identificação de processos de aquisição semelhantes e empresas para licitantes (Nai *et al.*, 2023).

Os resultados do estudo destacam a aplicabilidade e o potencial das técnicas de aprendizado de máquina e sistemas de recomendação na melhoria dos processos de aquisição e na previsão de possíveis reclamações nos tribunais administrativos. Essas técnicas podem contribuir para a eficiência e transparência no setor público, permitindo que as autoridades identifiquem possíveis anomalias nos processos de aquisição e tomem medidas preventivas para evitar reclamações futuras. Além disso, o estudo demonstra a importância de explorar conjuntos de dados jurídicos abertos para melhorar os sistemas, procedimentos e serviços do setor público (Nai *et al.*, 2023).

Oliveira *et al.* (2023) identificaram indícios de fraudes por meio da análise das interações entre os participantes das licitações, como empresas licitantes e seus sócios, modelando essas interações como uma rede social. Essa metodologia foi exitosa em reduzir um grande volume de dados em um indicador, que possibilitou ranquear as contratações onde o risco de fraude era maior. Dessa forma, o trabalho de auditoria dos especialistas do Ministério Público de Minas Gerais pode ser mais bem orientado para análise daqueles casos onde o risco de fraude seja maior.

Gallego *et al.* (2021) identificaram a capacidade demonstrada pelos modelos de aprendizado de máquina em prever condutas ilegais em contratos de aquisição pública,

fornecendo pontuações de risco que podem ser utilizadas para priorizar contratos que exigem monitoramento mais rigoroso. Além disso, o estudo identificou variáveis-chave, como tamanho e duração dos contratos, atrasos na implementação e padrões setoriais, que estão significativamente associadas à corrupção e ineficiência.

Com isso, os resultados desse estudo oferecem sugestões/propostas interessantes para a formulação de políticas direcionadas e reformas regulatórias para mitigar o risco de corrupção na aquisição pública. O trabalho também destaca a importância da transparência e responsabilidade no processo de aquisição, propondo o uso de plataformas baseadas na *web* para registrar e relatar transações públicas. Ademais, a combinação de métodos de aprendizado de máquina com técnicas tradicionais de inferência causal é apresentada como uma abordagem promissora para fortalecer as políticas de combate à corrupção e aprimorar a integridade dos processos de aquisição pública.

Decarolis e Giorgiantonio (2020) enfatizam a importância de melhorar a coleta de dados e a transparência em contratações públicas como meios eficazes de combate à corrupção. Eles recomendam a adaptação dos indicadores de sinais vermelhos de acordo com as características específicas de cada região e setor, dada a variabilidade nos padrões de corrupção observados. Além disso, destacam o potencial do uso de ferramentas analíticas avançadas, como o aprendizado de máquina, na administração pública para detectar e prevenir a corrupção de forma mais eficaz.

Utilizando modelos de aprendizado de máquina, eles demonstraram que certos indicadores, como a unicidade do licitante e a falta de competição, são eficazes na previsão de corrupção. Observaram também variações significativas desses sinais entre diferentes regiões e tipos de contratos na Itália, sugerindo a necessidade de abordagens regionais específicas na detecção de corrupção (Decarolis; Giorgiantonio, 2020).

Decarolis e Giorgiantonio (2020) indicam que para melhorar os processos de aquisição e reduzir a corrupção em licitações públicas devem ser identificados os indicadores de risco, deve haver o aprimoramento da competição entre empresas e a implementação de práticas anticorrupção. Essas medidas podem ajudar a fortalecer a integridade e a transparência nas licitações públicas, contribuindo para a redução da corrupção. Além disso, foi identificada uma forte correlação entre certos sinais vermelhos e o aumento do risco de corrupção em contratações públicas.

Henrique, Sobreiro e Kimura (2020) compararam as principais técnicas de aprendizado de máquina, incluindo *Random Forest*, para classificar contratos públicos brasileiros com relação ao risco de não conformidade. Eles também analisaram regressão logística, K-vizinhos

mais próximos (*K-Nearest Neighbours*, KNN), análise discriminante e máquina de vetores de suporte (*Support Vector Machine*, SVM). Empresas que receberam quaisquer sanções graves foram classificadas como de alto risco. As variáveis independentes utilizadas foram: dados de aquisições, quantidades de itens, lances em negociações eletrônicas, valor médio aprovado no processo de licitação, número de órgãos com os quais contratos foram concluídos, número de entidades federais com as quais órgãos contrataram com a empresa, número de itens vencedores no processo de licitação, desclassificações/irregularidades sofridas pelas empresas, número médio de empregados e seus salários médios, renda recebida pelos sócios, número de atividades comerciais registradas junto à SRF, idade da empresa, impostos pagos às autoridades, tamanho da empresa (MESB e NPO), soma dos valores negociados em contratos na condição favorecida, número de sócios e empregados que também eram servidores públicos ou com algum tipo de função remunerada, e quaisquer doações para campanhas políticas. Os dados foram obtidos das bases Comprasnet e Sistema de Cadastramento Unificado de Fornecedores (SICAF), Relação Anual de Informações Sociais (RAIS) e no Tribunal Superior Eleitoral (TSE). O modelo utilizou 500 árvores de decisão.

De forma similar, Sá, Pessanha e Alves (2024) também compararam os principais algoritmos de classificação, incluindo *Random Forest*, juntamente com regressão logística, KNN, redes neurais e SVM, para identificar a propensão de fraude nos processos de contratação do Governo do Estado do Rio de Janeiro. A variável dependente utilizada também foi sanção sofrida pelo fornecedor. As variáveis independentes foram: CNPJ do fornecedor, tamanho empresa (Microempresa ou Empresa de Pequeno Porte), quantidade e valor dos contratos registrados para o fornecedor, número de licitações em que o fornecedor participou, total de licitações que o fornecedor venceu, quantidade de compras diretas registradas para o fornecedor, valor de compras diretas registradas para o fornecedor, capital social, total de CNAEs, distância entre a sede do fornecedor e o local da realização do serviço, anos de criação da empresa, quantidade de empresas em que o fornecedor possui sócios em comum, situação na Receita (ativa ou inapta/suspensa), quantidade de contratos firmados com o fornecedor cujos números iniciais falharam no teste de Benford, média das idades dos sócios da empresa. Os dados foram extraídos do Portal da Transparência do Estado do Rio de Janeiro e do Portal da Transparência da União. O modelo foi configurado limitando-se a profundidade máxima da árvore para 8.

Ambos os trabalhos de Henrique, Sobreiro e Kimura (2020) e Sá, Pessanha e Alves (2024) obtiveram bons resultados para o *Random Forest*, mas não calcularam as principais métricas de qualidade do modelo de classificação, como precisão, *recall* e *F1 Score*. De

maneira similar, os trabalhos acadêmicos de Faria (2023), Lopes (2019) e Silva (2022) também compararam modelos de classificação, incluindo o *Random Forest*, na detecção de fraudes em licitações na administração pública federal brasileira. Os resultados de Faria (2023) e Silva (2022) foram conflitantes com os resultados de Lopes (2019). O modelo de *Random Forest* teve a melhor performance nos estudos de Faria (2023) e Silva (2022), enquanto obteve a pior performance nos estudos de Lopes (2019).

Reconhecendo o potencial da técnica para os órgãos de fiscalização brasileiros, este trabalho busca avaliar a eficácia do algoritmo *Random Forest* na detecção de fraudes em licitações usando os dados do Portal Transparência da Controladoria-Geral da União.

3 METODOLOGIA

3.1 Coleta de dados

Para conduzir esta pesquisa, os dados foram obtidos por meio do Portal Transparência, vinculado à Controladoria-Geral da União, durante o período de 01/01/2022 a 31/12/2022. Foi necessário acessar quatro planilhas correspondentes a cada mês, disponíveis na seção de Licitações e Contratos, para a obtenção das informações, conforme preconizado por Gallego, Rivero e Martínez (2021). Os dados coletados e o local no site podem ser visualizados no Quadro 3.

Quadro 3 – Dados coletados Portal Transparência da Controladoria-Geral da União

VARIÁVEL	DESCRIÇÃO	PLANILHA	ABA
Valor Licitação	Valor monetário	Licitação	Licitações
Nome UG	Comprador		
Objeto	Processo de aquisição		
Modalidade Compra	Processo de aquisição		
Data Resultado Compra	Data de compra		
UF	Estado		
Número Licitação	Valor numérico da licitação		
Código UG	Código do comprador		
Nome Vencedor	Vencedor da licitação	Item Licitação	
Número Licitação	Valor numérico da licitação		
Código UG	Código do comprador		
Descrição Item Compra	Item negociado	Participantes Licitação	
Número Licitação	Valor numérico da licitação		
Código UG	Código do comprador		
Data Assinatura Contrato	Data	Compras	Contratos
Data Início Vigência	Data		
Data Fim Vigência	Data		
Data Publicação DOU	Data		
Data Assinatura Contrato	Data		
Número Licitação	Valor numérico da licitação		
Código UG	Código do comprador		

Fonte: Elaboração própria.

Após a obtenção dos dados mencionados, procedeu-se à combinação das variáveis "Número de Licitação" e "Código UG", resultando em uma nova variável denominada "Número Código", para a consolidação dos dados em um único quadro. Tal procedimento se fez essencial devido à identidade do número de licitação entre diversas entidades. Assim, ao

agrupar o número da licitação com o código da entidade, obtivemos uma numeração única para cada processo.

Dessa forma, foram obtidas 16 variáveis independentes empregadas na pesquisa, as quais estão relacionadas no Quadro 4. Na literatura, observa-se uma ampla diversidade nas variáveis utilizadas para estudar fraudes em licitações. Contudo, a maior parte dos estudos não foca explicitamente em explicar os motivos pelos quais as empresas se tornam desqualificadas. Além disso, a seleção de variáveis independentes frequentemente não se baseia em uma lógica estrita de causalidade. Em vez disso, muitas pesquisas optam por abordagens exploratórias ou descritivas, que visam identificar padrões de comportamento sem necessariamente estabelecer relações causais diretas entre as variáveis, como no trabalho de Henrique, Sobreiro e Kimura (2020).

Nesse sentido, o trabalho de Gallego, Rivero e Martínez (2021) se destaca ao propor um conjunto de variáveis independentes básicas, comum a vários outros estudos (ver referências Quadro 4), detalhando uma interpretação para o modelo e para a importância das variáveis. Esse conjunto inclui as variáveis (valor, entidades, empresas, localização, data, tipo de construção) apontadas por peritos criminais como informações que geralmente levam a uma boa estimativa sobre o risco de aquisição de fraude (LIMA *et al.*, 2020). Assim, este trabalho utilizou principalmente o conjunto de variáveis preditoras proposto por Gallego, Rivero e Martínez (2021), aplicadas no contexto brasileiro. Os padrões de risco em potencial associados às variáveis são detalhados a seguir:

- **Orçamento:** Comparação entre o orçamento planejado e o valor final do contrato pode indicar sobrepreços. Valores muito acima do esperado podem sugerir conluio ou corrupção.
- **Tipo/Nível da entidade:** Entidades governamentais de diferentes níveis ou tipos podem ter diferentes padrões de risco de fraude. O nível de escrutínio público e regulatório pode variar significativamente entre diferentes níveis governamentais. Por exemplo, entidades locais podem ter menos rigor em seus processos de licitação em comparação com entidades federais.
- **Objetivo da licitação**: O propósito específico pode estar associado a certos riscos de fraude. Grandes projetos de infraestrutura, por exemplo, podem ser mais vulneráveis a fraudes devido ao seu grande valor financeiro. Ou, nos últimos anos vários projetos do setor de transportes estavam associados a escândalos de corrupção.

- **Tipo de processo:** Diferentes processos (pregão, concorrência pública, convite, etc.) têm diferentes níveis de transparência e risco associado. Processos menos transparentes podem ser mais suscetíveis a fraudes.
- **Tipo de fornecedor:** Fornecedores com histórico de irregularidades ou aqueles recém-criados podem aumentar o risco de fraude.
- **Dia/Mês/Ano que o contrato foi assinado:** Datas próximas ao fim de mandatos políticos ou datas específicas podem ser usadas para apressar contratações sem o devido processo legal.
- **Período de execução:** Projetos com períodos de execução anormalmente longos ou curtos podem indicar má gestão ou esforços para desviar fundos.
- **Período de espera:** Indicador de transparência e publicidade sinalizando a diligência e pontualidade das autoridades ao publicar informações sobre contratos. Um período de espera muito curto entre o anúncio e a decisão pode indicar falta de concorrência adequada.
- **Intervalo entre a data de assinatura e a data de concessão:** Intervalos muito curtos podem sugerir que o fornecedor já estava predestinado a ganhar, independentemente das propostas concorrentes.
- **Intervalo entre a data de início da execução e a data de assinatura:** Intervalos muito curtos podem indicar preparação insuficiente, enquanto intervalos longos podem ser tentativas de adiar a execução sem justificativas válidas.
- **Distância para a eleição mais próxima:** Contratos assinados muito próximos a eleições podem estar tentando influenciar resultados eleitorais ou aproveitar a mudança de administração para passar despercebidos. Particularmente no ano 2022, ocorreram eleições no Brasil para Presidente da República, governadores dos estados e do Distrito Federal, senadores, deputados federais, estaduais e distritais. O pleito foi dividido em dois turnos nos dias 2 e 30 de outubro de 2022. Essa eleição se destacou historicamente pela intensa polarização política, marcada pela disputa entre os candidatos Jair Bolsonaro e Luiz Inácio Lula da Silva.
- **Categoria do produto:** Certas categorias de produtos podem ser mais propensas a fraudes, especialmente se envolverem grandes somas de dinheiro ou se forem de natureza mais técnica e difícil de avaliar.
- **Estado:** Variações regionais em termos de governança e transparência podem afetar o risco de fraude. Estados com histórico de corrupção podem ter mais incidências de

fraudes em licitações. Além disso, a distância entre a sede registrada do fornecedor e o local de realização do serviço podem indicar fraude.

Quadro 4 – Variáveis organizadas por tipo, seguindo o formato padrão da gestão pública

VARIÁVEIS INDEPENDENTES	Referências
Orçamento	Gallego, Rivero e Martínez (2021) Lima *et al.* (2020)
Tipo da entidade	Aldana, Falcón-Cortés e Larralde (2022) Gallego, Rivero e Martínez (2021) Lima *et al.* (2020)
Nível da entidade	
Objetivo da licitação	Gallego, Rivero e Martínez (2021)
Tipo de processo	Aldana, Falcón-Cortés e Larralde (2022) Gallego, Rivero e Martínez (2021)
Tipo de fornecedor	Aldana, Falcón-Cortés e Larralde (2022) Sá, Pessanha e Alves (2024)
Dia que o contrato foi assinado	Gallego, Rivero e Martínez (2021) Lima *et al.* (2020)
Mês que o contrato foi assinado	
Ano que o contrato foi assinado	
Período de execução	Aldana, Falcón-Cortés, Larralde (2022) Gallego, Rivero e Martínez (2021).
Período de espera	Gallego, Rivero e Martínez (2021)
Intervalo entre a data de assinatura e a data de concessão	Gallego, Rivero e Martínez (2021)
Intervalo entre a data de início da execução e a data de assinatura	Gallego, Rivero e Martínez (2021)
Distância para a eleição mais próxima	Aldana, Falcón-Cortés e Larralde (2022) Gallego, Rivero e Martínez (2021) Henrique, Sobreiro e Kimura (2020) Velasco *et al.* (2021)
Categoria do produto	Gallego, Rivero e Martínez, (2021) Lima *et al.* (2020)
Estado	Abreu, Pereira e Gomes-Jr (2024) Gallego, Rivero e Martínez (2021) Sá; Pessanha e Alves (2024)
VARIÁVEL DEPENDENTE	
Recebeu penalidade registrada na controladoria geral	Gallego, Rivero e Martínez (2021) Henrique, Sobreiro e Kimura (2020) Sá; Pessanha e Alves (2024)

As variáveis destacadas em negrito foram as utilizadas no modelo final.

Fonte: Elaboração própria.

A variável dependente foi definida como as instituições que receberam penalidade da Controladoria-Geral no ano de 2022. Sua escolha foi baseada nos trabalhos de Gallego, Rivero e Martínez (2021), Henrique, Sobreiro e Kimura (2020) e Sá, Pessanha e Alves (2024). A utilização do registro de penalidade da Controladoria-Geral como variável dependente para

indicar fraude em licitações pode não capturar todos os casos de fraude existentes. De fato, pode haver situações de fraude que não resultam na aplicação de penalidade devido a diversos fatores, como limitações nas investigações, falta de provas conclusivas ou até mesmo atrasos no processo de fiscalização. No entanto, a escolha dessa variável dependente foi baseada na necessidade de um indicador objetivo e verificável de comportamento fraudulento, conforme adotado em estudos anteriores na literatura. Além disso, a aplicação de penalidade é um forte indicativo de violações significativas e detectadas, o que a torna uma métrica confiável para os casos em que a fraude é efetivamente identificada e penalizada. Aprimoramentos posteriores do modelo deverão incorporar outras variáveis (como investigações em andamento e denúncias) que indiquem irregularidades ampliando assim o espectro de detecção de fraudes para além das penalidades já confirmadas.

Analisando a Tabela 1 é possível analisar os dados descritivos das variáveis preditoras a serem utilizadas neste estudo.

Tabela 1 – Dados descritivos das variáveis

VARIÁVEIS	OBSERVAÇÕES	VALORES EXCLUSIVOS	VALOR MAIS COMUM	FREQUÊNCIA
Valor licitação	109196	52069	0	6195
Nome UG	109196	2005	Universidade Federal do Rio Grande do Sul	1510
Objeto	109196	98010	Informação Protegida por sigilo nos termos da...	1012
Modalidade compra	109196	10	Dispensa de Licitação	59614
Data resultado compra	109196	345	15/12/2022	886
UF	109196	28	RJ	22505
Nome vencedor	950270	54509	J. J. VITALLI	7620
Data assinatura contrato	35613	1056	01/09/2022	486
Data início vigência	35613	1111	01/09/2022	676
Data fim vigência	35300	2162	31/12/2022	1194
Data publicação DOU	35613	270	26/12/2022	286
Descrição assinatura contrato	35613	1056	01/09/2022	486
Descrição item compra	4442683	10926	Manutenção/Reforma Predial	72103

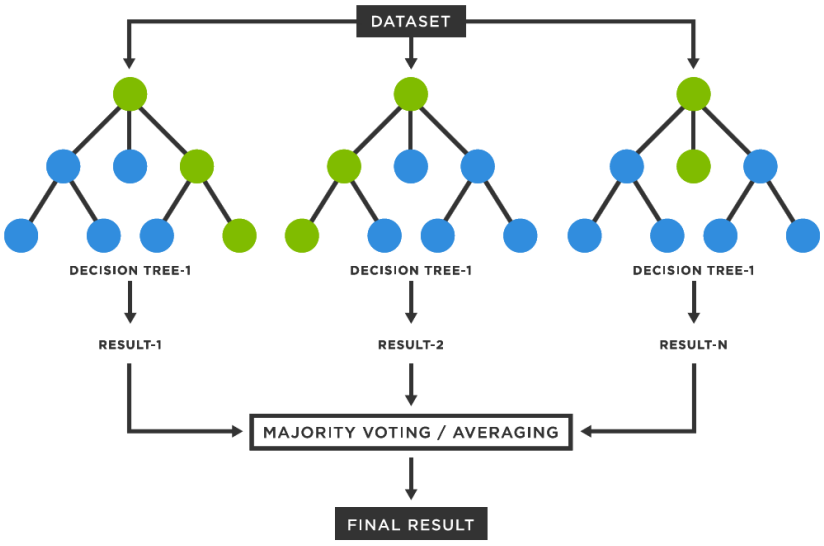
Fonte: Elaboração própria.

3.2 Random Forest

Observa-se que as técnicas de Inteligência Artificial têm demonstrado avanços significativos na detecção de fraudes em licitações (Decarolis; Giorgiantonio, 2020; Gallego *et al.*, 2021; Nai *et al.*, 2023; Oliveira *et al.*, 2023). Diante desse cenário, optou-se por empregar uma dessas técnicas para a previsão no presente estudo, especificamente o *Random Forest* (RF – ou Floresta Aleatória em português).

A *Random Forest* (Figura 1) é formada por múltiplas árvores de decisão, analogamente a uma floresta. O processo de criação das árvores de decisão começa com a geração aleatória de subconjuntos de dados a partir dos originais, chamados de amostra *bootstrap*. Em seguida, a aleatoriedade é novamente empregada para selecionar as variáveis preditoras (atributos), formando os nós de decisão. Os nós são definidos com base nas características que melhor separam os dados de acordo com o resultado desejado, estabelecendo pontos de corte nas variáveis independentes de forma a maximizar a diferenciação entre as categorias. Métricas como impureza de Gini, ganho de informação ou erro quadrático médio (MSE) podem ser utilizadas para avaliar a qualidade da divisão. A separação recursiva dos dados forma nós (majoritariamente) puros contendo dados de apenas uma classe, chamados nós folha. Esse procedimento é repetido para formar várias árvores de decisão, cada uma com estruturas de características distintas, resultando em modelos diferentes. Por fim, para obter uma predição final, o algoritmo percorre cada árvore e determina o resultado de cada variável, selecionando o resultado mais frequente através de um processo de "votação".

Figura 1 – Esquema do modelo *Random Forest*



Fonte: Gunay (2023).

O desenvolvedor do modelo, Breiman (2001) destaca as vantagens da técnica da *Random Forest*, salientando sua precisão, capacidade de lidar com *outliers* e ruídos, e facilidade de implementação. Ele ressalta que a técnica fornece estimativas adequadas para variáveis internas, como erro, força, correlação e importância das variáveis, enquanto evita problemas de *overfitting* (falha na generalização), garantindo confiabilidade na previsão de resultados futuros.

3.3 Algoritmo

O modelo *Random Forest* foi implementado em Python 3, utilizando a classe *RandomForestClassifier* do módulo *sklearn.ensemble* da biblioteca *Scikit-Learn* (Scikit-Learn, 2024). O código implementado, contido no Apêndice e publicado na plataforma Code Ocean (https://doi.org/10.24433/CO.1983915.v1), segue os passos do algoritmo descrito a seguir detalhando-se as principais classes, métodos e parâmetros utilizados (Morais *et al.*, 2024).

1. **Importação de Bibliotecas**: Importou-se Pandas para manipulação de dados, *Matplotlib* e *Seaborn* para plotagem, *Imbalanced-learn* para tratar classes desequilibradas, e as funções relevantes do *Scikit-Learn* para modelagem e avaliação.
2. **Carregamento de Dados**: Os dados foram carregados a partir de um arquivo CSV.
3. **Preparação dos Dados**: Separou-se as variáveis independentes (x) e dependentes (y - Multas).
4. **Balanceamento dos Dados:** após a verificação da distribuição das classes, os dados foram balanceados com *RandomUnderSampler,* do submódulo *imblearn.under_sampling* da biblioteca *imbalanced-learn*, para equilibrar as classes sub-representadas e sobre-representadas, reduzindo o número de exemplos na classe majoritária.
5. **Divisão dos Dados**: Os dados foram divididos em conjunto de treinamento (*x_train, y_train* - 70%) e teste (*x_test* e *y_test* - 30%), seguindo a metodologia de Henrique, Sobreiro e Kimura (2020) e Sá, Pessanha e Alves (2024).
6. **Criação e Treinamento do Modelo:** A classe *RandomForestClassifier* instanciada foi configurada com os parâmetros *n_estimators*=100 e *random_state*=42. *n_estimators* configura o número de árvores na floresta. Normalmente, quanto maior o número de árvores, melhor é a performance do modelo, mas também aumenta o custo computacional. Henrique, Sobreiro e Kimura (2020) utilizaram 500 árvores, enquanto Sá, Pessanha e Alves (2024) não especificaram o número de árvores empregadas. Se utilizou o valor padrão, *n_estimators* = 100 a partir da versão do *Scikit-Learn* 0.22.

(lançada em Dezembro de 2019). Já o *random_state* controla a aleatoriedade do *bootstrap* das amostras e as características a serem consideradas ao construir cada árvore. Sá, Pessanha e Alves (2024) utilizou *random_state* = 50. O treinamento do modelo *Random Forest* foi implementado por meio do método *fit* a partir do conjunto de treinamento (*x_train, y_train).*

7. **Avaliação do Modelo**: Avaliou-se o modelo usando o conjunto de teste para prever e comparar as previsões com os valores reais. A predição do conjunto de teste foi feita pelo método *predict* que realiza a classificação em uma matriz de teste (*x_test*) retorna o vetor de rótulos previstos (*y_pred*). A precisão do modelo foi avaliada por meio da matriz de confusão (método *confusion_matrix*) e do relatório de classificação (método *classification_report*). Também, foi feito a validação cruzada estratificada (*StratifiedKFold*) para verificar a eficácia do modelo em termos de sua capacidade de generalizar para novos dados, que não foram usados durante o treinamento do modelo. A estratificação significa que a classe garante que cada conjunto de dados de treino e teste contém aproximadamente a mesma porcentagem de amostras de cada classe alvo, que é crucial para manter um conjunto de dados balanceado em cada parte (*fold*), especialmente quando os dados são desbalanceados. O conjunto de dados foi divido em 5 partes (*n_splits* = 5).
8. **Importância das Características**: Analisou-se quais características são mais importantes para a previsão de fraudes. Para a análise da importância das variáveis, utilizou-se o método *feature_importances_*, a qual retorna valores em porcentagem para classificar quais foram os principais indicadores.
9. **Curva ROC:** Para plotar a curva ROC, calculou-se as probabilidades de classe (*y_scores*) para a amostra de teste por meio do método *predict_proba.* A curva de ROC foi gerada a partir das dos rótulos reais (*y_test*) e das probabilidades preditas, pelo método *roc_curve.*

As análises foram feitas inicialmente por mês, devido a restrições computacionais (processador com 2 núcleos e RAM 4GB) para processar o volume de dados. Posteriormente, com um computador mais potente (processador com 12 núcleos e RAM 32GB) foi possível expandir a análise para o intervalo anual, permitindo uma avaliação mais abrangente.

3.4 Análises dos resultados

A performance do modelo de classificação implementado foi avaliada analisando as matrizes de confusão, os erros tipo I e II, acurácia, precisão, sensibilidade (*recall*), pontuação F1 (*F1 Score*) e curva ROC. Essas são as principais métricas de avaliação amplamente utilizadas e descritas em Buathong, Sieng-Ek e Jarupunphol (2023), Butaru *et al.* (2016), Dalianis (2018), Henrique, Sobreiro e Kimura (2020), Hossin e Sulaiman (2015), Sá, Pessanha e Alves (2024), e Yulianto, Sukarno e Suwastika (2019). Os resultados foram comparados com os estudos mencionados no referencial teórico.

O Quadro 5 mostra a matriz de confusão para um problema de classificação binário com a Classe 0 para 'Irregularidade não detectada' e Classe 1 para 'Irregularidade detectada'. A matriz apresenta a contagem de eventos e previsões, de forma que é possível observar a qualidade do modelo preditivo. Permite, ainda, visualizar de forma clara os diferentes tipos de erros (tipo I e tipo II) e os resultados corretos do modelo. Isso facilita a análise da precisão do modelo em prever fraudes como também identificar processos que cumprem os pressupostos da legalidade, ou seja, que não se verifica evidências de fraudes.

Quadro 5 – Matriz de Confusão, a qual apresentará os resultados das previsões e a contagem de acertos e erros que são usados para avaliar a qualidade preditiva

Real versus previsto	**Previsto como Fraude**	**Previsto como Não Fraude**
Fraude ocorreu	Verdadeiro Positivo (TP)	Falso Negativo (FN) *Erro Tipo II*
Fraude Não ocorreu	Falso Positivo (FP) *Erro Tipo I*	Verdadeiro Negativo (TN)

Fonte: Elaboração própria.

De maneira mais específica, o erro tipo I (dado pela Equação 1) ocorre quando a fraude não foi detectada no processo licitatório, mas é apontada erroneamente pelo modelo de previsão. Errar neste caso é considerado menos oneroso, uma vez que se assemelha a um sinal de alerta para verificação mais cuidadosa e, portanto, parte integrante (e quase natural) das atividades dos órgãos fiscalizadores. Já o erro do tipo II (Equação 2) ocorre quando a fraude é constada, mas não é prevista pelo modelo.

$$Erro\ do\ Tipo\ I = \frac{FP}{FP + TN} \tag{1}$$

$$Erro\ do\ Tipo\ II = \frac{FN}{FN + TP} \tag{2}$$

A acurácia (Equação 3) representa a média geral de acertos do modelo. No entanto, ela é uma medida insuficiente, especialmente para classes desbalanceadas. Uma análise mais aprofundada da eficácia do modelo foi feita analisando também precisão e sensibilidade.

$$Acurácia = \frac{TN + TP}{TP + FP + TN + FN} \tag{3}$$

Precisão e sensibilidades (Equações 4) são métricas complementares. A escolha de qual métrica enfatizar depende do contexto específico da aplicação e do equilíbrio desejado entre detectar positivos e evitar falsos positivos. Na detecção de fraudes em licitações, a sensibilidade (*recall*) é frequentemente considerada mais crítica. Falsos negativos, ou seja, fraudes que não são detectadas, estão associadas a perdas financeiras, danos a reputação e favorece as redes de corrupção. Priorizar a sensibilidade ajuda a garantir que a maioria das fraudes seja detectada, minimizando a chance de fraudes passarem despercebidas. Embora menos crítica, a precisão ainda é importante, especialmente em contextos onde o volume de transações é alto. Altas taxas de falsos positivos podem levar a uma sobrecarga operacional, aumentando os custos com investigações desnecessárias e potencialmente prejudicando a experiência das empresas/fornecedores, caso transações legítimas sejam frequentemente bloqueadas ou atrasadas. Aumentar a precisão reduz o número de falsos positivos, o que pode ajudar a manter a eficiência operacional.

Assim, a Pontuação F1 é particularmente valiosa porque calcula a média harmônica entre precisão e sensibilidade (*recall*), permitindo a identificação do modelo que oferece o melhor custo-benefício. A pontuação F1 varia de 0 a 1, onde 1 é o melhor valor possível e 0 é o pior. Um valor de 1 indica perfeição na classificação, com precisão e sensibilidade perfeitas, o que significa que todos os positivos verdadeiros e nenhum falso positivo foram identificados. Já um valor de 0 indica que o modelo falhou completamente em identificar positivos verdadeiros, ou que todos os resultados positivos são falsos positivos.

$$Precisão = \frac{TP}{TP + FP} \qquad Sensibilidade = \frac{TP}{TP + FN} \tag{4}$$

$$F1\_Score = 2.\frac{Precisão \times Sensibilidade}{Precisão + Sensibilidade} \tag{5}$$

Além disso, a técnica de *Random Forest* possui uma propriedade interessante que permite examinar quais as variáveis preditoras foram mais importantes para realizar as previsões. Assim, esta foi também uma parte da análise que pode ser consideravelmente útil

para que seus usuários tenham mais atenção em determinados sinalizadores e, assim aumentar sua eficiência e aprimorar o processo de fiscalização nos processos de licitação.

A curva *Receiver Operating Characteristic* (ROC) permite avaliar graficamente o desempenho do classificador, relacionando sensibilidade e especificidade. A eficiência do classificador é observada na medida em que a curva se aproxima ao classificador ideal: taxa de falso positivo igual a zero e a taxa de verdadeiro positivo igual a 1, correspondente ao ponto superior esquerdo do gráfico, coordenadas (0,1) e a área sob a curva ROC (AUC, do inglês *Area Under Curve*) idealmente próxima de 1. Assim, o classificador é efetivo se AUC >0,5 e a curva ROC estiver acima da linha tracejada diagonal, a qual representa um classificador aleatório.

4 RESULTADOS E DISCUSSÕES

Modelos de classificação permitem prever a ocorrência de fraudes, baseados em padrões pré-identificados em um conjunto de dados rotulado, e apontar as principais variáveis indicadoras de irregularidades, direcionando os esforços de investigação. Dentre essas técnicas, as características do modelo *Random Forest* são particularmente adequadas para esta aplicação, devido à sua robustez, precisão, capacidade de manejar a complexidade dos dados (grande volume, alta dimensionalidade e valores ausentes), além de sua simplicidade de implementação. Destaca-se também por fornecer uma medida de importância das variáveis. Vários estudos comparativos entre os principais algoritmos comprovaram sua eficiência no contexto de detecção de fraudes em licitações brasileiras (Henrique; Sobreiro; Kimura, 2020; Sá; Pessanha; Alves, 2024) e outros ainda indicaram que era o melhor modelo de classificação (Faria, 2023; Silva, 2022).

O modelo *Random Forest* foi implementado utilizando dados do Portal Transparência da Controladoria-Geral da União, do período de 01/01/2022 a 31/12/2022. Foi utilizado um conjunto de 11 variáveis independentes: orçamento, objetivo da licitação, tipo de processo, dia que o contrato foi assinado, período de execução, período de espera, intervalo entre a data de assinatura e a data de concessão, intervalo entre a data de início da execução e a data de assinatura, distância para a eleição mais próxima, categoria do produto, estado. Para melhoria do desempenho computacional foram excluídas as variáveis independentes iniciais (Quadro 4) de menor relevância para o modelo: tipo da entidade, nível da entidade, tipo de fornecedor, mês que o contrato foi assinado e ano que o contrato foi assinado. A variável dependente foi definida como multas recebidas pelas empresas. Os dados foram divididos em 70% para treinamento e 30% para teste. O classificador foi implementado com 100 árvores.

Os resultados do modelo *Random Forest* para o conjunto de variáveis selecionadas, considerando o ano de 2022, constam nas Tabelas 2 a 5. Os dados foram analisados mensalmente e anualmente através da matriz de confusão (Tabelas 2 e 3) e das medidas de acurácia, precisão, *recall* e *F1 Score* (Tabelas 4 e 5).

As matrizes de confusão obtidas por meio do modelo encontram-se nas Tabela 2 e Tabela 3. Observa-se que para a análise anual, o erro do tipo I (quando o modelo identifica um falso positivo) e o erro do tipo II (quando o modelo deixa de identificar um negativo) não afetam significativa o modelo, considerando que a acurácia foi de 0,78 e o *F1 Score* foi maior que 0,74. No que se relaciona às matrizes mensais, a média geral da acurácia e do *F1 Score* foram superiores a 0,7. Ao pormenorizar a análise para cada mês, temos situações como o caso

do mês de julho. Neste caso, as métricas indicam que o modelo tem uma boa sensibilidade (*recall*), identificando corretamente a maioria dos casos positivos, mas a precisão é menor (0,66), sugerindo que há uma quantidade significativa de falsos positivos. A acurácia geral é razoável, mas pode haver espaço para melhorias, especialmente em relação à precisão.

Tabela 2 – Matriz de confusão mensal (amostra de 10%)

Janeiro	Predito positivo	Predito negativo
Real positivo	590	54
Real negativo	0	595

Fevereiro	Predito positivo	Predito negativo
Real positivo	2927	2702
Real negativo	548	4964

Março	Predito positivo	Predito negativo
Real positivo	4603	2849
Real negativo	698	6752

Abril	Predito positivo	Predito negativo
Real positivo	2202	458
Real negativo	17	2658

Maio	Predito positivo	Predito negativo
Real positivo	17155	6785
Real negativo	4462	19609

Junho	Predito positivo	Predito negativo
Real positivo	4854	1124
Real negativo	76	5927

Julho	Predito positivo	Predito negativo
Real positivo	1694	1665
Real negativo	119	3221

Agosto	Predito positivo	Predito negativo
Real positivo	535	149
Real negativo	17	618

Setembro	Predito positivo	Predito negativo
Real positivo	1851	1404
Real negativo	665	2565

Outubro	Predito positivo	Predito negativo
Real positivo	14357	7035
Real negativo	4391	16933

Novembro	Predito positivo	Predito negativo
Real positivo	28604	9667
Real negativo	20546	17922

Dezembro	Predito positivo	Predito negativo
Real positivo	4457	849
Real negativo	523	4742

Fonte: Elaboração Própria.

Tabela 3 – Matriz de confusão anual (amostra de 10%)

2022	Predito positivo	Predito negativo
Real positivo	114501	60487
Real negativo	18011	156264

Fonte: Elaboração Própria.

A Tabela 5 mostra os resultados do modelo para uma amostra de 10% dos dados, considerando as informações de todo o período, ou seja, o ano de 2022. A acurácia média foi de 0,78, ligeiramente superior à média mensal (ver Tabela 4). Já o *F1 Score* do modelo

apresentou valores de 0,74 para irregularidade não detectada e de 0,8 para irregularidade detectada.

As informações consignadas na Tabela 4, foram geradas com amostra de 10% dos dados. A acurácia média foi de 0,76, sendo que o maior valor foi de 0,96 em Janeiro e o menor valor foi de 0,61 para o mês de Novembro. Dessa forma, pode-se afirmar que o modelo acertou as previsões realizadas em 76% das vezes.

Os valores de acurácia foram semelhantes aos obtidos por Henrique, Sobreiro e Kimura (2020) igual 75,58% e inferiores aos obtidos por Sá, Pessanha e Alves (2024) igual a 95,83%. Vale ressaltar que, embora utilizamos apenas 11 variáveis independentes, um menor número em comparação com as 17 variáveis usadas por Henrique, Sobreiro e Kimura (2020), os resultados obtidos para o *Random Forest* foram praticamente equivalentes. Por outro lado, Sá, Pessanha e Alves (2024) definiram a profundidade máxima da árvore, parâmetro *max_depth* = 8. Quando este parâmetro não é definido, os nós são expandidos até que todas as folhas sejam puras ou até que todas as folhas contenham menos amostras do que o necessário para serem divididas. A definição da profundidade da árvore pode explicar o resultado excepcional obtido no trabalho de Sá, Pessanha e Alves (2024). Quando a profundidade das árvores é menor, o modelo se torna mais generalista e menos propenso a memorizar padrões específicos do conjunto de treinamento (*overfitting*). Portanto, trabalhos futuros deverão explorar a influência do parâmetro *max_depth* no desempenho do modelo *Random Forest* implementado.

O menor valor obtido para *F1 Score* foi de 0,54 para irregularidade detectada no mês de Novembro. Registra-se que este foi o único valor menor que 0,6 para este coeficiente. Em novembro, nota-se uma queda significativa na precisão para casos não irregulares e um aumento no *recall* para a mesma classe, indicando que o modelo passou a identificar mais eficazmente os casos legítimos como tais, mas com um aumento correspondente em falsos positivos. Para casos de irregularidades, houve uma redução considerável no *recall* (de 1,00 para 0,47), o que significa que o modelo deixou de identificar muitos casos reais de irregularidades. De fato, nota-se uma mudança no padrão do modelo de setembro a novembro em relação aos meses precedentes.

Tabela 4 – Resultados do modelo *Random Forest* por mês (amostra de 10%)

Mês	Situação	Precision	*Recall*	F1-Score	Support
Janeiro	Irregularidade não detectada (0)	1	0,92	0,96	644
	Irregularidade detectada (1)	0,92	1	0,96	595
	Acurácia			0,96	1239
Fevereiro	Irregularidade não detectada (0)	0,84	0,52	0,64	5629
	Irregularidade detectada (1)	0,65	0,9	0,75	5512
	Acurácia			0,71	11141
Março	Irregularidade não detectada (0)	0,87	0,62	0,72	7452
	Irregularidade detectada (1)	0,7	0,91	0,79	7450
	Acurácia			0,76	14902
Abril	Irregularidade não detectada (0)	0,99	0,83	0,9	2660
	Irregularidade detectada (1)	0,85	0,99	0,92	2675
	Acurácia			0,91	5335
Maio	Irregularidade não detectada (0)	0,79	0,72	0,75	23940
	Irregularidade detectada (1)	0,74	0,81	0,78	24071
	Acurácia			0,77	48011
Junho	Irregularidade não detectada (0)	0,98	0,81	0,89	5978
	Irregularidade detectada (1)	0,84	0,99	0,91	6003
	Acurácia			0,9	11981
Julho	Irregularidade não detectada (0)	0,93	0,5	0,66	3359
	Irregularidade detectada (1)	0,66	0,96	0,78	3340
	Acurácia			0,73	6699
Agosto	Irregularidade não detectada (0)	0,97	0,78	0,87	684
	Irregularidade detectada (1)	0,81	0,97	0,88	635
	Acurácia			0,87	1319
Setembro	Irregularidade não detectada (0)	0,74	0,57	0,64	3255
	Irregularidade detectada (1)	0,65	0,79	0,71	3230
	Acurácia			0,68	6485
Outubro	Irregularidade não detectada (0)	0,77	0,67	0,72	21392
	Irregularidade detectada (1)	0,71	0,79	0,75	21324
	Acurácia			0,73	42716
Novembro	Irregularidade não detectada (0)	0,58	0,75	0,65	38271
	Irregularidade detectada (1)	0,65	0,47	0,54	38468
	Acurácia			0,61	76739
Dezembro	Irregularidade não detectada (0)	0,89	0,84	0,87	5306
	Irregularidade detectada (1)	0,85	0,9	0,87	5265
	Acurácia			0,87	10571
Média	Irregularidade não detectada (0)	0,88	0,735	0,735	94630
	Irregularidade detectada (1)	0,725	0,905	0,785	94497
	Acurácia			0,765	189127

Fonte: Elaboração própria.

Tabela 5 – Resultados do modelo *Random Forest* anual (amostra de 10%)

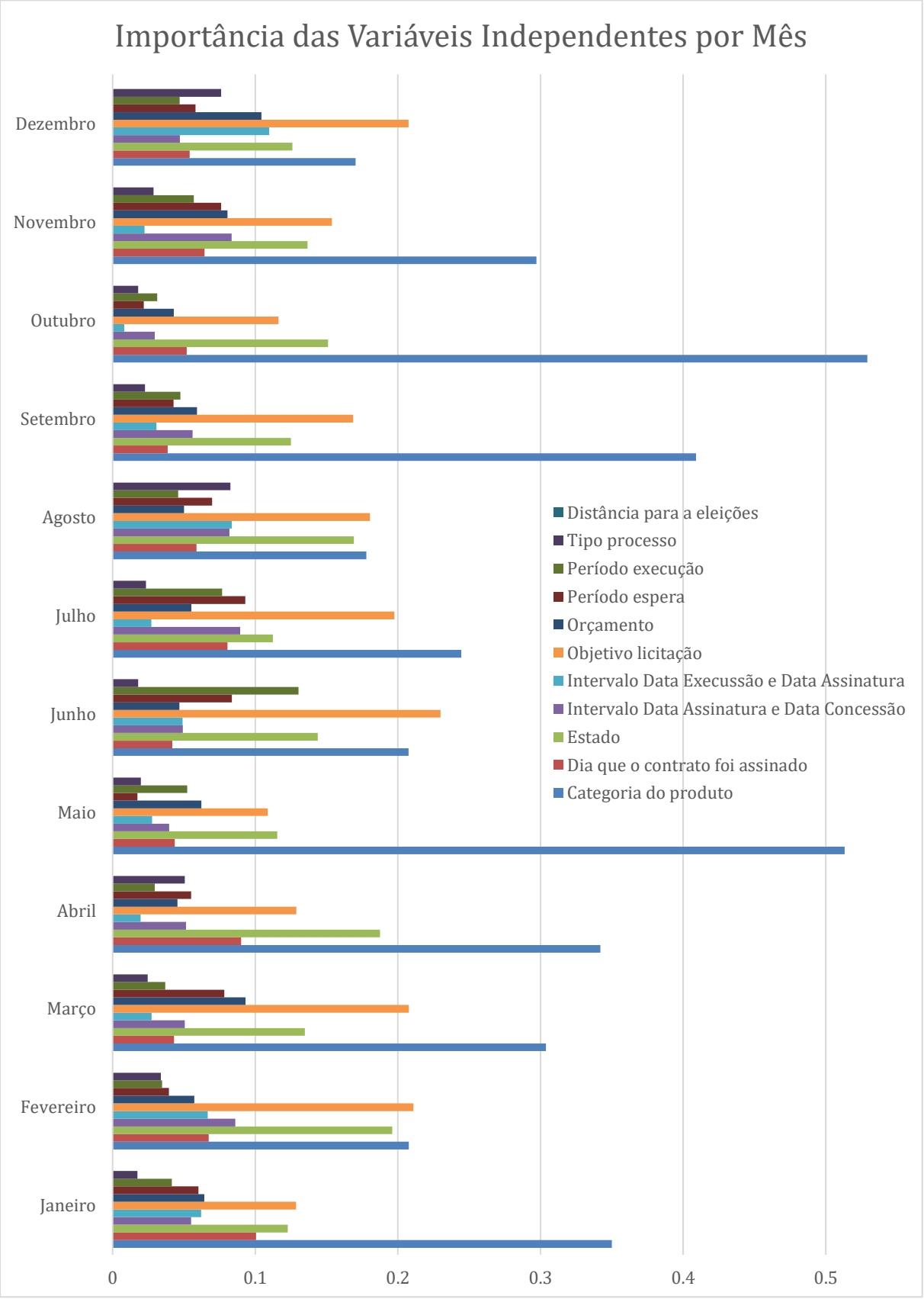
Ano	Situação	Precision	*Recall*	F1-Score	Support
2022	Irregularidade não detectada (0)	0,86	0,65	0,74	174988
	Irregularidade detectada (1)	0,72	0,9	0,8	174275
	Acurácia			0,78	349263

Fonte: Elaboração própria.

O último trimestre do ano é tradicionalmente um período de intensa atividade em licitações públicas no Brasil. Muitas agências buscam utilizar os orçamentos restantes para garantir que os fundos não sejam desperdiçados ou reduzidos no próximo ciclo fiscal. No ano de 2022 isso pode ainda ter sido intensificado pelas eleições presidenciais, marcadas pela forte polarização política. O resultado das eleições no final de outubro e a mudança de governo pode ter influenciado uma mudança no padrão dos dados de licitação, refletindo-se em uma piora do modelo de previsão observada especialmente no mês de novembro. Outra possível explicação que pode ter influenciado na baixa performance do modelo neste mês, foi o volume de dados utilizado consideravelmente superior correspondente a 32% dos dados totais, o que pode ter causado um subajuste do modelo.

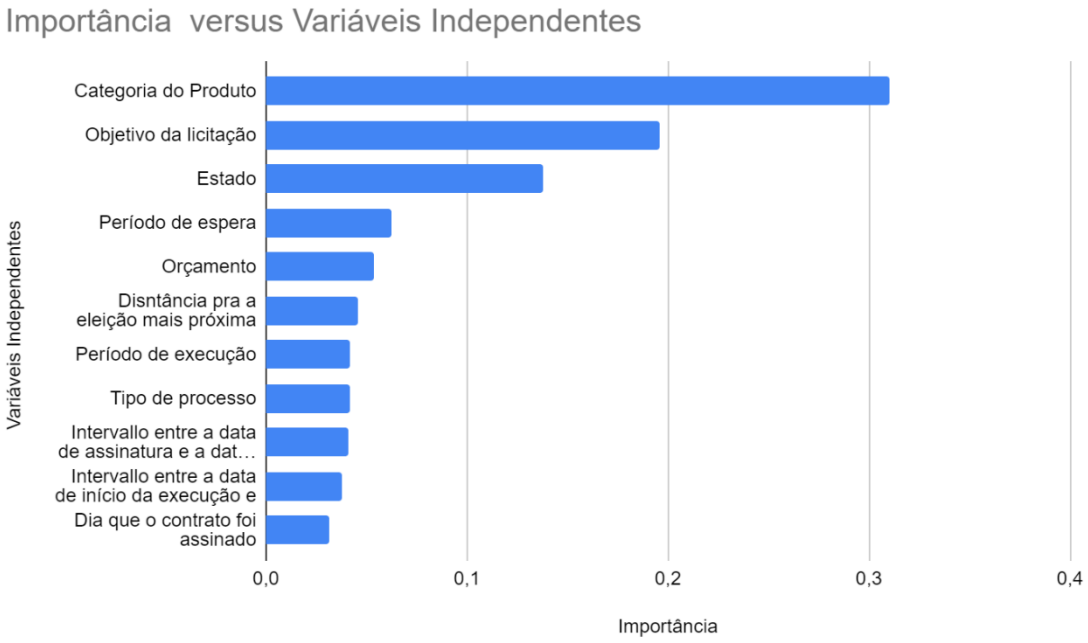
Outro resultado obtido do modelo *Random Forest* é uma lista com a importância das variáveis independentes (Figura 2 e Figura 3). Para a análise anual, Categoria do Produto, Objetivo da Licitação e Estado são as variáveis mais contribuem para a precisão do modelo, sendo que somadas, respondem por 64,40%. Excluindo-se as três variáveis com maior contribuição para o modelo, observa-se um certo balanceamento entre a importância das demais variáveis, sendo que apresentou o melhor resultado obteve 0,06 e a que obteve o pior resultado foi 0,03. Deste segundo grupo as variáveis mais importantes são Orçamento (0,06), o Período de Espera (0,05) e a Distância para a eleição mais próxima (0,05).

Figura 2 – Importância das variáveis independentes por mês (amostra de 10%)



Fonte: Elaboração própria.

Figura 3 – Importância das variáveis independentes anual (amostra de 10%)



Fonte: Elaboração própria.

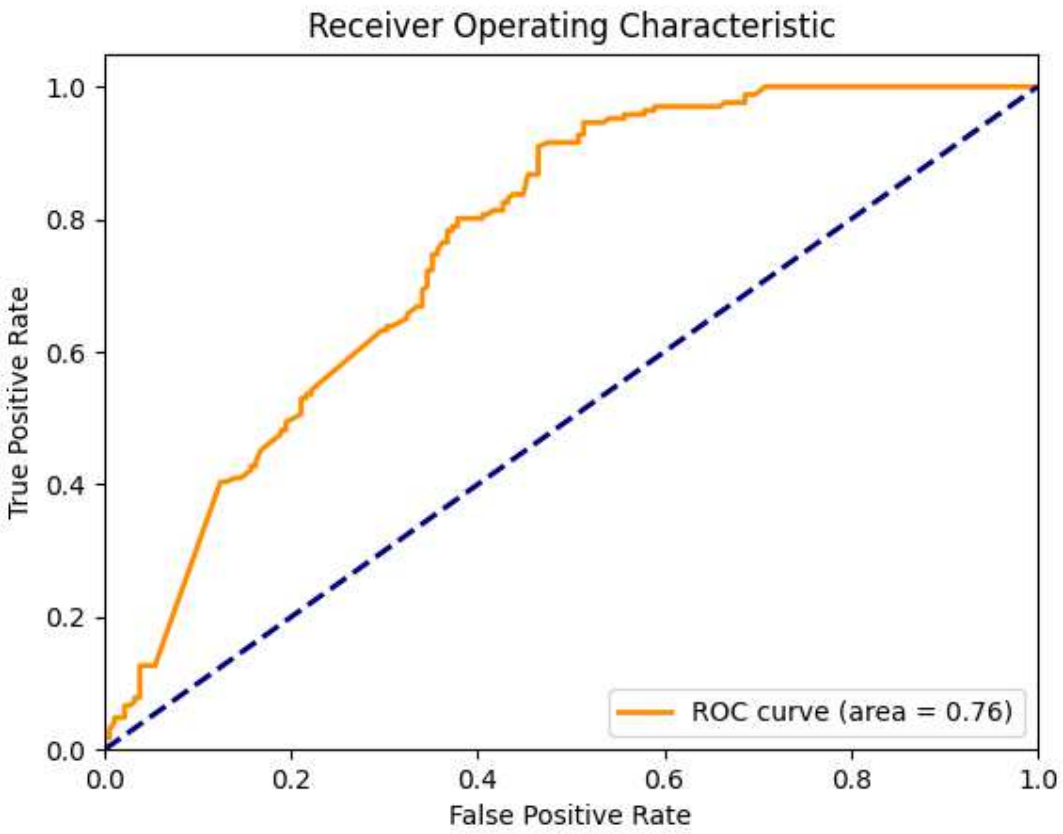
Embora esse estudo tenha utilizado variáveis do trabalho de Gallego, Rivero e Martínez (2021), eles utilizaram modelos de classificação diferentes, comparando o lasso e GBM, com dados da Colômbia. Eles obtiveram como variáveis independentes chaves as características típicas dos contratos como seu tamanho (medido pelo orçamento estimado e duração), atrasos na implementação, tempo antes da próxima eleição, e padrões geográficos e específicos do setor. Embora no geral a importância das variáveis tenha sido similar entre os dois estudos, em Gallego, Rivero e Martínez (2021) o Orçamento foi a variável mais relevante, enquanto esse estudo obteve a Categoria de Produto como mais significativa. Isso pode indicar que contextos geográficos diferentes com fatores econômicos, políticos e culturais específicos podem afetar a relevância das variáveis. Ou a diferença pode estar simplesmente relacionada as diferenças nos modelos utilizados, sendo algo que deve ser aprofundado na escolha do método, sugerindo que determinados modelos de classificação podem ser melhores para detectar certos contextos.

Sá, Pessanha e Alves (2024) utilizando *Random Forest* com dados do Rio de Janeiro, obteve resultados semelhantes ao de Gallego, Rivero e Martínez (2021), onde a principal variável foi o ‘Valor dos contratos registrados para o fornecedor do período’, seguido por ‘Quantidade de contratos registrados para o fornecedor do período’ e capital social.

A análise mensal (Figura 2) mostrou haver variações na importância das variáveis independentes ao longo do ano. Por exemplo, a variável Orçamento que veio em quarto em importância na análise anual, com exceção dos meses de Agosto e Dezembro, foi menos importante em todos os outros meses. A variável 'Distância para a eleição mais próxima' não exibiu relevância em nenhum mês estudado. Isso se deve ao fato de que, dentro de cada mês, a distância temporal até a próxima eleição é constante para todos os dados analisados, eliminando variações que poderiam evidenciar sua importância no modelo.

A área abaixo da curva (AUC - do inglês *area under the curve*) alcançada para uma amostra de teste com 0,01% dos dados anuais foi de 76%, ligeiramente superior ao obtido por Henrique, Sobreiro e Kimura (2020) (74,92%). A curva ROC obtida é mostrada na Figura 4.

Figura 4 – Curva ROC para amostra de 0,01% dos dados anuais



Fonte: Elaboração própria.

5 CONCLUSÃO

A informatização dos processos de licitação trouxe maior transparência e eficiência, reduzindo custos operacionais e acelerando o processo de contratação. Além disso, promoveu uma concorrência mais ampla. No entanto, esse avanço também impõe aos órgãos de auditoria a necessidade de adotar novas ferramentas para a detecção de fraudes, adaptando-se as inovações tecnológicas no setor e ao volume massivo de dados envolvido. Assim, existe uma tendência crescente de aplicação de técnicas de inteligência artificial e aprendizado de máquinas na fiscalização dos sistemas eletrônicos de contratações públicas.

O classificador *Random Forest* implementado teve um desempenho eficiente com *F1 Score* mensal médio de 78,5% e anual de 80%. Destacou-se pela capacidade de detectar a maioria dos casos reais de fraude, com *recall* mensal médio e anual de 90%. As variáveis de maior importância para o modelo foram Categoria do Produto, Objetivo da Licitação e Estado para ambas as análises, mensal e anual. No entanto, a importância dessas variáveis flutuou significativamente ao longo dos meses.

Observou-se uma mudança de padrão significativa entre os meses, que deve ser investigada mais profundamente. Para casos de irregularidades, houve uma redução considerável nos parâmetros de qualidade com valores máximo em Janeiro e mínimo em Novembro (variação no *recall* de 1,00 para 0,47, respectivamente, e variação no *F1 Score* de 0,96 para 0,54, respectivamente). Apesar que a variável ‘Distância para a eleição mais próxima’ ser somente a sexta em nível de importância, acredita-se que as eleições presidenciais de 2022 tenham influenciado significativamente na mudança de padrão.

A pesquisa foi limitada por restrições computacionais. O grande volume de dados aumentou significativamente o tempo de processamento, exigindo a divisão dos dados por mês para realizar a análise. O processamento anual demandou um computador com pelo menos 32GB de memória RAM e, mesmo assim, levou aproximadamente 1 dia para concluir o processamento dos dados. Ademais, somente a variável dependente de multas recebidas pelas empresas não captura todos os casos de fraudes existentes. Investigações em andamento ou denúncias são possíveis variáveis dependentes a serem incorporadas no modelo em trabalhos futuros para ampliar o espectro de detecção de fraudes. A importância de outras variáveis independentes também deverá ser avaliada. Por fim, análises futuras deverão incluir outros anos para comparação de desempenho e refinamento do modelo.

O modelo *Random Forest* desenvolvido tem potencial de aplicação prática além da academia, oferecendo aos gestores públicos uma ferramenta dinâmica eficaz na detecção de

possíveis desvios de conduta no processo de licitação. Mais ainda, a tendência crescente de informatização dos processos de licitação torna indispensável a utilização de modelos de aprendizagem de máquina, como o implementado, para a detecção de fraudes. A ferramenta será fundamental para direcionar a tomada de decisões dos auditores. A alta capacidade de detectar a maioria dos casos reais de fraude permite aos auditores priorizem as investigações de empresas ou licitações com alta probabilidade de irregularidades, reduzindo o tempo gasto em casos que seriam falsos positivos. Ademais, as variáveis mais importantes identificadas pelo modelo, como Categoria do Produto, Objetivo da Licitação, e Estado, fornecem aos auditores uma visão clara sobre os indicadores mais relevantes que apontem irregularidades. Com isso, eles podem direcionar seus esforços para monitorar essas variáveis de forma mais detalhada. A importância flutuante das variáveis ao longo do ano também sinaliza que os auditores precisam planejar revisões mais intensas e ajustar o foco de sua investigação com base em padrões sazonais ou eventos externos (como as eleições). Por exemplo, em meses onde certas categorias de produtos se tornam mais relevantes, o auditor pode aumentar a fiscalização dessas licitações. A identificação de padrões de fraudes ao longo do tempo também pode direcionar o desenvolvimento de sistemas de alerta. Ao observar que fraudes aumentam em certos períodos ou regiões, os auditores podem programar verificações automáticas mais rigorosas para esses casos, criando uma estratégia de fiscalização proativa. Por fim, o *Random Forest* é um modelo dinâmico e flexível que permite a atualização constante das variáveis mais importantes, garantindo que o auditor se adapte rapidamente a novos padrões de fraude.

Essa identificação mais proativa, eficiente e precisa, leva a uma redução significativa dos custos de auditoria. Para implementação do sistema é necessária uma infraestrutura computacional com no mínimo 32GB de RAM. No entanto, o investimento inicial é consideravelmente menor em comparação com os recursos humanos que seriam necessários para processar manualmente um volume tão grande de dados.

5.1 Artigo Publicado

Previsão de Fraude em Licitações no Brasil

Cadernos de Finanças Públicas, v. 24, n. 03.

https://publicacoes.tesouro.gov.br/index.php/cadernos/article/view/256

Autores: Vinícius Souza Morais (UFU)

Daniel Vitor Tartari Garruti (UFU)

Flávio Luiz de Moraes Barboza (UFU/FAGEN)

Pablo Rogers (UFU/FAGEN)

REFERÊNCIAS

ABREU, B. M.; PEREIRA, T. H. S.; GOMES-JR, L. Detecção de Fraudes em Licitações Públicas: Uma Comparação de Modelos de Detecção de Anomalias. In: ESCOLA REGIONAL DE BANCO DE DADOS (ERBD), 2024, Farroupilha/RS. Anais [...]. Porto Alegre: **Sociedade Brasileira de Computação,** p. 81-90, 2024. https://doi.org/10.5753/erbd.2024.238821

ADOBOR, H.; YAWSON, R. The promise of artificial intelligence in combating public corruption in the emerging economies: A conceptual framework. **Science and Public Policy**, v. 50, n. 3, p. 355-370, 2023. https://doi.org/10.1093/scipol/scac068

ALDANA, A.; FALCÓN-CORTÉS, A.; LARRALDE, H. A machine learning model to identify corruption in México's public procurement contracts. **ArXiv**, v. abs/2211.01478, 2022. https://doi.org/10.48550/arXiv.2211.01478

AMORIM, V. A. J. **Licitações e contratos administrativos: teoria e jurisprudência.** Brasília: Senado Federal, 2020. Disponível em: https://bibliotecadigital.stf.jus.br/xmlui/bitstream/handle/123456789/2461/1182434.pdf?sequence=1&isAllowed=y. Acesso em: 24 jul. 2024. ISBN : 9788570188670.

BIAGINI, E. Ferramenta de IA já evitou gastos de R$ 11 bi em licitações suspeitas, diz ministro da CGU. **Agência Gov**, 2024. Disponível em: https://agenciagov.ebc.com.br/noticias/ferramenta-de-ia-da-cgu-ja-economizou-r-11-bilhoes-em-gastos-publicos-em-licitacoes-suspeitas-diz-ministro. Acesso em: 02 ago. 2024.

BUATHONG, W.; SIENG-EK, P.; JARUPUNPHOL, P. Measuring the performance of machine learning forecasting models to support Bitcoin investment decisions. **Journal of Data Science and Intelligent Systems**, v. 2, n. 2, p. 100-112, 2023. https://doi.org/10.47852/bonviewJDSIS3202677

BUTARU, F.; CHEN, Q.; CLARK, B.; DAS, S.; LO, A. W.; SIDDIQUE, A. Risk and risk management in the credit card industry. **Journal of Banking & Finance**, v. 72, p. 218-239, 2016. https://doi.org/10.1016/j.jbankfin.2016.07.015

BRASIL. Constituição (1988). Constituição da República Federativa do Brasil de 1988. Promulgada em 05 de outubro de 1988. Brasília, DF: Senado Federal, Centro Gráfico, 1988.

BRASIL. Controladoria-Geral da União. Manual de Responsabilização de Entes Privados. Brasília, DF: CGU, abril de 2022. Disponível em: https://repositorio.cgu.gov.br/handle/1/68182. Acesso em: 21 jul. 2024.

BRASIL. Lei nº 8.666, de 21 de junho de 1993. Estabelece normas para licitações e contratos da Administração Pública e dá outras providências. Diário Oficial da União, Brasília, DF, 22 jun. 1993. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm. Acesso em: 21 jul. 2024.

BRASIL. Lei nº 9.784, de 29 de janeiro de 1999. Regula o processo administrativo no âmbito da Administração Pública Federal. Diário Oficial da União, Brasília, DF, 1 fev. 1999. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l9784.htm. Acesso em: 21 jul. 2024.

BRASIL. Lei nº 10.520/2002, 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, 17 jul. 2002. Disponível em: https://www.planalto.gov.br/ccivil_03/LEIS/2002/L10520.htm?origin=instituicao. Acesso em: 21 jul. 2024.

BRASIL. Lei nº 12.462, de 4 de agosto de 2011. Institui o Regime Diferenciado de Contratações Públicas – RDC; altera a Lei nº 10.683, de 28 de maio de 2003, que organiza a Presidência da República e os Ministérios, e dá outras providências. Diário Oficial da União, Brasília, DF, 5 ago. 2011. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12462.htm. Acesso em: 21 jul. 2024.

BRASIL. Lei nº 12.846, de 1º de agosto de 2013. Dispõe sobre a responsabilização administrativa e civil de pessoas jurídicas pela prática de atos contra a administração pública, nacional ou estrangeira, e dá outras providências. Diário Oficial da União, Brasília, DF, 2 ago. 2013. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2013/lei/l12846.htm. Acesso em: 21 jul. 2024.

BRASIL. Lei nº 13.303, de 30 de junho de 2016. Dispõe sobre o estatuto jurídico da empresa pública, da sociedade de economia mista e de suas subsidiárias, no âmbito da União, dos Estados, do Distrito Federal e dos Municípios. Diário Oficial da União, Brasília, DF, 1 jul. 2016. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/lei/l13303.htm. Acesso em: 21 jul. 2024.

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: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm. Acesso em: 21 jul. 2024.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, p. 5-32, 2001. https://doi.org/10.1023/A:1010933404324

CARDOSO, R. F. Um estudo sobre os resultados da utilização da Bolsa Eletrônica de Compras no governo do estado de São Paulo. **Revista do Serviço Público** (RSP), v. 55, n. 4, 2004. Disponível em: https://repositorio.enap.gov.br/jspui/bitstream/1/1542/1/2004%20Vol.55%2cn.4%20Cardoso.pdf. Acesso em: 21 jul. 2024.

CHEVITARESE ALVES, M. V.; DUFLOTH, S. C. Portais Eletrônicos de Compras da Administração Pública: Contribuição para Avaliação da Governança Eletrônica no Brasil. **Revista Gestão & Tecnologia**, v. 8, n. 1, p. 1–19, 2010. https://doi.org/10.20397/2177-6652/2008.v8i1.209

CONTADOR, J. C.; CARDOSO, R. F. Governo eletrônico do Estado de São Paulo e a avaliação da Bolsa Eletrônica de Compras. **Revista de Administração Fazendária**, São Paulo, v. 1, n.1, p. 85-115, 2005. Disponível em: http://iefe.sefaz.ma.gov.br/wp-content/uploads/2013/03/v01_n01_art03_New.pdf. Acesso em: 21 jul. 2024.

CONTROLADORIA-GERAL DA UNIÃO. Alice - Analisador de Licitações, Contratos e Editais. Disponível em: https://www.gov.br/cgu/pt-br/assuntos/auditoria-e-fiscalizacao/alice. Acesso em: 02 ago. 2024.

COSTA, C. C. M.; TERRA, A. C. P. **Compras públicas: para além da economicidade.** Brasília: Escola Nacional de Administração Pública (ENAP); Sociedade Brasileira de Administração Pública (SBAP), 2019. Disponível em: https://repositorio.enap.gov.br/jspui/bitstream/1/4277/1/1_Livro_Compras%20p%C3%BAblicas%20para%20al%C3%A9m%20da%20economicidade.pdf. Acesso em: 05 ago. 2024.

COSTA, L. L.; BACHA, C. A.; OLIVEIRA, G. P.; SILVA, M. O.; TEIXEIRA, M. C.; BRANDÃO, M. A.; LACERDA, A.; PAPPA, G. L. *et al.* Identificação de licitações suspeitas de fraude por meio de trilhas de auditagem. **iSys-Brazilian Journal of Information Systems**, v. 16, n. 1, p. 13–1, 2023. https://doi.org/10.5753/isys.2023.3013

COSTA, R. E.; HOLLNAGEL, H. C.; BUENO, R. L. P. Compras Governamentais: Panorama Atual e Desafios. **Revista Científica Hermes**, v. 23, p. 51-75, 2019. https://doi.org/10.21710/rch.v23i0.459

DALIANIS, H. Evaluation metrics and evaluation. In: **Clinical text mining**. Cham: Springer, 2018. https://doi.org/10.1007/978-3-319-78503-5_6

DECAROLIS, F.; GIORGIANTONIO, C. Corruption red flags in public procurement: new evidence from Italian calls for tenders. **EPJ Data Science**, v. 11, n. 1, p. 16, 2022. https://doi.org/10.1140/epjds/s13688-022-00325-x

DE PAULA, M. M.; GOBBES, M. S. Medidas de prevenção e controle de fraudes no pregão eletrônico. **Universitas: Revista FANORPI de Divulgação Científica**, v. 02, n. 10, p. 61-79, 2024. Disponível em: https://fanorpi.com.br/universitas/index.php/revista/article/view/280/269 . Acesso em: 21 jul. 2024.

FARIA, G. A. **Fraudes em compras governamentais: Detecção com Aprendizado de Máquina.** 2023. Trabalho de Conclusão de Curso (Graduação em Administração Pública) - Escola de Administração Pública, Centro de Ciências Jurídicas e Políticas, Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

FERNANDES, C. C. C. Compras Públicas no Brasil: vertentes de inovação, avanços e dificuldades no período recente. **Administração Pública E Gestão Social**, v. 4, n. 11, p. 1–19, 2019. https://doi.org/10.21118/apgs.v4i11.7262

GADOUR, M. M. E. Corruption in public procurement: Can e-procurement and artificial intelligence make a difference in Africa? **QScience Connect,** v. 2024, n. 1-Thesis, p. 2, 2024. https://doi.org/10.5339/connect.2024.spt.2

GALLEGO, J.; RIVERO, G.; MARTÍNEZ, J. Preventing rather than punishing: An early warning model of malfeasance in public procurement. **International Journal of Forecasting**, v. 37, n. 1, p. 360-377, 2021. http://dx.doi.org/10.1016/j.ijforecast.2020.06.006

GUNASEGARAN, M.; BASIRUDDIN, R.; RIZAL, A. M. Detecting and Preventing Fraud in e-procurement of Public Sector: A Review, Synthesis and Opportunities for Future Research**. International Journal of Academic Research in Business and Social Sciences**, v. 13, n.1, p. 1444 – 1463, 2023. https://doi.org/10.6007/IJARBSS/v13-i1/15970

HENRIQUE, B. M.; SOBREIRO, V. A.; KIMURA, H. Contracting in Brazilian public administration: A machine learning approach. **Expert Systems**, v. 37, n. 5, 2020. https://doi.org/10.1111/exsy.12550

HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. **International Journal of Data Mining & Knowledge Management Process**, v. 5, p. 01-11, 2015. https://doi.org/10.5121/ijdkp.2015.5201

LIMA, M. C.; SILVA, R. C.; MENDES, F. L. S.; CARVALHO, L. R.; ARAUJO, A.; VIDAL, F. B. Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach. In: COHN, T.; HE, Y.; LIU, Y. (Ed.). Findings of the Association for Computational Linguistics: EMNLP 2020. Online: **Association for Computational Linguistics**, p. 1580-1588, 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.143

LOPES, M. A. **Aplicação de aprendizado de máquina na detecção de fraudes públicas.** 2019. Dissertação (Mestrado em Administração) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2019. https://doi.org/10.11606/D.12.2020.tde-10022020-174317

MARIN, A. M.; SANTOS, I. C. dos; DAZZI, R.; FERNANDES, A. Aplicação de Aprendizado de Máquina para Classificação de Produtos no Processo de e-procurement. In: Anais do XIV Computer on the Beach - COTB'23. **Anais [...].** Itajaí, SC: Universidade do Vale do Itajaí, 2023. https://doi.org/10.14210/cotb.v14.p492-494

MAVIDIS, A.; FOLINAS, D. From Public E-Procurement 3.0 to E-Procurement 4.0; A Critical Literature Review. **Sustainability**, v. 14, n. 18, 2022. https://doi.org/10.3390/su141811252

MOHUNGOO, I.; BROWN, I.; KABANDA, S. A. Systematic Review of Implementation Challenges in Public E-Procurement. In: Responsible Design, Implementation and Use of Information and Communication Technology, 2020. **Anais [...].** Lecture Notes in Computer Science(), vol 12067. Cham: Springer, 2020. https://doi.org/10.1007/978-3-030-45002-1_5

MORAIS, V. S.; ROGERS, P.; GARRUTI, D.; BARBOZA, F. Previsão de fraude em licitações no Brasil. **Code Ocean**, 2024. https://doi.org/10.24433/CO.1983915.v1

NAI, R.; SULIS, E.; MEO, R. Public procurement fraud detection and artificial intelligence techniques: a literature review. In: Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management. **Anais [...].** CEUR-WS, 2022. https://iris.unito.it/bitstream/2318/1888876/1/KM4LAW22_preprint.pdf

OECD. **Public procurement.** 2024. Disponível em: https://www.oecd.org/en/topics/public-procurement.html. Acesso em: 05 ago. 2024.

OLIVEIRA, G. P.; MENDES, B. M. A.; BRAZ, C. S.; COSTA, L. L.; SILVA, M. O.; BRANDÃO, M. A.; LACERDA, A.; PAPPA, G. L. Ranqueamento de Licitações Públicas a partir de Alertas de Fraude. In: Anais do XII Brazilian Workshop on Social Network Analysis and Mining. **Anais [...].** SBC, 2023. p. 1-12. https://doi.org/10.5753/brasnam.2023.232105

PAES, B. H.; SELMINI, A. M. Machine Learning para detecção de transações financeiras fraudulentas. In: SEMINÁRIOS EM ADMINISTRAÇÃO (SEMEAD), 24., 2021, São Paulo. **Anais [...].** São Paulo: Faculdade de Economia, Administração, Contabilidade e Atuária da Universidade de São Paulo, 2021. Disponível em: https://login.semead.com.br/24semead/anais/resumo.php?cod_trabalho=203. Acesso em: 22 jul. 2024.

PANIS, A.; ISIDRO, A. da S. F.; CARNEIRO, D. K. de O.; MONTEZANO, L.; RESENDE JUNIOR, P. C.; SANO, H. Inovação em compras públicas: Atividades e resultados no caso do robô Alice da Controladoria-Geral da União. **Cadernos Gestão Pública e Cidadania, São Paulo,** v. 27, n. 86, p. 1–19, 2022. https://doi.org/10.12660/cgpc.v27n86.83111

PINTO, V. R. R. Um breve histórico sobre inovações em compras e licitações públicas no Brasil. **Brazilian Journal of Development**, v. 6, n. 8, p. 63378–63397, 2020. https://doi.org/10.34117/bjdv6n8-680

REJEB, A.; REJEB, K.; APPOLLONI, A.; KAYIKCI, Y.; IRANMANESH, M. The landscape of public procurement research: A bibliometric analysis and topic modelling based on Scopus. **Journal of Public Procurement**, v. 23, n. 2, 2023. https://doi.org/10.1108/JOPP-06-2022-0031

RIBEIRO, C. G.; INÁCIO, E. O mercado de compras governamentais brasileiro (2006-2017): Mensuração e análise. Brasília: Instituto de Pesquisa Econômica Aplicada (IPEA), 2019. Texto para Discussão, No. 2476. Disponível em: https://hdl.handle.net/10419/211431. Acesso em: 02 ago. 2024.

RODRIGUES, B. C.; REIS, P. R. da C. Partes interessadas internas e desempenho em contratações públicas na perspectiva das teorias dos stakeholders e dos custos de transação. **Cadernos Gestão Pública e Cidadania**, São Paulo, v. 28, p. e88342 , 2023. https://doi.org/10.34117/bjdv6n8-680

SÁ, T. A.; PESSANHA, J. F. M.; ALVES, F. J. S. Métodos de classificação supervisionada aplicados à identificação de fraudes de fornecedores. **CONTABILOMETRIA - Brazilian Journal of Quantitative Methods Applied to Accounting, Monte Carmelo**, v. 11, n. 2, p.125-146, 2024.

SAMPAIO, A. da H.; FIGUEIREDO, P. S.; LOIOLA, E. Compras públicas no Brasil: Indícios de fraudes usando a lei de Newcomb-Benford. **Cadernos Gestão Pública e Cidadania, São Paulo,** v. 27, n. 86, p. 1–20, 2022. https://doi.org/10.12660/cgpc.v27n86.82760

SCIKIT-LEARN. **RandomForestClassifier.** Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble. RandomForestClassifier. Acesso em: 02 ago. 2024.

SECOM TCU. **Chatbot do TCU fornece certidões pelo whatsapp.** 2020. Disponível em: https://portal.tcu.gov.br/imprensa/noticias/chatbot-do-tcu-fornece-certidoes-pelo-whatsapp.htm. Acesso em: 02 ago. 2024.

SICILIANI, L.; TACCARDI, V.; BASILE, P.; DI CIANO, M.; LOPS, P. AI-based decision support system for public procurement. **Information Systems,** v. 119, 2023. https://doi.org/10.1016/j.is.2023.102284

SILVA, G. **Prevenção de irregularidades em contratos públicos: uma análise utilizando técnicas de Machine Learning.** 2022. 91 f., il. Trabalho de Conclusão de Curso (Bacharelado em Administração) — Universidade de Brasília, Brasília, 2022.

SOUZA, K. R.; SILVA, **Combate a Corrupção em Licitações Públicas,** 2021. Disponível em: https://www.3rcapacita.com.br/curso/combate-a-corrupcao-em-licitacoes-publicas-65-horas. Acesso em: 16 set. 2020.

SUPREMO TRIBUNAL FEDERAL. **Projeto Victor avança em pesquisa e desenvolvimento para identificação dos temas de repercussão geral.** 2021. Disponível em: https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=471331&ori=1. Acesso em: 02 ago. 2024.

SUPREMO TRIBUNAL FEDERAL**. STF finaliza testes de nova ferramenta de Inteligência Artificial.** 2023a. Disponível em: https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=507120&ori=1. Acesso em: 02 ago. 2024.

SUPREMO TRIBUNAL FEDERAL. **STF realiza seminário sobre Inteligência Artificial nesta segunda-feira (17).** 2023b. Disponível em: https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=505698&ori=1. Acesso em: 02 ago. 2024.

TORRES-BERRU, Y.; LOPEZ-BATISTA, V. F.; ZHINGRE, L. C. A data mining approach to detecting bias and favoritism in public procurement. **Intelligent Automation & Soft Computing**, v. 36, n. 3, p. 3501-3516, 2023. https://doi.org/10.32604/iasc.2023.035367

TRIBUNAL DE CONTAS DA UNIÃO. **Fiscalização de tecnologia da informação – O que é Governo digital?** Disponível em: https://portal.tcu.gov.br/fiscalizacao-de-tecnologia-da-informacao/atuacao/governo-digital/. Acesso em: 21 jul. 2024.

TRIBUNAL DE CONTAS DA UNIÃO. **Guia de uso de inteligência artificial generativa no Tribunal de Contas da União (TCU).** Disponível em: https://portal.tcu.gov.br/data/files/42/F7/91/4B/B59019105E366F09E18818A8/Guia%20de%20uso%20de%20IA%20generativa%20no%20TCU.pdf. Acesso em: 02 ago. 2024.

TRIBUNAL DE CONTAS DA UNIÃO. **Licitações & Contratos: Orientações e Jurisprudência do TCU.** 5. ed. Brasília: TCU, Secretaria-Geral da Presidência, 2023.

U.S. BUREAU OF ECONOMIC ANALYSIS. Shares of gross domestic product: Government consumption expenditures and gross investment [A822RE1A156NBEA]. In: Federal Reserve Bank of St. Louis. Disponível em: https://fred.stlouisfed.org/series/A822RE1A156NBEA. Acesso em: 05 ago. 2024.

VELASCO, R. B.; CARPANESE, I.; INTERIAN, R.; PAULO NETO, O.C.G.; RIBEIRO, C.C. A decision support system for fraud detection in public procurement. **International Transactions Inoperational Research**, v. 28, p. 27-47, 2021. https://doi.org/10.1111/itor.12811

YAMAJI, D. M.; VIEIRA, S. F. A.; FERRER, M. F. Marketplaces: um estudo comparativo do portal de compras do Governo Federal com experiências internacionais. **Práticas De Administração Pública**, v. 6, n. 3, p. 67–87, 2024. https://doi.org/10.5902/2526629274756

YULIANTO, A.; SUKARNO, P.; SUWASTIKA, N. A. Improving AdaBoost-based Intrusion Detection System (IDS) performance on CIC IDS 2017 dataset. **Journal of Physics: Conference Series**, v. 1192, 2019. https://doi.org/10.5902/2526629274756

XU, Y. *et al.* Artificial intelligence: A powerful paradigm for scientific research. **The Innovation**, v. 2, n. 4, p. 100179, 2021. https://doi.org/10.1016/j.xinn.2021.100179

ZAGO, M. F. **Poder de compra estatal como instrumento de políticas públicas?** Brasília: Escola Nacional de Administração Pública, 2018.

APÊNDICE

Código fonte: Organizar dados Licitação

```
import pandas as pd
import dask.dataframe as dd

##Acesso ao Drive
from google.colab import drive
drive.mount('/content/drive')

### Lendo os dados
#-------------------------------------------------
#Configurando os caminhos das pastas onde estão os dados
periodo = '202201'
pastal = f'/content/drive/MyDrive/previsão
licitação/licitações/'+periodo+'_Licitacoes/'
pastac = f'/content/drive/MyDrive/previsão licitação/compras/'+periodo+'_Compras/'

# Lendo o nome dos arquivos
caminho_empenho202201 = '202201_EmpenhosRelacionados.csv'
caminho_item202201 = '202201_ItemLicitação.csv'
caminho_licitacao202201 = '202201_Licitação.csv'
caminho_participantes202201 = '202201_ParticipantesLicitação.csv'
caminho_compras202201 = '202201_Compras.csv

# Lendo os arquivos
empenho202201 = pd.read_csv(pastal + caminho_empenho202201, encoding='latin1',
delimiter=';',low_memory=False)
item202201 = pd.read_csv(pastal + caminho_item202201, encoding='latin1',
delimiter=';',low_memory=False)
licitacao202201 = pd.read_csv(pastal + caminho_licitacao202201, encoding='latin1',
delimiter=';',low_memory=False)
participantes202201 = pd.read_csv(pastal + caminho_participantes202201,
encoding='latin1', delimiter=';',low_memory=False)
compras202201 = pd.read_csv(pastac + caminho_compras202201, encoding='latin1',
delimiter=';',low_memory=False)

#-------------------------------------------------------------
# Repete seção anterior alterando período e nomes dos arquivos correspondente a
# cada mês
#-------------------------------------------------------------

### Concatenando os dados
# Concatenar ao longo das linhas (eixo 0)
```

```
empenho = pd.concat([empenho202201, empenho202202, empenho202203, empenho202204,
empenho202205,
                     empenho202206, empenho202207, empenho202208, empenho202209,
empenho202210,
                     empenho202211, empenho202212], axis=0)

item = pd.concat([item202201, item202202, item202203, item202204, item202205,
                     item202206, item202207, item202208, item202209, item202210,
                     item202211, item202212], axis=0)

licitacao = pd.concat([licitacao202201, licitacao202202, licitacao202203,
licitacao202204, licitacao202205,
                     licitacao202206, licitacao202207, licitacao202208,
licitacao202209, licitacao202210,
                     licitacao202211, licitacao202212], axis=0)

participantes = pd.concat([participantes202201, participantes202202,
participantes202203, participantes202204, participantes202205,
                     participantes202206, participantes202207, participantes202208,
participantes202209, participantes202210,
                     participantes202211, participantes202212], axis=0)

compras = pd.concat([compras202201, compras202202, compras202203, compras202204,
compras202205,
                     compras202206, compras202207, compras202208, compras202209,
compras202210,
                     compras202211, compras202212], axis=0)

### Criando uma nova coluna nos dataframes
# Criar uma nova coluna 'Número_código'
empenho['Número_código'] = empenho['Número Licitação'].astype(str) +
empenho['Código UG'].astype(str)

# Criar uma nova coluna 'Número_código'
item['Número_código'] = item['Número Licitação'].astype(str) + item['Código
UG'].astype(str)
# Criar uma nova coluna 'Número_código'
licitacao['Número_código'] = licitacao['Número Licitação'].astype(str) +
licitacao['Código UG'].astype(str)

# Criar uma nova coluna 'Número_código'
participantes['Número_código'] = participantes['Número Licitação'].astype(str) +
participantes['Código UG'].astype(str)
```

```
# Criar uma nova coluna 'Número_código'
compras['Número_código'] = compras['Número Licitação'].astype(str) +
compras['Código UG'].astype(str)

### Visualizando os dataframes
empenho.reset_index(inplace=True)
empenho
empenho.to_csv('/content/drive/MyDrive/previsão       licitação/dados_empenho.csv',
sep=';', encoding='utf-8')

item.reset_index(inplace=True)
item
item.to_csv('/content/drive/MyDrive/previsão licitação/dados_item.csv', sep=';',
encoding='utf-8')

licitacao.reset_index(inplace=True)
licitacao
licitacao.to_csv('/content/drive/MyDrive/previsão licitação/dados_licitacao.csv',
sep=';', encoding='utf-8')

participantes.reset_index(inplace=True)
participantes
participantes.to_csv('/content/drive/MyDrive/previsão
licitação/dados_participantes.csv', sep=';', encoding='utf-8')

compras.reset_index(inplace=True)
compras
compras.to_csv('/content/drive/MyDrive/previsão licitação/dados_compras.csv',
sep=';', encoding='utf-8')

### Pegando apenas as colunas que me interessam
# Lista de colunas desejadas
colunas_licitacao = ['Valor Licitação', 'Nome UG', 'Objeto', 'Modalidade Compra',
'Data Resultado Compra', 'UF', 'Número_código']

# Criar um novo DataFrame com apenas as colunas desejadas
licitacao2 = licitacao[colunas_licitacao]
licitacao2
licitacao2.to_csv('licitacao2.csv', sep=';', encoding='utf-8')

# Lista de colunas desejadas
colunas_item = ['Nome Vencedor', 'Número_código']

item2= item[colunas_item]
item2
```

```
item2.to_csv('item2.csv', sep=';', encoding='utf-8')

# Lista de colunas desejadas
colunas_compras = ['Data Assinatura Contrato', 'Data Início Vigência', 'Data Fim
Vigência', 'Data Publicação DOU', 'Data Assinatura Contrato', 'Número_código']

compras2 = compras[colunas_compras]
compras2.columns
compras2.to_csv('compras2.csv', sep=';', encoding='utf-8')

compras2.columns

# Lista de colunas desejadas
colunas_participantes = ['Descrição Item Compra', 'Número_código']

participantes2 = participantes[colunas_participantes]
participantes2
participantes2.to_csv('participantes2.csv', sep=';', encoding='utf-8')

### Estatística descritiva
licitacao2.describe(include='all')
item2.describe()
compras2.describe()
participantes2.describe()

### Juntando as bases
# Merge
dados = pd.merge(licitacao2, item2, on='Número_código', how='inner')

dados.columns

dados.to_csv(pastal + 'dados.csv', encoding='utf-8')

dados1 = pd.merge(dados, compras2, on='Número_código', how='inner')

dados1.to_csv('/content/drive/MyDrive/previsão licitação/dados1.csv', sep=';',
encoding='utf-8')
participantes2.to_csv('/content/drive/MyDrive/previsão
licitação/participantes2.csv', sep=';', encoding='utf-8')

dados2 = pd.merge(dados1, participantes2, on='Número_código', how='inner')
```

Código Fonte: Organizar dados Compras

```
import pandas as pd

# Lendo o nome dos arquivos

caminho_apostilamento202201 = '202201_Apostilamento.csv'
caminho_compras202201 = '202201_Compras.csv'
caminho_item202201 = '202201_ItemCompra.csv'
caminho_termo202201 = '202201_TermoAditivo.csv'

# Lendo os arquivos

apostilamento202201 = pd.read_csv(caminho_apostilamento202201, encoding='latin1',
delimiter=';')
compras202201 = pd.read_csv(caminho_compras202201, encoding='latin1',
delimiter=';')
item202201 = pd.read_csv(caminho_item202201, encoding='latin1', delimiter=';')
termo202201 = pd.read_csv(caminho_termo202201, encoding='latin1', delimiter=';')

### Repetir para os demais meses alterando variáveis e nomes de arquivos
### correspondentes
```

Código Fonte: Combinar dados Licitações e Compras

```
# Importando os dados
import pandas as pd
from google.colab import drive
drive.mount('/content/drive')

# Leia o arquivo CSV com separação por ponto e vírgula
#empenho = pd.read_csv('/content/drive/MyDrive/previsão
licitação/dados_empenho.csv', sep=';')

#item = pd.read_csv('/content/drive/MyDrive/previsão licitação/dados_item.csv',
sep=';')

licitacao = pd.read_csv('/content/drive/MyDrive/previsão
licitação/dados_licitacao.csv', sep=';')

#participantes = pd.read_csv('/content/drive/MyDrive/previsão
licitação/dados_participantes.csv', sep=';')

compras = pd.read_csv('/content/drive/MyDrive/previsão
licitação/dados_compras.csv', sep=';')

## Pegando apenas as colunas que interessa

# Lista de colunas desejadas
colunas_licitacao = ['Valor Licitação', 'Nome UG', 'Objeto', 'Modalidade Compra',
'Data Resultado Compra', 'UF', 'Número_código']

# Criar um novo DataFrame com apenas as colunas desejadas
licitacao = licitacao[colunas_licitacao]
licitacao.to_csv('licitacao2.csv', sep=';', encoding='utf-8')

# Lista de colunas desejadas
colunas_item = ['Nome Vencedor', 'Número_código']

item= item[colunas_item]
item.to_csv('item2.csv', sep=';', encoding='utf-8')

# Lista de colunas desejadas
colunas_compras = ['Data Assinatura Contrato', 'Data Início Vigência', 'Data Fim
Vigência', 'Data Publicação DOU', 'Data Assinatura Contrato', 'Número_código']

compras = compras[colunas_compras]
compras.to_csv('compras2.csv', sep=';', encoding='utf-8')
```

```
# Lista de colunas desejadas
colunas_participantes = ['Descrição Item Compra', 'Número_código']

participantes = participantes[colunas_participantes]
participantes.to_csv('participantes2.csv', sep=';', encoding='utf-8')

## Merge

# Merge
dados = pd.merge(licitacao, item, on='Número_código', how='inner')

dados = pd.merge(dados, compras, on='Número_código', how='inner')
```

Código fonte: Organizar dados por mês

```
# Importando as bibliotecas
import pandas as pd # para uso de dataframe

##Acesso ao Drive
from google.colab import drive
drive.mount('/content/drive')

# Janeiro

#Configurando os caminhos das pastas onde estão os dados
periodo = '202201'
pastal = f'/content/drive/MyDrive/previsão
licitação/licitações/'+periodo+'_Licitacoes/'
pastac = f'/content/drive/MyDrive/previsão licitação/compras/'+periodo+'_Compras/'

# Lendo o nome dos arquivos
caminho_empenho202201 = '202201_EmpenhosRelacionados.csv'
caminho_item202201 = '202201_ItemLicitação.csv'
caminho_licitacao202201 = '202201_Licitação.csv'
caminho_participantes202201 = '202201_ParticipantesLicitação.csv'
caminho_compras202201 = '202201_Compras.csv'

# Lendo os arquivos
empenho202201 = pd.read_csv(pastal + caminho_empenho202201, encoding='latin1',
delimiter=';',low_memory=False)
item202201 = pd.read_csv(pastal + caminho_item202201, encoding='latin1',
delimiter=';',low_memory=False)
licitacao202201 = pd.read_csv(pastal + caminho_licitacao202201, encoding='latin1',
delimiter=';',low_memory=False)
participantes202201 = pd.read_csv(pastal + caminho_participantes202201,
encoding='latin1', delimiter=';',low_memory=False)
compras202201 = pd.read_csv(pastac + caminho_compras202201, encoding='latin1',
delimiter=';',low_memory=False)

# Criar uma nova coluna 'Número_código'
empenho202201['Número_código'] = empenho202201['Número Licitação'].astype(str) +
empenho202201['Código UG'].astype(str)

# Criar uma nova coluna 'Número_código'
item202201['Número_código'] = item202201['Número Licitação'].astype(str) +
item202201['Código UG'].astype(str)

# Criar uma nova coluna 'Número_código'
```

```
licitacao202201['Número_código'] = licitacao202201['Número Licitação'].astype(str)
+ licitacao202201['Código UG'].astype(str)

# Criar uma nova coluna 'Número_código'
participantes202201['Número_código'] = participantes202201['Número
Licitação'].astype(str) + participantes202201['Código UG'].astype(str)

# Criar uma nova coluna 'Número_código'
compras202201['Número_código'] = compras202201['Número Licitação'].astype(str) +
compras202201['Código UG'].astype(str)

# Lista de colunas desejadas
colunas_licitacao = ['Valor Licitação', 'Nome UG', 'Objeto', 'Modalidade Compra',
'Data Resultado Compra', 'UF', 'Número_código']
colunas_item = ['Nome Vencedor', 'Número_código']
colunas_compras = ['Data Assinatura Contrato', 'Data Início Vigência', 'Data Fim
Vigência', 'Data Publicação DOU', 'Data Assinatura Contrato', 'Número_código']
colunas_participantes = ['Descrição Item Compra', 'Número_código', 'Código
Participante']

# Criar um novo DataFrame com apenas as colunas desejadas
licitacao202201 = licitacao202201[colunas_licitacao]
item202201 = item202201[colunas_item]
compras202201 = compras202201[colunas_compras]
participantes202201 = participantes202201[colunas_participantes]

# Merge
dados1 = pd.merge(licitacao202201, item202201, on='Número_código', how='inner')

dados1 = pd.merge(dados1, compras202201, on='Número_código', how='inner')

dados1 = pd.merge(dados1, participantes202201, on='Número_código', how='inner')

dados1
# salvando para csv /content/drive/MyDrive/previsão licitação/rf_geral
dados1.to_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosjaneiro.csv', sep=';')

### Repetir para os demais meses alterando variáveis e nomes de arquivos
### correspondentes
```

Código fonte: Arrumando os Dados Previsão

```
import pandas as pd

#Acesso ao Drive
from google.colab import drive
drive.mount('/content/drive')

# Lendo os dados

# Lendo os dados csv
dados1 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosjaneiro.csv', sep=';', encoding='utf-8')
dados2 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosfevereiro.csv', sep=';', encoding='utf-8')
dados3 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosmarco.csv', sep=';', encoding='utf-8')
dados4 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosabril.csv', sep=';', encoding='utf-8')
dados5 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosmaio.csv', sep=';', encoding='utf-8')
dados6 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosjunho.csv', sep=';', encoding='utf-8')
dados7 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosjulho.csv', sep=';', encoding='utf-8')
dados8 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosagosto.csv', sep=';', encoding='utf-8')
dados9 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadossetembro.csv', sep=';', encoding='utf-8')
dados10 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosoutubro.csv', sep=';', encoding='utf-8')
dados11 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosnovembro.csv', sep=';', encoding='utf-8')
dados12 = pd.read_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dadosdezembro.csv', sep=';', encoding='utf-8')

# Arrumando os DF

## Janeiro
dados1

# Definir as colunas do DataFrame
colunas_dados1 = [
    "Orçamento",
    "Objetivo da licitação",
```

```
    "Tipo de processo",
    "Dia que o contrato foi assinado",
    "Período de execução",
    "Período de espera",
    "Intervalo entre a data de assinatura e a data de concessão",
    "Intervalo entre a data de início da execução e a data de assinatura",
    "Distância para a eleição mais próxima",
    "Categoria do produto",
    "Estado",
    "Código participante"
]

# Criar o DataFrame
dadosjan = pd.DataFrame(columns=colunas_dados1)

# Preencher a coluna
dadosjan['Orçamento'] = dados1['Valor Licitação']
dadosjan['Objetivo da licitação'] = dados1['Objeto']
dadosjan['Tipo de processo'] = dados1['Modalidade Compra']

# Converter a coluna 'Data Assinatura Contrato' para o tipo datetime
dados1['Data Assinatura Contrato'] = pd.to_datetime(dados1['Data Assinatura
Contrato'], format='%d/%m/%Y')
dados1['Data Início Vigência'] = pd.to_datetime(dados1['Data Início Vigência'],
format='%d/%m/%Y')
dados1['Data Fim Vigência'] = pd.to_datetime(dados1['Data Fim Vigência'],
format='%d/%m/%Y')
dados1['Data Resultado Compra'] = pd.to_datetime(dados1['Data Resultado Compra'],
format='%d/%m/%Y')
dados1['Data Publicação DOU'] = pd.to_datetime(dados1['Data Publicação DOU'],
format='%d/%m/%Y')

# Extrair apenas o dia e preencher a coluna 'Dia que o contrato foi assinado' de
'dadosjan'
dadosjan['Dia que o contrato foi assinado'] = dados1['Data Assinatura
Contrato'].dt.day

# Calcular o período de execução (diferença entre as datas)
dadosjan['Período de execução'] = (dados1['Data Fim Vigência'] - dados1['Data
Início Vigência']).dt.days

# Calcular o período de espera (diferença entre as datas)
dadosjan['Período de espera'] = (dados1['Data Publicação DOU'] - dados1['Data
Resultado Compra']).dt.days
```

```
# Preencher a coluna 'Intervalo entre a data de assinatura e a data de concessão'
de 'dadosjan' co
dadosjan['Intervalo entre a data de assinatura e a data de concessão'] =
(dados1['Data Resultado Compra'] - dados1['Data Assinatura Contrato']).dt.days

# Preencher a coluna 'Intervalo entre a data de início da execução e a data de
assinatura'
dadosjan['Intervalo entre a data de início da execução e a data de assinatura'] =
(dados1['Data Início Vigência'] - dados1['Data Assinatura Contrato']).dt.days

# Preencher a coluna 'Distância para a eleição mais próxima' com o valor 7
dadosjan['Distância para a eleição mais próxima'] = 7

dadosjan['Categoria do produto'] = dados1['Descrição Item Compra']
dadosjan['Estado'] = dados1['UF']
dadosjan['Código participante'] = dados1['Código Participante']

dadosjan

dadosjan.to_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dados_previsao/dadosjaneiro.csv', sep=';')

### Repetir para os demais meses alterando variáveis e nomes de arquivos
### correspondentes
```

Código fonte: Dados Finais

```
import pandas as pd

dados1 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosjaneiro.csv', sep=';', encoding='utf-8')
dados2 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosfevereiro.csv', sep=';', encoding='utf-8')
dados3 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosmarco.csv', sep=';', encoding='utf-8')
dados4 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosabril.csv', sep=';', encoding='utf-8')
dados5 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosmaio.csv', sep=';', encoding='utf-8')
dados6 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosjunho.csv', sep=';', encoding='utf-8')
dados7 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosjulho.csv', sep=';', encoding='utf-8')
dados8 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosagosto.csv', sep=';', encoding='utf-8')
dados9 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadossetembro.csv', sep=';', encoding='utf-8')
dados10 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosoutubro.csv', sep=';', encoding='utf-8')
dados11 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosnovembro.csv', sep=';', encoding='utf-8')
dados12 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosdezembro.csv', sep=';', encoding='utf-8')

# var dependente
vd = pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf/multa.csv',
sep=';', encoding='latin1')
```

```
# Suponha que você tenha um DataFrame chamado 'dados1'
# Novo nome das colunas
novos_nomes_colunas = [
    "Orçamento",
    "Objetivo da licitação",
    "Tipo de processo",
    "Dia que o contrato foi assinado",
    "Período de execução",
    "Período de espera",
    "Intervalo entre a data de assinatura e a data de concessão",
    "Intervalo entre a data de início da execução e a data de assinatura",
    "Distância para a eleição mais próxima",
    "Categoria do produto",
    "Estado",
    "Código participante"]  # Substitua com os nomes desejados

# Passo 1: Converter a coluna "DATA INÍCIO SANÇÃO" para o formato de data
vd['DATA INÍCIO SANÇÃO'] = pd.to_datetime(vd['DATA INÍCIO SANÇÃO'],
format='%d/%m/%Y')

dados1 = dados1.drop('Unnamed: 0', axis=1)

# Substituir os nomes das colunas
dados1.columns = novos_nomes_colunas

dados1
vd

# Acima foi analisado contratos assinados em janeiro, logo preciso pegar apenas os
que sofreram sansão a partir de fevereiro de 2022

# Passo 2: Definir a data limite
data_limite = pd.Timestamp('2022-02-01')

# Passo 3: Filtrar o DataFrame
vd_filtrado = vd[vd['DATA INÍCIO SANÇÃO'] >= data_limite]

# Passo 1: Obter os valores únicos da coluna "CPF OU CNPJ DO SANCIONADO" de 'vd'
valores_sancionados = set(vd['CPF OU CNPJ DO SANCIONADO'].dropna().unique())

# Passo 2: Criar a coluna "Multas" em 'vis' com base na presença dos valores
dados1['Multas'] = dados1['Código participante'].apply(lambda x: 1 if x in
valores_sancionados else 0)

# Exibir o DataFrame 'vis' com a nova coluna
```

```
dados1
### Repetir para os demais meses alterando variáveis e nomes de arquivos
### correspondentes

# RF

import numpy as np
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, cross_val_score,
StratifiedKFold
from sklearn.metrics import classification_report, confusion_matrix
from imblearn.under_sampling import RandomUnderSampler

## Janeiro

# Carregar o DataFrame 'dados1'
# dados1 = pd.read_csv('caminho/para/seu/arquivo.csv')  # Descomente e ajuste o
caminho conforme necessário

# Dividir o DataFrame em duas partes com base na classe 'Multas'
dados1_0 = dados1[dados1['Multas'] == 0]
dados1_1 = dados1[dados1['Multas'] == 1]

# Amostrar 0,01% de cada grupo
frac = 0.0001
amostra_0 = dados1_0.sample(frac=frac, random_state=42)
amostra_1 = dados1_1.sample(frac=frac, random_state=42)

# Concatenar as amostras para obter o DataFrame final
dados1 = pd.concat([amostra_0, amostra_1])

### Repetir para os demais meses alterando variáveis e nomes de arquivos
### correspondentes


# Concatenar todos os DataFrames
dados = pd.concat([dados1, dados2, dados3, dados4, dados5, dados6, dados7, dados8,
dados9, dados10, dados11, dados12], ignore_index=True)

dados.to_csv('/content/drive/MyDrive/previsão
licitação/rf_geral/dados_previsao/dadosgeral.csv', sep=';')
```

Código fonte: *Random Forest* Mensal

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, cross_val_score,
StratifiedKFold
from sklearn.metrics import classification_report, confusion_matrix
from imblearn.under_sampling import RandomUnderSampler
import os

# selecionar dados mês
dados1 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/dadosagosto.csv', sep=';', encoding='utf-8')
# var dependente
vd = pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf/multa.csv',
sep=';', encoding='latin1')

# Suponha que você tenha um DataFrame chamado 'dados1'
# Novo nome das colunas
novos_nomes_colunas = [
    "Orçamento",
    "Objetivo da licitação",
    "Tipo de processo",
    "Dia que o contrato foi assinado",
    "Período de execução",
    "Período de espera",
    "Intervalo entre a data de assinatura e a data de concessão",
    "Intervalo entre a data de início da execução e a data de assinatura",
    "Distância para a eleição mais próxima",
    "Categoria do produto",
    "Estado",
    "Código participante"]  # Substitua com os nomes desejados

# Passo 1: Converter a coluna "DATA INÍCIO SANÇÃO" para o formato de data_limite
vd['DATA INÍCIO SANÇÃO'] = pd.to_datetime(vd['DATA INÍCIO SANÇÃO'],
format='%d/%m/%Y')

dados1 = dados1.drop('Unnamed: 0', axis=1)

# Substituir os nomes das colunas
dados1.columns = novos_nomes_colunas
```

```
# Acima foi analisado contratos assinados em janeiro, logo preciso pegar apenas os
que sofreram sansão a partir de fevereiro de 2022

# Passo 2: Definir a data limite
################################################################################
data_limite = pd.Timestamp('2022-09-01') # Modificar segundo mês analisado

# Passo 3: Filtrar o DataFrame
vd_filtrado = vd[vd['DATA INÍCIO SANÇÃO'] >= data_limite]

# Passo 1: Obter os valores únicos da coluna "CPF OU CNPJ DO SANCIONADO" de 'vd'
valores_sancionados = set(vd['CPF OU CNPJ DO SANCIONADO'].dropna().unique())

# Passo 2: Criar a coluna "Multas" em 'vis' com base na presença dos valores
dados1['Multas'] = dados1['Código participante'].apply(lambda x: 1 if x in
valores_sancionados else 0)

# Exibir o DataFrame 'vis' com a nova coluna
dados1

# Carregar o DataFrame 'dados1'
# dados1 = pd.read_csv('caminho/para/seu/arquivo.csv')  # Descomente e ajuste o
caminho conforme necessário

# Dividir o DataFrame em duas partes com base na classe 'Multas'
dados1_0 = dados1[dados1['Multas'] == 0]
dados1_1 = dados1[dados1['Multas'] == 1]

# Amostrar 10% de cada grupo
frac = 0.1
amostra_0 = dados1_0.sample(frac=frac, random_state=42)
amostra_1 = dados1_1.sample(frac=frac, random_state=42)

# Concatenar as amostras para obter o DataFrame final
dados1 = pd.concat([amostra_0, amostra_1])

# Separar variáveis independentes e dependentes
X = dados1.drop(columns=['Código participante', 'Multas'])
y = dados1['Multas']

# Converter colunas numéricas (substituir vírgulas por pontos)
for col in X.select_dtypes(include=['object']).columns:
    try:
        X[col] = X[col].str.replace(',', '.').astype(float)
```

```
    except ValueError:
        continue

# Codificar colunas categóricas
X = pd.get_dummies(X, drop_first=True)

# Excluir linhas com valores NaN
X = X.dropna()
y = y[X.index]

# Verificar a distribuição das classes
print("Distribuição das classes:")
print(y.value_counts())

# Balancear os dados usando RandomOverSampler
ros = RandomUnderSampler(random_state=42)
X_resampled, y_resampled = ros.fit_resample(X, y)

# Dividir os dados em conjuntos de treino e teste
X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled,
test_size=0.3, random_state=42)

# Treinar o modelo de Random Forest
clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(X_train, y_train)

# Predição no conjunto de teste
y_pred = clf.predict(X_test)

# Calcular e exibir métricas
print("Classification Report:")
print(classification_report(y_test, y_pred))

# Exibir a matriz de confusão
conf_matrix = confusion_matrix(y_test, y_pred)
print('Confusion Matrix:')
print(conf_matrix)

# Validação cruzada
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
cross_val_scores = cross_val_score(clf, X_resampled, y_resampled, cv=skf,
scoring='f1')
print(f'Cross-Validation F1 Scores: {cross_val_scores}')
print(f'Mean Cross-Validation F1 Score: {np.mean(cross_val_scores)}')
```

```
# Criar DataFrame com os resultados das previsões
result_df = X_test.copy()
result_df['Actual'] = y_test
result_df['Predicted'] = y_pred
result_df

# Obter importâncias das features
feature_importances = clf.feature_importances_

# Obter os nomes das features
feature_names = X.columns

# Criar um DataFrame com os nomes das features e suas importâncias
feature_importance_df = pd.DataFrame({'Feature': feature_names, 'Importance':
feature_importances})

# Extrair os nomes das variáveis originais
# Supondo que os nomes das variáveis originais estão antes do primeiro '_' nos
nomes das colunas
feature_importance_df['OriginalFeature'] =
feature_importance_df['Feature'].apply(lambda x: x.split('_')[0])

# Agrupar pelas variáveis originais e somar as importâncias
aggregated_importance_df =
feature_importance_df.groupby('OriginalFeature').agg({'Importance':
'sum'}).reset_index()

#-------------------------------------------------------------------------
# SALVAR CSV
# Cria uma nova coluna com dados aleatórios ou dados específicos

nome_arquivo = 'importancia.csv'

# Verifica se o arquivo existe
if
os.path.exists('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_pr
evisao/importancia.csv'):
    # Lê o arquivo CSV existente
    df =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/importancia.csv')

else:
    # Se o arquivo não existir, cria um DataFrame vazio com uma estrutura inicial
    # Você pode definir outras colunas iniciais se necessário
```

```
    df = aggregated_importance_df ['OriginalFeature']

df.to_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previsa
o/importancia.csv', index=False)

# Define o nome da nova coluna baseado no número de colunas existentes
# Isso garante que cada nova coluna terá um nome único
#nova_coluna_nome = f'Nova_Coluna_{len(df.columns) + 1}'
nova_coluna_nome = 'Dezembro'

#print(aggregated_importance_df['Importance']  )
#df = aggregated_importance_df['Importance']
df[nova_coluna_nome] = aggregated_importance_df ['Importance'].astype(float)


# Salva o DataFrame atualizado de volta ao arquivo CSV

df.to_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previsa
o/importancia.csv', index=False)

#-------------------------------------------------------------------------

# Ordenar o DataFrame pela importância
aggregated_importance_df = aggregated_importance_df.sort_values(by='Importance',
ascending=False)

# Exibir as importâncias das variáveis originais
print("Importância das Variáveis Independentes:")
print(aggregated_importance_df)

# Criar a figura e os eixos
plt.figure(figsize=(12, 8))
#ns.barplot(x='Importance', y='OriginalFeature', data=aggregated_importance_df)

# Adicionar títulos e rótulos
plt.title('Importância das Variáveis Independentes - Novembro')
plt.xlabel('Importância')
plt.ylabel('Variável')

# Mostrar o gráfico
plt.show()
```

Código fonte: *Random Forest* Anual

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, cross_val_score,
StratifiedKFold
from sklearn.metrics import classification_report, confusion_matrix
from imblearn.under_sampling import RandomUnderSampler
from sklearn.metrics import roc_curve, auc

dados1 =
pd.read_csv('C:/Users/vsouz/OneDrive/Desktop/previsaolicitacao/rf_geral/dados_previ
sao/resultadosfinais.csv', sep=';', encoding='utf-8')
#dados1 = pd.read_csv('D:/Vini/dadosgeral_10.csv', sep=';', encoding='utf-8')

# var dependente
#vd = pd.read_csv('C:\Users\vsouz\OneDrive\Desktop\previsaolicitacao\rf\multa.csv',
sep=';', encoding='latin1')

dados1 = dados1.drop('Unnamed: 0', axis=1)
dados1

# Separar variáveis independentes e dependentes
X = dados1.drop(columns=['Código participante', 'Multas'])
y = dados1['Multas']

# Converter colunas numéricas (substituir vírgulas por pontos)
for col in X.select_dtypes(include=['object']).columns:
    try:
        X[col] = X[col].str.replace(',', '.').astype(float)
    except ValueError:
        continue

# Codificar colunas categóricas
X = pd.get_dummies(X, drop_first=True)

# Excluir linhas com valores NaN
X = X.dropna()
y = y[X.index]

# Verificar a distribuição das classes
```

```
print("Distribuição das classes:")
print(y.value_counts())

# Balancear os dados usando RandomOverSampler
ros = RandomUnderSampler(random_state=42)
X_resampled, y_resampled = ros.fit_resample(X, y)

# Dividir os dados em conjuntos de treino e teste
X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled,
test_size=0.3, random_state=42)

# Treinar o modelo de Random Forest
clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(X_train, y_train)

# Predição no conjunto de teste
y_pred = clf.predict(X_test)

# Calcular e exibir métricas
print("Classification Report:")
print(classification_report(y_test, y_pred))

# Exibir a matriz de confusão
conf_matrix = confusion_matrix(y_test, y_pred)
print('Confusion Matrix:')
print(conf_matrix)

# Validação cruzada
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
cross_val_scores = cross_val_score(clf, X_resampled, y_resampled, cv=skf,
scoring='f1')
print(f'Cross-Validation F1 Scores: {cross_val_scores}')
print(f'Mean Cross-Validation F1 Score: {np.mean(cross_val_scores)}')

# Criar DataFrame com os resultados das previsões
result_df = X_test.copy()
result_df['Actual'] = y_test
result_df['Predicted'] = y_pred
result_df

# Obter importâncias das features
feature_importances = clf.feature_importances_

# Obter os nomes das features
feature_names = X.columns
```

```
# Criar um DataFrame com os nomes das features e suas importâncias
feature_importance_df = pd.DataFrame({'Feature': feature_names, 'Importance':
feature_importances})

# Extrair os nomes das variáveis originais
# Supondo que os nomes das variáveis originais estão antes do primeiro '_' nos
nomes das colunas
feature_importance_df['OriginalFeature'] =
feature_importance_df['Feature'].apply(lambda x: x.split('_')[0])

# Agrupar pelas variáveis originais e somar as importâncias
aggregated_importance_df =
feature_importance_df.groupby('OriginalFeature').agg({'Importance':
'sum'}).reset_index()

# Ordenar o DataFrame pela importância
aggregated_importance_df = aggregated_importance_df.sort_values(by='Importance',
ascending=False)

# Exibir as importâncias das variáveis originais
print("Importância das Variáveis Independentes:")
print(aggregated_importance_df)

# Criar a figura e os eixos
plt.figure(figsize=(12, 8))
sns.barplot(x='Importance', y='OriginalFeature', data=aggregated_importance_df,
palette='viridis')

# Adicionar títulos e rótulos
plt.title('Importância das Variáveis Independentes')
plt.xlabel('Importância')
plt.ylabel('Variável')

# Mostrar o gráfico
plt.show()

result_df.to_csv('c:/dados_previsao/resultadosfinais.csv', sep=';')

# Calculando as probabilidades
y_scores = clf.predict_proba(X_test)[:, 1]

# print("------------------")
# print(X_test)
```

```
print("------------------")
print(result_df)

print("------------------")

df = pd.DataFrame({
    'Prob': y_scores
})

# Calculando a ROC
fpr, tpr, thresholds = roc_curve(y_test, y_scores)
roc_auc = auc(fpr, tpr)

# Plotando a Curva ROC
plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area = %0.2f)' %
roc_auc)
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.show()


# Criando o ambiente do gráfico
sns.set_style("white")
plt.figure(figsize=(10, 10))

# Gráfico de Dispersão
g = sns.scatterplot(x="Predicted",y="Actual", data=result_df)
plt.xlabel('Predicted labels')
plt.ylabel('True labels')
plt.show()

# Plotar a matriz de confusão
# plt.figure(figsize=(8, 6))
# sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues', cbar=False)
# plt.xlabel('Predicted labels')
# plt.ylabel('True labels')
# plt.title('Confusion Matrix')
# plt.show()
result_df.to_csv('D:/Vini/Result.csv', index=False)
```