

CÂMARA DOS DEPUTADOS
CENTRO DE FORMAÇÃO, TREINAMENTO E
APERFEIÇOAMENTO
PROGRAMA DE PÓS-GRADUAÇÃO
MESTRADO PROFISSIONAL EM PODER LEGISLATIVO



Rubens Vasconcellos Terra Neto

**APRENDIZADO DE MÁQUINA NO COMBATE AO SOBREPREGO DAS COMPRAS
PÚBLICAS:
uma aplicação de algoritmos computacionais na análise de preços de contratações do
Senado Federal**

Brasília, DF
2023

Rubens Vasconcellos Terra Neto

**APRENDIZADO DE MÁQUINA NO COMBATE AO SOBREPREÇO DAS COMPRAS
PÚBLICAS:
uma aplicação de algoritmos computacionais na análise de preços de contratações do
Senado Federal**

Trabalho de conclusão de curso (modalidade **relatório técnico**) apresentado como requisito parcial para a obtenção do grau de **Mestre** no Curso de Mestrado Profissional do Programa de Pós-Graduação do Centro de Formação, Treinamento e Aperfeiçoamento (Cefor) da Câmara dos Deputados, na área de concentração **Poder Legislativo**, linha de pesquisa **Gestão Pública no Poder Legislativo**.

Orientador: Prof. Dr. Fabiano Peruzzo Schwartz

Brasília, DF
2023

Termo de Consentimento

Conforme previsto na Lei n.º 13.709/2018, o(a) autor(a) autoriza a divulgação do texto completo deste Trabalho de Conclusão de Curso do Mestrado Profissional em Poder Legislativo no sítio eletrônico da Câmara dos Deputados e a sua reprodução total ou parcial para fins acadêmicos e científicos, estando ciente de que, após a divulgação, o conteúdo será de livre acesso ao público.

Terra Neto, Rubens Vasconcellos.

Aprendizado de máquina no combate ao sobrepreço das compras públicas: uma aplicação de algoritmos computacionais na análise de preços de contratações do Senado Federal / Rubens Vasconcellos Terra Neto. – 2023.

101 f.

Orientador: Fabiano Peruzzo Schwartz.

Impresso por computador.

Dissertação (mestrado profissional) – Câmara dos Deputados, Centro de Formação, Treinamento e Aperfeiçoamento (Cefor), 2023.

1. Inteligência Artificial. 2. Setor público, contratação. 3. Brasil. Congresso Nacional. Senado Federal. 4. Poder Legislativo I. Título.

CDU 328(81)

Bibliotecária: Débora Machado de Toledo – CRB1: 1303

CÂMARA DOS DEPUTADOS
Centro de Formação, Treinamento e Aperfeiçoamento
Programa de Pós-Graduação
Mestrado Profissional em Poder Legislativo



Rubens Vasconcellos Terra Neto

**APRENDIZADO DE MÁQUINA NO COMBATE AO SOBREPREÇO DAS COMPRAS
PÚBLICAS:
uma aplicação de algoritmos computacionais na análise de preços de contratações do
Senado Federal**

Trabalho de conclusão de curso (modalidade **relatório técnico**) apresentado como requisito parcial para a obtenção do grau de **Mestre** no Curso de Mestrado Profissional do Programa de Pós-Graduação do Centro de Formação, Treinamento e Aperfeiçoamento (Cefor) da Câmara dos Deputados, na área de concentração **Poder Legislativo**, linha de pesquisa **Gestão Pública no Poder Legislativo**.

Trabalho **aprovado** pela seguinte Banca Examinadora, designada pela Coordenação do Programa de Pós-Graduação:

Prof. Dr. Fabiano Peruzzo Schwartz
Presidente da Banca – Câmara dos Deputados

Prof. Dr. Mauro Moura Severino
Membro interno – Câmara dos Deputados

Dr. Leandro Carísio Fernandes
Membro externo – Tribunal de Contas da União

Brasília, DF, 15 de dezembro de 2023.

*Dedico este trabalho à minha família, que
sempre me incentivou.*

AGRADECIMENTOS

Agradeço ao meu orientador professor Fabiano Peruzzo Schwartz, pela sabedoria com que me guiou nesta trajetória e pelo incentivo em momentos difíceis da pesquisa.

Aos professores Mauro Moura Severino e Leandro Carísio Fernandes, pelas relevantes críticas feitas para o aprimoramento do trabalho durante a qualificação e a defesa.

Aos meus colegas de mestrado, que dividiram comigo a busca constante de conhecimento e aprimoramento profissional.

Aos professores do curso cuja dedicação foi fator preponderante para a qualidade do curso.

A todo o pessoal da COPOS que tanto se dedicou por nós do corpo discente.

Gostaria de deixar registrado também, o meu reconhecimento à minha família, pois acredito que sem o apoio deles seria muito difícil vencer esse desafio.

Enfim, a todos os que por algum motivo contribuíram para a realização desta pesquisa.

“As máquinas de previsão não fornecem julgamentos. Apenas os humanos o fazem, porque apenas os humanos podem expressar as recompensas relativas de realizar ações diferentes.”

(Ajay Agrawal)

“Um computador mereceria ser chamado de inteligente se pudesse enganar um humano fazendo-o acreditar que era humano.”

(Alan Turing)

RESUMO

Quando o assunto é compras públicas, o caminho da legalidade exige a realização de processo licitatório. A pesquisa de preços é um dos principais artefatos que compõem o processo licitatório, e tem como finalidade estimar o montante de dinheiro que poderá ser gasto na avença. A aferição desse montante tem como objetivos a reserva do valor no orçamento e o balizamento de preços para garantir que a Administração Pública esteja pagando um preço justo e compatível com o de mercado. Essa iniciativa promove a transparência dos atos do governo ao divulgar informações sobre os processos licitatórios, gerando atas em todas as fases e abrindo prazo para os recursos de habilitação e julgamento. Dessa forma, contribui para a detecção de anomalias de preços (ou sobrepreços) em novas compras. O presente estudo relata pesquisa de mestrado que se dedicou a explorar e processar os dados de contratações públicas nos últimos 2 anos. O principal objetivo consistiu no desenvolvimento e aplicação de algoritmos baseados em Aprendizagem de Máquina (*Machine Learning*) capazes de indicar, a partir de estimativas de probabilidade, possíveis indícios de sobrepreços durante a fase de pesquisa de preços em contratações. A construção da base de dados empregada no desenvolvimento se utilizou dos dados abertos de compras governamentais que mantém registro das compras e contratações firmadas pelo Poder Executivo e por todas as instituições que utilizam o sistema Comprasnet. Os dados recuperados são referentes ao período de janeiro/2022 até março/2023 e estão disponíveis em repositório público. Um primeiro protótipo de detecção de sobrepreço de compras foi desenvolvido com base na biblioteca denominada *Python Outlier Detection* (PyOD), que apresenta 14 algoritmos de detecção de anomalia não-supervisionados. Para a escolha dos algoritmos e definição dos parâmetros foram feitos testes em duas categorias de materiais. Optou-se por utilizar no modelo os algoritmos *Empirical-Cumulative-distribution-based Outlier Detection* (ECOD), Análise de Componentes Principais (PCA) e Fator de *Outlier* Local (LOF) com um percentual estimado de 10% de grau de contaminação da base de dados com anomalias. Foram realizados testes com todas as combinações possíveis entre estes 3 algoritmos e optou-se por criar um modelo utilizando os 3 agrupados através da utilização da biblioteca *Scalable Unsupervised Outlier Detection* (SUOD) que é uma estrutura de aceleração para treinamento e previsão de detectores heterogêneos não supervisionados em larga escala, facilitando a utilização simultânea de mais de um modelo de algoritmo. Para a avaliação final da ferramenta desenvolvida foi utilizada uma terceira categoria de material e os resultados obtidos foram de 95,24% de Acurácia com Precisão de 74,29% e *Recall* de 76,47%.

Palavras-chave: Inteligência Artificial; Setor Público; Processo de Contratação; Senado Federal; Poder Legislativo.

ABSTRACT

When it comes to public acquisitions, the path to legality requires a bidding process. Price research is one of the main artifacts that make up the bidding process, and aims to estimate the amount of money that can be spent on the agreement. The purpose of measuring this amount is to reserve the amount in the budget and set prices to ensure that the Public Administration is paying a fair price compatible with the market. This initiative promotes the transparency of government actions by disclosing information about bidding processes, generating minutes at all stages and opening deadlines for qualification and judgment appeals. In this way, it contributes to the detection of price anomalies (or overpricing) in new purchases. The present study reports a master's research that was dedicated to exploring and processing public procurement data in the last 2 years. The main objective consisted in the development and application of algorithms based on Machine Learning capable of indicating, based on probability estimates, possible indications of overpricing during the price research phase in acquisitions. The construction of the database used in the development used open data from government purchases that keep records of purchases and contracts signed by the Power Executive and by all institutions that use the Comprasnet system. The recovered data refer to the period from January/2022 to March/2023 and are available in a public repository. A first shopping overprice detection prototype was developed with based on the library called Python Outlier Detection (PyOD), which presents 14 algorithms unsupervised anomaly detection. For the choice of algorithms and definition of parameters tests were performed on two categories of materials. It was decided to use in the model Empirical-Cumulative-distribution-based Outlier Detection (ECOD) algorithms, Principal Components Analysis (PCA) and Local Outlier Factor (LOF) with an estimated percentage of 10% degree of database contamination with anomalies. Tests were carried out with all possible combinations between these 3 algorithms and it was decided to create a model using the 3 grouped through the use of the Scalable Unsupervised Outlier library Detection (SUOD) which is an accelerating framework for training and predicting detectors unsupervised heterogeneous applications on a large scale, facilitating the simultaneous use of more of an algorithm model. For the final evaluation of the developed tool, a third category of material was used and the results obtained were 95.24% Accuracy with Precision of 74.29% and Recall of 76.47%.

Keywords:Artificial Intelligence; Public sector; Acquisition process; Brazilian Federal Senate; Legislative Branch.

LISTA DE FIGURAS

Figura 1 – Modelo de dados do Sistema Integrado de Administração e Serviços Gerais (SIASG)	35
Figura 2 – A Robô Rosie	40
Figura 3 – O Painel Jarbas	41
Figura 4 – Matriz de confusão	43
Figura 5 – Modelo da ferramenta de detecção de sobrepreço	56
Figura 6 – Tela de Seleção de Material	61
Figura 7 – Tela de Avaliação da Pesquisa	61
Figura 8 – Matriz de confusão do SUOD - Material Livro	67
Figura 9 – Aba Fonte de pesquisa	96
Figura 10 – Aba Mapa de Cotações	97
Figura 11 – Aba Mapa de Cotações (total)	98
Figura 12 – Aba Planilha Estimativa de Despesas	99

LISTA DE QUADROS

Quadro 1 – Número de registros recuperados dos dados abertos do SIASG	48
Quadro 2 – Características do repositório	48
Quadro 3 – Descrição dos atributos do repositório	48
Quadro 4 – Avaliação dos Algoritmos de detecção de anomalias sem exclusão de extremos	51
Quadro 5 – Avaliação dos Algoritmos de detecção de anomalias com exclusão de extremos	51
Quadro 6 – Avaliação do Algoritmo no material Água por percentual de contaminação	63
Quadro 7 – Avaliação do Algoritmo no material Microcomputador por percentual de contaminação	64
Quadro 8 – Avaliação dos Algoritmos nos materiais Água e Microcomputador com 10% de contaminação	65
Quadro 9 – Avaliação das combinações dos Algoritmos nos materiais Água com 10% de contaminação	65
Quadro 10 – Avaliação das combinações dos Algoritmos no material Microcomputador com 10% de contaminação	66
Quadro 11 – Avaliação dos Algoritmos no material Livro	66
Quadro 12 – Algoritmos do PyOD	83
Quadro 13 – Métodos de Consultas Básicas de Contratos	87
Quadro 14 – Métodos de Informações Detalhadas de Contratos	87
Quadro 15 – Métodos de Consultas Básicas de Fornecedores	88
Quadro 16 – Métodos de Informações Detalhadas de Fornecedores	89
Quadro 17 – Métodos de Consultas Básicas de Licitações	90
Quadro 18 – Métodos de Informações Detalhadas das Licitações	90
Quadro 19 – Métodos de Consultas Básicas de Materiais	91
Quadro 20 – Métodos de Informações Detalhadas de Materiais	92

LISTA DE ABREVIATURAS E SIGLAS

ALICE	acrônimo para Analisador de Licitações, Contratos e Editais
AM	aprendizagem de máquina
APF	administração pública federal
API	<i>Application Programming Interface</i>
CEAP	Cota para o Exercício da Atividade Parlamentar
CGU	Controladoria-Geral da União
IA	Inteligência Artificial
IOT	Internet da Coisas
IP	<i>Internet Protocol</i>
ML	<i>Machine Learning</i>
NIST	<i>National Institute of Standards and Technology</i>
OGD	<i>open government data</i>
SIASG	Sistema Integrado de Administração e Serviços Gerais
SISG	Sistema de Serviços Gerais
STF	Supremo Tribunal Federal
TCU	Tribunal de Contas da União
UnB	Universidade de Brasília
URL	<i>Uniform Resource Locator</i> (Localizador Uniforme de Recursos, em tradução livre)

SUMÁRIO

1	INTRODUÇÃO	25
1.1	Contextualização	25
1.2	Problema de pesquisa	26
1.3	Justificativa	27
1.4	Objetivos	28
1.4.1	Objetivo Geral	28
1.4.2	Objetivos específicos	28
1.5	Escopo do relatório	28
1.6	Organização do Relatório	29
2	FUNDAMENTAÇÃO TEÓRICA	31
2.1	Contratações na Administração Pública Federal	31
2.1.1	Licitações públicas	31
2.1.2	A corrupção nas compras públicas	32
2.2	Dados abertos governamentais, Transparência e Controle Social	34
2.2.1	Dados Abertos Governamentais	34
2.2.2	Transparência e Controle Social	36
2.3	Inteligência Artificial, seu uso na Administração Pública Federal e métricas de avaliação dos modelos	37
2.3.1	Inteligência Artificial	38
2.3.2	Uso da Inteligência Artificial no controle dos gastos públicos	39
2.3.3	Métricas de avaliação dos modelos	42
2.4	Recapitulando	44
3	METODOLOGIA E DESENVOLVIMENTO	45
3.1	Metodologia	45
3.2	Bases de Dados Existentes	46
3.3	Ferramenta para obtenção e pré-processamento dos dados necessários ao Treinamento da Inteligência Artificial	46
3.4	Repositório de dados de contratações	48
3.5	Técnicas, Métodos e Boas práticas na detecção de anomalias	49
3.5.1	<i>Empirical-Cumulative-distribution-based Outlier Detection</i> (ECOD)	52
3.5.2	<i>Copula-Based Outlier Detection</i> (COPOD)	52
3.5.3	Floresta de Isolamento ou <i>Isolation Florest</i> (iForest)	52

3.5.4	Fator de <i>Outlier</i> Local (LOF)	53
3.5.5	<i>Histogram-based Outlier Score</i> (HBOS)	53
3.5.6	Análise de Componentes Principais ou <i>Principal Component Analysis</i> (PCA)	54
3.5.7	Técnicas utilizadas para desenvolvimento do modelo	55
3.6	Modelo do Aplicativo de detecção de sobrepreço	55
3.6.1	Processo “Selecionar a Categoria da Avaliação”	55
3.6.2	Processo “Recuperar itens da categoria selecionada”	56
3.6.3	Processo “Retirar os extremos de valor unitário”	57
3.6.4	Processo “Treinar o modelo de detecção de anomalias”	58
3.6.5	Processo “Solicitar ao Usuário os dados a serem avaliados”	59
3.6.6	Processo “Calcular a possibilidade de ser uma anomalia”	59
3.6.7	Processo “Apresentar os resultados”	60
3.7	Protótipo da ferramenta de avaliação de indícios de sobrepreços disponibilizada no Senado Federal	60
3.7.1	Atualizador de dados diário	60
3.7.2	Tela de Seleção de Material	60
3.7.3	Tela de avaliação e resultados	61
3.8	Considerações finais do capítulo	62
4	RESULTADOS E ANÁLISES	63
5	CONCLUSÕES E CONSIDERAÇÕES FINAIS	69
5.1	Limitações	70
5.2	Trabalhos Futuros	70
	REFERÊNCIAS	73
	Glossário	79
	ANEXOS	81
	ANEXO A – ALGORITMOS DO PYOD	83
	ANEXO B – API DE COMPRAS GOVERNAMENTAIS	87
B.1	Contratos a partir de 2021	87
B.2	Fornecedores	88
B.3	Licitações	89
B.4	Materiais	91

ANEXO C –CHECKLIST DE VERIFICAÇÃO DE PESQUISA DE PREÇOS UTILIZADA NO SENADO FEDERAL	93
ANEXO D –PLANILHA DE PREÇOS UTILIZADA NO SENADO FEDERAL	95
ÍNDICE	101

1 INTRODUÇÃO

Neste capítulo é apresentada uma breve contextualização sobre o tema da pesquisa, sua motivação, o problema abordado, os objetivos e a justificativa. Ao final encontram-se detalhados o escopo e a organização deste relatório.

1.1 Contextualização

Na administração pública federal (APF) as contratações são precedidas de processos licitatórios que, segundo a Lei 14.133/21, tem os seguintes objetivos:

- I - assegurar a seleção da proposta apta a gerar o resultado de contratação mais vantajoso para a Administração Pública, inclusive no que se refere ao ciclo de vida do objeto;
 - II - assegurar tratamento isonômico entre os licitantes, bem como a justa competição;
 - III - evitar contratações com sobrepreço ou com preços manifestamente inexequíveis e superfaturamento na execução dos contratos;
 - IV - incentivar a inovação e o desenvolvimento nacional sustentável.
- (BRASIL, 2021, Art. 11).

A preparação de um processo licitatório demanda da equipe interna de um órgão público a elaboração de estudos técnicos preliminares, projetos básicos, pesquisas de preços entre outros procedimentos formais inerentes às licitações. Porém toda a regulamentação criada para poder evitar a corrupção onerou o gestor público com a confecção de mais e mais artefatos visando garantir a lisura do processo. Este aumento da burocracia tem provocado em muitos casos inação e entraves nos processos de aquisição e contratação. Cabe destacar que dentre as fases do processo licitatório, uma das que traz mais dificuldade ao gestor público é a pesquisa de preços. Não são poucas as vezes onde são levantados processos irregulares com sobrepreços, conluio de empresas e propinas, como citado por [Silveira \(2021\)](#).

Os órgãos de controle, responsáveis por apurar as fraudes nos processos licitatórios, têm se utilizado de novas tecnologias como *Big Data* e Inteligência Artificial (IA) na detecção de irregularidades nos processos de aquisição. Órgãos como o Tribunal de Contas da União (TCU) e a Controladoria-Geral da União (CGU) já fazem uso desses recursos, com destaque para a ferramenta de IA ALICE ¹, como se pode ver em [Panis \(2020\)](#), [Araujo, Zullo e Torres \(2020\)](#) e [Oliveira \(2018\)](#).

Segundo o *National Institute of Standards and Technology (NIST)*², “*Big Data* consiste em extensos conjuntos de dados – principalmente nas características de volume, variedade, velocidade e/ou variabilidade – que requerem uma arquitetura escalável para armazenamento, manipulação e análise” (NIST, 2015).

¹ acrônimo para Analisador de Licitações, Contratos e Editais

² É uma agência governamental americana não regulatória que foi criada para impulsionar a inovação e promover a competitividade industrial nas áreas de ciência, engenharia e tecnologia

Principalmente com o advento das Redes Sociais e da Internet das Coisas (IOT)³ ocorreu um aumento exponencial dos dados gerados e acessíveis via internet que podem ser manipulados, analisados e interpretados para gerar informações. A capacidade humana de analisar e correlacionar tamanha quantidade de dados é limitada, o que impulsionou o desenvolvimento da tecnologia da Inteligência Artificial, que utiliza esta grande quantidade de dados para encontrar, por meio de métodos estatísticos, padrões e correlações.

A tecnologia da Inteligência Artificial tem se tornado mais acessível e com o advento da Lei de Acesso à Informação (BRASIL, 2011) e dos Dados Abertos Governamentais a possibilidade de se utilizar esses dados e tecnologias para auxiliar o gestor público começa a se tornar uma realidade.

1.2 Problema de pesquisa

Toda contratação pública é feita por meio de processo licitatório e uma das etapas do planejamento desta contratação é a pesquisa de preços. Cabe ao gestor público realizar esta fase com a maior acurácia possível, já que uma pesquisa de preços mal formulada é uma das causas de improbidades e irregularidades causadoras de dano ao erário na administração pública conforme Borges Júnior (2020).

Cabe destacar que a utilização de tecnologias nos órgãos de controle está muito mais aprimorada do que as tecnologias disponíveis para o gestor, como se nota no artigo de Peci e Braga (2021):

Enquanto os órgãos de controle contam com recursos humanos especializados, com bases de dados sofisticados e com acesso privilegiado a dados sigilosos de transações financeiras e outras informações cruciais para compreender complexos cenários marcados pela corrupção, esta mesma capacidade não está disponível para o gestor público, o responsável direto pela tomada de decisões de políticas públicas. (PECI; BRAGA, 2021)

Neste mesmo artigo, Peci e Braga (2021) explicam os motivos para esta diferença, que entre outros foi o fortalecimento dos órgãos de controle e o acesso privilegiado a dados sigilosos em virtude de suas atribuições. A burocracia desses órgãos é especializada exatamente no controle e desenvolveu ferramentas próprias utilizando Inteligência Artificial para correlacionar os dados disponíveis e encontrar indícios de irregularidades nos processos licitatórios. Nas ferramentas desenvolvidas, além de dados sigilosos, foram utilizados os dados disponíveis no Sistema Integrado de Administração e Serviços Gerais (SIASG) do Governo Federal (BRASIL, 2021).

Neste contexto, o problema de pesquisa recai sobre a necessidade de se prover ao gestor público tecnologia de apoio à etapa de pesquisa de preços das contratações. O presente estudo utiliza a tecnologia de aprendizagem de máquina (AM) para avaliar as contratações públicas,

³ do termo em inglês *Internet of Things*, que se refere a interligação de objetos e sensores usados no cotidiano à Internet

de forma a indicar a probabilidade de ocorrer sobrepreço. O sobrepreço é definido pela Lei nº 14.133/2021 como “preço orçado para licitação ou contratado em valor expressivamente superior aos preços referenciais de mercado” (BRASIL, 2021, art. 6, LVI) Para o propósito de definir os preços referenciais de mercado são utilizados os dados abertos das contratações, em especial os Dados Abertos do SIASG.

Pretende-se elaborar um banco de dados a ser utilizado no treinamento de algoritmos de IA de forma a poder-se estimar a probabilidade de sobrepreço na pesquisa.

Ao final do estudo, espera-se poder responder a seguinte questão: “Qual é o grau de confiabilidade que se pode conseguir com a utilização de aprendizagem de máquina na detecção de sobrepreço nas pesquisas de preço do Senado Federal?”. Entende-se por confiabilidade a combinação dos indicadores de acurácia, precisão e *recall*, que são explicados nas seções seguintes.

1.3 Justificativa

Fortini e Sherman (2017, p. 42) afirmam que o Brasil entrou definitivamente no movimento mundial de combate à corrupção e reconhecem que o campo das licitações e contratos públicos são especialmente vulneráveis aos desvios de conduta que drenam o erário público, e destacam, ainda, que é necessário que “a Administração abandone a postura reativa, caracterizada pelo agir após a ocorrência do dado, e de fato implemente as vias preventivas de combate às condutas corruptas” (FORTINI; SHERMAN, 2017, p. 38).

Para poder adotar uma postura preventiva, é preciso adotar procedimentos que identifiquem possibilidades de irregularidades antes que efetivamente o dano ocorra. Muitas pesquisas sobre IA têm sido realizadas para identificar irregularidades em diversos processos. A operação Serenata de Amor, que fiscaliza os reembolsos apresentados às Cotas para Exercício de Atividade Parlamentar tem atraído a atenção de diversos pesquisadores, como se pode constatar nos trabalhos de Oliveira (2018), Silva (2018), Nohara e Colombo (2019) e Lima (2019) entre outros.

Também a CGU tem se utilizado de um robô chamado “ALICE, acrônimo para Analisador de Licitações, Contratos e Editais, com o objetivo de identificar automaticamente indícios de irregularidades nas licitações, pelo uso de Inteligência Artificial (IA)” (PANIS, 2020, p.16).

Até mesmo o Poder Judiciário, segundo Junquilho e Maia Filho (2021), tem a sua Inteligência Artificial, o Projeto Victor, parceria entre o Supremo Tribunal Federal (STF) e a Universidade de Brasília (UnB), utilizado para identificar, nos processos que chegam ao STF, a presença do requisito da repercussão geral.

Porém, ainda não está à disposição do gestor público uma ferramenta capaz de auxiliá-lo na elaboração do projeto básico das contratações, e indicar a probabilidade de a pesquisa de preços elaborada estar condizente com o valor justo do objeto. A importância deste estudo está intrinsecamente relacionada ao aumento de segurança que pode ser dado ao gestor público nas suas tarefas de planejamento da contratação.

A criação e disponibilização ao gestor público de ferramenta que o ajude a elaborar projetos com maior acurácia, tendem a minorar a preocupação e a inação de inúmeros gestores, decorrente do temor de ser auditado por um órgão de controle, que está muito mais preparado.

1.4 Objetivos

1.4.1 Objetivo Geral

Desenvolver um sistema automatizado baseado em algoritmos de aprendizagem de máquina capaz de indicar possíveis indícios de sobrepreços durante a fase de pesquisa de preços em contratações do Senado Federal.

1.4.2 Objetivos específicos

- Levantar Bases de Dados existentes sobre as contratações;
- Desenvolver uma ferramenta para obtenção e pré-processamento dos dados necessários ao Treinamento de algoritmos de aprendizagem;
- Disponibilizar repositório público dos dados de contratações públicas utilizado pela ferramenta;
- Levantar as técnicas, métodos e boas práticas mais recentes utilizados nas detecções de fraudes em compras;
- Treinar e utilizar um modelo de Aprendizagem de Máquina na detecção de sobrepreço nas contratações públicas do banco de dados preparado;
- Prover uma ferramenta de avaliação de indícios de sobrepreços para uso efetivo no Senado Federal;
- Avaliar o grau de confiabilidade da ferramenta na predição de sobrepreço.

1.5 Escopo do relatório

O escopo se delimita ao processamento de dados de contratações públicas, com vistas a identificar indícios de sobrepreço em futuras compras do Senado Federal, servindo como ponto de partida para um projeto maior que objetiva prover maior segurança e serenidade ao gestor público na elaboração de pesquisas de preço.

1.6 Organização do Relatório

O relatório está organizado em 5 capítulos nos quais são apresentados os fundamentos teóricos, os procedimentos adotados, os resultados alcançados e as conclusões. A seguir, descreve-se o conteúdo de cada um dos capítulos:

- O Capítulo 1 apresenta a Introdução, o contexto, o problema de pesquisa, a justificativa, os objetivos que se buscam alcançar, o escopo e a organização do relatório;
- O Capítulo 2 apresenta os fundamentos teóricos necessários ao desenvolvimento da pesquisa, que incluem os conceitos de compras públicas, licitações, pesquisa de preços, dados abertos, transparência, inteligência artificial e sua aplicabilidade na Administração Pública Federal;
- O Capítulo 3 descreve a metodologia e relata o desenvolvimento da pesquisa, apresentando as bases de dados existentes, como foi feita a coleta de dados, como foi desenvolvida a ferramenta para obtenção e pré-processamento dos dados, o repositório utilizado para o treinamento da Inteligência Artificial, as técnicas empregadas neste treinamento, o protótipo de ferramenta para avaliação de indícios de sobrepreços e o grau de confiabilidade alcançado;
- O Capítulo 4 apresenta em detalhes os resultados obtidos com a ferramenta, o seu grau de confiabilidade e uma análise desses resultados;
- O Capítulo 5 apresenta as conclusões de todo o trabalho desenvolvido, quais foram as limitações e contribuições da pesquisa e da ferramenta desenvolvida, bem como propõe caminhos e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste estudo foi necessária uma revisão de literatura em três principais aspectos: contratações na APF, dados abertos governamentais e Inteligência Artificial. Apresenta-se uma revisão inicial sobre estes assuntos nas sessões seguintes.

2.1 Contratações na Administração Pública Federal

2.1.1 Licitações públicas

Quando se fala em compras públicas, a regra é a realização de procedimento licitatório. Estes procedimentos até o ano de 2021 estavam previstos na Lei nº 8.666/1993 (BRASIL, 1993) e na Lei nº 10.520/2002 (BRASIL, 2002), que foram substituídas pela Lei nº 14.133, de 1º de abril de 2021 (BRASIL, 2021). O artigo 193 da Lei nº 14.133/2021 prevê que a Administração poderá optar até 30 de dezembro de 2023 por licitar de acordo com o regime atual ou com os anteriores, bastando para tanto a indicação expressa no edital, sendo vedada a aplicação combinada dos regimes. Por este motivo, a pesquisa ainda referencia tanto as leis anteriores como a lei atual.

A pesquisa de preços, definida na Lei nº 8.666/1993 no artigo 15, e na Lei nº 14.133/2021 no artigo 23, é um dos principais artefatos que compõem o processo licitatório, e tem como finalidade estimar o montante de dinheiro que poderá ser gasto na avença. Aferir este montante tem como objetivos a reserva do valor no orçamento e o balizamento de preços para garantir que a Administração está pagando um preço justo e de mercado. Borges Júnior ainda destaca outras funções:

- informar o preço estimado e justo que a Administração está disposta a contratar;
- definir a modalidade licitatória;
- identificar jogo de planilhas;
- conferir maior segurança na análise da exequibilidade da proposta ou de itens da proposta;
- impedir a contratação acima do preço praticado no mercado;
- servir de parâmetro objetivo para julgamento das ofertas apresentadas;
- garantir a seleção da proposta mais vantajosa para a administração.

(BORGES JÚNIOR, 2020).

Na nova Lei nº 14.133/2021, já nas definições constantes do artigo 6º, destacam-se o sobrepreço e o superfaturamento:

LVI - sobrepreço: preço orçado para licitação ou contratado em valor expressivamente superior aos preços referenciais de mercado, seja de apenas 1 (um) item, se a licitação ou a contratação for por preços unitários de serviço, seja do valor global do objeto, se a licitação ou a contratação for por tarefa, empreitada por preço global ou empreitada integral, semi-integrada ou integrada;

LVII - superfaturamento: dano provocado ao patrimônio da Administração, caracterizado, entre outras situações, por:

- a) medição de quantidades superiores às efetivamente executadas ou fornecidas;
- b) deficiência na execução de obras e de serviços de engenharia que resulte em diminuição da sua qualidade, vida útil ou segurança;
- c) alterações no orçamento de obras e de serviços de engenharia que causem desequilíbrio econômico-financeiro do contrato em favor do contratado;
- d) outras alterações de cláusulas financeiras que gerem recebimentos contratuais antecipados, distorção do cronograma físico-financeiro, prorrogação injustificada do prazo contratual com custos adicionais para a Administração ou reajuste irregular de preços; (BRASIL, 2021, art. 6)

Demonstra ainda no seu artigo 11 a preocupação com os custos da aquisição, como se pode destacar no inciso III – “III - evitar contratações com sobrepreço ou com preços manifestamente inexequíveis e superfaturamento na execução dos contratos;” (BRASIL, 2021, art. 11). Apesar dos artefatos na nova lei não terem o mesmo detalhamento que existe na Lei 8.666/1993, o artigo 23 deixa claro que o valor estimado da contratação deve ser compatível com os valores praticados pelo mercado.

O SIASG, instituído pelo art. 7º do Decreto nº1.094, de 23 de março de 1994, é o sistema informatizado de apoio às atividades operacionais do Sistema de Serviços Gerais (SISG). Sua finalidade é integrar os órgãos da Administração Pública Federal direta, autárquica e fundacional (BRASIL, 2020).

O SIASG é o sistema onde são realizadas as operações das compras governamentais dos órgãos integrantes do SISG. O sistema inclui: divulgação e a realização das licitações; emissão de notas de empenho; registro dos contratos administrativos; catalogação de materiais e serviços; e cadastro de fornecedores (BRASIL, 2020).

Os órgãos que não integram o SISG podem utilizar o SIASG, integralmente ou em módulos específicos, por meio de adesão formal para uso do sistema, mediante assinatura de termo de adesão (BRASIL, 2020).

2.1.2 A corrupção nas compras públicas

Segundo Fortini e Motta (2016), a Transparência Internacional aponta que as licitações e contratações públicas são vulneráveis à corrupção. Ressaltam ainda que a prática da corrupção é um problema social universal, apesar de o conceito da corrupção não o ser, já que depende da opção política de cada país.

Silveira (2021) destaca que “A corrupção é um fenômeno que ocorre em todo o mundo, atingindo empresas ou o setor público” e que as fraudes mais comuns em processos licitatórios são:

o superfaturamento, o jogo de planilha, o direcionamento da licitação, a inexigibilidade da licitação, a dispensa de licitação, as fraudes na modalidade pregão, a corrupção dos servidores públicos, o acordo entre empresas, a entrega de material de qualidade inferior ao previsto no edital, as empresas fantasmas, a falsificação de documentos, a simulação de licitação e, por fim, o preço inexequível.(SILVEIRA, 2021).

Dentre estas fraudes mais comuns citadas por [Silveira \(2021\)](#) pode-se perceber que algumas delas podem ser detectadas ou mitigadas por meio de um bom planejamento da contratação e uma pesquisa de preços adequada.

Como citado anteriormente, a Lei Federal nº 14.133/2021, definiu sobrepreço e superfaturamento. Mas como identificar o que realmente é um sobrepreço. Segundo [Lima \(2022\)](#):

A dificuldade do intérprete reside na compreensão do que é “valor expressivamente superior”, bem como no dimensionamento do mercado no qual serão apurados os preços referenciais. De modo geral, em tempos de normalidade, três fatores devem ser considerados na análise de um preço pago pela Administração Pública, para efeito de caracterização de sobrepreço:

- a) o momento temporal em que a aquisição é realizada;
- b) a quantidade de bens ou serviços objeto da contratação (economia de escala); e
- c) as condicionantes logísticas que afetam a entrega do bem ou serviço pelo contratado ao contratante.

Todas essas variáveis são maximizadas em situações de calamidade pública como, por exemplo, na área da saúde, quando, no intervalo de poucos dias, pode ocorrer aumento na demanda de determinados insumos ou equipamentos e desabastecimento de outros, gerando significativas flutuações nos preços de referência.

Segundo [Martins \(2020\)](#) o jogo de planilha consiste em alterações quantitativas dentro do orçamento contratual realizadas por meio de acréscimos, decréscimos, supressões ou inclusões de serviços e materiais, ou mesmo de variações de preços, que modifiquem o equilíbrio econômico-financeiro da avença, sem justificativa adequada e que cause dano ao erário. Chama a atenção também que este tipo de irregularidade muitas vezes está ligada a informações privilegiadas sobre quais itens terão seus quantitativos alterados ao longo da execução do contrato, podendo a licitante atribuir custos elevados a itens a serem aditados e diminuídos para os itens que serão suprimidos.

Segundo [Vilhena et al. \(2017\)](#) a formação de cartéis em licitações públicas é um ajuste entre concorrentes que visa a prejudicar a competitividade do certame e obter o maior lucro possível sobre a Administração Pública sem a interferência de concorrência. As empresas participantes do cartel se beneficiam indevidamente mediante a obtenção de lucros adicionais resultantes da ausência de competição efetiva nos certames licitatórios e da cobrança de preço acima do valor normal de mercado. [Mondo \(2019\)](#) chama a atenção que a formação de cartéis também utiliza outras técnicas como propostas de cobertura, onde através de propostas de valores superiores ou com documentação incompleta, validam a proposta do participante que o cartel definiu que seria o vencedor. O cartel acaba dividindo o mercado através destas práticas.

Preço inexecutável segundo a Lei 8.666/13 ([BRASIL, 1993](#), artigo 48) é aquele que não demonstra sua viabilidade de execução por meio de dados e documentos que comprovem que seus custos e coeficientes de produtividade são compatíveis com o objeto contratado. Aparentemente a apresentação de um preço inexecutável não causaria nenhum dano ao erário, porém, segundo [Rosilho \(2011\)](#) esta definição da lei reforça a tese de que se procurou manipular os procedimentos licitatórios em prol de um grupo em específico, neste caso, buscando impedir a participação de aventureiros nos procedimentos licitatórios que tenderiam a derrubar os preços

das propostas. O que vai ao encontro de Schramm (2018) que chama a atenção da utilização deste artifício no jogo de planilha que se destina a ocultar inconsistências nos preços unitários apresentados pela empresa, normalmente inexecutáveis, sob o véu de uma oferta global razoável e aparentemente vantajosa. Por esse mecanismo, a empresa arbitra valores irrisórios para aqueles itens de menor juízo de exequibilidade da proposta e sujeitando a Administração Pública à contratação superfaturada.

2.2 Dados abertos governamentais, Transparência e Controle Social

Com o advento dos dados abertos governamentais e da transparência, tornou-se acessível à população o ferramental necessário para realizar um controle social mais efetivo. Nas próximas seções serão definidos os conceitos de Dados Abertos Governamentais, Transparência e Controle Social, e como se relacionam.

2.2.1 Dados Abertos Governamentais

Possamai (2016) define Dados Abertos Governamentais como:

Dados abertos governamentais (*open government data* (OGD)) são dados públicos, publicados na Web em formato aberto, estruturado e compreensível logicamente, de modo que qualquer pessoa possa livremente acessar, reutilizar, modificar e redistribuir, para quaisquer finalidades, estando sujeito a, no máximo, exigências de creditar a sua autoria e compartilhar sob a mesma licença.

O TCU (2015) define que são dados abertos governamentais os que atendem a oito requisitos listados abaixo:

1. são completos;
2. são primários;
3. estão atualizados;
4. são acessíveis;
5. são processáveis por máquina;
6. não é necessária a identificação do interessado para acessá-los;
7. são disponibilizados em formatos não proprietários;
8. são livres de licenças.

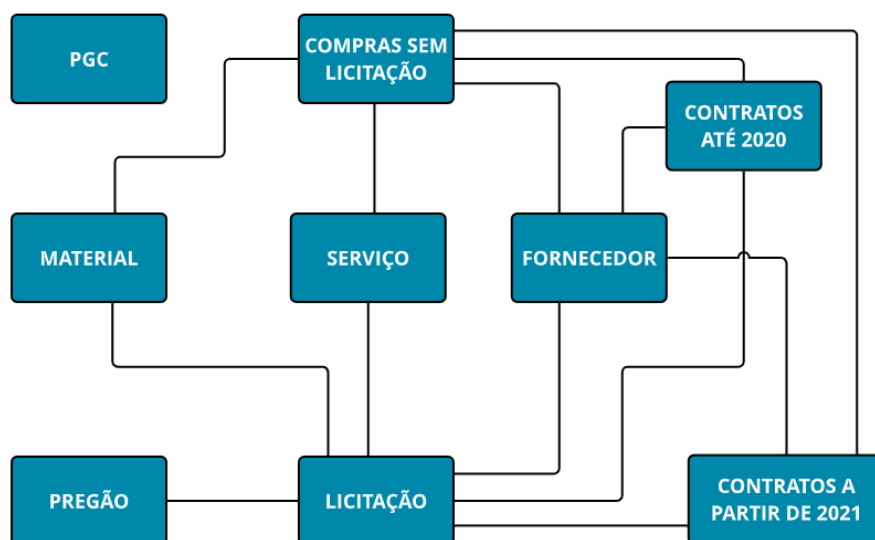
Dentre os dados abertos governamentais disponíveis hoje, dois serão de suma importância neste estudo, os Dados Abertos do SIASG e os Dados Abertos do Senado Federal.

Os Dados Abertos do SIASG são os dados disponibilizados pelo sistema que operacionaliza as compras do Governo Federal. Como informado no *site* da API¹ de Compras Governamentais ele é composto por diversos módulos, os quais relacionamos abaixo:

o catálogo de materiais e serviços (CATMAT / CATSER), o cadastramento e divulgação da licitação (SIDECA, Divulgação), as intenções de registros de preços (IRP), o cadastramento dos fornecedores (SICAF), a realização das licitações (Compras governamentais, Sessão Pública, RDC), o resultado das licitações (SISPP, SISRP), os empenhos de pagamentos (SISME) e o registro e gestão dos contratos (SICON). (BRASIL, 2021).

Por meio da utilização dos dados abertos do sistema se tem acesso aos dados dos fornecedores, catálogo de materiais, catálogo de serviços, licitações, contratos, compras sem Licitação e ao plano anual de contratações (representado no modelo pela sigla PGC de Planejamento e Gerenciamento de Contratações). O modelo de dados do sistema é mostrado na Figura 1.

Figura 1 – Modelo de dados do SIASG



Fonte: *Site* da API de Compras Governamentais (BRASIL, 2021b)

O acesso aos dados é feito através de URLs², e retorna os recursos desejados em diversos formatos, incluindo JSON³ que será o formato utilizado pela pesquisa. A API ainda possui recursos de classificação de conteúdo e possibilidade de escolha dos parâmetros a serem utilizados na busca.

Já os Dados Abertos do Senado Federal (SENADO FEDERAL, 2022) são disponibilizados na Internet em um portal de dados abertos que se divide em dois grandes grupos, Administrativo e Legislativo.

¹ API é um conjunto de rotinas e padrões de programação para acesso a um aplicativo de software ou plataforma baseado na Web. A sigla API refere-se ao termo em inglês *Application Programming Interface* que significa em tradução para o português “Interface de Programação de Aplicativos” (NASCIMENTO, 2014)

² URL significa *Uniform Resource Locator* (Localizador Uniforme de Recursos, em tradução livre). É um termo utilizado para descrever o endereço de um recurso na Internet. (DUTRA, 2023)

³ JSON, é um acrônimo de JavaScript Object Notation, é um formato compacto, de padrão aberto independente, de troca de dados simples e rápida entre sistemas que utiliza texto legível a humanos, no formato atributo-valor (natureza auto-descritiva) (JSON, 2022)

O Grupo Administrativo se subdivide nos seguintes temas:

- a) Gestão de Pessoas;
- b) Orçamento do Senado;
- c) Senadores;
- d) Contratações.

O Grupo Legislativo se subdivide nos seguintes temas:

- a) Projetos e Matérias;
- b) Senadores;
- c) Plenário;
- d) Composição;
- e) Comissões;
- f) Legislação.

Os dados abertos de interesse deste estudo estão no Grupo Administrativo e podem ser encontrados dentro do item "Contratações", subitem "Licitações e Contratos". No subitem "Licitações e Contratos" podem ser encontrados dados relativos a:

- a) Aditivos;
- b) Atas de Registros de Preços - ARP;
- c) Contratos;
- d) Empresas Contratadas;
- e) Itens;
- f) Licitações;
- g) Notas de Empenho com Força de Contrato.

2.2.2 Transparência e Controle Social

Mendes, Oleiro e Quintana (2008) destacam que, nos últimos anos, a Administração Pública passou por diversas transformações, entre elas a migração de um modelo burocrático para o gerencial, voltado não apenas a aspectos formalísticos, mas também na agregação de valor aos serviços disponíveis ao cidadão. As mudanças incluem maior participação popular, mas que para isso, “a administração pública precisa apresentar soluções pragmáticas, como transparência, responsabilidade (accountability) e controles eficazes” (MENDES; OLEIRO; QUINTANA, 2008).

Os autores procuraram demonstrar que o inimigo nº 1 da corrupção é a transparência na gestão pública, sendo esta uma das vertentes de combate à mesma. Chamam atenção ao fato que a sociedade brasileira passou a se organizar e, em consequência, exigir maior transparência e *accountability* na gestão da *res publica*, fortalecendo o controle social.

Nessa trajetória, não basta que informações sejam disponibilizadas ao cidadão por meio de dados abertos, mas que essas espelhem fidedignamente os atos de gestão praticados pela administração.

Pinho e Gouveia (2019) destacaram que, para que a sociedade possa utilizar os serviços e dados disponibilizados, é necessária a divulgação, orientação e educação da população sobre como interagir com o governo. Cada cidadão tem o dever de ser útil e contribuir para o combate à corrupção. O momento é oportuno, pois o uso da tecnologia da informação de forma adequada amplia a transparência e a participação social. A Internet se apresenta como uma aliada, possibilitando à população o controle social.

Entretanto Batista, Rocha e Santos (2020) procuraram demonstrar que o compromisso institucional com a transparência pública diminui a corrupção e a má gestão governamental com base no argumento geral de que os governantes se comportam melhor quando são observados. Existe uma expectativa teórica e prática de que a transparência reduza a corrupção e melhore a performance do governo. Porém, os resultados do trabalho indicaram que a transparência não diminui a ocorrência de irregularidades na gestão local, nem reduz a má gestão e a corrupção. Os autores verificaram contudo que:

Sobre as variáveis que apresentam resultado estatisticamente significativo, encontramos que a qualidade da burocracia diminui o número de irregularidades, má gestão e corrupção. Ou seja, quanto maior a escolaridade dos funcionários na administração local, menor o número de irregularidades identificadas (BATISTA; ROCHA; SANTOS, 2020).

Neste sentido, não se pode dizer que a transparência no governo, de forma mais ampla, não importa para o combate à corrupção e para a melhoria da gestão pública, mas é preciso ter em mente que os efeitos das políticas de transparência tendem a ser graduais, indiretos e difusos, e que ainda existe um longo caminho a se percorrer. Pode-se porém destacar que o envolvimento concreto da burocracia pode resultar em melhorias de gestão e diminuição da corrupção como ressaltado por Batista, Rocha e Santos (2020), por isso, a divulgação de dados abertos possibilita que o controle social seja exercido também pela própria burocracia, o que será demonstrado nas próximas seções.

2.3 Inteligência Artificial, seu uso na Administração Pública Federal e métricas de avaliação dos modelos

A Inteligência Artificial está evoluindo cada vez mais ao longo do tempo e sendo utilizada em várias áreas do conhecimento. Nesta seção se definirá o que vem a ser Inteligência Artificial. Serão explicitados alguns de seus usos dentro da APF e que métricas podem ser utilizadas para avaliar os modelos.

2.3.1 Inteligência Artificial

Segundo Raschka, Patterson e Nolet (2020), a IA é um subcampo da ciência da computação que está focada no projeto de programas de computador e máquinas capazes de executar tarefas em que os humanos normalmente são bons. Destacam que no meio do século passado o aprendizado de máquina ou *Machine Learning* (ML) emergiu como uma das áreas da IA. Apesar de ainda estar profundamente entrelaçada com as pesquisas de IA, o aprendizado de máquina está sendo considerado um campo científico com foco no projeto de modelos de computador e algoritmos que podem realizar tarefas específicas, muitas vezes envolvendo reconhecimento de padrões, sem a necessidade de serem explicitamente programados.

A utilização de aprendizado de máquina tem aumentado dia a dia e diversos estudos tem utilizado esta técnica para prever preços, como por exemplo Wolstad (2020).

O Aprendizado de Máquina pode ser dividido em quatro tipos de acordo com as técnicas utilizadas. Os tipos são: supervisionado, não-supervisionado, semi-supervisionado e por reforço.

Géron (2021) define estes tipos da seguinte forma:

- a) Aprendizado supervisionado é a técnica onde são utilizados no treinamento dados rotulados, ou seja, além dos dados propriamente ditos, são fornecidos também os resultados esperados.
- b) Aprendizado não-supervisionado é a técnica onde os dados de treinamento não são rotulados e o sistema tenta aprender sem orientação.
- c) Aprendizado semi-supervisionado é a técnica de se utilizar algumas instâncias rotuladas e uma grande quantidade de instâncias não rotuladas, já que rotular os dados consome tempo e dinheiro.
- d) Aprendizado por reforço é a técnica onde o sistema de aprendizado, chamado de agente, assiste o ambiente, seleciona e executa ações obtendo recompensas ou penalidades aprendendo sozinho qual é a melhor estratégia. Muito utilizado em jogos.

Como pode-se perceber, a maioria das técnicas estão relacionadas a rotulagem. Quanto à questão dos rótulos, Chandola, Banerjee e Kumar (2009) explicam que eles são o que definem uma instância como normal ou anômala e que deve-se observar que a obtenção de dados rotulados que sejam precisos e representativos de todos os tipos de comportamento é muitas vezes proibitivamente caro. Destacam que a rotulagem geralmente é feita manualmente por um especialista humano e, portanto, requer um esforço substancial para obter o conjunto de dados de treinamento rotulado, e que esta definição está intrinsecamente ligada ao conhecimento do especialista.

Na pesquisa serão utilizadas técnicas de aprendizado não-supervisionado mas especificamente algoritmos para detecção de anomalias, que indicarão os dados que divergem do esperado.

Detecção de anomalias, também conhecida como detecção de *outliers* ou detecção de novidades, é definida por Pang *et al.* (2021) como o processo de detecção de instâncias de dados

que se desviam significativamente da maioria das instâncias de dados.

De acordo com Eega (2021) qualquer método que identifique os *outliers* em um conjunto de dados, ou seja, que determine os que não pertencem ao conjunto, é conhecido como detecção de anomalia. Segundo ele, as anomalias podem indicar atividade de rede inesperada, revelar um sensor com defeito, destacar dados que precisam ser limpos antes da análise ou mesmo apontar uma fraude.

Zhao, Nasrullah e Li (2019) apresentaram uma ferramenta *open-source* para *Python* para detecção de anomalias com o nome de *Python Outlier Detection* (PyOD) que implementava 20 diferentes algoritmos. Sua versão atual, segundo seu repositório (ZHAO, 20–), inclui mais de 40 algoritmos de detecção, desde o clássico LOF até alguns mais recentes como ECOD. Desde 2017, tem sido utilizado com sucesso em numerosas pesquisas acadêmicas e produtos comerciais. No Quadro 12 do Anexo A estão listados os algoritmos individuais de detecção disponíveis hoje.

Alguns desses modelos serão avaliados e testados durante a execução da pesquisa, e os testes e resultados serão detalhados na seção 3.5

Quanto às etapas utilizadas para desenvolver o modelo, na visão de Kuhn e Johnson (2013 apud SPEDICATO; DUTANG; PETRINI, 2018) independentemente do processo de construção do modelo, os seguintes passos serão seguidos:

- a) Pré-processamento dos Dados: Tarefa que consiste em limpar os dados, selecionar os preditores, verificar a necessidade de utilização de transformações e selecionar as variáveis que serão utilizadas no estágio de modelagem.
- b) Separação dos dados: Neste passo, os dados são divididos em dois conjuntos, um que será utilizado no treinamento e na validação do modelo e um outro conjunto que será utilizado no teste. É muito importante esta separação, pois a utilização de um mesmo conjunto no treinamento e no teste causa um viés do modelo conhecido como *overfitting* que ocorre quando um modelo tem um bom desempenho no conjunto testado, mas apresenta um desempenho significativamente menor em dados não utilizados anteriormente.
- c) Ajustes de hiperparâmetros: A maioria das famílias de modelos necessita que um ou mais parâmetros de ajuste sejam definidos com antecedência para adequar o modelo de maneira exclusiva. Nesta etapa se realizará a definição destes parâmetros de ajuste.
- d) Seleção do Modelo: Neste passo é realizada uma avaliação de qual modelo dentre os testados tem o melhor desempenho em um conjunto de teste, sendo o modelo que melhor pode inferir o resultado em dados não utilizados.

2.3.2 Uso da Inteligência Artificial no controle dos gastos públicos

Nos últimos anos tem aumentado a utilização da Inteligência Artificial em diversos ramos de atividades. O controle dos gastos públicos não é uma exceção, e pode-se destacar dois grandes

projetos: a “Operação Serenata de Amor” e a ferramenta “ALICE (Analisador de Licitações, Contratos e Editais da CGU)”.

Segundo o *site* do projeto, a “Operação Serenata de Amor” é

um projeto de tecnologia que usa inteligência artificial para auditar contas públicas e auxiliar no controle social. A ideia surgiu do cientista de dados Irio Musskopf, como forma de participar ativamente do processo democrático, fiscalizando os gastos públicos (OPEN KNOWLEDGE BRASIL, 2015-).

Segundo Coutinho e Freitas (2021) o site foi desenvolvido com dados públicos abertos e possibilita que qualquer cidadão fiscalize a utilização da Cota para o Exercício da Atividade Parlamentar (CEAP), que é destinada aos deputados federais para reembolso de diversas despesas, como alimentação, transporte, hospedagem e atividades educacionais e culturais.

O projeto criou uma IA, denominada Rosie, que analisa os gastos reembolsados, identificando suspeitas e incentivando o controle social das despesas públicas por meio de *posts* em redes sociais como *Facebook* e *Twitter*. No *Twitter* foi criado o perfil @RosieDaSerenata, ilustrado na Figura 2, que também está presente no *Facebook* (OPEN KNOWLEDGE BRASIL, 20-a).

Figura 2 – A Robô Rosie



Fonte: (OPEN KNOWLEDGE BRASIL, 2015-)

Além disso, a fim de facilitar o entendimento dos cidadãos, foi criado o Jarbas, um painel que facilita a consulta de informações a respeito dos reembolsos das despesas dos parlamentares, ilustrado Na Figura 3.

No painel Jarbas (OPEN KNOWLEDGE BRASIL, 20-b) são listados todos os pedidos de reembolsos, que podem ser filtrados pelos seguintes critérios: por reembolso suspeito, por nota fiscal digitalizada, por pagamento direto ao fornecedor, por estado, por ano, por mês, por tipo do documento fiscal e por sub-cota.

O segundo dos grandes projetos destacados é a ALICE - Analisador de Licitações, Contratos e Editais da CGU, que Panis (2020) analisou em seu trabalho:

Em 2014, por exemplo, a Controladoria-Geral da União (CGU), iniciou o desenvolvimento da ferramenta Alice, acrônimo para Analisador de Licitações, Contratos e Editais, com o objetivo de identificar automaticamente indícios de irregularidades

Figura 3 – O Painei Jarbas

Jarbas Dashboard					
Início > Câmara dos Deputados - Cota para Exercício da Atividade Parlamentar > Reembolsos					
Selecione reembolso para visualizar					
<div> <input type="text"/> <input type="button" value="Buscar"/> </div>					
REEMBOLSO	SOCIAL	NOME DO PARLAMENTAR	ANO	SUBQUOTA TRANSLATED	FORNECEDOR
7441613		Pedro Lucas Fernandes	2022	Hospedagem ,exceto do parlamentar no distrito federal	OLIVEIRA & FILHOS LTDA 09.594.488/0001-05
7441611		Heitor Freire	2022	Combustíveis e lubrificantes	POSTO FIVE STARS 00.327.248/0001-61
0		Juninho do Pneu	2022	Serviços postais	CORREIOS - SEDEX CONVENCIONAL 00.000.000/0000-07
0		Da Vitoria	2022	Serviços postais	CORREIOS - SEDEX CONVENCIONAL 00.000.000/0000-07

FILTRO

- Por reembolso suspeito
 - Todos
 - Sim
 - Não
- Por nota fiscal digitalizada
 - Todos
 - Sim
 - Não
- Por pagamento direto ao fornecedor

Fonte: (OPEN KNOWLEDGE BRASIL, 2015-)

nas licitações, pelo uso de Inteligência Artificial (IA). Essa ferramenta entra no site do Comprasnet e coleta arquivos e dados de todas as licitações e de todas as atas de realização de pregão publicadas para identificar irregularidades em licitações e pregões eletrônicos da administração pública federal a partir do texto do edital. Esta inovação tem possibilitado a avaliação tempestiva e automatizada de editais de licitação e atas de pregão, com a identificação de indícios de irregularidades, fraudes, desvios e desperdícios de recursos públicos, possibilitando ações de controle mais eficientes e efetivas.

Costa e Bastos (2020) chamam a atenção que em maio de 2016 foi concretizada uma parceria entre o TCU e a CGU para que o sistema ALICE pudesse ser implementado no controle externo.

Os autores destacam que ALICE é um sistema que busca diariamente possíveis inconsistências nos editais de licitação e atas de pregão eletrônico publicados no Portal de Compras do Governo Federal – Comprasnet. Sua análise dos editais é feita no mesmo dia de sua publicação e mensagens contendo os apontamentos e riscos detectados são encaminhados aos auditores do TCU.

Costa e Bastos (2020) ressaltam a importância da IA para o incremento na eficácia das análises de forma ampla e tempestiva, de milhões de documentos, com vistas a detectar correlações e apontar alertas, alcançando resultados que não seriam possíveis sem a utilização desta tecnologia.

Narram ainda os autores que o sucesso alcançado por ALICE acabou impulsionando o surgimento de mais Robôs no TCU, destacando os seguintes:

- MONICA (Monitoramento Integrado para o Controle de Aquisições) é um painel que contempla informações relativas às aquisições efetuadas pela esfera federal. As informa-

ções do Monica são dispostas por unidades administrativas de serviços gerais - UASGs, por fornecedores e por materiais/serviços adquiridos. Na utilização desse painel são aplicados filtros para obtenção de dados específicos e também são efetuadas análises mais aprofundadas, com visão analítica e exportação de dados para o sistema Microsoft Excel.

- SOFIA (Sistema de Orientação sobre Fatos e Indícios para o Auditor) é uma ferramenta que provê informações ao auditor no momento da elaboração de relatórios. Por meio desse sistema é feita, a partir de um botão no editor de textos *Word*, a revisão nos relatórios de auditoria e instruções em geral, além de ser efetuada busca de correlação das informações neles constantes, captando as informações associadas aos CNPJs indicados no documento e verificando se já foram aplicadas sanções àquelas empresas ou se elas já foram responsabilizadas em outros processos em trâmite no TCU, ou, ainda, elenca os contratos já pactuados por essas empresas com órgãos ou entidades da Administração Pública Federal, entre outras informações.
- ADELE (Análise de Disputa em Licitações Eletrônicas), que traz um painel da dinâmica de cada pregão eletrônico, sendo efetuados filtros que permitem que sejam analisados todos os lances de modo cronológico e todas as informações acerca das empresas participantes (composição societária, ramo de atuação etc.), além de possibilitar a identificação da utilização por mais de uma licitante de um mesmo endereço *Internet Protocol* (IP), o que caracterizaria um conluio.
- ÁGATA (Aplicação para Geração de Análise Textual Acelerada), desenvolvida pela Secretaria de Gestão de Informações para o Controle Externo do TCU, se baseia em algoritmos de aprendizado de máquina para refinar os alertas emitidos pela ALICE.
- CARINA (Crawler e Analisador de Registros da Imprensa Nacional). Em abril de 2020 os auditores federais de controle externo começaram a utilizar esta ferramenta que, diariamente, rastreia possibilidades de inconsistências nas informações de aquisições governamentais extraídas de publicações no Diário Oficial da União, de maneira similar à testagem que ALICE faz nos editais publicados no Portal de Compras do Governo Federal.

2.3.3 Métricas de avaliação dos modelos

Para avaliar os modelos que serão construídos na aplicação serão necessárias métricas. Segundo Géron (2021) avaliar um classificador é bem mais complicado que avaliar um regressor, e para sua avaliação ele apresenta a matriz de confusão e as métricas acurácia, precisão, *recall* e f_1 score.

Para se montar uma matriz de confusão é necessário apurar a quantidade dos seguintes valores:

- Verdadeiro positivo (TP, do inglês *True positive*): quando o método diz que o valor é positivo e, ao verificar a resposta, vê-se que o valor era realmente positivo;
- Verdadeiro negativo (TN, do inglês *True negative*): quando o método diz que o valor é negativo e, ao verificar a resposta, vê-se que o valor era realmente negativo;
- Falso positivo (FP): quando o método diz que o valor é positivo, mas ao verificar a resposta, vê-se que o valor era negativo;
- Falso negativo (FN): quando o método diz que o valor é negativo, mas ao verificar a resposta, vê-se que o valor era positivo;

Os valores são alocados em uma matriz 2x2 conforme ilustrado na Figura 4.

Figura 4 – Matriz de confusão

Valores reais	Normal	Verdadeiro Negativo TN	Falso Positivo FP
	Anomalia	Falso Negativo FN	Verdadeiro Positivo TP
		Normal	Anomalia
		Valores preditos	

Fonte: elaboração própria

A matriz de confusão, segundo (GÉRON, 2021), fornece muita informação, mas pode-se precisar de métricas mais concisas. Neste caso se pode utilizar as métricas de acurácia, precisão, recall e f_1 score.

Acurácia é uma métrica que avalia o percentual de acertos. Pode ser obtida pela razão entre a quantidade de acertos e o total de registros da amostra. Seu objetivo é indicar quão certo está o modelo e é calculada pela Equação 2.1:

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

Precisão é uma métrica que avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores identificados como positivos. Indica a qualidade do modelo em identificar os resultados positivos e é calculado pela Equação 2.2:

$$Precisão = \frac{TP}{TP + FP} \quad (2.2)$$

Recall, Revocação ou Sensibilidade, avalia a capacidade do método de detectar com sucesso resultados classificados como positivos. Ela indica o percentual dos positivos identificados dentro do universo total de valores realmente positivos e pode ser obtida pela Equação 2.3:

$$Recall = \frac{TP}{TP + FN} \quad (2.3)$$

F₁score é uma média harmônica calculada com base na Precisão e no *Recall*. Ela pode ser obtida com base na Equação 2.4:

$$F_1score = 2 \times \frac{(Precisão \times Recall)}{(Precisão + Recall)} \quad (2.4)$$

2.4 Recapitulando

Neste capítulo foi apresentada a fundamentação teórica que embasa o trabalho. Falou-se sobre a necessidade de utilização de processos licitatórios nas compras governamentais e a importância da pesquisa de preços. Foram detalhados o que são Dados Abertos Governamentais, com destaque para os dados do SIASG e da área de contratações do Senado Federal, ressaltando a importância dos dados abertos para a transparência e o controle social. O capítulo conclui com a definição de Inteligência Artificial, seus modelos, sua utilização na APF e métricas de avaliação dos modelos.

3 METODOLOGIA E DESENVOLVIMENTO

Neste capítulo será descrita a metodologia e o desenvolvimento da pesquisa, relatando os procedimentos adotados na elaboração da ferramenta. Serão apresentadas as Bases de Dados existentes, a ferramenta para obtenção e pré-processamento dos dados, a descrição do repositório público dos dados disponibilizado, as técnicas, métodos e boas práticas mais recentes utilizados na detecções de fraudes em compras e o modelo de Inteligência Artificial utilizado na detecção de sobrepreço nas contratações do banco de dados preparado. Ao final será apresentada a ferramenta de avaliação de indícios de sobrepreços e seu grau de confiabilidade.

3.1 Metodologia

Nesta pesquisa são treinados algoritmos de Aprendizagem de Máquina para avaliar se há indícios de sobrepreços em pesquisas de preço de contratações públicas. A pesquisa se iniciou com levantamentos documentais onde são coletados o embasamento teórico tanto na área de contratações, como na área de Inteligência Artificial, mais especificamente AM, além de serem levantadas as bases de dados existentes sobre contratações.

A segunda fase realizada foi a de coleta de dados, que se concentrou principalmente na utilização de dados abertos disponibilizados pela Administração Pública quanto às contratações, em especial os Dados Abertos do SIASG. Foi realizada nesta fase uma análise quantitativa dos dados, com a intenção de identificar dentre os objetos contratados o que apresenta uma massa significativa para servir de subsídio ao treinamento e avaliação da IA. Ao final desta fase, foi disponibilizado o repositório de dados de contratações utilizado para o treinamento dos algoritmos de AM.

Adotou-se a abordagem de pesquisa exploratória para se conhecerem as técnicas, os métodos e as boas práticas utilizados nas detecções de fraudes em compras, o que serviu de base para a pesquisa experimental, na qual são identificados, durante o treinamento dos modelos, quais apresentam os melhores resultados para a solução do problema. Nesta fase do processo, foi definido, dentre os modelos de AM disponíveis, os mais apropriados ao conjunto de dados específico.

O treinamento de AM é um experimento recursivo, onde em cada iteração é aprimorada a acurácia do modelo por meio de refinamentos e ajustes sucessivos. É uma parte da pesquisa aplicada, uma vez que esta busca gerar conhecimentos a partir da prática em problemas específicos. Tem ainda um caráter explicativo, pois procura identificar os fatores que determinam ou contribuem para a ocorrência do fenômeno.

Considerou-se para a entrega final o protótipo da ferramenta e o relatório técnico que descreve seu desenvolvimento e avaliação. Tanto os bancos de dados quanto o *software* desenvolvido

foram disponibilizados em repositório público no *site* desta pesquisa.¹

3.2 Bases de Dados Existentes

Durante as pesquisas de bases de dados existentes, duas se destacaram: a base de dados do SIASG e a base de dados de contratações do Senado Federal.

Como informado na seção 2.2, o SIASG mantém registro das compras e contratações firmadas pelo Poder Executivo da Administração Pública Federal e por todas as instituições que utilizam o sistema Comprasnet, entre elas o Senado Federal.

O acesso aos dados é feito através da utilização da API de compras governamentais (BRASIL, 2021), que para se ter acesso é necessário conhecer o endereço ou URL². O Anexo B contém a descrição das chamadas e parâmetros desta API que são de interesse da pesquisa.

Já as pesquisas de preços no Senado Federal utilizam o *checklist* apresentado no Anexo C para verificar se todos os dados estão corretos, e sua elaboração utiliza a planilha de preços presente no Anexo D. Quanto aos dados das contratações, já que o Senado Federal utiliza o Compranet, eles já se encontram na base do SIASG, motivo pelo qual se utilizou para obtenção dos dados apenas a API de compras governamentais.

3.3 Ferramenta para obtenção e pré-processamento dos dados necessários ao Treinamento da Inteligência Artificial

A coleta inicial dos dados começou com a tentativa de se recuperar todas as licitações cadastradas no sistemas Comprasnet a partir da utilização da API de dados abertos com o seguinte comando:

<http://compras.dados.gov.br/licitacoes/v1/licitacoes.json?offset=0>

Quando da primeira execução, o total de licitações era de 1.355.292 (hum milhão, trezentos e cinquenta e cinco mil, duzentas e noventa e duas). A cada chamada a função retorna no máximo 500 (quinhentos) registros, o que demandava 2.711 chamadas apenas para recuperar as licitações. Durante a execução inicial do programa de recuperação de dados, pôde-se perceber que diversos erros ocorriam e a carga destes dados era interrompida.

Buscou-se reduzir a quantidade de dados a serem recuperados para agilizar o processo. Para isto, foram utilizados parâmetros, como por exemplo limitar os dias das licitações por meio da data da publicação, como se pode observar no exemplo a seguir que foi utilizado para tentar retornar todas as licitações do ano de 2022:

¹ <https://github.com/terraneo/IA-PP-Mestrado>

² URL significa *Uniform Resource Locator* (Localizador Uniforme de Recursos, em tradução livre). É um termo utilizado para descrever o endereço de um recurso na Internet. (DUTRA, 2023)

`http://compras.dados.gov.br/licitacoes/v1/licitacoes.json?data_publicacao_min=2022-01-01 & data_publicacao_max=2022-12-31`

Este procedimento não funcionou, gerando inúmeras vezes um erro interno do servidor (Erro 500) com a descrição *java.lang.NullPointerException*. Optou-se em incluir mais um parâmetro que é a UASG, ou seja, foi executada uma chamada para cada unidade por ano, por meio de um comando como o do exemplo abaixo, que retorna todas as licitações do ano de 2022 para o Centro de Ciências Agrárias da UFES:

`http://compras.dados.gov.br/licitacoes/v1/licitacoes.json? uasg = 153050 & data_publicacao_min= 2022-01-01 & data_publicacao_max= 2022-12-31`

Estavam ativas 13.381 unidades, e executou-se a recuperação das licitações de todas as unidades. No entanto, percebeu-se que nenhuma unidade retornava mais do que 500 licitações em um ano e ao analisar um desses arquivos, verificou-se que na verdade a licitação aparecia repetida, pois existia um registro para cada item de uma licitação. Neste momento constatou-se que a chamada da API de consulta de licitações apresenta um erro de codificação, que quando a soma dos itens das licitações que atendem aos parâmetros é maior do que 500, somente os primeiros 500 registros são apresentados, e não é possível paginar.

Desta forma optou-se por criar um procedimento para a leitura de todos os arquivos anuais e quando um deles apresentava um total de 500 registros, executava-se uma chamada para cada mês. Por exemplo, a chamada a seguir retorna as licitações do mês de março:

`http://compras.dados.gov.br/licitacoes/v1/licitacoes.json? uasg = 153050 & data_publicacao_min = 2022-03-01 & data_publicacao_max = 2022-03-31`

Este procedimento não foi suficiente, sendo necessário criar um algoritmo para quebrar um mês em cada um dos dias, como no exemplo a seguir:

`http://compras.dados.gov.br/licitacoes/v1/licitacoes.json? uasg = 154054 & data_publicacao = 2022-04-28`

Mesmo dividindo o período de apuração em dias para algumas UASGs, isso não foi suficiente para determinados dias. Foi utilizada então, a classificação do material como novo parâmetro, como no exemplo abaixo da Fundação Universidade Federal/MS:

`http://compras.dados.gov.br/licitacoes/v1/licitacoes.json?uasg=154054 & data_publicacao=2022-04-28 & item_material_classificacao=6505`

3.4 Repositório de dados de contratações

Após a fase de coleta de dados foi feita uma primeira análise quanto à quantidade de itens que se conseguiu recuperar das licitações e contratos do sistema SIASG, conforme apresentado no Quadro 1.

Quadro 1 – Número de registros recuperados dos dados abertos do SIASG

Entidade	quantidade
Contratos	51.162
Licitações	42.993
Itens de contratos	123.824
Itens de licitações	861.733

Fonte: Elaboração própria

Avaliando os dados disponíveis das licitações e contratações foi realizada uma seleção das características de interesse da pesquisa e elaborado o repositório a ser utilizado no treinamento da IA. Todos os itens que não possuíam o valor unitário ou total foram desprezados, o que fez com que caísse substancialmente a quantidade de itens, de aproximadamente 985.557 (novecentos e oitenta e cinco mil, quinhentos e cinquenta e sete) para o total de 65.569 (sessenta e cinco mil, quinhentos e sessenta e nove).

As características do repositório de dados de contratações públicas estão descritas no Quadro 2 e os respectivos atributos no Quadro 3.

Quadro 2 – Características do repositório

Característica	Valor
Número de registros (instâncias)	65.569
Número de atributos	10
Número de valores nulos das características	0
número de atributos numéricos	4
número de atributos categóricos	6

Fonte: Elaboração própria

Quadro 3 – Descrição dos atributos do repositório

Atributo	Descrição	Tipo	Categórico
licitacao_contrato	Determina se é um item de licitação ou contrato. 0 se contrato, e número do item se licitação	Inteiro	Sim
id	identificador da licitação ou contrato	Inteiro	Sim

Continua na próxima página

Quadro 3 – Continuação

Atributo	Descrição	Tipo	Categórico
data	Data da assinatura do contrato ou da publicação da licitação	Data	Não
catmat_id	identificador da categoria do material do item	Inteiro	Sim
quantidade	quantidade do item adquirido	Inteiro	Não
unidade	unidade de aquisição do item na licitação	Texto	Sim
valor_unitario	valor unitário do item	Numérico	Não
valor_total	valor total do item	Numérico	Não
uf_uasg	Unidade da federação da UASG que adquiriu o item	Texto	Sim
uf_fornecedor	Unidade da federação do fornecedor do item	Texto	Sim

Fonte: Elaboração própria

Os dados da pesquisa foram disponibilizados em um repositório público hospedado no *GitHub* (TERRA NETO, 2023) ³

3.5 Técnicas, Métodos e Boas práticas na detecção de anomalias

A primeira dificuldade encontrada foi determinar o algoritmo a ser utilizado para a detecção de sobrepreço. Como os dados do Comprasnet não apontam se ocorreu ou não sobrepreço nas compras e a análise compra a compra para poder rotular como sobrepreço é uma tarefa muito onerosa, a primeira definição foi a de que seriam utilizados algoritmos de Aprendizado Não Supervisionado. Como existem diversos modelos descreve-se a seguir como se deu a definição do modelo.

Géron (2021) apresenta em seu livro os seguintes algoritmos para detecção de anomalias: Mistura de Gaussianas (GMM), *Principal Component Analysis* (PCA), Fast-MCD, Floresta de Isolamento (iForest), Fator de *Outlier Local*(LOF) e *One-Class Support Vector Machine* (OCSVM). Incluindo esses, encontram-se disponíveis no PyOD 14 (quatorze) algoritmos não supervisionados, como apresentados no Quadro 12 do Anexo A. Para poder selecionar o modelo se verificou que Han *et al.* (2022) avaliou os algoritmos presentes na biblioteca PyOD e os classificou. A avaliação foi feita sob três perspectivas: nível de supervisão, tipos de anomalias e nível de ruído e corrupção dos dados.

No nível de supervisão os algoritmos foram divididos em três categorias: não supervisionado, supervisionado e semi-supervisionado.

Quanto ao tipo de anomalia foram divididos em:

- a) Local: são as anomalias que diferem das instâncias vizinhas mais próximas;

³ No endereço <https://github.com/terraneto/IA-PP-Mestrado>

- b) Global: são as anomalias que diferem dos dados normais gerados a partir de uma distribuição uniforme;
- c) Anomalias de dependência: referem-se às amostras que não seguem a estrutura de dependência que dados normais seguem, ou seja, as características de entrada de anomalias de dependência são assumidas como sendo independentes um do outro; e
- d) Anomalias agrupadas: também conhecidas como anomalias de grupo ou *clustered*, são as que exibem características semelhantes.

Foram avaliados ainda quanto à robustez frente ao nível de ruído e corrupção de dados.

Os problemas avaliados dos dados foram:

- a) Anomalias duplicadas: quando no conjunto de dados certas anomalias se repetem, a detecção destas anomalias fica prejudicada em diversos algoritmos, motivo pelo qual o autor avaliou o comportamento dos algoritmos em relação a este problema;
- b) Características irrelevantes: segundo [Han et al. \(2022\)](#) os dados tabulares podem conter recursos irrelevantes causados por ruído de medição ou unidades de medição inconsistentes, onde essas dimensões ruidosas podem esconder as características de dados de anomalias e, assim, tornar o processo de detecção mais difícil, motivo pelo qual este item foi avaliado; e
- c) Erros de anotação: erros que acontecem na rotulação dos dados como anomalia ou não, o que contamina o desempenho do algoritmo. Como este erro não afeta métodos não supervisionados, pois eles não usam nenhum rótulo, não será um dos critérios de escolha.

Na avaliação conduzida por [Han et al. \(2022\)](#), chegou-se a conclusão de que nenhum método não supervisionado é estatisticamente melhor do que o outro, mas cada um dos métodos responde diferentemente aos tipos de anomalias existentes nos dados.

Como não está claro o tipo de anomalia que existe nos dados e quais os problemas existentes, partiu-se então para a avaliação dos algoritmos em um material selecionado da base. Foi escolhido o catmat 445485 cuja descrição é “ÁGUA MINERAL NATURAL, TIPO SEM GÁS MATERIAL EMBALAGEM PLÁSTICO TIPO EMBALAGEM RETORNÁVEL”. Nestes dados foi introduzido um campo anomalia onde constava a avaliação manual feita para cada um dos registros.

Para a avaliação do método foram utilizadas as medidas de acurácia, precisão, *recall* e F_1 score e os valores de Falso Positivo e Falso negativo⁴. A avaliação foi feita em duas situações, na primeira foram testados todos os valores e na segunda, foram excluídos os extremos. Foram definidos como extremos os 2,5% menores valores e os 2,5% maiores valores, com a intenção de diminuir o impacto de possíveis anomalias no resultado, mas garantindo que pelo menos 95% dos valores seriam avaliados. Os resultados estão expressos no Quadro 4 e no Quadro 5. Procurou-se alcançar em cada algoritmo a maior acurácia com o menor valor de falsos negativos.

⁴ Essas métricas foram explicadas na subseção 2.3.3.

Quadro 4 – Avaliação dos Algoritmos de detecção de anomalias sem exclusão de extremos

Algoritmo	Contaminação	Acurácia	Precisão	Recall	Falso Negativo	Falso Positivo
ECOD	0,10	92,39	95,00	57,57	28	2
COPOD	0,08	89,34	87,50	42,42	38	4
IForest	0,10	86,80	67,50	40,91	39	13
LOF	0,03	86,29	100,00	18,18	54	0
HBOS	0,04	86,29	92,86	19,70	53	1
LODA	0,03	86,29	100,00	18,18	54	0
PCA	0,03	85,79	91,67	16,67	55	1
CBLOF	0,03	85,79	91,67	16,67	55	1
KNN	0,03	85,79	91,67	16,67	55	1
GMM	0,03	85,79	91,67	16,67	55	1
SOD	0,03	85,28	83,33	15,15	56	2
DeepSVDD	0,03	84,77	75,00	13,64	57	3
COF	0,02	84,26	75,00	9,09	60	2
OCSVM	0,01	82,23	0,00	0,00	66	4

Fonte: Elaboração própria

Quadro 5 – Avaliação dos Algoritmos de detecção de anomalias com exclusão de extremos

Algoritmo	Contaminação	Acurácia	Precisão	Recall	Falso Negativo	Falso Positivo
LOF	0,06	94,17	70,00	50,00	14	6
ECOD	0,02	93,88	100,00	25,00	21	0
COPOD	0,02	93,29	85,71	21,43	22	1
PCA	0,03	92,71	63,64	25,00	21	4
HBOS	0,04	92,42	57,14	28,57	20	6
IForest	0,01	92,42	75,00	10,71	25	1
LODA	0,01	91,84	50,00	7,14	26	2
GMM	0,02	91,55	42,86	10,71	25	4
CBLOF	0,01	91,25	25,00	3,57	27	3
KNN	0,01	91,25	25,00	3,57	27	3
SOD	0,01	91,25	25,00	3,57	27	3
OCSVM	0,01	90,96	0,00	0,00	28	3
COF	0,01	90,67	0,00	0,00	28	4
DeepSVDD	0,01	90,67	0,00	0,00	28	4

Fonte: Elaboração própria

Decidiu-se então utilizar no desenvolvimento da ferramenta os cinco algoritmos que tiveram melhor acurácia em cada uma das duas situações. Porém como alguns dos algoritmos se repetiram, acabaram sendo selecionados os seguintes seis algoritmos: ECOD, COPOD, IForest, LOF, HBOS, e PCA, que serão explicados nas próximas seções.

3.5.1 *Empirical-Cumulative-distribution-based Outlier Detection (ECOD)*

Li *et al.* (2022) propuseram o ECOD, que significa *Empirical-Cumulative-distribution-based Outlier Detection*. O ECOD é um algoritmo simples, mas eficaz, usado para detecção de *outliers* não supervisionados.

O algoritmo estima a distribuição subjacente dos dados de entrada de maneira não paramétrica, calculando a distribuição cumulativa empírica por dimensão dos dados. Em seguida, ele usa essas distribuições empíricas para estimar as probabilidades por dimensão para cada ponto de dados. Por fim, ele calcula uma pontuação atípica de cada ponto de dados agregando probabilidades estimadas entre as dimensões.

A grande vantagem do ECOD é que ele é livre de parâmetros e fácil de interpretar. Uma implementação Python escalonável e fácil de usar do ECOD está disponível na biblioteca PyOD.

3.5.2 *Copula-Based Outlier Detection (COPOD)*

Li *et al.* (2020) propuseram o método Copula-Based Outlier Detection (COPOD) que é um algoritmo para detecção de *outliers* baseado na estimativa de probabilidades marginais usando cópula. Cópula é uma função de distribuição que descreve a estrutura de dependência entre variáveis aleatórias.

Para obter previsões de anomalias do COPOD, há duas opções, (1) definir um limite nas pontuações atípicas, onde qualquer linha com uma pontuação que exceda o limite é uma anomalia ou (2) selecionar as pontuações de anomalias do percentil superior ou n-ésimo superior. Com a utilização do PyOD, não existe a necessidade de realizar esta escolha, já que os limites são definidos automaticamente com base na contaminação fornecida.

O COPOD tem ainda as seguintes vantagens para a pesquisa: é determinístico sem hiperparâmetros evitando os desafios em seleção de hiperparâmetros e possíveis vieses; é um dos melhores algoritmos de detecção de *outliers*, segundo os autores, pontuando 2,7% a mais em Precisão Média do que o segundo detector com melhor desempenho.

3.5.3 *Floresta de Isolamento ou Isolation Florest (iForest)*

Liu, Ting e Zhou (2009) definem *Isolation Forest* como um método baseado em modelo para detecção de anomalias que isola anomalias em vez de criar perfis de instâncias normais. Ele funciona selecionando aleatoriamente um recurso e, em seguida, selecionando aleatoriamente um valor dividido entre os valores máximo e mínimo do recurso selecionado. Este processo é repetido recursivamente até que todas as instâncias estejam isoladas em seu próprio nó folha.

Espera-se que as anomalias exijam menos divisões para serem isoladas do que as instâncias normais e, portanto, elas têm um comprimento de caminho menor na árvore. O comprimento do caminho é usado como uma medida de pontuação de anomalia, onde caminhos mais curtos indicam pontuações de anomalia mais altas. O método é projetado para ser eficiente e escalável, com uma complexidade de tempo linear e um baixo requisito de memória.

Entre suas limitações destaca-se:

- Pode não funcionar bem para conjuntos de dados com um número muito pequeno de anomalias, pois foi projetado para isolar anomalias em vez de criar perfis de instâncias normais;
- Pode não ser adequado para conjuntos de dados com um grande número de atributos altamente correlacionados, pois se baseia no princípio de subamostragem para criar árvores de isolamento.

3.5.4 Fator de *Outlier* Local (LOF)

Segundo Pang *et al.* (2021) LOF é um algoritmo de aprendizado de máquina usado para detecção de anomalias que mede o grau de isolamento de uma instância em relação aos seus vizinhos. Ele calcula a densidade das instâncias em torno de uma determinada instância e a compara com a densidade de seus vizinhos. Se a densidade da instância for significativamente menor que a densidade de seus vizinhos, ela será considerado um *outlier*. LOF é um algoritmo baseado em densidade e pode lidar com distribuições de dados complexos e dados de alta dimensão⁵.

Para Pang *et al.* (2021) o LOF tem algumas limitações, que incluem:

- É sensível à escolha do parâmetro k , que determina o número de vizinhos a considerar. Um pequeno valor de k pode resultar em alta sensibilidade a *outliers*, enquanto um grande valor de k pode resultar em baixa sensibilidade a *outliers* locais;
- Pode não funcionar bem com conjuntos de dados com densidades variadas ou *clusters* de tamanhos e formas diferentes;
- Pode não ser adequado para dados de alta dimensão, pois a métrica de distância se torna menos significativa em espaços de alta dimensão;
- Pode ser computacionalmente caro para grandes conjuntos de dados, pois requer o cálculo da distância entre cada instância e seus vizinhos.

3.5.5 *Histogram-based Outlier Score* (HBOS)

A técnica *Histogram-based Outlier Score*, ou HBOS, é utilizada para detectar anomalias em dados com base na pontuação de *outliers* em histograma (CHANDOLA; BANERJEE;

⁵ Dados de alta dimensão são dados com grande quantidade de variáveis.

KUMAR, 2009). Segundo os autores, a idéia básica por trás do HBOS é criar um histograma dos dados e, em seguida, usar a altura das caixas no histograma como uma medida da probabilidade de uma nova instância pertencer a essa caixa. A pontuação discrepante de uma instância é então calculada como a soma das alturas inversas das caixas às quais a instância pertence. Quanto maior a pontuação de *outlier*, mais provável é que a instância seja uma anomalia. O HBOS é uma técnica rápida e escalável que pode ser usada para dados de alta dimensão.

Para Goldstein e Dengel (2012) HBOS é um algoritmo de detecção de anomalias não supervisionado usado para encontrar *outliers* em conjuntos de dados sem treinamento prévio. Ele assume que os atributos no conjunto de dados são independentes uns dos outros, o que o torna mais rápido do que as abordagens multivariadas, mas menos preciso. O HBOS pontua recordes em tempo linear e pode detectar *outliers* globais de forma tão confiável quanto os algoritmos de última geração, mas pode não ter um bom desempenho em problemas de *outliers* locais.

Entre suas limitações destaca-se:

- O algoritmo HBOS assume independência de recursos, o que significa que pode não funcionar bem em conjuntos de dados em que os atributos sejam dependentes.
- O HBOS pode não ter um bom desempenho em problemas de *outliers* locais, onde os *outliers* são agrupados em regiões específicas do conjunto de dados.

3.5.6 Análise de Componentes Principais ou *Principal Component Analysis* (PCA)

Segundo Elkhadir, Chougali e Benattou (2015), a Análise de Componentes Principais, do inglês *Principal Component Analysis*, é uma técnica usada para redução de dimensionalidade, o que significa reduzir o número de atributos em um conjunto de dados, mantendo as informações mais importantes. Em outras palavras, o PCA ajuda a simplificar dados complexos identificando padrões e relacionamentos entre variáveis. Ele faz isso transformando os dados originais em um novo conjunto de variáveis chamadas componentes principais, que são combinações lineares dos recursos originais. O primeiro componente principal captura a maior variação nos dados, seguido pelo segundo e assim por diante. O PCA é comumente usado em aprendizado de máquina e análise de dados para melhorar a eficiência e a precisão dos algoritmos.

Géron (2021) explica que a detecção de anomalias depende do erro de reconstrução. Uma vez identificados os componentes principais podemos reconstruir os dados originais a partir dos dados transformados sem perda de dados. Da mesma forma, escolhendo apenas componentes principais que explicam a maioria da variância, devemos ser capazes de recriar uma aproximação dos dados originais. O erro gerado durante a reconstrução ao gerar os dados originais é chamado de erro de reconstrução. Ao se comparar o erro de reconstrução de uma instância normal com o erro de reconstrução de uma anomalia, a última será geralmente bem maior. O autor aponta ainda que essa é uma abordagem de detecção de anomalias simples e muitas vezes bastante eficiente.

Para Elkhadir, Chougali e Benattou (2015), a PCA possui algumas limitações que devem ser consideradas ao usá-la para redução de dimensionalidade ou detecção de anomalias:

- A PCA assume que os dados estão linearmente relacionados, o que significa que pode não funcionar bem para conjuntos de dados com relacionamentos não lineares entre variáveis.
- A PCA pode ser sensível a *outliers*, o que pode afetar a precisão dos resultados.
- A PCA pode nem sempre reter todas as informações importantes nos dados originais, especialmente se os primeiros componentes principais capturarem a maior parte da variação.
- A PCA pode ser computacionalmente cara para grandes conjuntos de dados, o que pode torná-lo impraticável para algumas aplicações.

3.5.7 Técnicas utilizadas para desenvolvimento do modelo

Para o agrupamento dos diversos modelos de detecção de anomalia foi utilizada a biblioteca *Scalable Unsupervised Outlier Detection* (SUOD) (ZHAO *et al.*, 2021).

Segundo Zhao *et al.* (2021), o SUOD

é uma estrutura de aceleração para treinamento e previsão de detectores heterogêneos não supervisionados em larga escala. Ele se concentra em três aspectos complementares para acelerar (redução de dimensionalidade para dados de alta dimensão, aproximação de modelo para modelos complexos e melhoria da eficiência de execução para desequilíbrio de carga de tarefas em sistemas distribuídos), enquanto controla a degradação do desempenho da detecção.

Zhao *et al.* (2021) chama a atenção de que a detecção de *outliers*, devido à falta de rótulos de verdade, leva os profissionais a precisarem construir um grande número de modelos não supervisionados que são heterogêneos (ou seja, diferentes algoritmos e hiperparâmetros) para posteriormente combinarem e analisarem o resultado em conjunto.

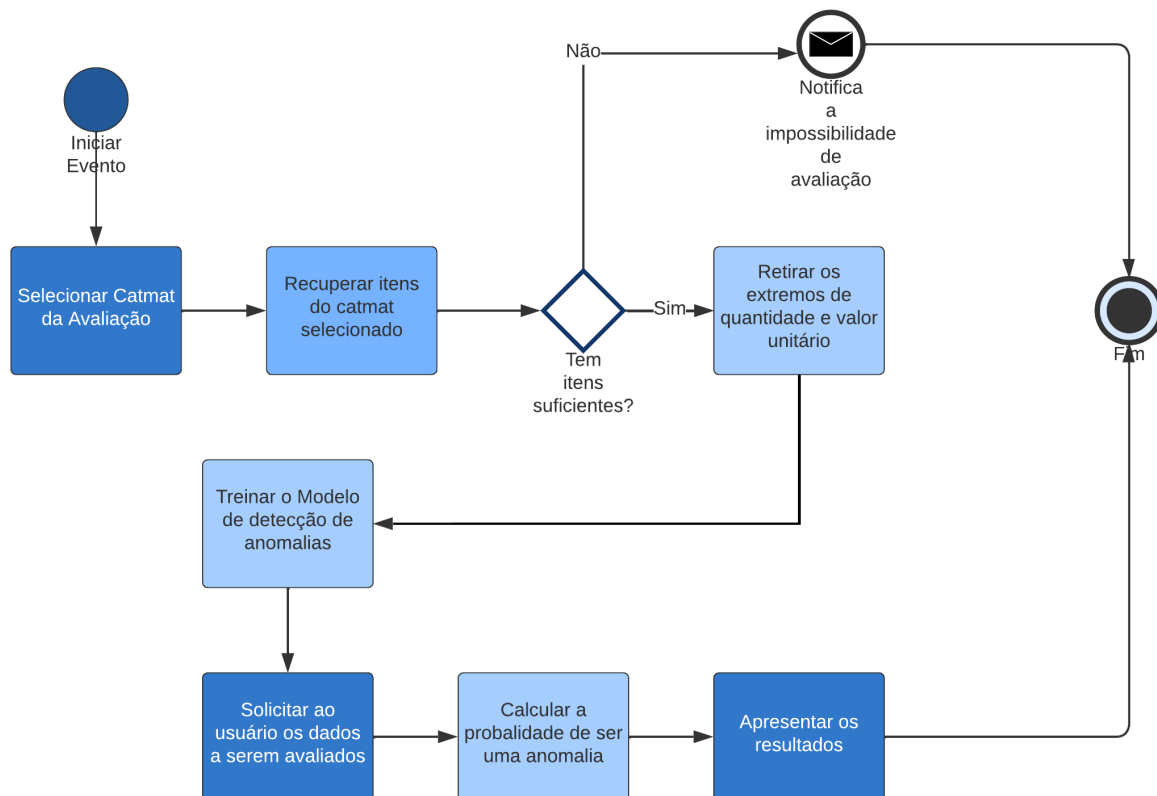
Neste sentido, foi utilizado o SUOD para agrupar os algoritmos HBOS, ECOD, COPOD, LOF, PCA e IForest, fazendo com que com um único treinamento fosse gerado um modelo que utilizasse todos eles.

3.6 Modelo do Aplicativo de detecção de sobrepreço

Escolhidos os algoritmos a serem utilizados, partiu-se para a definição do modelo a ser utilizado na ferramenta de detecção de sobrepreço. A Figura 5 ilustra o diagrama dos processos da ferramenta. Nas próximas seções, será detalhado cada um dos processos.

3.6.1 Processo “Selecionar a Categoria da Avaliação”

Neste ponto do modelo, será criada uma tela de interação com o usuário, onde o mesmo poderá escolher a categoria (atributo Catmat) do item que ele deseja avaliar.

Figura 5 – Modelo da ferramenta de detecção de sobrepreço

Fonte: Elaboração própria

As opções de categorias serão recuperadas dos dados que foram importados dos Dados Abertos de Compras Governamentais

3.6.2 Processo “Recuperar itens da categoria selecionada”

Neste processo serão recuperados todos os itens da categoria selecionada de contratos e licitações de no máximo um ano atrás. Não serão utilizados itens mais antigos, por não servirem ao propósito de validação de pesquisa de preços. O código deste processo está listado abaixo no Código-Fonte 1

Código-Fonte 1 – Função recuperar_itens_catmat

```

1 #####
2 # Funcao recuperar_itens_catmat
3 # Objetivo: Retornar todos os registros de um determinado material
4 #           desde uma data especificada ate o dia de hoje
5 # Parametros: catmat - codigo do material a ser recuperado
6 #           data - Data a partir da qual os registros serao
7 #           selecionados
8 # Retorno: dataframe pandas com todos os registros selecionados

```

```

9 #####
10 def recuperar_itens_catmat(catmat, data):
11
12     # Cria a conexao com o servidor de banco de dados
13     sqlEngine = create_engine('mysql+pymysql://user:password@server/siasg',
14                               pool_recycle=3600)
15     dbConnection = sqlEngine.connect()
16
17     # Le o banco de dados para buscar os registros que atendem os
18     # parametros informados
19     df = pd.read_sql(
20         "SELECT quantidade, valor_unitario FROM siasg.itens2 where
21         catmat_id=" + str(catmat) + " and data > '" + data
22         + "'", dbConnection);
23     return df

```

Caso a quantidade de itens selecionados daquela categoria seja muito baixa, a ferramenta dará a mensagem “Parâmetros inválidos ou erro no cálculo” ao usuário. Esta mensagem irá ocorrer toda vez que devido à uma limitação técnica der algum erro no treinamento do modelo.

3.6.3 Processo “Retirar os extremos de valor unitário”

Ao analisar os dados baixados das licitações, percebeu-se que alguns dados tem valores muito discrepantes, provavelmente decorrentes de erros de digitação. Para evitar que estes dados venham a comprometer a confiabilidade do modelo, optamos por excluir os itens cujo valor unitário esteja abaixo do percentil 2,5 ou superior ao percentil 97,5. Conforme apresentado no Código-Fonte 2 abaixo.

Código-Fonte 2 – Função retirar_extremos

```

1 #####
2 # Funcao retirar_extremos
3 # Objetivo: Retirar da base de treinamento os 2,5% maiores valores
4 #           e os 2,5% menores valores, pela possibilidade de serem
5 #           erros ou distorcoes. Nao serao retirados mais do que
6 #           5% dos registros.
7 # Parametros: df - dataframe pandas com os dados a serem avaliados
8 # Retorno: dataframe pandas com todos os registros ajustados
9 #####
10 def retirar_extremos(df):
11     # calcula o valor unitario onde 97,5% dos registros estao abaixo.
12     maior = df['valor_unitario'].quantile(0.975)
13
14     # calcula o valor unitario onde 2,5% dos registros estao abaixo.
15     menor = df['valor_unitario'].quantile(0.025)
16
17     # Seleciona todos os registros que estao entre o menor e o maior valor

```

```

18     dfajustado = df.loc[(df["valor_unitario"] > menor) & (df["
    valor_unitario"] < maior)]
19
20     # Caso o tamanho do dataframe ajustado seja menor do que 95% dos
    registros, desconsidera o ajuste
21     if len(dfajustado) < (len(df) // (100 / 95)):
22         dfajustado = df
23
24     # retorna o dataframe ajustado
25     return dfajustado

```

Não serão excluídos os dados se a exclusão eliminar mais do que 5% dos dados. Este processo é o primeiro pré-processamento que será realizado nos dados.

3.6.4 Processo “Treinar o modelo de detecção de anomalias”

Como citado anteriormente, foi utilizado o SUOD para agrupar os algoritmos HBOS, ECOD, COPOD, LOF, PCA e IForest, fazendo com que um único treinamento gerasse um modelo que utilizasse todos eles. A função do protótipo inicial da ferramenta encontra-se detalhada no Código-Fonte 3.

Código-Fonte 3 – Função treina_modelo

```

1 #####
2 # Funcao treina_modelo
3 # Objetivo: Treina o modelo de deteccao de anomalias
4 # Parametros: df - dataframe pandas com os dados a
5 #               serem utilizados no treinamento
6 #               contamination - % estimado de valores
7 #               anomalos nos dados.
8 # Retorno: clf - modelo treinado utilizando o SUOD
9 #####
10 # Importa bibliotecas do PyOD com os algoritmos de deteccao de anomalias
11 from pyod.models.iforest import IForest
12 from pyod.models.pca import PCA
13 from pyod.models.hbos import HBOS
14 from pyod.models.ecod import ECOD
15 from pyod.models.copod import COPOD
16 from pyod.models.lof import LOF
17 from pyod.models.suod import SUOD
18
19 def treina_modelo(df,contamination):
20     # Lista de detectores
21     detector_list = [
22         ECOD(contamination=contamination),
23         COPOD(contamination=contamination),
24         PCA(n_components=2,n_selected_components=1,
    contamination=contamination),

```

```

25         IForest(n_estimators=3, contamination=contamination),
26         LOF(n_neighbors=21, contamination=contamination),
27         HBOS(n_bins=90, contamination=contamination)
28     ]
29
30     # configurar o SUOD com a lista de parametros e detectores
31     clf = SUOD(base_estimators=detector_list, n_jobs=2, combination='
32     maximization', contamination=contamination,
33               verbose=False)
34
35     # treinar o modelo
36     clf.fit(df)
37     return clf

```

Tendo sido elaborado o modelo, o mesmo foi aplicado a dois materiais distintos, um onde ocorria a exclusão dos extremos e outro onde não ocorria. Através da análise dos resultados dos testes foi verificada a validade da exclusão dos extremos e o grau de contaminação mais adequado a ser utilizado. Os resultados obtidos e as análises serão apresentadas no Capítulo 4.

Tendo sido determinada a contaminação a ser utilizada na ferramenta, aferiu-se o desempenho do modelo e de cada um dos algoritmos. Com essa aferição foi possível descartar alguns algoritmos que não performavam adequadamente.

Para poder definir o modelo final a ser utilizado pela ferramenta foram realizados testes combinando os algoritmos restantes em todas as combinações possíveis, inclusive o próprio algoritmo isoladamente.

Tendo então sido definido o modelo, foi feito o teste final de sua performance em outro material para evitar o *overfitting*.

3.6.5 Processo “Solicitar ao Usuário os dados a serem avaliados”

Nesta parte do processo será solicitado ao usuário que entre com os dados que deseja avaliar. Podendo informar a quantidade e valor unitário obtido na sua pesquisa de preços.

3.6.6 Processo “Calcular a possibilidade de ser uma anomalia”

Tendo sido informados no processo anterior a quantidade e valor obtido na pesquisa, os mesmos são informados para a função de avaliação detalhada no Código-Fonte 4.

Código-Fonte 4 – Função avalia_dados

```

1 #####
2 # Funcao avalia_dados
3 # Objetivo: Treina o modelo de deteccao de anomalias
4 # Parametros: clf - modelo treinado
5 #             quantidade - quantidade utilizada na
6 #             pesquisa de precos

```

```

7 #         valor - valor obtido na pesquisa de
8 #         precos
9 # Retorno: predicacao - resultado da avaliacao
10 #####
11 def avalia_dados(clf, quantidade, valor):
12     # cria um dataframe de apenas 1 linha com os valores recebidos
13     dfteste=pd.DataFrame({'quantidade':quantidade,'valor_unitario':valor},
14                           index=[0])
15     # Calcula a predicacao
16     predicacao=clf.predict(dfteste)
17     # Retorna o resultado
18     return predicacao

```

3.6.7 Processo “Apresentar os resultados”

No processo de apresentar os resultados, serão mostradas na tela as seguintes informações:

- início de sobrepreço ou não;
- o valor mínimo, máximo e a mediana para o item selecionado;
- a lista de licitações que tiveram valores menores que a pesquisa informada com o respectivo valor e ganhador, se for o caso.

3.7 Protótipo da ferramenta de avaliação de indícios de sobrepreços disponibilizada no Senado Federal

Nesta seção, serão apresentadas as partes que compõe a ferramenta de avaliação de indícios de sobrepreços: o atualizador de dados diário, a tela de seleção de categoria e a tela de avaliação e resultados.

3.7.1 Atualizador de dados diário

O atualizador diário utiliza a API do Comprasnet para buscar os dados das licitações mais recentes e atualizar a base utilizada para treinamento do modelo.

3.7.2 Tela de Seleção de Material

A tela do protótipo seguirá o modelo apresentado na Figura 6 abaixo.

O *combobox* de seleção do material será carregado com todos as categorias disponíveis na base de itens do repositório.

Será apresentada a descrição do material conforme o que consta no catálogo de materiais da API de Compras Governamentais.

Figura 6 – Tela de Seleção de Material

A interface apresenta o cabeçalho "SENADO FEDERAL" e "IA para Pesquisa de Preços". O menu de navegação inclui: IAPP, Home, Listagens, Preparação do Repositório, Licitações, Preparação de Dados e Avaliação da Pesquisa de Preços. O título da seção é "Selecionar Material".

Selecione o Material: 462729 - GELATINA ALIMENTÍCIA, APRESENTAÇÃO PÓ S

Data inicial de preços para validação: 09/06/2022

Descrição: GELATINA ALIMENTÍCIA, APRESENTAÇÃO PÓ SABOR VARIADO ORIGEM ANIMAL

Botões: Avaliar, Voltar

Senado Federal - Praça dos Três Poderes - Brasília DF - CEP 70165-900 - Fone: (61)3303-4141

Fonte: elaboração própria

3.7.3 Tela de avaliação e resultados

Após a seleção do material, a tela de avaliação de resultados, conforme ilustrada na Figura 7, apresentará as medidas estatísticas sobre a categoria selecionada e irá receber o valor e quantidade da pesquisa.

Figura 7 – Tela de Avaliação da Pesquisa

A interface apresenta o cabeçalho "SENADO FEDERAL" e "IA para Pesquisa de Preços". O menu de navegação inclui: IAPP, Home, Listagens, Preparação do Repositório, Licitações, Preparação de Dados e Avaliação da Pesquisa de Preços. O título da seção é "Avaliação de Pesquisa de Preços".

Material selecionado = 462729 - GELATINA ALIMENTÍCIA, APRESENTAÇÃO PÓ SABOR VARIADO ORIGEM ANIMAL

Dados avaliados desde de 09/06/2022

Valor unitário está entre R\$ 3,18 e R\$ 89,97 com média= R\$ 12,14 e mediana= R\$ 11,92

Quantidade adquirida está entre 20 e 1700 com média= 158,97 e mediana= 60,00

Número de registros do item = 34

Quantidade a ser adquirida: 20

Valor encontrado na pesquisa: 12

Botões: Avaliar, Voltar

Resultado da avaliação: Valor aceitável

Senado Federal - Praça dos Três Poderes - Brasília DF - CEP 70165-900 - Fone: (61)3303-4141

Fonte: elaboração própria

Poderão ser testados diversos valores e quantidades a critério do usuário.

3.8 Considerações finais do capítulo

Durante este capítulo foram detalhados as metodologias empregadas e os procedimentos de desenvolvimento da ferramenta. Foram vistas as dificuldades encontradas para coleta dos dados, para a seleção do modelo a ser utilizado e o algoritmo utilizado no treinamento da ferramenta de predição de sobrepreço. No próximo capítulo, serão apresentados os resultados obtidos, o grau de confiabilidade da ferramenta e a análise dos resultados.

4 RESULTADOS E ANÁLISES

Para a avaliação do modelo inicial elaborado foram realizados teste em dois materiais: Água Mineral Natural (catmat=445485) e Microcomputador (catmat=469793), que serão referenciados a partir deste ponto como Água e Microcomputador.

O primeiro passo para a avaliação dos modelos era determinar o parâmetro grau de contaminação, que indica o percentual de anomalias que se encontram no conjunto de dados, e tem influencia em todos os algoritmos. Para poder determinar este parâmetro testou-se o modelo utilizando diferentes graus de contaminação.

Os resultados obtidos durante a avaliação do modelo no material Água estão detalhados no Quadro 6.

Quadro 6 – Avaliação do Algoritmo no material Água por percentual de contaminação

Con-tami-nação	Acurácia	Precisão	Recall	F1-Score	Falso Negativo	Falso Positivo	Verdadeiro Positivo
0,01	88,50	100,0	8,51	15,69	43	0	4
0,02	89,57	100,0	17,02	29,09	39	0	8
0,03	88,77	72,73	17,02	27,59	39	3	8
0,04	88,77	69,23	19,15	30,00	38	4	9
0,05	91,44	89,47	36,17	51,52	30	2	17
0,06	86,36	40,91	19,15	26,09	38	13	9
0,07	87,70	51,85	29,79	37,84	33	13	14
0,08	88,77	58,62	36,17	44,74	30	12	17
0,09	88,24	54,29	40,43	46,34	28	16	19
0,10	87,17	48,65	38,30	42,86	29	19	18
0,11	87,70	51,35	40,43	45,24	28	18	19
0,12	85,83	43,18	40,43	41,76	28	25	19
0,13	85,56	42,22	40,43	41,30	28	26	19
0,14	85,03	41,51	46,81	44,00	25	31	22
0,15	83,69	36,54	40,43	38,38	28	33	19

Fonte: Elaboração própria.

Na Quadro 6 o grau de contaminação que teve a melhor acurácia foi o de 5%, destaca-se que durante a execução houve a exclusão dos extremos, retirando outros 5%. Apesar da acurácia estar em um bom percentual entre 83,69% e 88,5%, em uma situação de detecção de anomalias onde a quantidade de anomalias é pequena em relação ao total dos registros este indicador sozinho não deve ser considerado. O *F1-Score* é um indicador que aponta um desempenho mais

global do algoritmo, já que se baseia na Precisão e no *Recall* simultaneamente. Neste sentido a contaminação de 5% foi a única com um *F1-Score* maior do que 50%.

O mesmo teste foi repetido para o material Microcomputador e o resultado pode ser verificado no Quadro 7.

Quadro 7 – Avaliação do Algoritmo no material Microcomputador por percentual de contaminação

Con-tami-nação	Acurácia	Precisão	Recall	F1-Score	Falso Negativo	Falso Positivo	Verdadeiro Positivo
0,01	97,00	100,00	11,11	20	8	0	1
0,02	95,88	25,00	11,11	15,38	8	3	1
0,03	97,00	57,14	44,44	50	5	3	4
0,04	96,25	45,45	55,56	50	4	6	5
0,05	95,13	35,71	55,56	43,48	4	9	5
0,06	94,76	33,33	55,56	41,67	4	10	5
0,07	93,63	27,78	55,56	37,04	4	13	5
0,08	93,26	26,32	55,56	35,71	4	14	5
0,09	92,13	22,73	55,56	32,26	4	17	5
0,10	94,01	36,00	100,0	52,94	0	16	9
0,11	92,88	32,14	100,0	48,65	0	19	9
0,12	91,39	28,12	100,0	43,9	0	23	9
0,13	90,26	25,71	100,0	40,91	0	26	9
0,14	89,51	24,32	100,0	39,13	0	28	9
0,15	88,39	22,50	100,0	36,73	0	31	9

Fonte: Elaboração própria.

No caso do Microcomputador, o melhor grau de contaminação foi o de 10%, tomando-se como base o *F1-score* acima de 50% e chamando a atenção para o *Recall* de 100% que indica que todas as anomalias foram identificadas. Não houve neste caso a exclusão dos extremos, pois durante a execução foi detectado que a exclusão diminuiria a base em mais de 5% e foram utilizados então os dados originais.

A partir da análise destes testes determinou-se que o grau de contaminação padrão da ferramenta seria de 10% e que não seria executada a exclusão de extremos. Observou-se que havia espaço para melhora da ferramenta pelo fato de o melhor *F1-Score* obtido ser de 52,94. Tendo sido escolhida a contaminação a ser aplicada, optou-se por avaliar cada um dos algoritmos nesta contaminação.

Quadro 8 – Avaliação dos Algoritmos nos materiais Água e Microcomputador com 10% de contaminação

Algoritmo	Água			Microcomputador		
	Acurácia	Recall	F1-Score	Acurácia	Recall	F1-Score
SUOD	86,55	39,39	49,52	93,63	100,00	51,43
ECOD	92,39	57,58	71,70	93,26	100,00	50,00
COPOD	87,31	42,42	52,83	88,76	11,11	6,25
IForest	83,50	30,30	38,10	88,39	11,11	6,06
LOF	80,20	21,21	26,42	95,51	100,00	60,00
HBOS	83,25	22,73	31,25	89,89	11,11	6,90
PCA	89,34	48,48	60,38	93,63	100,00	51,43

Fonte: Elaboração própria.

No Quadro 8 estão os resultados da avaliação dos algoritmos em ambos os materiais no grau de contaminação 10%. Pelos resultados obtidos percebe-se que o desempenho dos algoritmos COPOD, IForest e HBOS no material Microcomputador foi muito baixo, já que o Recall está em apenas 11,11% indicando que o nível de detecção das anomalias foi inexpressivo, e o desempenho apesar de melhor no material Água também não foi bom. Optou-se então por retirá-los do modelo.

O desempenho do modelo SUOD não foi melhor do que os algoritmos ECOD, PCA e COPOD em relação ao material Água. Acreditando-se que este desempenho poderia estar ligado ao desempenho dos algoritmos COPOD, IForest e HBOS partiu-se para testar o desempenho das várias combinações entre os algoritmos ECOD, LOF e PCA e os algoritmos isoladamente para buscar um melhor desempenho.

Quadro 9 – Avaliação das combinações dos Algoritmos nos materiais Água com 10% de contaminação

Algoritmo	Acurácia	Precisão	Recall	F1-Score
(ECOD+LOF+PCA)	92,39	95	57,58	71,7
(ECOD+PCA)	92,39	95	57,58	71,7
ECOD	92,39	95	57,58	71,7
(LOF+PCA)	89,34	80	48,48	60,38
PCA	89,34	80	48,48	60,38
(ECOD+LOF)	88,83	77,5	46,97	58,49
LOF	80,2	35	21,21	26,42

Fonte: Elaboração própria.

No Quadro 9 encontram-se os resultados obtidos com o Material Água, pode-se perceber que a retirada dos algoritmos proporcionou uma performance melhor e que o modelo do SUOD

com os algoritmos ECOD, LOF e PCA e o modelo com os algoritmos ECOD e PCA, juntamente com o algoritmo ECOD sozinho tiveram o melhor desempenho com uma acurácia de 92,39%, um F1-Score de 71,7 e um Recall de 57,58.

Quadro 10 – Avaliação das combinações dos Algoritmos no material Microcomputador com 10% de contaminação

Algoritmo	Acurácia	Precisão	Recall	F1-Score
LOF	95,51	42,86	100	60
(ECOD+LOF)	94,38	37,50	100	54,55
PCA	93,63	34,62	100	51,43
(ECOD+LOF+PCA)	93,26	33,33	100	50
(ECOD+PCA)	93,26	33,33	100	50
ECOD	93,26	33,33	100	50
(LOF+PCA)	92,88	32,14	100	48,65

Fonte: Elaboração própria.

No Quadro 10 encontram-se os resultados obtidos com o Material Microcomputador, aqui também pode-se perceber que a retirada dos algoritmos proporcionou uma performance melhor. Destaca-se que todos os algoritmos tiveram um Recall de 100% indicando que todas as anomalias foram detectadas. A melhor acurácia, precisão e F1-Score variam em função do número de falsos positivos identificados, já que com um Recall de 100% não existem falsos negativos.

Avaliando estes resultados optou-se por utilizar na ferramenta os algoritmos ECOD, LOF e PCA. Esta foi a opção escolhida em virtude do bom desempenho nos dois casos e com a possibilidade de que com mais de um tipo de algoritmo pode-se detectar mais tipos de anomalias.

Tendo sido definido o grau de contaminação a ser utilizado e a conformação dos algoritmos da ferramenta de detecção passou-se a realização dos testes com um novo material para verificar o desempenho em dados até então desconhecidos. O Material escolhido foi o “LIVRO DIDÁTICO, GRAU ENSINO SUPERIOR / UNIVERSITÁRIO DEFINIÇÃO COLEÇÃO INTERDISCIPLINAR CONTEÚDO CIÊNCIAS EXATAS E DA TERRA FORMATO IMPRESSO (catmat = 481567)”, referenciado como Livro a partir de agora. No Quadro 11 e na Figura 8 podemos ver os resultados da avaliação.

Quadro 11 – Avaliação dos Algoritmos no material Livro

Algoritmo	Acurácia	Precisão	Recall	F1-Score	Falso Negativo	Falso Positivo
SUOD	95,24	74,29	76,47	75,36	8	9
ECOD	95,24	74,29	76,47	75,36	8	9

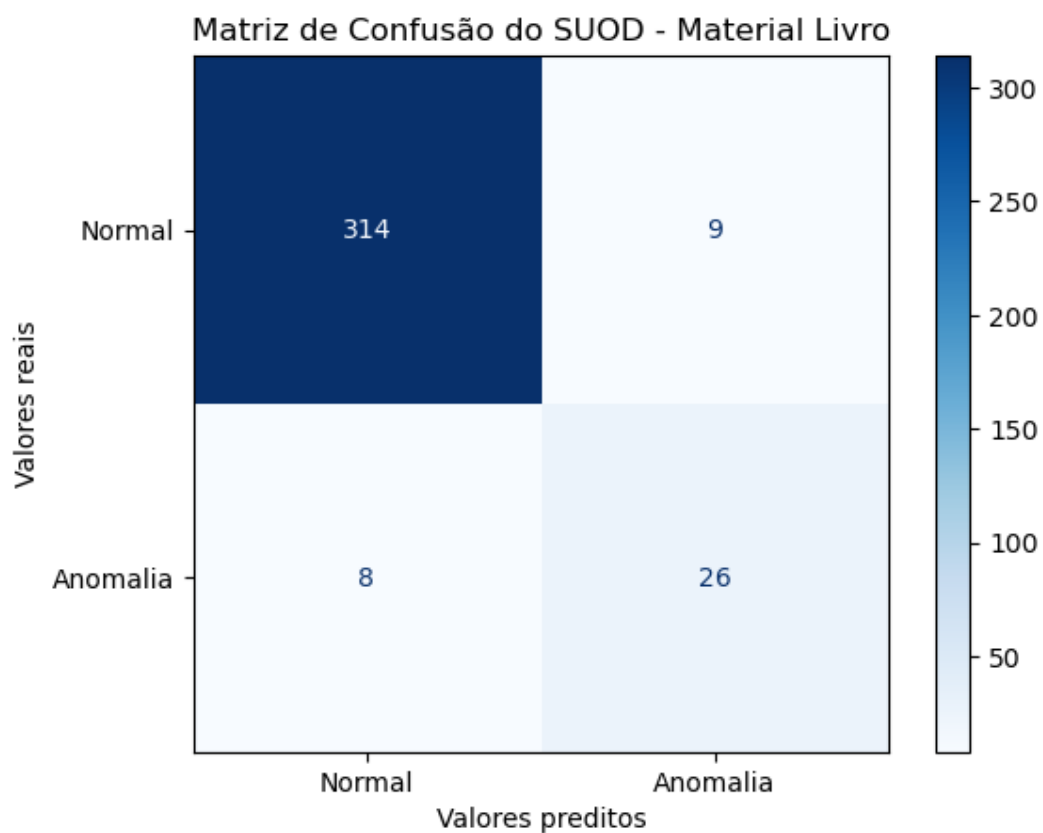
Continua na próxima página

Quadro 11 – Continuação

	Acurácia	Precisão	Recall	F1-Score	Falso Negativo	Falso Positivo
LOF	91,60	55,56	58,82	57,14	14	16
PCA	85,71	25,71	26,47	26,09	25	26

Fonte: Elaboração própria.

Figura 8 – Matriz de confusão do SUOD - Material Livro



Fonte: elaboração própria

Observa-se no Quadro 11 que a acurácia do modelo para o caso do Livro ficou em 95,24%. A precisão de 74,29% indica que o algoritmo conseguiu determinar os sobrepreços de maneira eficaz, conclusão confirmada pelos resultados de recall de 75,36%, o que significa que foi possível identificar os sobrepreços existentes em no mínimo 75,36% dos casos. A tabela ainda apresenta os demais algoritmos separadamente e pode-se perceber que o algoritmo SUOD que agrupa os demais teve o melhor desempenho somente equiparado ao algoritmo ECOD, o que demonstra que alcançamos uma boa equalização dos componentes da ferramenta.

A Figura 8 apresenta a matriz de confusão da avaliação final que demonstra que o algoritmo dos 357 registros identificou corretamente 340. Apenas 8 anomalias não foram iden-

tificadas corretamente e 9 itens seriam identificados como anomalias levando à uma avaliação mais aprofundada.

Esses resultados são satisfatórios, uma vez que os gestores, atualmente, não dispõem de qualquer ferramental de auxílio na elaboração da pesquisa de preços. Logo, um indicativo que aponta para uma probabilidade de certeza em torno de 75% e uma probabilidade de sobrepreço de no mínimo 74,29% é um importante balizamento para a avaliação.

5 CONCLUSÕES E CONSIDERAÇÕES FINAIS

Este trabalho de pesquisa teve como objetivo desenvolver um sistema automatizado baseado em algoritmos de Aprendizagem de Máquina (Machine Learning) capaz de indicar possíveis indícios de sobrepreços durante a fase de pesquisa de preços em Contratações do Senado Federal. Nesta pesquisa foi possível desenvolver um modelo inicial com capacidade de apontar indícios de sobrepreço nas contratações do Senado Federal, que certamente servirá de embrião para um projeto maior que venha a dar maior segurança e serenidade na elaboração de pesquisas de preço.

Para alcançar este objetivo foi realizado um levantamento das Bases de Dados existentes sobre às contratações, mais especificamente as contratações públicas. Neste sentido os dados abertos do Sistema Integrado de Administração de Serviços Gerais – SIASG, sistema onde são realizadas as operações das compras governamentais dos órgãos integrantes do SISG é o Banco de Dados mais relevante e foi o escolhido para servir de subsídio para a pesquisa.

Tendo sido então determinada a origem dos dados a serem utilizados na pesquisa, passou-se ao objetivo de desenvolver uma ferramenta para obtenção e pré-processamento dos dados necessários ao Treinamento de algoritmos de aprendizagem. Neste ponto foi onde foram encontradas as maiores dificuldades, já que a API apresentou diversos problemas e erros de design que fizeram com que fosse necessário criar métodos para contornar suas deficiências.

De posse dos dados, passamos ao objetivo de disponibilizar repositório público dos dados de contratações públicas utilizado pela ferramenta, para tanto, foi criado um repositório no Site GitHub e disponibilizado de forma pública. Neste repositório se encontram os dados de licitações e contratos desde 01/01/2022.

Uma vez organizado o repositório, passou-se a tarefa de levantar as técnicas, métodos e boas práticas mais recentes utilizados na detecções de fraudes em compras. Durante as pesquisas realizadas neste levantamento, optou-se por utilizar a biblioteca PyOD, que é uma biblioteca em *Python* especializada em detecção de anomalias. A biblioteca também já disponibiliza o SUOD, que é uma estrutura de aceleração para treinamento e previsão de detectores heterogêneos não supervisionados em larga escala, facilitando a utilização simultânea de mais de um modelo de algoritmo. Destaca-se também que a biblioteca PyOD tem disponíveis os algoritmos mais recentes como por exemplo o ECOD de 2022, que foi um dos algoritmos que apresentou melhor performance nos testes realizados.

Com a utilização da biblioteca PyOD e do SUOD foi possível treinar e utilizar um modelo de Aprendizagem de Máquina na detecção de sobrepreço nas contratações públicas do banco de dados preparado, realizou-se diversos testes. Primeiramente testou-se todos os algoritmos não supervisionados existentes na biblioteca para selecionar aqueles que se encaixavam melhor aos dados disponíveis. Após a seleção, diversas simulações foram feitas para escolher a melhor conformação de algoritmos e determinar os parâmetros a serem utilizados. Chegou-se a conclusão

de que a utilização do SUOD com os algoritmos ECOD, LOF e PCA eram os mais adequados para o problema em questão e que utilizar a contaminação em 10% proporcionava um bom resultado.

Tendo sido elaborado o modelo, foi possível desenvolver uma ferramenta de avaliação de indícios de sobrepreços para uso efetivo no Senado Federal. O protótipo que foi desenvolvido pode selecionar o material desejado e qual o período de tempo das amostras, de forma a somente utilizar preços que estejam dentro da validade para elaboração de um documento de pesquisa de preços. A ferramenta já apresenta resultados e pode ser utilizada efetivamente como um auxílio aos gestores.

Finalmente, ao avaliar o grau de confiabilidade da ferramenta na predição de sobrepreço foi encontrada uma confiabilidade de 75% nos testes realizados. Tendo em vista que no presente momento não existe uma ferramenta semelhante, há um ganho substancial na sua utilização. Porém, apesar de uma confiança de pelo menos 75% na predição realizada, o número de falsos negativos é alto, o que indica que muitos itens que em nossa avaliação pessoal deveriam ser encarados como suspeitos, não foram identificados, deixando espaço para um aperfeiçoamento da ferramenta.

5.1 Limitações

Nos testes efetuados escolheu-se os materiais que tinham mais registros no repositório o que possibilitou realizar um maior número de testes. Contudo os dados abertos do SIASG possuem muitos registros sem o valor unitário ou total da compra, o que se pode perceber pelo fato de que foram recuperados mais de 984.000 itens de contratos e licitações e pouco mais de 65.000 compuseram o repositório.

Existe ainda o fato de que os dados abertos não são disponibilizados a medida que ocorrem as licitações, dependendo de processos internos para sua liberação ao público em geral.

Qualquer pesquisa de Ciência de Dados e IA depende da quantidade e qualidade dos dados, e pode-se perceber que existem ainda diversos fatores limitantes em relação a qualidade e quantidade dos dados de compras públicas.

Outro fator limitante para a entrega de uma ferramenta mais aprimorada é o tempo reduzido para a conclusão de um trabalho de mestrado, não tendo sido possível colocar a aplicação dentro de um projeto piloto e avaliar os seus resultados em um ambiente de produção.

5.2 Trabalhos Futuros

Existem diversas possibilidades de evolução da ferramenta, dentre estas possibilidades seria ideal iniciar com a implementação de um projeto piloto onde fossem avaliados os resultados da ferramenta em situações reais pelos gestores do Senado Federal.

Outra possibilidade é a retroalimentação da ferramenta, sendo possível incluir na sua

base de treinamento as propostas de preços apresentadas durante a elaboração e planejamento das licitações do Senado Federal. A alimentação destas propostas poderia ser realizada por meio da leitura da planilha de preços que é elaborada para compor os processos de aquisição.

A ferramenta poderia também ser incluída em um projeto maior de elaboração de pesquisas de preços, onde pudessem ser documentadas as providências tomadas no caso de um alerta de sobrepreço, gerando uma base de conhecimento, além de ao final ser emitido o artefato de pesquisa de preços validado.

REFERÊNCIAS

- ARAÚJO, Valter Shuenquener de; ZULLO, Bruno Almeida; TORRES, Maurílio. Big data, algoritmos e inteligência artificial na administração pública: reflexões para a sua utilização em um ambiente democrático. **A&C - Revista de Direito Administrativo & Constitucional**, v. 20, n. 80, p. 241, 2020. Disponível em: <http://www.revistaaec.com/index.php/revistaaec/article/view/1219>. Acesso em: 23 mar. 2022. Citada 1 vez na página 25.
- BATISTA, Mariana; ROCHA, Virginia; SANTOS, José Luiz Alves dos. Transparência, corrupção e má gestão: uma análise dos municípios brasileiros. **Revista de Administração Pública**, FapUNIFESP (SciELO), São Paulo, SP, v. 54, n. 5, p. 1382-1401, out. 2020. ISSN 1982-3134. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122020000501382&tlng=pt. Acesso em: 12 maio 2023. Citada 3 vezes na página 37(3x).
- BORGES JÚNIOR, Renildo Aguis. A pesquisa de preços e seu papel fundamental nas licitações públicas. 2020. Disponível em: <https://jus.com.br/artigos/79447/a-pesquisa-de-precos-e-seu-papel-fundamental-nas-licitacoes-publicas>. Acesso em: 1 ago. 2021. Citada 2 vezes nas páginas 26 e 31.
- BRASIL. **Lei nº 8.666, de 21 de junho de 1993**. Regulamenta o art. 37, inciso XXI, da Constituição Federal, institui normas para licitações e contratos da Administração Pública e dá outras providências. Brasília, DF: Presidência da República, 1993. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/L8666compilado.htm. Acesso em: 7 maio 2019. Citada 2 vezes nas páginas 31 e 33.
- BRASIL. **Lei nº 10.520, de 17 de julho de 2002**. Institui, no âmbito da União, Estados, Distrito Federal e Municípios, nos termos do art. 37, inciso XXI, da Constituição Federal, modalidade de licitação denominada pregão, para aquisição de bens e serviços comuns, e dá outras providências. Brasília, DF: Presidência da República, 2002. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/2002/110520.htm. Acesso em: 25 ago. 2022. Citada 1 vez na página 31.
- BRASIL. **Lei nº 12.527, de 18 de novembro de 2011**. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Brasília, DF: Presidência da República, 2011. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/112527.htm. Acesso em: 25 mar. 2023. Citada 1 vez na página 26.
- BRASIL. **Compras - Portal de Compras do Governo Federal**. Brasília/DF: gov.br, 2020. Disponível em: <https://www.gov.br/compras/pt-br/sistemas/conheca-o-compras/compras/compras>. Acesso em: 5 out. 2023. Citada 3 vezes na página 32(3x).
- BRASIL. **Lei nº 14.133, de 1º de abril de 2021**. Lei de Licitações e Contratos Administrativos. Brasília, DF: Presidência da República, 2021. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14133.htm. Acesso em: 21 mar. 2022. Citada 5 vezes nas páginas 25, 27, 31 e 32(2x).

BRASIL. Ministério da Economia. **API de compras governamentais**. Brasília/DF: Ministério da Economia, 2021. Disponível em: <http://compras.dados.gov.br/docs/home.html>. Acesso em: 21 mar. 2022. Citada 10 vezes nas páginas 26, 35, 46, 87, 88, 89(2x), 90, 91 e 92.

CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. Anomaly detection: A survey. **ACM Computing Surveys**, ACM, v. 41, n. 3, p. 15-, 2009. Disponível em: <https://dl.acm.org/doi/10.1145/1541880.1541882>. Acesso em: 20 abr. 2023. Citada 2 vezes nas páginas 38 e 54.

COSTA, Marcos Bemquerer; BASTOS, Patrícia Reis Leitão. Alice, Monica, Adele, Sofia, Carina e Ágata: o uso da inteligência artificial pelo Tribunal de Contas da União. **Controle Externo: Revista do Tribunal de Contas do Estado de Goiás**, n. 3, 2020. Disponível em: <https://revcontext.tce.go.gov.br/index.php/context/article/download/59/57/344>. Acesso em: 15 out. 2022. Citada 2 vezes na página 41(2x).

COUTINHO, Eduardo D.; FREITAS, Angilberto S. Valor público via tecnologias desenvolvidas com dados governamentais abertos: o caso Operação Serenata de Amor. **RAM. Revista de Administração Mackenzie**, FapUNIFESP (SciELO), v. 22, n. 6, 2021. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122020000501382&tlng=pt. Acesso em: 22 mar. 2022. Citada 1 vez na página 40.

DUTRA, Daniel. **O que é URL? Entenda o que significa o endereço de sites da Internet**. [S. l.]: TechTudo, 2023. Disponível em: <https://www.techtudo.com.br/guia/2023/05/o-que-e-url-entenda-o-que-significa-o-endereco-de-sites-da-internet-edsoftwares.ghtml>. Acesso em: 5 out. 2023. Citada 3 vezes nas páginas 35, 46 e 79.

EEGA, Varnika. A review on machine learning models used for anomaly detection. **SSRN Electronic Journal**, v. 4, n. 8, p. 9–15, 2021. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3904483. Acesso em: 22 mar. 2022. Citada 1 vez na página 39.

ELKHADIR, Zyad; CHOUGDALI, Khalid; BENATTOU, Mohammed. Intrusion detection system using pca and kernel pca methods. *In: MEDITERRANEAN CONFERENCE ON INFORMATION & COMMUNICATION TECHNOLOGIES*, 1., 2015, Saïdia, Morocco. **Proceedings [...]**. [S. l.]: Springer, Cham, 2015. p. 489-497. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-30298-0_50. Acesso em: 20 abr. 2023. Citada 2 vezes nas páginas 54 e 55.

FORTINI, Cristiana; MOTTA, Fabrício. Corrupção nas licitações e contratações públicas: sinais de alerta segundo a transparência internacional. **A&C - Revista de Direito Administrativo & Constitucional**, Revista de Direito Administrativo and Constitucional, v. 16, n. 64, p. 93, 2016. P. 27-44. Citada 1 vez na página 32.

FORTINI, Cristiana; SHERMAN, Ariane. Governança pública e combate à corrupção: novas perspectivas para o controle da administração pública brasileira. **Interesse Público - IP**, v. 19, n. 102, p. 27–44, 2017. Disponível em: <https://www.editoraforum.com.br/wp-content/uploads/2017/11/governanca-combate-corrupcao.pdf>. Acesso em: 21 mar. 2022. Citada 2 vezes na página 27(2x).

GÉRON Aurélien. **Mãos à Obra: Aprendizado de máquina com scikit-learn, keras e tensorflow**. 2ª ed. Rio de Janeiro/RJ: Alta Books Editora, 2021. 640 p. ISBN 978-85-5081-548-0. Citada 5 vezes nas páginas 38, 42, 43, 49 e 54.

GOLDSTEIN, Markus; DENGEL, Andreas. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *In: KI 2012: ADVANCES IN ARTIFICIAL INTELLIGENCE: 35TH ANNUAL GERMAN CONFERENCE ON AI*, 35., 2012, Saarbrücken, Alemanha. **Proceedings** [...]. Saarbrücken, Alemanha: Springer/Sci-Tech/Trade, 2012. Disponível em: https://www.researchgate.net/publication/231614824_Histogram-based_Outlier_Score_HBOS_A_fast_Unsupervised_Anomaly_Detection_Algorithm. Acesso em: 20 abr. 2023. Citada 1 vez na página 54.

HAN, Songqiao *et al.* Adbench: Anomaly detection benchmark. *In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, 35., 2022, New Orleans, Louisiana, USA. **Proceedings** [...]. New York, USA: Curran Associates, Inc., 2022. v. 35, p. 32142–32159. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2022/file/cf93972b116ca5268827d575f2cc226b-Paper-Datasets_and_Benchmarks.pdf. Acesso em: 12 abr. 2023. Citada 3 vezes nas páginas 49 e 50(2x).

JSON. *In: WIKIPÉDIA*, a enciclopédia livre. Flórida, EUA: Wikimedia Foundation, 2022. Disponível em: <https://pt.wikipedia.org/w/index.php?title=JSON&oldid=64592661>. Acesso em: 5 out. 2023. Citada 2 vezes nas páginas 35 e 79.

JUNQUILHO, Tainá Aguiar; MAIA FILHO, Mamede Said. Inteligência Artificial no Poder Judiciário: lições do Projeto Victor. v. 8, n. 48, 2021. Disponível em: <https://revista.unitins.br/index.php/humanidadeseinovacao/article/view/5615>. Acesso em: 23 mar. 2022. Citada 1 vez na página 27.

LI, Zheng *et al.* COPOD: Copula-based outlier detection. *In: 2020 IEEE INTERNATIONAL CONFERENCE ON DATA MINING (ICDM)*, 2020, Sorrento, Itália. **Proceedings** [...]. Washington, DC, USA: IEEE, 2020. Disponível em: <https://arxiv.org/abs/2009.09463>. Acesso em: 12 maio 2023. Citada 1 vez na página 52.

LI, Zheng *et al.* ECOD: Unsupervised outlier detection using empirical cumulative distribution functions. **IEEE Transactions on Knowledge and Data Engineering**, Institute of Electrical and Electronics Engineers (IEEE), 2022. Disponível em: <https://arxiv.org/abs/2201.00382>. Acesso em: 11 maio 2023. Citada 1 vez na página 52.

LIMA, Luiz Henrique. **Como identificar sobrepreço e superfaturamento?** Brasília/DF: Associação dos Membros dos Tribunais de Contas do Brasil, 2022. Disponível em: <https://atrimon.org.br/como-identificar-sobrepreco-e-superfaturamento/>. Acesso em: 10 set. 2023. Citada 1 vez na página 33.

LIMA, Wendell da Cunha. Dados abertos governamentais no contexto da ciência cidadã: o caso da operação serenata de amor. *In: IX ENCONTRO IBÉRICO DA ASOCIACIÓN DE EDUCACIÓN E INVESTIGACIÓN EN CIENCIA DE LA INFORMACIÓN DE IBEROAMÉRICA Y EL CARIBE (EDICIC)*, 9., 2019. **Anais** [...]. [S. l.]: Asociación de Educación e Investigación en Ciencia de la Información de Iberoamérica y el Caribe (EDICIC), 2019. Disponível em: <http://hdl.handle.net/10316/95874>. Acesso em: 21 mar. 2022. Citada 1 vez na página 27.

LIU, Fei Tony; TING, Kai; ZHOU, Zhi-Hua. Isolation forest. *In: ICDM '08. EIGHTH IEEE INTERNATIONAL CONFERENCE ON DATA MINING*, 8., 2009. **Proceedings** [...]. [S. l.]: IEEE, 2009. p. 413-422. Disponível em: https://www.researchgate.net/publication/224384174_Isolation_Forest. Acesso em: 20 abr. 2023. Citada 1 vez na página 52.

MARTINS, Eulália Maria Braga. **Fraudes nos processos de licitações e procedimentos preventivos**. 2020. Trabalho de Conclusão de Curso (Direito) – Centro Universitário UNIFACIG, 2020. Disponível em: <http://pensaracademico.facig.edu.br/index.php/repositoriottcc/article/view/2484>. Acesso em: 14 out. 2022. Citada 1 vez na página 33.

MENDES, Roselaine da Cruz; OLEIRO, Walter Nunes; QUINTANA, Alexandre Costa. A contribuição da contabilidade e auditoria governamental para uma melhor transparência na gestão pública em busca do combate contra a corrupção. **Repositório Institucional da FURG**, 2008. Disponível em: <http://repositorio.furg.br/handle/1/5434>. Acesso em: 22 mar. 2022. Citada 2 vezes na página 36(2x).

MONDO, Bianca Vaz. **Métodos de detecção de fraude e corrupção em contratações públicas**. São Paulo/SP: Transparência Brasil, 2019. 76 p. Disponível em: <https://www.transparencia.org.br/downloads/publicacoes/MetodosDetecçãodeFraude.pdf>. Acesso em: 3 set. 2022. Citada 1 vez na página 33.

NASCIMENTO, Anderson. **O que é API?** [S. l.]: Canaltech, 2014. Disponível em: <https://canaltech.com.br/software/o-que-e-api>. Acesso em: 5 out. 2023. Citada 2 vezes nas páginas 35 e 79.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. NIST big data interoperability framework: Volume 1, definitions. National Institute of Standards and Technology, 2015. Disponível em: <https://doi.org/10.6028/NIST.SP.1500-1r2>. Acesso em: 12 out. 2022. Citada 1 vez na página 25.

NOHARA, Irene Patrícia; COLOMBO, Bruna Armonas. Tecnologias cívicas na interface entre direito e inteligência artificial: Operação serenata de amor para gostosuras ou travessuras? **A&C - Revista de Direito Administrativo & Constitucional**, Revista de Direito Administrativo & Constitucional, v. 19, n. 76, p. 83, 2019. Disponível em: <http://www.revistaaec.com/index.php/revistaaec/article/download/1100/807>. Acesso em: 22 mar. 2022. Citada 1 vez na página 27.

OLIVEIRA, Cristiano Cesar da Silva. **O uso de Inteligência Artificial para Controle Social da Administração Pública: Uma Análise da Operação Serenata de Amor**. 2018. Trabalho de Conclusão de Curso (Especialização em Gestão Pública) – Universidade Federal de São João del-Rei, 2018. Disponível em: <http://hdl.handle.net/123456789/267>. Acesso em: 21 mar. 2022. Citada 2 vezes nas páginas 25 e 27.

OPEN KNOWLEDGE BRASIL. **Operação Serenata de Amor**. São Paulo/SP: Open Knowledge Brasil, 20–. Disponível em: <https://www.facebook.com/operacaoSerenataDeAmor>. Acesso em: 3 abr. 2023. Citada 1 vez na página 40.

OPEN KNOWLEDGE BRASIL. **Painel Jarbas**. São Paulo/SP: Open Knowledge Brasil, 20–. Disponível em: https://jarbas.serenata.ai/dashboard/chamber_of_deputies/reimbursement. Acesso em: 3 abr. 2023. Citada 1 vez na página 40.

OPEN KNOWLEDGE BRASIL. **Operação Serenata de Amor**. São Paulo/SP: Open Knowledge Brasil, 2015–. Disponível em: <https://serenata.ai>. Acesso em: 15 out. 2022. Citada 3 vezes nas páginas 40(2x) e 41.

PANG, Guansong *et al.* Deep learning for anomaly detection: A review. **ACM Computing Surveys**, ACMPUB27New York, NY, USA, v. 54, n. 2, p. 1-38, mar. 2021. Disponível em:

<https://doi.org/10.1145%2F3439950>. Acesso em: 16 out. 2022. Citada 3 vezes nas páginas 38 e 53(2x).

PANIS, Amanda da Cunha. **Inovação em compras públicas: estudo de caso do robô ALICE da Controladoria-Geral da União (CGU)**. 2020. Dissertação (Mestrado em Administração) – Universidade de Brasília, 2020. Disponível em: <https://repositorio.unb.br/handle/10482/38639>. Acesso em: 21 mar. 2022. Citada 3 vezes nas páginas 25, 27 e 40.

PECI, Alketa; BRAGA, Marcus Vinicius de Azevedo. Corrupção e capacidades assimétricas da gestão e do controle. **Estadão**, São Paulo/SP, 19/04/2021 abr. 2021. Disponível em: [https://repositorio.ufsc.br/bitstream/handle/123456789/222445/\[corrupÃ§Ã£o\] corrupÃ§Ãoe capacidades assimÃªtricas da gestÃ£o do controle-estadao.pdf?sequence=1&isAllowed=y](https://repositorio.ufsc.br/bitstream/handle/123456789/222445/[corrupÃ§Ã£o] corrupÃ§Ãoe capacidades assimÃªtricas da gestÃ£o do controle-estadao.pdf?sequence=1&isAllowed=y). Acesso em: 21 mar. 2022. Citada 3 vezes na página 26(3x).

PINHO, Maria Nazare Goncalves; GOUVEIA, Luis Borges. O uso do governo digital pelo controle social no combate à corrupção pública brasileira. **Revista Controle - Doutrina e Artigos**, v. 17, n. 2, p. 206–237, 2019. ISSN 1980-086X. Disponível em: <https://dialnet.unirioja.es/servlet/articulo?codigo=7671481>. Acesso em: 22 mar. 2022. Citada 1 vez na página 37.

POSSAMAI, Ana Júlia. **Dados abertos no governo federal brasileiro: desafios de transparência e interoperabilidade**. 2016. 313 f. Tese (Doutorado em Ciência Política) – Universidade Federal do Rio Grande do Sul. Instituto de Filosofia e Ciências Humanas. Programa de Pós-Graduação em Ciência Política., Porto Alegre/RS, 2016. Disponível em: <http://hdl.handle.net/10183/156363>. Acesso em: 25 abr. 2022. Citada 1 vez na página 34.

ROSILHO, André Janjácómo. **Qual é o modelo legal das licitações no Brasil?: as reformas legislativas federais no sistema de contratações públicas**. 2011. Dissertação de mestrado (Direito) – Fundação Getúlio Vargas, São Paulo, 2011. Disponível em: <https://bibliotecadigital.fgv.br/dspace/handle/10438/8824>. Acesso em: 20 abr. 2022. Citada 1 vez na página 33.

SCHRAMM, Fernanda Santos. **O Compliance como instrumento de combate à corrupção no âmbito das contratações públicas**. 2018. Dissertação (Mestrado em Direito) – Universidade Federal de Santa Catarina, 2018. Disponível em: <https://repositorio.ufsc.br/handle/123456789/190091>. Acesso em: 22 mar. 2022. Citada 1 vez na página 34.

SENADO FEDERAL. **Dados Abertos**. [S. l.: s. n.], 2022. Disponível em: <https://www12.senado.leg.br/dados-abertos>. Acesso em: 29 abr. 2022. Citada 5 vezes nas páginas 35, 96, 97, 98 e 99.

SILVA, Amanda Vieira e. **Dados governamentais abertos à luz da accountability : um estudo da Operação Serenata de Amor**. 2018. Trabalho de Conclusão de Curso (Bacharelado em Ciência Política) – Universidade de Brasília, Brasília, 2018. Disponível em: <https://bdm.unb.br/handle/10483/22555>. Acesso em: 21 mar. 2022. Citada 1 vez na página 27.

SILVEIRA, Suêldes Matias. **A inteligência de fontes abertas na prevenção e combate a corrupção em processo licitatório para aquisição de equipamentos de engenharia**. 2021. Trabalho de Conclusão de Curso (Especialização) – Curso Gestão, Assessoramento e Estado-Maior – Escola de Formação Complementar do Exército, 2021. Disponível em: <https://bdex.eb.mil.br/jspui/handle/123456789/9529>. Acesso em: 23 abr. 2022. Citada 4 vezes nas páginas 25, 32(2x) e 33.

SOBREAJUSTE. In: WIKIPÉDIA, a enciclopédia livre. Flórida, EUA: Wikimedia Foundation, 2021. Disponível em: <https://pt.wikipedia.org/w/index.php?title=Sobreajuste&oldid=60665169>. Acesso em: 5 out. 2023. Citada 1 vez na página 79.

SPEDICATO, Giorgio Alfredo; DUTANG, Christophe; PETRINI, Leonardo. Machine Learning Methods to Perform Pricing Optimization. A Comparison with Standard GLMs. **Variance**, Casualty Actuarial Society, v. 12, n. 1, p. 69-89, 2018. Disponível em: <https://hal.archives-ouvertes.fr/hal-01942038>. Acesso em: 21 mar. 2022. Citada 2 vezes na página 39(2x).

TERRA NETO, Rubens Vasconcellos. **Repositório IA-PP-MESTRADO do TCC de Mestrado Profissional da Câmara dos Deputados**. Brasília/DF: Rubens Vasconcellos Terra Neto, 2023. Disponível em: <https://github.com/terraneto/IA-PP-Mestrado>. Acesso em: 3 abr. 2023. Citada 1 vez na página 49.

TRIBUNAL DE CONTAS DA UNIÃO. **5 motivos para a abertura de dados na Administração Pública**. [S. l.: s. n.], 2015. Disponível em: <https://portal.tcu.gov.br/5-motivos-para-a-abertura-de-dados-na-administracao-publica.htm>. Acesso em: 25 abr. 2022. Citada 1 vez na página 34.

VILHENA, Eduardo Juntolli *et al.* Técnicas econométricas e seu papel inovador no cálculo do sobrepreço: o caso da lava jato. n. 138, 2017. ISSN 2594-6501. Disponível em: <https://portal.tcu.gov.br/biblioteca-digital/tecnicas-econometricas-e-seu-papel-inovador-no-calculo-do-sobrepreco-o-caso-da-lava-jato.htm>. Acesso em: 14 out. 2022. Citada 1 vez na página 33.

WOLSTAD, Henrik I W. **Machine learning as a tool for improved housing price prediction : the applicability of machine learning in housing price prediction and the economic implications of improvement to prediction accuracy**. 2020. Dissertação (Mestrado) – Norwegian School of Economics, 2020. Disponível em: <https://openaccess.nhh.no/nhh-xmlui/handle/11250/2739783>. Acesso em: 21 mar. 2022. Citada 1 vez na página 38.

ZHAO, Yue. **Python Outlier Detection (PyOD)**. [S. l.: s. n.], 20—. Disponível em: <https://github.com/yzhao062/pyod>. Acesso em: 27 mar. 2023. Citada 2 vezes nas páginas 39 e 85.

ZHAO, Yue *et al.* Suod: Accelerating large-scale unsupervised heterogeneous outlier detection. In: MACHINE LEARNING AND SYSTEMS. **Proceedings** [...]. [S. l.: s. n.], 2021. Citada 3 vezes na página 55(3x).

ZHAO, Yue; NASRULLAH, Zain; LI, Zheng. Pyod: A python toolbox for scalable outlier detection. **Journal of Machine Learning Research**, v. 20, n. 96, p. 1-7, 2019. Disponível em: <http://jmlr.org/papers/v20/19-011.html>. Acesso em: 27 mar. 2023. Citada 1 vez na página 39.

GLOSSÁRIO

API

API é um conjunto de rotinas e padrões de programação para acesso a um aplicativo de software ou plataforma baseado na Web. A sigla API refere-se ao termo em inglês *Application Programming Interface* que significa em tradução para o português “Interface de Programação de Aplicativos (NASCIMENTO, 2014) .

JSON

JSON, é um acrônimo de *JavaScript Object Notation*, é um formato compacto, de padrão aberto independente, de troca de dados simples e rápida entre sistemas que utiliza texto legível a humanos, no formato atributo-valor (natureza auto-descritiva).(JSON, 2022).

Overfitting

Overfitting, Sobre-ajuste ou Sobreajuste é um termo usado em estatística para descrever quando um modelo estatístico se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados (SOBREAJUSTE, 2021) .

URL

URL significa *Uniform Resource Locator* (Localizador Uniforme de Recursos, em tradução livre). É um termo utilizado para descrever o endereço de um recurso na Internet. (DUTRA, 2023) .

ANEXOS

ANEXO A – ALGORITMOS DO PYOD

Quadro 12 – Algoritmos do PyOD

Tipo	Sigla	Algoritmo	Ano
Probabilistic	ECOD	Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions	2022
Probabilistic	ABOD	Angle-Based Outlier Detection	2008
Probabilistic	FastABOD	Fast Angle-Based Outlier Detection using approximation	2008
Probabilistic	COPOD	COPOD: Copula-Based Outlier Detection	2020
Probabilistic	MAD	Median Absolute Deviation (MAD)	1993
Probabilistic	SOS	Stochastic Outlier Selection	2012
Probabilistic	QMCD	Quasi-Monte Carlo Discrepancy outlier detection	2001
Probabilistic	KDE	Outlier Detection with Kernel Density Functions	2007
Probabilistic	Sampling	Rapid distance-based outlier detection via sampling	2013
Probabilistic	GMM	Probabilistic Mixture Modeling for Outlier Analysis	
Linear Model	PCA	Principal Component Analysis (the sum of weighted projected distances to the eigenvector hyperplanes)	2003
Linear Model	KPCA	Kernel Principal Component Analysis	2007
Linear Model	MCD	Minimum Covariance Determinant (use the mahalanobis distances as the outlier scores)	1999
Linear Model	CD	Use Cook's distance for outlier detection	1977
Linear Model	OCSVM	One-Class Support Vector Machines	2001
Linear Model	LMDD	Deviation-based Outlier Detection (LMDD)	1996
Proximity-Based	LOF	Local Outlier Factor	2000
Proximity-Based	COF	Connectivity-Based Outlier Factor	2002
Proximity-Based (Incremental)	COF	Memory Efficient Connectivity-Based Outlier Factor (slower but reduce storage complexity)	2002
Proximity-Based	CBLOF	Clustering-Based Local Outlier Factor	2003
Proximity-Based	LOCI	LOCI: Fast outlier detection using the local correlation integral	2003
Proximity-Based	HBOS	Histogram-based Outlier Score	2012

Continua na próxima página

Quadro 12 – Continuação

Tipo	Sigla	Algoritmo	Ano
Proximity-Based	kNN	k Nearest Neighbors (use the distance to the kth nearest neighbor as the outlier score)	2000
Proximity-Based	AvgKNN	Average kNN (use the average distance to k nearest neighbors as the outlier score)	2002
Proximity-Based	MedKNN	Median kNN (use the median distance to k nearest neighbors as the outlier score)	2002
Proximity-Based	SOD	Subspace Outlier Detection	2009
Proximity-Based	ROD	Rotation-based Outlier Detection	2020
Outlier Ensembles	IForest	Isolation Forest	2008
Outlier Ensembles	INNE	Isolation-based Anomaly Detection Using Nearest-Neighbor Ensembles	2018
Outlier Ensembles	FB	Feature Bagging	2005
Outlier Ensembles	LSCP	LSCP: Locally Selective Combination of Parallel Outlier Ensembles	2019
Outlier Ensembles	XGBOD	Extreme Boosting Based Outlier Detection (Supervised)	2018
Outlier Ensembles	LODA	Lightweight On-line Detector of Anomalies	2016
Outlier Ensembles	SUOD	SUOD: Accelerating Large-scale Unsupervised Heterogeneous Outlier Detection (Acceleration)	2021
Neural Networks	AutoEncoder	Fully connected AutoEncoder (use reconstruction error as the outlier score)	
Neural Networks	VAE	Variational AutoEncoder (use reconstruction error as the outlier score)	2013
Neural Networks	Beta-VAE	Variational AutoEncoder (all customized loss term by varying gamma and capacity)	2018
Neural Networks	SO_GAAL	Single-Objective Generative Adversarial Active Learning	2019
Neural Networks	MO_GAAL	Multiple-Objective Generative Adversarial Active Learning	2019
Neural Networks	DeepSVDD	Deep One-Class Classification	2018
Neural Networks	AnoGAN	Anomaly Detection with Generative Adversarial Networks	2017

Continua na próxima página

Quadro 12 – Continuação

Tipo	Sigla	Algoritmo	Ano
Neural Networks	ALAD	Adversarially learned anomaly detection	2018
Graph-based	R-Graph	Outlier detection by R-graph	2017
Graph-based	LUNAR	LUNAR: Unifying Local Outlier Detection Methods via Graph Neural Networks	2022

Fonte: Repositório do PyOD no GitHub ([ZHAO, 20–](#))

ANEXO B – API DE COMPRAS GOVERNAMENTAIS

B.1 Contratos a partir de 2021

Módulo Contratos a partir de 2021

Possibilita a obtenção de dados sobre os contratos a partir de 2021 realizadas pelo Governo Federal.

Quadro 13 – Métodos de Consultas Básicas de Contratos

Método	Descrição
contratos	Retorna uma relação de contratos cadastrados.
cronogramas	Retorna uma relação de cronogramas cadastrados.
despesas_acessorias	Retorna uma relação de despesas acessórias cadastrados.
empenhos	Retorna uma relação de empenhos cadastrados.
faturas	Retorna uma relação de faturas cadastrados.
garantias	Retorna uma relação de garantias cadastrados.
historicos	Retorna uma relação de históricos cadastrados.
itens_compras_contratos	Retorna uma relação de itens de contratos cadastrados.
prepostos	Retorna uma relação de prepostos cadastrados.
responsaveis	Retorna uma relação de responsáveis cadastrados.
terceirizados	Retorna uma relação de terceirizados cadastrados.

Fonte: API de Compras Governamentais (BRASIL, 2021)

Quadro 14 – Métodos de Informações Detalhadas de Contratos

Método	Descrição
contrato	Fornece dados detalhados de um contrato selecionado.
cronograma	Fornece dados detalhados de um cronograma selecionado.
despesa_acessoria	Fornece dados detalhados de uma despesa acessória selecionada.
empenho	Fornece dados detalhados de um empenho selecionado.
fatura	Fornece dados detalhados de uma fatura selecionada.
garantia	Fornece dados detalhados de uma garantia selecionada.
historico	Fornece dados detalhados de um histórico selecionado.
item_compras_contratos	Fornece dados detalhados de um item de contrato selecionado.
preposto	Fornece dados detalhados de um preposto selecionado.

Continua na próxima página

Quadro 14 – Continuação

Método	Descrição
responsavel	Fornece dados detalhados de um responsável selecionado.
terceirizado	Fornece dados detalhados de um terceirizado selecionado.

Fonte: API de Compras Governamentais (BRASIL, 2021)

B.2 Fornecedores

Módulo de fornecedores

Possibilita a obtenção de dados relativos ao Cadastro de Fornecedores.

Quadro 15 – Métodos de Consultas Básicas de Fornecedores

Método	Descrição
ambitos_ocorrendia	Fornece uma lista com informações relacionadas aos âmbitos de ocorrência.
cnaes	Fornece uma lista com informações relacionadas aos códigos da Classificação Nacional de Atividade Econômica - CNAE.
fornecedores	Retorna lista de fornecedores de acordo com os parâmetros informados inicialmente.
linhas_fornecimento	Retorna dados sobre linhas de fornecimento e fornecedores relacionados.
municipios	Retorna dados sobre municípios de acordo com parâmetros informados, com possibilidade de relacionar os fornecedores e unidades cadastradoras do município selecionado.
naturezas_juridicas	Retorna lista de naturezas jurídicas através de parâmetros informados, com possibilidade de consultar dados de fornecedores relacionados com uma natureza jurídica selecionada.
ocorrencias_fornecedores	Retorna lista com ocorrências registradas por fornecedor, de acordo com parâmetros informados.
portes_empresa	Fornece lista de portes de fornecedores com possibilidade de consultar dados de fornecedores relacionados com um porte selecionado.
prazos_ocorrendia	Fornece uma lista com informações relacionadas aos prazos de ocorrência.
ramos_negocio	Fornece lista com ramos de negócio de fornecedores, com possibilidade de consultar dados de fornecedores relacionados com um ramo selecionado.

Continua na próxima página

Quadro 15 – Continuação

Método	Descrição
tipos_ocorrendia	Retorna lista com tipos de ocorrências.

Fonte: API de Compras Governamentais (BRASIL, 2021)

Quadro 16 – Métodos de Informações Detalhadas de Fornecedores

Método	Descrição
ambito_ocorrendia	Fornece uma descrição detalhada relacionada a um âmbito de ocorrência informado.
cnae	Fornece descrição detalhada do Código Nacional de Atividade Econômica – CNAE do fornecedor selecionado.
fornecedor_pf	Fornece dados detalhados de um fornecedor Pessoa Física cadastrado no SICAF.
fornecedor_pj	Fornece dados detalhados de um fornecedor Pessoa Jurídica cadastrado no SICAF.
linha_fornecimento	Fornece uma descrição detalhada da linha de fornecimento de produtos e serviços.
municipio	Fornece a descrição do município da localidade do fornecedor.
natureza_juridica	Fornece a descrição da natureza jurídica do fornecedor.
ocorrendia_fornecedor	Fornece dados relativos às ocorrências sofridas pelo fornecedor e registradas no SICAF.
porte_empresa	Fornece dados sobre o porte do fornecedor.
prazo_ocorrendia	Fornece uma descrição detalhada relacionada a um prazo de ocorrência informado.
ramo_negocio	Fornece uma descrição detalhada relacionada a um ramo de negócio informado.
tipo_ocorrendia	Fornece dados detalhados sobre os tipos de ocorrências relacionadas com fornecedores e que estejam cadastradas no SICAF.

Fonte: API de Compras Governamentais (BRASIL, 2021)

B.3 Licitações

Módulo de Licitações

Possibilita a obtenção de dados sobre as Licitações realizadas pelo Governo Federal.

Quadro 17 – Métodos de Consultas Básicas de Licitações

Método	Descrição
irps	Retorna uma relação de intenção de registro de preços cadastrados.
itens_irp	Retorna uma relação de itens de uma intenção de registro de preços cadastrada.
participantes_item_irp	Retorna uma relação de UASGs participantes de um item de uma intenção de registro de preços.
licitacoes	Fornece dados sobre licitações cadastradas.
itens_licitacao	Fornece dados sobre itens de uma licitação cadastrada.
modalidades_licitacao	Retorna uma lista de Modalidades de Licitação.
orgaos	Retorna uma lista de Órgãos.
precos_praticados	Retorna uma lista de dados sobre os preços praticados nas licitações, obtidos através do SISPP.
itens_preco_praticado	Retorna uma lista de dados sobre os itens de preços praticados nas licitações, obtidos através do SISPP.
registros_preco	Fornece uma lista de dados sobre registros dos preços.
itens_registro_preco	Fornece uma lista de dados sobre itens de registros dos preços.
fornecedores_item_registro_preco	Fornece uma lista de dados sobre fornecedores de item de registros dos preços.
renegociacoes_fornecedor_item_registro_preco	Fornece uma lista de dados sobre renegociações de fornecedor de item de registros dos preços.
uasgs	Fornece uma lista com informações relacionadas às UASGs.
rdds	Fornece uma lista com informações relacionadas às RDCs.

Fonte: API de Compras Governamentais (BRASIL, 2021)

Quadro 18 – Métodos de Informações Detalhadas das Licitações

Método	Descrição
irp	Fornece dados detalhados de uma intenção de registro de preços selecionada.
item_irp	Fornece dados detalhados de um item de intenção de registro de preços selecionada.
participante_item_irp	Fornece dados detalhados de um participante de item de intenção de registro de preços selecionado.
licitacao	Retorna dados detalhados de uma licitação selecionada.

item_licitacao	Retorna dados detalhados de um item de licitação selecionada.
modalidade_licitacao	Retorna dados detalhados de uma modalidade de licitação.
orgao	Retorna dados detalhados de um órgão.
preco_praticado	Retorna dados detalhados sobre os preços praticados.
item_preco_praticado	Retorna dados detalhados sobre um item de preços praticados.
registro_preco	Retorna dados detalhados sobre os registros de preço.
item_registro_preco	Retorna dados detalhados sobre um item de registros de preço.
fornecedor_item_registro_preco	Retorna dados detalhados sobre um fornecedor de item de registros de preço.
renegociacao_fornecedor_item_registro_preco	Retorna dados detalhados sobre uma renegociação de fornecedor de item de registros de preço.
uasg	Fornece uma descrição detalhada relacionada a uma UASG informado.
rdc	Fornece uma descrição detalhada relacionada a um RDC informado.

B.4 Materiais

Módulo de Materiais

Possibilita a obtenção de dados sobre os Materiais cadastrados no Catálogo de Materiais - CATMAT.

Quadro 19 – Métodos de Consultas Básicas de Materiais

Método	Descrição
classes	Retorna lista de classes de materiais cadastrados no CATMAT de acordo com o grupo pré-selecionado.
grupos	Retorna lista de grupos de materiais cadastrados no CATMAT.
pdms	Retorna lista de PDM – Padrão Descritivo de Materiais cadastrados no CATMAT de acordo com o grupo e classe pré-selecionados.
materiais	Retorna lista de itens de materiais cadastrados no CATMAT de acordo com o grupo, classe e PDM.


Fonte: API de Compras Governamentais (BRASIL, 2021)

Quadro 20 – Métodos de Informações Detalhadas de Materiais

Método	Descrição
classe	Fornece dados detalhados sobre uma classe selecionada.
grupo	Fornece dados detalhados sobre um grupo selecionado.
pdm	Fornece dados detalhados sobre um PDM selecionado.
material	Fornece dados detalhados sobre um item de material selecionado.

Fonte: API de Compras Governamentais (BRASIL, 2021)

**ANEXO C – CHECKLIST DE VERIFICAÇÃO DE PESQUISA DE
PREÇOS UTILIZADA NO SENADO FEDERAL**

<div><div><div>SENADO FEDERAL</div><div></div></div><div>LEMBRETES DE VERIFICAÇÃO (CHECKLIST) - PESQUISA DE PREÇOS</div><div>Secretaria de Administração de Contratações – SADCON Coordenação de Controle e Validação de Processos - COCVAP</div></div>					
A COCVAP com intuito de promover maior celeridade na ratificação da Pesquisa de Preços, disponibiliza como ferramenta facilitadora a listagem de "Lembretes de Verificação" (Checklist) com algumas questões simples e capazes de colaborar com os Órgãos Técnicos no momento de finalizarem a Pesquisa de Preços.					
CRITÉRIOS QUANTITATIVOS E QUALITATIVOS					
Estabelecidos de acordo com o Ato da Diretoria-Geral Nº 9/2015 e outros normativos utilizados pela Administração Pública.					
CRITÉRIOS NORMATIVOS	LEMBRETES DE VERIFICAÇÃO				
	a) Foi anexado ao processo o Documento de Oficialização de Demanda (DOD), o Termo de Referência (TR) e, no caso de obras e serviços de engenharia, o Projeto Básico (PB)?				
	b) O Documento de Oficialização de Demanda (DOD), o Termo de Referência (TR) ou o Projeto Básico (PB), foram elaborados com todas as informações previstas nos incisos dos arts. 10 e 11 do Ato da Diretoria-Geral nº 9/2015?				
	c) A pesquisa de preços foi realizada de acordo com as informações do último aditivo contratual, se for um caso de prorrogação de contrato, ou caso trate de uma nova contratação, conforme a última versão do Termo de Referência (TR) ou do Projeto Básico (PB)? Verificar quantitativos, unidades de medida, especificações, entre outros aspectos presentes no último aditivo contratual (caso de prorrogação de contrato) ou, caso trate de uma nova contratação, na última versão do Termo de Referência (TR) ou do Projeto Básico (PB).				
	d) Consta no Termo de Referência (TR) ou no Projeto Básico (PB) o número sequencial do item do Plano de Contratações e o processo, autuado no SIGAD, está associado a esse item no Sistema de Gestão da Estratégia e Projetos - GEP?				
	e) Os documentos anexos aos autos do processo estão legíveis e devidamente assinados?				
	Observação: Caso já tenha elaborado o Edital de Licitação, verificar se os itens pesquisados correspondem à última versão do documento.				
ITEM VERIFICADO	LEMBRETES DE VERIFICAÇÃO				
1. CONTRATAÇÕES PÚBLICAS	1.1 - Quantas contratações públicas (ex. preços de contratos, certames licitatórios etc.) foram consideradas similares pelo Órgão Técnico para cada item do objeto?				
	1.2 - Quantas contratações públicas estão válidas (180 dias)?				
	1.3 - Foi utilizada contratação para o mesmo objeto ou para objeto semelhante firmado pelo Senado Federal?				
	1.4 - Caso se utilizem apenas amostras de fontes públicas ou se trate de uma importação, o Órgão Técnico considerou aspectos macroeconômicos em sua estimativa (ex.: inflação, mudanças tecnológicas, câmbio, tributação, frete, encargos etc.)?				
	Observação: Caso não tenha conseguido 3 (três) amostras de preços, apresente justificativa, conforme ATO Nº 9/2015 - Art. 12 § 3º.				
ITEM VERIFICADO	LEMBRETES DE VERIFICAÇÃO				
2. MÍDIA ESPECIALIZADA***	2.1 - Quantos preços de mídias especializadas foram encontrados pelo Órgão Técnico?				
	2.2 - Quantas estimativas provenientes de sítios eletrônicos de domínio amplo**** contém data e hora de acesso? (ex. Lojas Virtuais)				
	Observação 1: Ressaltamos, conforme ATO Nº 9/2015 - Art. 12 § 6º que "não serão admitidas amostras de preços obtidas em sítios de leilão ou de intermediação de vendas" (ex. Buscapé, Mercado Livre, Bondfaro, Zoom etc.).				
	Observação 2: Caso não tenha conseguido 3 (três) amostras de preços, apresente justificativa, conforme ATO Nº 9/2015 - Art. 12 § 3º.				
ITEM VERIFICADO	LEMBRETES DE VERIFICAÇÃO				
3. EMPRESAS E FORNECEDORES	3.1 - Foram consultadas as demais empresas, além das vencedoras, que participaram dos pregões com objeto similar e realizados pela Administração Pública?				
	3.2 - Quantos fornecedores (representantes, fabricantes, distribuidores, revendedores etc.) foram consultados pelo Órgão Técnico?				
	3.3 - Quantas solicitações / reiterações foram enviadas para as empresas / fornecedores pelo Órgão Técnico?				
	3.4 - Quantas estimativas foram consideradas pelo Órgão Técnico para cada item do objeto?				
	3.5 - As estimativas refletem às especificações atualizadas do objeto? Verificar quantitativos, unidades de medida, especificações, entre outros aspectos presentes no último aditivo contratual (caso de prorrogação de contrato) ou, caso trate de uma nova contratação, na última versão do Termo de Referência (TR) ou do Projeto Básico (PB).				
	3.6 - Quantas estimativas informam "Marca" e "Modelo" do objeto (se houver a necessidade)?				
	3.7 - Quantas estimativas de preços estão válidas (com data até 180 dias anteriores a esta pesquisa, ou não declaradas fora do prazo de validade pelo fornecedor)?				
	3.8 - Quantas propostas foram recebidas de fornecedores distintos (inclusive de grupos econômicos diferentes) com sócios, endereços, telefones, representantes/analistas de vendas diferentes?				
	3.9 - Foi realizada a consulta ao cadastro das empresas no sítio da Receita Federal?				
	3.10 - Foi realizada a consulta ao cadastro de fornecedores no SICAF (ex. para atestar se tem sócios, endereços, telefones, representantes/analistas de vendas diferentes)?				
	3.11 - Quantas propostas foram obtidas de fornecedores caracterizados como ME/EPP?				
	3.12 - O Órgão Técnico se manifestou quanto à existência de objeção à aplicação do tratamento exclusivo às ME/EPP a que se refere a Lei Complementar nº 123/2006, essa atualizada pela Lei Complementar nº 147/2014?				
	Observação 1: Caso não tenha conseguido 3 (três) amostras de preços, apresente justificativa, conforme ATO Nº 9/2015 - Art. 12 § 3º.				
	Observação 2: No caso de fornecedor exclusivo, comprovada a exclusividade por documentação atualizada, deverá ser feita a busca de pelo menos 3 (três) contratações semelhantes realizadas por outros órgãos da Administração Pública com o mesmo fornecedor para embasar a vantagemidade do preço sugerido ao Senado Federal.				
	ITEM VERIFICADO	LEMBRETES DE VERIFICAÇÃO			
4. SISTEMAS	4.1 - Caso o processo trate da contratação de obras e serviços de engenharia de que trata o Decreto nº 7.983/2013, todos os itens tiveram preços encontrados nas tabelas SINAPI, SICRO ou TCPO (Pini)?				
	4.2 - Caso o processo trate de contratação para fornecimento de combustível(is), os preços foram obtidos do Sistema de Levantamento de Preços da ANP?				
	4.3 - Caso o processo trate da compra de medicamento(s), os preços de todos os itens foram obtidos na Lista de Preços de Medicamentos para Compras Públicas da CMED / ANVISA?				
	Observação: Caso não tenha conseguido 3 (três) amostras de preços, apresente justificativa, conforme ATO Nº 9/2015 - Art. 12 § 3º.				
MAPA / PLANILHA DE PREÇOS - ADG Nº 9/2015		MÉTODO ESTATÍSTICO		SIM	
Art. 11, inciso II, alínea n Caberá ao Órgão Técnico, ao receber o DOD do Órgão Solicitante elaborar a estimativa de custo e respectiva planilha de composição.		MEDIANA		APRESENTOU JUSTIFICATIVA? (Outro Método Estatístico)	NÃO
		MÉDIA			
		OUTRO MÉTODO			
Art. 12 § 7º O preço ou valor de referência será, preferencialmente, calculado pela média ou mediana dos preços pesquisados, podendo ser utilizado outro método que dê ao valor de referência a representação adequada do valor de mercado, desde que não seja superior à média ou mediana.		O MAPA DE PREÇOS ESTÁ ASSINADO? (Deverá ter a assinatura do Servidor responsável)		SIM	NÃO
Legenda: *** Mídia Especializada: Preços de referência publicados por veículos especializados, por exemplo, Revista Técnica, PC & CIA, Lojas Virtuais (legalmente estabelecidas), Tabela FIPE (veículos), W Imóveis (aluguéis e vendas) etc. **** Sítios eletrônicos de domínio amplo: site presente no mercado nacional de comércio eletrônico ou de fabricante do produto, detentor de boa credibilidade no ramo de atuação, desde que seja uma empresa legalmente estabelecida (ex. emissão de nota fiscal). Sempre que possível, a Pesquisa deve recair em sites seguros, detentores de certificados que venha a garantir que estes são confiáveis e legítimos. OBSERVAÇÃO (ADG Nº 9/2015 Art. 18 §§ 1º e 2º): O Órgão Técnico, após concluir a pesquisa de preços, deverá submetê-la à ratificação pela SADCON. Atualmente, a competência pela ratificação está a cargo da COCVAP, a qual validar o cumprimento dos requisitos legais, jurisprudenciais e regulamentares na pesquisa realizada. Se houver alguma inconsistência na pesquisa realizada, por falha ou pelo não cumprimento de determinações legais, jurisprudenciais ou regulamentares, a SADCON/COCVAP deverá apontá-la, cabendo ao Órgão Técnico sanar o que for apontado.					

ANEXO D – PLANILHA DE PREÇOS UTILIZADA NO SENADO FEDERAL

A planilha de preços utilizada pelo Senado Federal para realização de pesquisa de preços possui 4 abas que são: Fonte de Pesquisa (ilustrada na Figura 9), Mapa de Cotações (ilustrada na Figura 10), Mapa de Cotações (total) (ilustrada na Figura 11) e Planilha Estimativa de Despesas (ilustrada na Figura 12). Apresentamos a seguir um modelo delas.

Figura 9 – Aba Fonte de pesquisa

A

B

C

D

E

F

G

H

I

J

K

L

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

FONTE DE PESQUISA DE PREÇOS

Objeto: XXXXXXXXXXXXX

Data: dd de mmmmmmm de aaaa

Processo: 00200.xxxxxxx/201x-xx

Empresas consultadas para cotação que APRESENTARAM propostas:

Nº	Data	CNPJ	Nome do Fornecedor e (ou) Empresa	DDD	Telefone	Fax	E-mail	Contato
1	00/00/00	00.000.000/0001-91	Empresa 1	61	telefone 1	-	-	Vendedor 1
2	00/00/00	00.000.000/0001-91	Empresa 2	61	telefone 2	-	-	Vendedor 2
3	00/00/00	00.000.000/0001-91	Empresa 3	61	telefone 3	-	-	Vendedor 3
4	00/00/00	00.000.000/0001-91	Empresa 4	61	telefone 4	-	-	Vendedor 4
5	00/00/00	00.000.000/0001-91	Empresa 5	61	telefone 5	-	-	Vendedor 5
6	00/00/00	00.000.000/0001-91	Empresa 6	61	telefone 6	-	-	Vendedor 6
7	00/00/00	00.000.000/0001-91	Empresa 7	61	telefone 7	-	-	Vendedor 7
8	00/00/00	00.000.000/0001-91	Empresa 8	61	telefone 8	-	-	Vendedor 8
9	00/00/00	00.000.000/0001-91	Empresa 9	61	telefone 9	-	-	Vendedor 9
10	00/00/00	00.000.000/0001-91	Empresa 10	61	telefone 10	-	-	Vendedor 10
11	00/00/00	00.000.000/0001-91	Empresa 11	61	telefone 11	-	-	Vendedor 11
12	00/00/00	00.000.000/0001-91	Empresa 12	61	telefone 12	-	-	Vendedor 12

xx empresas consultadas para cotação NÃO APRESENTARAM propostas

Fonte de Pesquisa

Mapa de Cotações

Mapa de Cotações (total)

Planilha Estimativa de Despesas

Fonte: (SENADO FEDERAL, 2022)

Figura 10 – Aba Mapa de Cotações

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2			MAPA DE COTAÇÕES										
3			Objeto: XXXXXXXXXXXXX										
4			Processo: 00200.xxxxxxx/201x-xx										
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													
20													
21													
22													
23													
24													

Item	Discriminação dos materiais (especificações)	Qtde.	Un.	Empresa 1	Empresa 2	Empresa 3
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
TOTAL GERAL				0,00	0,00	0,00

Legenda:
N.C. Empresa não apresentou cotação para o item.
N.A. Item não atende às especificações.

Quantidade de preços cotados	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	3

Mapa de Cotações

Planilha Estimativa de Despesas

Fonte: (SENADO FEDERAL, 2022)

Figura 11 – Aba Mapa de Cotações (total)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2	MAPA DE COTAÇÕES												
3	Objeto: xxxxxxxxxxxxxx												
4	Processo: 00200.xxxxxxx/201x-xx												
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													
20													
21													
22													
23													
24													

MAPA DE COTAÇÕES												
Item	Discriminação dos materiais (especificações)	Qtde.	Un.	Preços dos fornecedores (R\$)			Qtde.	Un.				
				Empresa 1	Empresa 2	Empresa 3						
1												
2												
3												
4												
5												
6												
7												
8												
9												
10												
TOTAL GERAL				0,00	0,00	0,00	0,00					
Legenda:												
N.C.	Empresa não apresentou cotação para o item.											
N.A.	Item não atende às especificações.											

Quantidade de preços cotados	0	0	0	0	0	0	0	0	0	0	3
------------------------------------	---	---	---	---	---	---	---	---	---	---	---

<>

Fonte de Pesquisa

Mapa de Cotações

Mapa de Cotações (total)

Planilha Estimativa de Despesas

Fonte: (SENADO FEDERAL, 2022)

ÍNDICE

P

Pergunta de pesquisa, 27

Pré-processamento dos Dados, 39

S

Separação dos dados, 39