

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



**Rubens Vasconcellos Terra Neto**

---

**APRENDIZAGEM DE MÁQUINA NO COMBATE AO SOBREPREÇO DAS COMPRAS  
PÚBLICAS:  
Análise de pesquisas de preços de contratações do Senado Federal**

---

Brasília, DF  
**2024**



**Rubens Vasconcellos Terra Neto**

**APRENDIZAGEM DE MÁQUINA NO COMBATE AO SOBREPREÇO DAS COMPRAS  
PÚBLICAS:**

**Análise de pesquisas 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

**2024**

### **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.

Aprendizagem de máquina no combate ao sobrepreço das compras públicas:  
Análise de pesquisas de preços de contratações do Senado Federal / Rubens  
Vasconcellos Terra Neto. – 2024.

124 f.

Orientador: Fabiano Peruzzo Schwartz.

Impresso por computador.

Dissertação (mestrado profissional) – Câmara dos Deputados, Centro de  
Formação, Treinamento e Aperfeiçoamento (Cefor), 2024.

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**



---

**Rubens Vasconcellos Terra Neto**

**APRENDIZAGEM DE MÁQUINA NO COMBATE AO SOBREPREÇO DAS COMPRAS  
PÚBLICAS:  
Análise de pesquisas 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, 05 de junho de 2024.



*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. O presente estudo relata pesquisa de mestrado que se dedicou a explorar e processar os dados de contratações públicas nos anos de 2022 e 2023. O principal objetivo consistiu no desenvolvimento de uma aplicação com utilização 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 mantêm 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 outubro/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 diversos algoritmos de detecção de anomalia não-supervisionados. Para a escolha dos algoritmos, definição dos parâmetros e seleção do modelo foram feitos testes em até oito categorias de materiais. Optou-se por utilizar no modelo o algoritmo *Copula-Based Outlier Detection* (COPOD). Para avaliação final foram utilizadas 13 categorias e o método de validação cruzada chamado de LeaveOneOut (LOOCV). O resultado final obtido foi um *Recall* médio de 99,54% com uma acurácia de 90,68%.

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 contract. 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. The present study reports a master's research that was dedicated to exploring and processing public contracting data in the years 2022 and 2023. The main objective was to development of an application using algorithms based on Machine Learning, capable of indicating, based on probability estimates, possible signs 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 maintain 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 October/2023 and are available in a public repository. A first prototype for detecting overpricing in purchases was developed based on the library called Python Outlier Detection (PyOD), which presents several algorithms unsupervised anomaly detection. To choose the algorithms, define the parameters and model selection tests were carried out on up to eight material categories. It was chosen for using the Copula-Based Outlier Detection (COPOD) algorithm in the model. For evaluation In the final, 13 categories and a cross-validation method called LeaveOneOut (LOOCV) were used. The final result obtained was an average recall of 99.54% with an accuracy of 90.68%.

**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) . . . . .	37
Figura 2 – Tipos de aprendizado de máquina . . . . .	41
Figura 3 – A Robô Rosie . . . . .	44
Figura 4 – O Painel Jarbas . . . . .	44
Figura 5 – Matriz de confusão . . . . .	47
Figura 6 – Passos de treinamento da validação cruzada K-fold (k=10) . . . . .	61
Figura 7 – Quatro mapas de contorno de pontuações de anomalia . . . . .	67
Figura 8 – Modelo da ferramenta de detecção de sobrepreço . . . . .	76
Figura 9 – Tela de seleção de material . . . . .	81
Figura 10 – Tela de avaliação da pesquisa . . . . .	82
Figura 11 – Tela de preços analisados . . . . .	83
Figura 12 – Histograma do Valor Unitário . . . . .	86
Figura 13 – Histograma da quantidade . . . . .	87
Figura 14 – Histograma da distância . . . . .	88
Figura 15 – Matriz de confusão (13 materiais com mais registros) . . . . .	93



## LISTA DE QUADROS

Quadro 1 – Utilização dos tipos de AM . . . . .	41
Quadro 2 – Número de registros recuperados dos dados abertos do Sistema Integrado de Administração e Serviços Gerais (SIASG) . . . . .	54
Quadro 3 – Distribuição de dados de contratos e licitações por ano . . . . .	54
Quadro 4 – Características do repositório . . . . .	55
Quadro 5 – Descrição dos atributos do repositório . . . . .	55
Quadro 6 – Materiais com a maior quantidade de registros . . . . .	57
Quadro 7 – Avaliação dos Algoritmos de detecção de anomalias . . . . .	64
Quadro 8 – Medidas de Estatística Descritiva por Atributo . . . . .	85
Quadro 9 – Registros avaliados como anomalias pela função de rotulagem <b>preco_anomalo</b>	86
Quadro 10 – Registros avaliados como anomalias pela função de rotulagem <b>quanti- dade_alta</b> . . . . .	87
Quadro 11 – Registros avaliados como anomalias pela função de rotulagem <b>distancia_alta</b>	88
Quadro 12 – Avaliação da normalidade dos Registros avaliados como anomalias pelas funções anteriores . . . . .	89
Quadro 13 – Os 15 modelos com maior <i>Recall</i> . . . . .	89
Quadro 14 – Algoritmos do PyOD . . . . .	113
Quadro 15 – Métodos de Consultas Básicas de Contratos . . . . .	117
Quadro 16 – Métodos de Informações Detalhadas de Contratos . . . . .	117
Quadro 17 – Métodos de Consultas Básicas de Fornecedores . . . . .	118
Quadro 18 – Métodos de Informações Detalhadas de Fornecedores . . . . .	119
Quadro 19 – Métodos de Consultas Básicas de Licitações . . . . .	120
Quadro 20 – Métodos de Informações Detalhadas das Licitações . . . . .	120
Quadro 21 – Métodos de Consultas Básicas de Materiais . . . . .	121
Quadro 22 – Métodos de Informações Detalhadas de Materiais . . . . .	122



## LISTA DE TABELAS

Tabela 1 – Comparativo dos modelos COPOD e KNN+Sampling . . . . .	90
Tabela 2 – Avaliação Final do Modelo . . . . .	91



## LISTA DE ABREVIATURAS E SIGLAS

ADELE	Análise de Disputa em Licitações Eletrônicas
AGATA	Aplicação para Geração de Análise Textual Acelerada
ALICE	acrônimo para Analisador de Licitações, Contratos e Editais
AM	aprendizagem de máquina
APF	administração pública federal
API	<i>Application Programming Interface</i>
ARP	ata de registro de preços
CARINA	Crawler e Analisador de Registros da Imprensa Nacional
CEAP	Cota para o Exercício da Atividade Parlamentar
CGU	Controladoria-Geral da União
CNPJ	Cadastro Nacional de Pessoa Jurídica
ECDF	função de distribuição cumulativa empírica
FN	falso negativo
FP	falso positivo
IA	Inteligência Artificial
IOT	Internet da Coisas
IP	<i>Internet Protocol</i>
JSON	acrônimo de <i>JavaScript Object Notation</i>
LF	função de rotulagem, do inglês <i>labeling function</i>
ML	<i>Machine Learning</i>
MONICA	Monitoramento Integrado para o Controle de Aquisições
NIST	<i>National Institute of Standards and Technology</i>
NN	<i>nearest neighbour</i>
OGD	<i>open government data</i>

PGC	Planejamento e Gerenciamento de Contratações
PyOD	<i>Python Outlier Detection</i>
SIASG	Sistema Integrado de Administração e Serviços Gerais
SISG	Sistema de Serviços Gerais
SOFIA	Sistema de Orientação sobre Fatos e Indícios para o Auditor
STF	Supremo Tribunal Federal
TCU	Tribunal de Contas da União
TN	verdadeiro negativo do inglês <i>true negative</i>
TP	verdadeiro positivo do inglês <i>true positive</i>
UASG	unidade administrativa de serviços gerais
UFES	Universidade Federal do Espírito Santo
UnB	Universidade de Brasília
URL	<i>Uniform Resource Locator</i> (Localizador Uniforme de Recursos, em tradução livre)



## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>27</b>
1.1	Contextualização	27
1.2	Problema de pesquisa	28
1.3	Justificativa	29
1.4	Objetivos	30
1.4.1	Objetivo Geral	30
1.4.2	Objetivos específicos	30
1.5	Escopo do relatório	30
1.6	Organização do Relatório	30
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>33</b>
2.1	Contratações na Administração Pública Federal	33
2.1.1	Licitações públicas	33
2.1.2	A corrupção nas compras públicas	34
2.2	Dados abertos governamentais, Transparência e Controle Social	36
2.2.1	Dados Abertos Governamentais	36
2.2.2	Transparência e Controle Social	38
2.3	Inteligência Artificial, seu uso na Administração Pública Federal e métricas de avaliação dos modelos	39
2.3.1	Inteligência Artificial	40
2.3.2	Uso da Inteligência Artificial no controle dos gastos públicos	43
2.3.3	Métricas de avaliação dos modelos	46
2.4	Recapitulando	49
<b>3</b>	<b>METODOLOGIA E DESENVOLVIMENTO</b>	<b>51</b>
3.1	Metodologia	51
3.2	Bases de Dados Existentes	52
3.3	Ferramenta para obtenção dos dados necessários ao Treinamento da Inteligência Artificial	52
3.4	Repositório de dados de contratações	54
3.5	Técnicas, Métodos e Boas práticas na detecção de anomalias	56
3.5.1	Definição da rotina de pré-processamento dos dados	59
3.5.2	Separação de dados para treinamento e teste	60
3.5.3	Ajustes de hiper-parâmetros	62
3.5.4	Seleção dos Modelos	63

3.5.4.1	Isolation using Nearest Neighbour Ensemble (iNNE)	65
3.5.4.2	<i>K-Nearest Neighbors</i> (KNN)	67
3.5.4.3	<i>Local Outlier Factor</i> (LOF)	69
3.5.4.4	Principal Component Analysis (PCA)	70
3.5.4.5	<i>Rapid Distance-Based Outlier Detection via Sampling</i> (Sampling)	71
3.5.4.6	<i>Empirical-Cumulative-distribution-based Outlier Detec-</i> <i>tion</i> (ECOD)	72
3.5.4.7	<i>Copula-Based Outlier Detection</i> (COPOD)	73
3.5.4.8	Técnicas utilizadas para desenvolvimento do modelo	74
3.5.5	Avaliação final do modelo	75
3.6	Modelo do Aplicativo de detecção de sobrepreço	76
3.6.1	Processo <b>Selecionar a categoria da avaliação</b>	76
3.6.2	Processo <b>Recuperar itens da categoria selecionada</b>	77
3.6.3	Processo <b>Pré-processamento dos dados</b>	77
3.6.4	Processo <b>Treinar o modelo de detecção de anomalias</b>	78
3.6.5	Processo <b>Solicitar ao usuário os dados a serem avaliados</b>	80
3.6.6	Processo <b>Calcular a possibilidade de ser uma anomalia</b>	80
3.6.7	Processo <b>Apresentar os resultados</b>	80
3.7	Protótipo da ferramenta de avaliação de indícios de sobrepreços disponibilizada no Senado Federal	81
3.7.1	Tela de seleção da categoria da avaliação	81
3.7.2	Tela de avaliação de resultados	81
3.7.3	Tela de preços analisados	82
3.8	Questionário	83
3.9	Considerações finais do capítulo	84
4	RESULTADOS E ANÁLISES	85
4.1	Avaliação qualitativa da rotulagem do Snorkel	85
4.2	Resultados da seleção dos algoritmos	89
4.3	Resultados do modelo final	91
4.4	Resumo do capítulo	93
5	CONCLUSÕES E CONSIDERAÇÕES FINAIS	95
5.1	Limitações	96
5.2	Trabalhos Futuros	97

REFERÊNCIAS	99
-------------	----

<b>APÊNDICES</b>	<b>107</b>
<b>APÊNDICE A –FUNÇÕES DE ROTULAGEM (LFS) UTILIZADAS NO SOFTWARE SNORKEL . . . . .</b>	<b>109</b>
<b>ANEXOS</b>	<b>111</b>
<b>ANEXO A –ALGORITMOS DO PYOD . . . . .</b>	<b>113</b>
<b>ANEXO B –API DE COMPRAS GOVERNAMENTAIS . . . . .</b>	<b>117</b>
<b>B.1 Contratos a partir de 2021 . . . . .</b>	<b>117</b>
<b>B.2 Fornecedores . . . . .</b>	<b>118</b>
<b>B.3 Licitações . . . . .</b>	<b>119</b>
<b>B.4 Materiais . . . . .</b>	<b>121</b>
<b>ANEXO C –CHECKLIST DE VERIFICAÇÃO DE PESQUISA DE PREÇOS UTILIZADA NO SENADO FEDERAL . . . .</b>	<b>123</b>



# 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 em que 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 <sup>1</sup>, como se pode ver em [Panis \(2020\)](#), [Araujo, Zullo e Torres \(2020\)](#) e [Oliveira \(2018\)](#).

Segundo o *National Institute of Standards and Technology (NIST)*<sup>2</sup>, “*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).

<sup>1</sup> acrônimo para Analisador de Licitações, Contratos e Editais

<sup>2</sup> 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)<sup>3</sup> 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 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, de forma a indicar uma estimativa da probabilidade de ocorrer sobrepreço. O sobrepreço é

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

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, que foram consolidados em um banco de dados para o 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 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 Junquillo 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 de 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 ferramental 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 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 anomalias;
- 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 a 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 \(2019\)](#) alerta que a formação de cartéis também utiliza outras técnicas como propostas de cobertura, na qual propostas de valores superiores ou com documentação incompleta são apresentadas, validando, assim, a proposta do participante que o cartel definiu que seria o vencedor. O cartel acaba dividindo o mercado por meio 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 para a utilização deste artifício no jogo de planilha, que se destina a ocultar inconsistências nos preços unitários apresentados pela empresa, normalmente inexequíveis, 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.

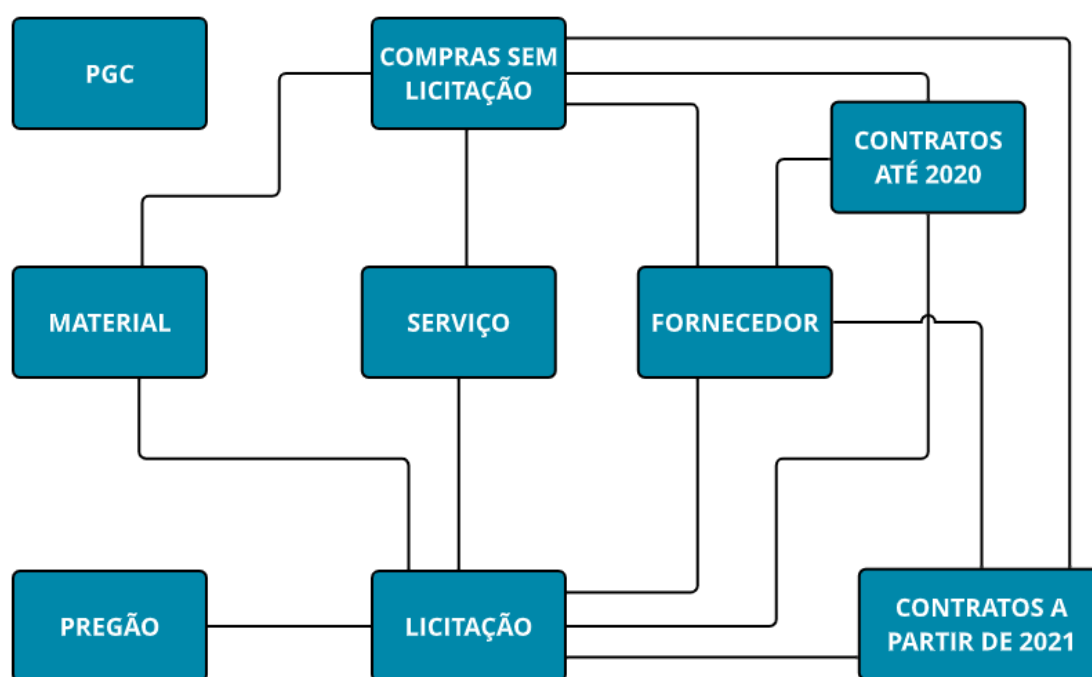
Dentre os dados abertos governamentais disponíveis hoje, dois serão de suma importância neste estudo, os Dados Abertos do SIASG e os Dados Abertos do Senado Federal.

Os Dados Abertos do SIASG são os dados disponibilizados pelo sistema que operacionaliza as compras do Governo Federal. Como informado no *site* da API<sup>1</sup> de Compras Governamentais ele é composto por diversos módulos, a saber:

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, pode-se ter 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 URLs<sup>2</sup>, e retorna os recursos desejados em diversos formatos, incluindo JSON<sup>3</sup> que será o formato utilizado pela pesquisa. A API ainda possui

<sup>1</sup> 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)

<sup>2</sup> 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)

<sup>3</sup> 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)

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 que se divide em dois grandes grupos, Administrativo e Legislativo.

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 abordado 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 será definido 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 AM ou *Machine Learning* emergiu como uma das áreas da IA. Apesar de ainda estar profundamente entrelaçada com as pesquisas de IA, o AM 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 AM tem aumentado dia a dia e diversos estudos tem utilizado esta técnica para prever preços, como abordado por Wolstad (2020).

O AM 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) 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) não-supervisionado: é a técnica onde os dados de treinamento não são rotulados e o sistema tenta aprender sem orientação.
- c) 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) por reforço: 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.

Um resumo dessas técnicas é apresentado na Figura 2.

**Figura 2** – Tipos de aprendizado de máquina

Fonte: Raphaell (2021)

Calanca, Matheus e Raphaell (2023) citam alguns exemplos de utilização para cada um dos tipos de AM apresentados no Quadro 1.

**Quadro 1** – Utilização dos tipos de AM

Tipo	Uso
supervisionado	usado para identificar padrões e classificar os dados, é muito utilizado em <i>softwares</i> de correio eletrônico para determinar o que é <i>spam</i> .
não-supervisionado	pode ser utilizado para marketing através da avaliação de grupos de produtos que o cliente costuma utilizar, ou até para detecção de fraudes, quando o dado difere do padrão conhecido.
semi-supervisionado	usado em sistemas de recomendação, como o do Netflix, onde os dados rotulados são as classificações dos usuários em relação aos filmes e séries assistidos, enquanto os dados não rotulados são os filmes e séries não avaliados por eles. Pelas avaliações o sistema aprende o gosto do usuário e aplica por similaridade aos outros filmes não rotulados para fazer a indicação. O sistema é aprimorado a cada avaliação feita pelo usuário.
por reforço	usado, por exemplo, em sistemas de IA que aprendem a jogar jogo de tabuleiro, como o AlphaGo, que aprendeu a jogar o jogo Go. O AlphaGo analisou milhões de jogos disponíveis para descobrir as melhores estratégias de vitória. Depois de jogar inúmeras vezes, aprimorou seu uso ao aprender com derrotas e vitórias.

Fonte: (CALANCA; MATHEUS; RAPHAELL, 2023)

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 cara. 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.

Nesta pesquisa os dados não são rotulados, o que leva a tendência de utilização das técnicas de aprendizado não-supervisionado ou semi-supervisionado. O objetivo consiste na detecção de sobrepreços, que neste caso são anomalias, ou seja, preços que divergem do esperado. Segundo a definição de [Pang et al. \(2021\)](#), a detecção de anomalias, também conhecida como detecção de *outliers* ou detecção de novidades, é exatamente 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 o autor, 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 em 2017 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 50 algoritmos de detecção, desde o clássico LOF até alguns mais recentes como ECOD e DIF. Desde 2017, o PyOD tem sido utilizado com sucesso em numerosas pesquisas acadêmicas e produtos comerciais. No Quadro 14 do Anexo A estão listados os algoritmos individuais de detecção disponíveis no momento da escrita deste trabalho. Alguns desses algoritmos são avaliados e testados durante a execução da pesquisa, e os testes e resultados são detalhados na seção 3.5

[Zhao \(2023a\)](#) na documentação do PyOD, quando explica sobre detecção de *outliers*<sup>4</sup> explica que a biblioteca PyOD foca nos tipos não-supervisionados e semi-supervisionados e aponta que, nos treinamentos de algoritmos não-supervisionados, os dados de treinamento contêm ambos os tipos de dados, tanto normais como anomalias, enquanto que nos algoritmos semi-supervisionados, os dados de treinamento devem conter apenas dados considerados normais. Como não se pode garantir que não haja sobrepreço nos dados obtidos do SIASG, optou-se por utilizar algoritmos não-supervisionados nesta pesquisa.

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 devem ser seguidos:

- a) Pré-processamento dos Dados: Tarefa que consiste em limpar os dados, selecionar

<sup>4</sup> *Outlier Detection* 101 no link [https://pyod.readthedocs.io/en/latest/relevant\\_knowledge.html](https://pyod.readthedocs.io/en/latest/relevant_knowledge.html)

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 para ser utilizado no treinamento e na validação do modelo e outro, 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, o que é feito nesta etapa.
- 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 3, que também está presente no *Facebook* (OPEN KNOWLEDGE BRASIL, 20–a).

**Figura 3 – 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 4.

**Figura 4 – O Painel 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-)

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 grande projeto destacado 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 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 os resultados são encaminhados aos auditores do TCU por meio de mensagens contendo os apontamentos e riscos detectados.

[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 informações 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 IP, o que caracterizaria um conluio.
- AGATA (Aplicação para Geração de Análise Textual Acelerada), desenvolvida pela Secretaria de Gestão de Informações para o Controle Externo do TCU, baseia-se 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

A avaliação de modelos de IA se dá por meio de métricas específicas. 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.

Uma matriz de confusão é uma tabela usada para avaliar o desempenho de um modelo de classificação. Ela apresenta a performance do modelo ao comparar suas previsões com os valores reais conhecidos do conjunto de dados. Para se montar uma matriz de confusão é necessário apurar a quantidade dos seguintes valores:

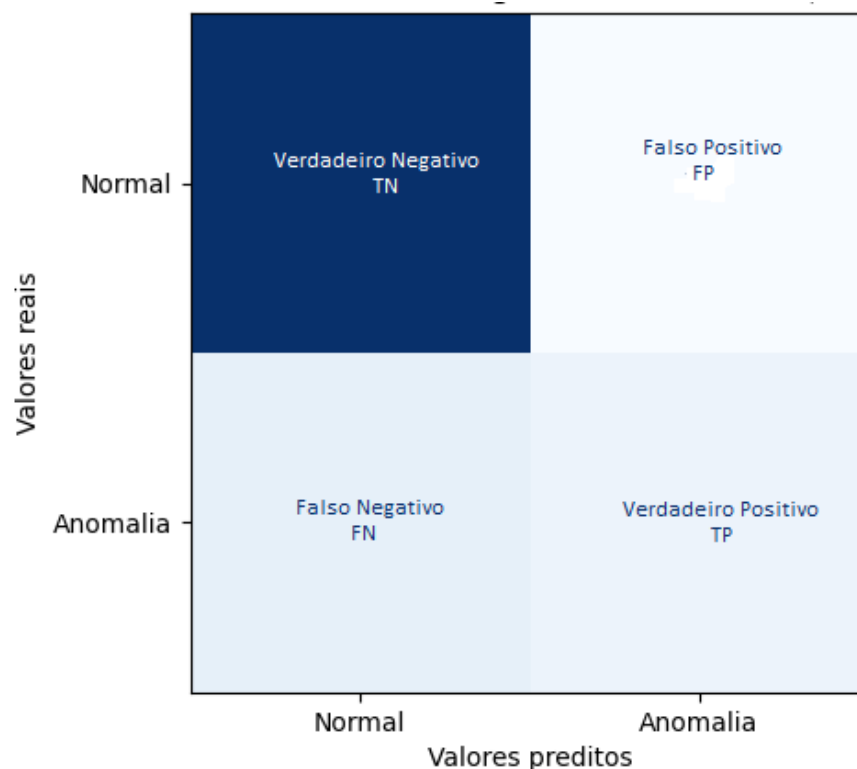
- Verdadeiro positivo (TP, do inglês *True positive*): são os casos em que o modelo prevê corretamente o valor positivo. No caso desta pesquisa, o valor positivo é uma anomalia, ou seja, um valor que diverge do esperado.
- Verdadeiro negativo (TN, do inglês *True negative*): são os casos em que o modelo prevê corretamente o valor negativo. No caso desta pesquisa, o valor negativo é um valor normal, ou uma não anomalia.



- Falso positivo (FP): são os casos em que o modelo prevê incorretamente o valor como positivo, ou seja, em que prevê uma anomalia quando, na verdade, é um valor normal.
- Falso negativo (FN): são os casos em que o modelo prevê incorretamente o valor como negativo, ou seja, em que prevê um valor normal quando, na verdade, é uma anomalia.

Essas métricas são alocadas na matriz de confusão 2x2, conforme ilustrado na Figura 5.

**Figura 5** – Matriz de confusão



Fonte: elaboração própria

A matriz de confusão, segundo (GÉRON, 2021), fornece muita informação e é uma ferramenta poderosa para entender as capacidades do modelo de classificação, especialmente quando há desequilíbrio entre as classes. Essa visualização permite uma análise detalhada dos erros cometidos pelo modelo e ajuda na identificação de áreas que precisam ser melhoradas. Apesar disso, pode-se precisar de métricas mais concisas. Neste caso, por meio da utilização dos valores identificados na matriz de confusão, pode-se calcular 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 de o método 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:

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

F<sub>1</sub>-Score é uma média harmônica calculada com base na Precisão e no *Recall*. Ela pode ser obtida com base na Equação 2.4:

$$F_1-Score = 2 \times \frac{(Precisão \times Recall)}{(Precisão + Recall)} \quad (2.4)$$

Em situações desbalanceadas, as classes minoritárias geralmente são mais importantes e é fundamental que o modelo consiga identificá-las corretamente. Um f<sub>1</sub>score próximo de 1 significa que o modelo tem um bom equilíbrio entre precisão e *recall*. Isso significa que o modelo está corretamente identificando a maioria dos exemplos positivos (alta revocação) e está minimizando o número de falsos positivos (alta precisão). Ao avaliar um modelo de detecção de anomalias, onde as anomalias são um grupo minoritário, é fundamental considerar as consequências dos falsos negativos e falsos positivos para o contexto específico da aplicação. O *recall* é crucial quando a detecção de todas as anomalias é essencial, como nesta pesquisa a intenção é identificar todos os possíveis sobrepreços, esta será a métrica principal de avaliação do modelo.

## 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 dos dados, a descrição do repositório público dos dados disponibilizado, as técnicas, métodos e boas práticas mais recentes utilizados na detecções de anomalias 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.

#### 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 foram coletados o embasamento teórico tanto na área de contratações, como na área de Inteligência Artificial, mais especificamente AM, além de terem sido 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 uma análise quantitativa dos dados, com a intenção de identificar dentre os objetos contratados os que apresentavam maiores quantidades de dados 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 (TERRA NETO, 2023).

Adotou-se a abordagem de pesquisa exploratória para se conhecerem as técnicas, os métodos e as boas práticas utilizados nas detecções de anomalias, 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.

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 (TERRA NETO, 2023).

### 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 URL<sup>1</sup>. 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 uma planilha de preços. Quanto aos dados das contratações, uma vez que o Senado Federal utiliza o Comprasnet, eles já se encontram na base do SIASG, motivo pelo qual foi utilizada apenas a API de compras governamentais para obtenção dos dados.

### 3.3 Ferramenta para obtenção 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 (um 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 isso, foram utilizados parâmetros específicos, como a data da publicação, para limitar o conjunto retornado a um período determinado. Por exemplo, a consulta de todas as licitações de 2022 foi escrita como:

**[http://compras.dados.gov.br/licitacoes/v1/licitacoes.json?data\\_publicacao\\_min=2022-01-01  
& data\\_publicacao\\_max=2022-12-31](http://compras.dados.gov.br/licitacoes/v1/licitacoes.json?data_publicacao_min=2022-01-01&data_publicacao_max=2022-12-31)**

---

<sup>1</sup> 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)

Contudo, este procedimento não funcionou, gerando inúmeras vezes um erro interno do servidor (Erro 500) com a descrição *java.lang.NullPointerException*. Optou-se, então, pela adição de mais um parâmetro denominado UASG, acrônimo de Unidade Administrativa de Serviço Geral. Assim, foi executada uma chamada para cada UASG e para cada ano. Por exemplo, o comando que retorna todas as licitações do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo (UFES) no ano de 2020 é dado por:

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

Foram identificadas 13.381 unidades ativas e efetuadas as recuperações de todas as licitações de cada unidade. 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, retornando somente os 500 primeiros registros, mesmo quando o total de licitações ultrapassasse esse número, não permitindo paginação.

Desta forma optou-se por criar um procedimento para a leitura de todos os arquivos anuais e, sempre que um deles apresentasse 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`**

Esse procedimento não foi suficiente, sendo necessário criar um algoritmo para quebrar um mês a 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 a seguir 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`**

Para a obtenção das licitações desejadas foi necessário criar um procedimento recursivo para poder fazer as quebras descritas acima quando necessário, já que a paginação não funcionava.

Em nenhuma das demais entidades da API este comportamento foi identificado, sendo então possível a paginação para a obtenção dos dados. Tendo sido ao final recuperados a quantidade de registros especificada no Quadro 2.

**Quadro 2** – Número de registros recuperados dos dados abertos do SIASG

<b>Entidade</b>	<b>Quantidade</b>
Classes	710
Contratos	75.565
Fornecedores	580.782
Grupos	79
Itens de contratos	275.050
Itens de licitações	124.002
Itens de preços praticados	829.897
Itens de pregões	23.045
Licitações	48.733
Materiais	211.427
Municípios	5.580
Pregões	44.299
Uasgs	31.440

Fonte: Elaboração própria

Os tipos de entidades recuperados são praticamente autoexplicativos mas, cabe destacar que, os materiais são divididos em grupos e dentro dos grupos temos as classes.

### 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 3.

**Quadro 3** – Distribuição de dados de contratos e licitações por ano

<b>Entidade</b>	<b>Quantidade</b>		
	<b>2022</b>	<b>2023</b>	<b>total</b>
Contratos	42.012	33.553	75.565
Licitações	38.316	10.417	48.733
Itens de materiais de contratos	62.382	44.427	106.809
Itens de materiais de licitações	588.128	95.203	683.331

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 e o valor total foram desprezados, assim como os que tinham ambos os valores iguais a zero. Este procedimento fez com que caísse substancialmente a quantidade de itens de materiais, de aproximadamente 790.140 (setecentos e noventa mil, cento e quarenta) para o total de 136.678 (cento e trinta e seis mil, seiscentos e setenta e oito).

As características do repositório de dados de contratações públicas estão descritas no Quadro 4 e os respectivos atributos no Quadro 5. Os dados da pesquisa foram disponibilizados em um repositório público hospedado no *GitHub* (TERRA NETO, 2023) <sup>2</sup>

**Quadro 4 – Características do repositório**

Característica	Valor
Número de registros (instâncias)	136.678
Número de atributos	11
Número de valores nulos das características	8 <sup>3</sup>
Número de atributos numéricos	9
Número de atributos numéricos não categóricos	4
Número de atributos não numéricos	2

Fonte: Elaboração própria

**Quadro 5 – Descrição dos atributos do repositório**

Atributo	Descrição	Tipo
licitacao_contrato	Determina se é um item de contrato (quando for 0) ou de licitação (quando apresentar um número do item)	Inteiro (categórico)
id	identificador da licitação ou contrato	Inteiro (categórico)
data	Data da assinatura do contrato ou da publicação da licitação	Data (categórico)
catmat_id	identificador da categoria do material do item	Inteiro (categórico)
quantidade	quantidade do item adquirido	Inteiro
unidade	unidade de aquisição do item na licitação	Texto (categórico)
valor_unitario	valor unitário do item	Numérico
valor_total	valor total do item	Numérico

*Continua na próxima página*

<sup>2</sup> No endereço <https://github.com/terraneto/IA-PP-Mestrado>

<sup>3</sup> Esses valores nulos se referem apenas ao atributo município do fornecedor

**Quadro 5** – Continuação

Atributo	Descrição	Tipo
municipio_uasg	Identificador do município onde se encontra a UASG que adquiriu o item	Inteiro (categórico)
municipio_fornecedor	identificador do município onde se encontra o fornecedor do item	Inteiro (categórico)
distancia_uasg_fornecedor	distância entre a UASG e o fornecedor em km calculada através da latitude e longitude entre os municípios da UASG e do fornecedor	Númerico

Fonte: Elaboração própria

Com relação aos atributos do Quadro 5, cabe explicar ainda que:

- Os atributos `licitacao_contrato` e `id` serão utilizados para identificar o item, e esta identificação servirá para apresentar os dados da licitação ou contrato que estão servindo de base para a avaliação;
- os itens `data`, `catmat_id` e `unidade` serão utilizados na seleção dos itens que entrarão na avaliação.
- o campo `valor total` é utilizado para atribuição do valor unitário, quando o mesmo não está informado;
- os campos `municipio_uasg` e `municipio_fornecedor` são utilizados no cálculo da distância entre a uasg e o fornecedor, já que na tabela de municípios temos a latitude e a longitude.
- os campos `quantidade`, `valor_unitario` e `distancia_uasg_fornecedor` serão utilizados no treinamento do modelo de detecção de sobrepreço.

### 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.

Na biblioteca PyOD encontram-se disponíveis vários algoritmos não supervisionados, apresentados no Quadro 14 do Anexo A. Na busca de uma forma de determinar a escolha dos algoritmos a serem utilizados, analisou-se a avaliação de [Han et al. \(2022\)](#) que os classificou 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) anomalia local: são as anomalias que diferem das instâncias vizinhas mais próximas;
- b) anomalia 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, sob três perspectivas:

- 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.

Na avaliação conduzida por [Han et al. \(2022\)](#), chegou-se à 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, decidiu-se realizar uma avaliação dos algoritmos sobre os dados de um material específico selecionado da base.

Para esta escolha, optou-se por utilizar os materiais com maior quantidade de registros. No Quadro 6 estão listados os 13 materiais com maior número de registros e suas quantidades.

**Quadro 6** – Materiais com a maior quantidade de registros

Material	Descrição	Qtd. Registros
104671	PECAS E COMPONENTES - AERONAVE/VEICULO ESPACIAL/SATELITE, PECAS E COMPONENTES - AERONAVE / VEICULO NOME	47.391

*Continua na próxima página*

Quadro 6 – Continuação

Material	Descrição	Qtd. Registros
150658	PEÇAS E ACESSÓRIOS FERRAMENTAS, PEÇAS E ACESSÓRIOS FERRAMENTAS NOME	17.409
445485	ÁGUA MINERAL NATURAL, TIPO SEM GÁS MATERIAL EMBALAGEM PLÁSTICO TIPO EMBALAGEM RETORNÁVEL	741
402920	BEBEDOURO ÁGUA GARRAÇÃO, MATERIAL PLÁSTICO ABS E CHAPA AÇO INOXIDÁVEL TIPO ELÉTRICO DE COLUNA CAPACIDADE 20 L VOLTAGEM 220 V CARACTERÍSTICAS ADICIONAIS BAIXO CONSUMO ENERGIA,TERMOSTATO REGULÁVEL,INMETRO	431
150877	PEÇAS / ACESSÓRIOS ARMAMENTO, PEÇAS / ACESSÓRIOS ARMAMENTO NOME	384
481567	LIVRO DIDÁTICO, GRAU ENSINO SUPERIOR / UNIVERSITÁRIO DEFINIÇÃO COLEÇÃO INTERDISCIPLINAR CONTEÚDO CIÊNCIAS EXATAS E DA TERRA FORMATO IMPRESSO	357
460872	PEÇAS E ACESSÓRIOS APARELHO AR CONDICIONADO, TIPO MOTOR COMPRESSOR CARACTERÍSTICAS ADICIONAIS APARELHO CONDICIONADOR DE AR APLICAÇÃO AR CONDICIONADO	347
469793	MICROCOMPUTADOR, MEMÓRIA RAM SUPERIOR A 8 GB NÚCLEOS POR PROCESSADOR SUPERIOR A 8 ARMAZENAMENTO HDD SEM DISCO HDD GB ARMAZENAMENTO SSD 110 A 300 MONITOR 21 A 29 POL COMPONENTES ADICIONAIS COM TECLADO E MOUSE SISTEMA OPERACIONAL PROPRIETÁRIO GARANTIA ON SITE SUPERIOR A 36 MESES GABINETE COMPACTO	277
461652	GÁS REFINO DE PETRÓLEO, TIPO GÁS LIQUEFEITO DE PETRÓLEO - GLP USO DOMÉSTICO	209
464381	FRUTA, TIPO BANANA PRATA / BANANA BRANCA APRESENTAÇÃO NATURAL	205
461506	GASOLINA, USO PARA AUTOMOTIVOS CLASSIFICAÇÃO COMUM ÍNDICE DE OCTANAGEM IAD 87 MIN	190
463795	LEGUME IN NATURA, TIPO MANDIOCA / AIPIM	159
150515	"LIVRO", LIVRO NOME	158

Fonte: Elaboração própria

Apesar de ser apenas o terceiro item com maior número de registros, optou-se pelo material de catmat 445485 cuja descrição é “ÁGUA MINERAL NATURAL, TIPO SEM GÁS MATERIAL EMBALAGEM PLÁSTICO TIPO EMBALAGEM RETORNÁVEL”, por ser um item adquirido por praticamente todos os órgãos da APF.

### 3.5.1 Definição da rotina de pré-processamento dos dados

Como explicado na subseção 2.3.1, o pré-processamento dos dados implica limpar os dados, selecionar os preditores, verificar a necessidade de utilização de transformações e selecionar as variáveis a serem utilizadas no estágio de modelagem. Neste trabalho de pesquisa, as tarefas de limpeza dos dados, seleção dos preditores e das variáveis foram efetuadas durante a construção do repositório, conforme os passos apresentados na seção 3.4. Nesta seção, discutem-se os procedimentos de transformação de variáveis.

Segundo Géron (2021), salvo raras exceções, os algoritmos de AM não funcionam bem quando os atributos numéricos de entrada tem escalas muito diferentes, o que se verifica entre as variáveis escolhidas nesta pesquisa. Portanto, é necessário criar uma rotina de pré-processamento dos dados para padronizar as escalas. O autor afirma que existem duas formas mais comuns de se realizar esta padronização, baseadas o escalonamento Min-Max ou padronização.

O escalonamento Min-Max, também conhecido como normalização Min-Max, é uma técnica de pré-processamento de dados usada para adequar todos os valores de cada característica a uma escala comum que varia de 0 a 1.

Segundo Pandey (2020) a fórmula para o escalonamento Min-Max é dada pela Equação 3.1:

$$X_{ajustado} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (3.1)$$

Onde:

$X_{ajustado}$  é o valor escalado,

$X_i$  é o valor original,

$X_{min}$  é o menor valor da característica,

$X_{max}$  é o maior valor da característica.

Pandey (2020) define ainda que, o escalonamento por padronização é uma técnica de pré-processamento de dados que transforma os valores de cada característica para que tenham uma média de 0 e um desvio padrão de 1. A padronização é dada pela Equação 3.2:

$$X_{ajustado} = \frac{X_i - \mu}{\sigma} \quad (3.2)$$

Onde:

$X_{ajustado}$  é o valor ajustado,

$X_i$  é o valor original,

$\mu$  é a média dos valores da característica,  
 $\sigma$  é o desvio padrão dos valores da característica.

Pandey (2020) chama a atenção que tanto o escalonamento Min-Max como a padronização não são robustos com relação a presença de anomalias, sugerindo a utilização do método de escalonamento robusto. Explica o autor, que o método é semelhante ao escalonamento Min-Max, mas usa o intervalo interquartil. A mediana e as escalas dos dados são removidas por esse algoritmo de escala de acordo com o intervalo de quantis. Seu ajuste se dá pela Equação 3.3:

$$X_{ajustado} = \frac{X_i - Q1}{Q3 - Q1} \quad (3.3)$$

Onde:

$X_{ajustado}$  é o valor ajustado,

$X_i$  é o valor original,

Q1 é o primeiro quartil dos valores da característica,

Q3 é o terceiro quartil dos valores da característica.

Segundo Géron (2021), o escalonamento Min-Max funciona melhor quando os dados serão utilizados em algoritmos de redes neurais, que é um dos tipos de algoritmos a ser testado, pois a escala fica entre 0 e 1. Mas por sua sensibilidade a *outliers*, optou-se também pela aplicação do escalonamento Min-Max na saída do escalonamento robusto. Esta estratégia pode ser útil quando se deseja primeiro reduzir o impacto de *outliers* e, em seguida, escalar os dados para um intervalo específico.

Foram utilizadas as funções da biblioteca *Scikit-Learn* (PEDREGOSA *et al.*, 2011) *sklearn.preprocessing.MinMaxScaler* e *sklearn.preprocessing.RobustScaler* na implementação da rotina de pré-processamento. Cada um dos algoritmos foi testado utilizando-se o escalonamento dos dados de quatro formas diferentes: o duplo escalonamento Robusto e Min-Max; apenas o escalonamento Robusto; apenas o escalonamento Min-Max; e sem utilização de escalonamentos. O método escolhido para cada um está expresso no Quadro 7, na subseção 3.5.4.

### 3.5.2 Separação de dados para treinamento e teste

Após a definição da rotina de pré-processamento dos dados, passou-se à rotina de separação de dados de treinamento e teste. Silva (2023) destaca que a separação dos dados em conjuntos de treinamento, validação e teste é uma etapa crucial no processo de modelagem de aprendizado de máquina. No entanto, existem vários problemas que podem ocorrer durante essa separação:

- a) Desbalanceamento de classes: Se os dados são desbalanceados, a separação aleatória

dos dados pode resultar em conjuntos de treinamento, validação ou teste que não representam adequadamente todas as classes.

- b) Viés de seleção: Se a separação dos dados não for feita de maneira aleatória, pode haver um viés de seleção. Isso pode levar a um modelo que funciona bem nos dados de treinamento, mas não generaliza bem para novos dados.
- c) Vazamento de dados: Se houver alguma sobreposição entre os conjuntos de treinamento, validação e teste, isso pode levar a um vazamento de dados. Isso significa que o modelo pode ter um desempenho artificialmente bom nos dados de teste porque já viu alguns desses dados durante o treinamento.
- d) Superajuste (*overfitting*) ou subajuste (*underfitting*): A separação inadequada dos dados pode levar a modelos que estão superajustados ou subajustados. Um modelo superajustado é aquele que se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados. Por outro lado, um modelo subajustado é aquele que não se ajusta bem nem aos dados de treinamento.

Durante a análise dos dados verificou-se que tanto as classes estão desbalanceadas, o que é natural em se tratando de detecção de anomalias, como a quantidade de registros existentes na maioria dos materiais é pequeno. [Alhamid \(2020\)](#) chama a atenção ao fato de que nos casos em que o conjunto de dados de treinamento for pequeno, a capacidade de dividi-los em treinamento, validação e teste afetará significativamente a precisão do treinamento. O autor sugere que se utilize então a Validação Cruzada para resolver este problema.

Em seu artigo [Alhamid \(2020\)](#) cita que a validação cruzada tem duas etapas principais: dividir os dados em subconjuntos (chamados de *fold*) e alternar o treinamento e a validação entre eles, como mostrado na Figura 6.

**Figura 6** – Passos de treinamento da validação cruzada K-fold (k=10)

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Fold-6	Fold-7	Fold-8	Fold-9	Fold-10
Passo 1	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Teste
Passo 2	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Teste	Treinamento
Passo 3	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Teste	Treinamento	Treinamento
Passo 4	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Teste	Treinamento	Treinamento	Treinamento
Passo 5	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Teste	Treinamento	Treinamento	Treinamento	Treinamento
Passo 6	Treinamento	Treinamento	Treinamento	Treinamento	Teste	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento
Passo 7	Treinamento	Treinamento	Treinamento	Teste	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento
Passo 8	Treinamento	Treinamento	Teste	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento
Passo 9	Treinamento	Teste	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento
Passo 10	Teste	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento

Fonte: Adaptado de ([ALHAMID, 2020](#))

Segundo [Alhamid \(2020\)](#), a técnica de divisão geralmente deve ter as seguintes propriedades:

- a) Cada *fold* tem aproximadamente o mesmo tamanho.
- b) Os dados podem ser selecionados aleatoriamente em cada *fold* ou estratificados.



- c) Todos os *folds* são usados para treinar o modelo, exceto um, que é usado para validação.
- d) Esse *fold* de validação deve ser alternado até que todos os *folds* se tornem um *fold* de validação uma vez e apenas uma vez.
- e) Recomenda-se que cada exemplo esteja contido em um e apenas um *fold*.

Uma das vantagens da validação cruzada é observar as previsões do modelo em relação a todas as instâncias do conjunto de dados. Ele garante que o modelo foi testado com todos os dados, sem testá-los simultaneamente. São esperadas variações em cada etapa da validação; portanto, calcular a média e o desvio padrão pode reduzir a informação a poucos valores de comparação (ALHAMID, 2020).

Tanto Alhamid (2020) quanto Géron (2021) ao descreverem a validação cruzada, destacam que ela não pode ser usada para conjuntos de dados desequilibrados porque os dados são divididos em *folds* com uma distribuição de probabilidade uniforme, o que pode ocasionar o problema do desbalanceamento de classes conforme informado por Silva (2023). Para solucionar esta questão, a melhor forma é utilizar o método de Validação Cruzada Estratificada, que é uma versão aprimorada da técnica de validação cruzada k-fold. Embora também divida o conjunto de dados em k *folds* iguais, cada *fold* tem a mesma proporção de instâncias de variáveis de destino que estão no conjunto de dados completo. Isso permite que funcione perfeitamente para conjuntos de dados desequilibrados (SHARMA, 2022).

Optou-se por utilizar a técnica de Validação Cruzada Estratificada e a biblioteca *Scikit-Learn* (PEDREGOSA *et al.*, 2011), por meio da função `sklearn.model_selection.StratifiedKFold`.

### 3.5.3 Ajustes de hiper-parâmetros

Após a separação dos dados é realizado o ajuste dos hiper-parâmetros de cada um dos modelos. Hiper-parâmetros são parâmetros de modelos que devem ser definidos antes de treiná-los. Para a definição existem diferentes técnicas que buscam otimizá-los resultando em uma melhor acurácia do modelo (VAZ, 2019).

Vaz (2019) trata de duas técnicas de ajuste de hiper-parâmetros: a *Grid Search* e a *Random Search*. Segundo o autor, a técnica de *Grid Search* testará todas as combinações possíveis dos hiper-parâmetros, exaustivamente, e selecionará os hiper-parâmetros que obtiverem o menor erro, sendo sua desvantagem um gasto computacional maior. Já a *Random Search* testa combinações aleatórias e os melhores resultados funcionam como um guia para a escolha dos próximos hiper-parâmetros. A desvantagem deste método é que isso poderá levar o resultado para o mínimo local e não para o mínimo global.

Optou-se por utilizar a técnica de *Grid Search* e a biblioteca *Scikit-Learn* (PEDREGOSA *et al.*, 2011), por meio da utilização da função `sklearn.model_selection.GridSearchCV`, que já



faz a utilização da Validação Cruzada Estratificada .

Cada um dos modelos testados neste trabalho possui um ou mais hiper-parâmetros avaliados pela técnica de *Grid Search*. O hiper-parâmetro comum a todos os modelos é a contaminação. O parâmetro contaminação é utilizado para informar ao algoritmo a proporção de anomalias que se espera encontrar nos dados. A definição correta deste parâmetro é um desafio, pois normalmente não sabemos a priori quantas anomalias estão presentes nos dados. Este parâmetro é utilizado para ajustar o valor limite para determinar se uma observação é ou não uma anomalia.

Cabe destacar que alguns algoritmos utilizam técnicas de divisão de dados em seu treinamento, e que para garantir que a divisão seja sempre a mesma na busca dos melhores hiper-parâmetros possuem um parâmetro chamado Random-State, cuja escolha é aleatória. Para o presente trabalho adotou-se o valor 69 em todos os algoritmos que utilizam este parâmetro.

### 3.5.4 Seleção dos Modelos

Após as fases anteriores para definir qual o melhor algoritmo e modelo a serem utilizados na aplicação, necessita-se comparar os modelos, por meio das métricas de avaliação explicadas na subseção 2.3.3. Contudo, para estimar as métricas é necessário que os dados sejam rotulados. Foi então utilizado o software Snorkel que permite a rotulagem de dados de forma automática. Segundo o site do software ([SNORKEL AI, 2020](#)), um dos principais componentes do Snorkel são as funções de rotulagem, do inglês labeling functions (LFs), que fazem uso de estratégias heurísticas para rotular dados. No entanto, é importante considerar suas limitações. A qualidade das LFs é crucial para o desempenho do modelo e se mal projetadas podem gerar rótulos de baixa qualidade. A criação de LFs eficazes requer um bom conhecimento do domínio e ainda demandam tempo para sua elaboração. Essas limitações não diminuem o valor do Snorkel como ferramenta para a criação de conjuntos de treinamento, mas é possível que ocorram erros na rotulação dos dados.

Neste trabalho, o Snorkel foi utilizado para a introdução de um campo anomalia onde consta a avaliação realizada pelo software para cada um dos registros. As LFs desenvolvidas para a rotulagem dos dados se encontram no Apêndice A. Foi escolhido também um material com um número menor de registros para avaliar qualitativamente a rotulagem executada

A avaliação do método se deu pelas medidas de *recall* e acurácia. O *recall* tem grande importância na nossa avaliação, pois indica o percentual de anomalias identificadas. Já a acurácia informa o percentual de identificações corretas dentre todas as realizadas. Nesta avaliação o escalonamento foi realizado conforme descrito na subseção 3.5.1 e o ajuste dos hiperparâmetros, conforme descrito na subseção 3.5.3. O resultado da avaliação se encontra no Quadro 7, bem como os melhores hiperparâmetros e o escalonamento de dados selecionado. A execução da avaliação foi realizada por meio do software *Jupyter Notebook* e os *notebooks* utilizados se encontram no repositório do *GitHub* ([TERRA NETO, 2023](#)) na pasta *Jupyter*. Nesta avaliação procurou-se alcançar em cada algoritmo o maior *recall* possível, o que geraria o maior número

de anomalias detectadas.

**Quadro 7** – Avaliação dos Algoritmos de detecção de anomalias

Algoritmo	Hiperparâmetros	Recall	Acurácia	Escalonamento
INNE	contamination=0.05, n_estimators=50	100,00	97,95	Robust
KNN	contamination=0.05, leaf_size=10, method=largest, n_neighbors=10	100,00	97,81	Sem escalonamento
LOF	contamination=0.05, leaf_size=1, n_neighbors=34	100,00	97,81	Sem escalonamento
PCA	contamination=0.05, n_components=3, n_selected_components=1	100,00	97,81	Sem escalonamento
Sampling	contamination=0.05, sub- set_size=10	100,00	97,81	Sem escalonamento
ECOD	contamination=0.09	100,00	93,84	Sem escalonamento
COPOD	contamination=0.12	100,00	90,82	Sem escalonamento
IForest	contamination=0.05, n_estimators=50,	90,48	97,26	Sem escalonamento
MCD	contamination=0.04	90,48	98,22	Sem escalonamento
DeepSVDD	contamination=0.12	90,48	90,27	Robust
OCSVM	contamination=0.12, ker- nel=rbf	90,48	90,27	Min-Max
COF	contamination=0.14, method=fast, n_neighbors=30	90,48	88,90	Min-Max
ALAD	contamination=0.15	90,48	87,26	Robust
HBOS	alpha=0.0001, contami- nation=0.13, n_bins=100, tol=0.00001	85,71	93,97	Min-Max
SOD	contamination=0.05, alpha=0.1	80,95	96,71	Sem escalonamento
LUNAR	contamination=0.1, n_neighbours=5, val_size=0.4	80,95	91,78	Robust

*Continua na próxima página*

Quadro 7 – Continuação

Algoritmo	Hiperparâmetros	Recall	Acurácia	Escalonamento
GMM	contamination=0.11	76,19	91,37	Sem escalonamento
KDE	contamination=0.05, leaf_size=10	71,43	96,16	Robust
DIF	contamination=0.05, n_estimators=1	66,67	95,89	Minmax
CD	contamination=0.05	52,38	95,89	Sem escalonamento
LODA	contamination=0.13, n_bins=70, n_random_cuts=50	47,62	95,21	Sem escalonamento
QMCD	contamination=0.05	33,33	93,97	Sem escalonamento
LMDD	contamination=0.05, n_iter=10	28,57	97,53	Sem escalonamento

Fonte: Elaboração própria

A partir deste resultado, decidiu-se pela adoção dos algoritmos com *recall* igual a 100% para o desenvolvimento da ferramenta de detecção de sobrepreço, sendo selecionados os seguintes: INNE, KNN, LOF, PCA, Sampling, ECOD e COPOD.

Nas próximas seções serão explicados cada um dos algoritmos selecionados e as suas principais características. Não é objetivo deste estudo entrar em aspectos muito técnicos dos modelos, mas, sim, aplicá-los e determinar empiricamente o mais adequado para o caso em questão.

#### 3.5.4.1 Isolation using Nearest Neighbour Ensemble (iNNE)

A documentação da biblioteca PyOD (ZHAO, 2023b) explica que o algoritmo iNNE usa o conjunto de vizinhos mais próximo para isolar anomalias. Ele particiona o espaço de dados em regiões usando uma subamostra e determina uma pontuação de isolamento para cada região. À medida que cada região se adapta à distribuição local, a pontuação de isolamento calculada é uma medida local relativa à vizinhança local, permitindo-lhe detectar anomalias globais e locais. iNNE possui complexidade de tempo linear e complexidade de espaço constante para lidar com eficiência com conjuntos de dados grandes e de alta dimensão com distribuições complexas.

Bandaragoda *et al.* (2018) propôs este método, que é uma abordagem de detecção de anomalias baseadas em isolamento. As definições principais do iNNE conforme apresentadas em seu artigo são:

- Mecanismo de Isolamento: O iNNE utiliza um mecanismo alternativo de isolamento que particiona o espaço de dados para isolar cada instância das demais em uma subamostra, determinando uma pontuação de isolamento para cada região de isolamento.

- Regiões de Isolamento: Cada região é uma hiperesfera definida com um centro representado por uma instância da subamostra, e seu limite é definido pela distância até o vizinho mais próximo (em inglês *nearest neighbour* (NN)), da instância central.

No PyOD (ZHAO, 2023a) o algoritmo possui quatro parâmetros:

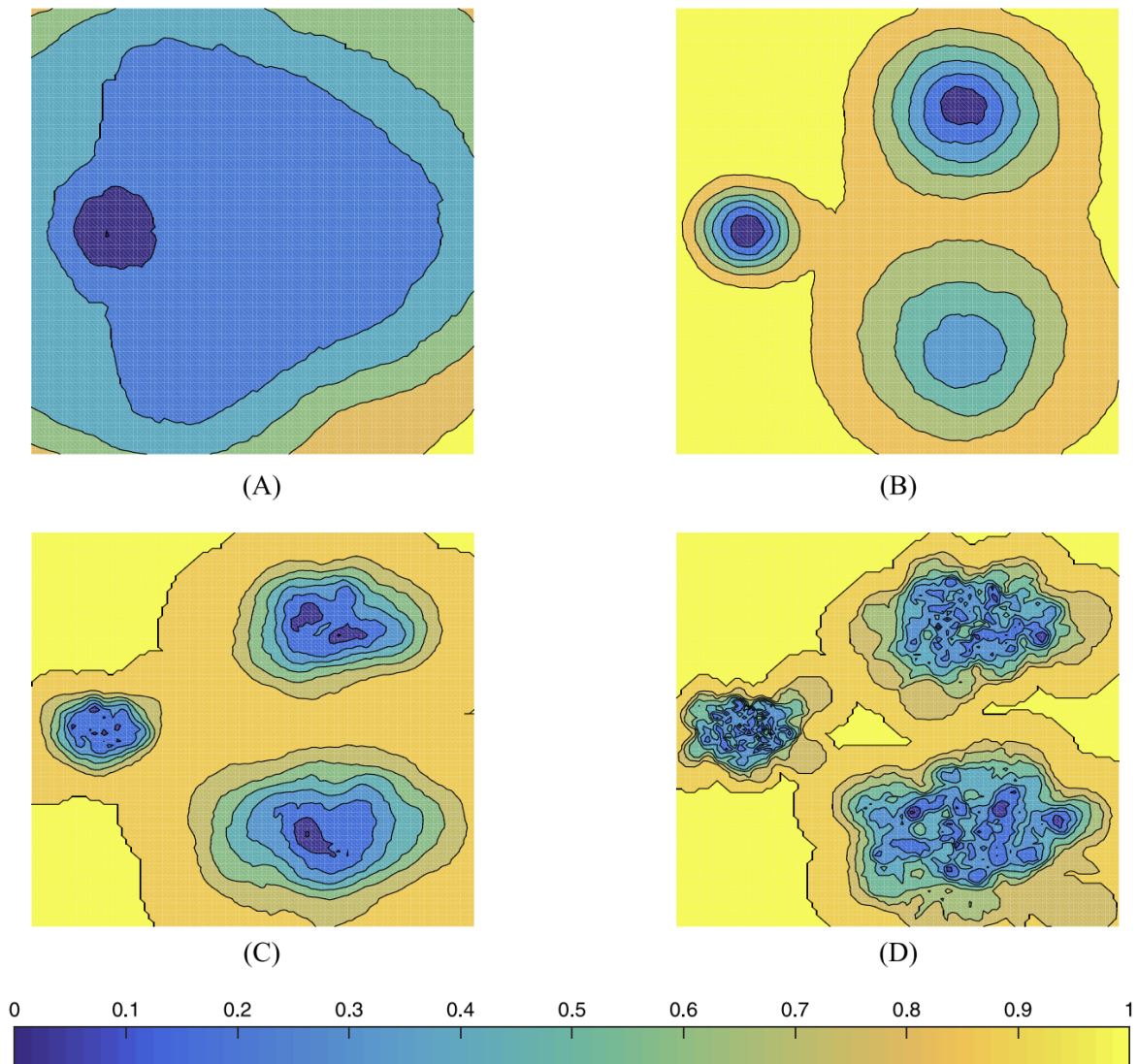
- a) `n_estimators`, que é o número de estimadores de base em um conjunto;
- b) `max_samples`, que é o número de amostras a serem extraídas de  $X$  para treinar cada estimador base;
- c) `contamination`, que é o percentual de anomalias no conjunto de dados;
- d) `random_state`, que é a semente usada pelo gerador de números aleatórios.

Em métodos de *ensemble* como o iNNE, o número de estimadores é um parâmetro importante porque determina quantas vezes o processo de isolamento será realizado, cada um com uma subamostra diferente dos dados. Isso afeta diretamente a robustez e a precisão do modelo na detecção de anomalias. Um número maior de estimadores geralmente proporciona um modelo mais estável e confiável, pois a decisão final é tomada com base em várias instâncias de isolamento, reduzindo a variância e o risco de sobreajuste. No entanto, também aumenta a complexidade computacional e o tempo de execução do modelo. Portanto, deve-se encontrar um equilíbrio entre o número de estimadores e os recursos computacionais disponíveis para treinamento e inferência.

O tamanho da amostra utilizado no iNNE afeta vários aspectos do modelo iNNE:

- Suavidade da Distribuição: Um tamanho de amostra menor resulta em uma distribuição de pontuação de anomalia mais suave, enquanto um tamanho maior leva a uma distribuição mais detalhada e irregular.
- Contaminação por Anomalias: Com amostras menores, a probabilidade de incluir anomalias é reduzida, tornando o modelo mais robusto.
- Desempenho do Ensemble: Um tamanho de amostra adequado permite que o iNNE atinja um desempenho ótimo, pois representa bem a distribuição dos dados normais, permitindo a detecção eficaz de anomalias locais e globais.
- Impacto na Detecção: O tamanho da amostra afeta diretamente a capacidade do iNNE de isolar e pontuar anomalias, especialmente em conjuntos de dados com muitos atributos irrelevantes ou em distribuições multimodais.

Na Figura 7 podemos ver quatro mapas de contorno de pontuações de anomalia desenhados para o conjunto de dados, utilizado por Bandaragoda *et al.* (2018) em seu artigo, empregando isolamento usando o iNNE com 4 valores diferentes de tamanho de amostras. Sendo o tamanho das amostras iguais a 2, 8, 64 e 256 para os quadros A, B, C e D respectivamente.

**Figura 7** – Quatro mapas de contorno de pontuações de anomalia

Fonte: (BANDARAGODA *et al.*, 2018)

Neste trabalho, durante o ajuste de hiperparâmetros deste algoritmo foram usados apenas os parâmetros `n_estimators` e `contamination`. Não foi realizado o ajuste do parâmetro `max_samples`, pois o mesmo possuía o valor `default=auto`, que já faz um ajuste automático do tamanho da amostra. Além disso, para todos os algoritmos foi definido que o parâmetro `random_state` teria o valor 69.

#### 3.5.4.2 *K-Nearest Neighbors* (KNN)

De acordo com Pang *et al.* (2021) KNN significa *K-Nearest Neighbors* e é um algoritmo de aprendizado de máquina usado para tarefas de classificação e regressão. No KNN, a saída é baseada na maioria das  $k$  instâncias mais próximas no espaço de recursos.

Os passos gerais para calcular o KNN são:

1. Definir K: é definido o número  $k$ , onde  $K$  é o número de vizinhos mais próximos a considerar.
2. Calcular Distâncias: é calculada a distância entre o ponto de teste e cada ponto de treinamento.
3. Identificar Vizinhos: determinar os  $k$  pontos de treinamento mais próximos do ponto de teste.
4. Votação para Classificação: Para classificação, a classe mais comum entre os vizinhos é atribuída ao ponto de teste.
5. Média para Regressão: Para regressão, a média dos valores dos vizinhos é calculada e atribuída ao ponto de teste.

Para a distância, a fórmula mais comum é a distância euclidiana, dada pela Equação 3.4.

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (3.4)$$

Onde  $(p)$  e  $(q)$  são dois pontos no espaço euclidiano e  $(n)$  é o número de dimensões do espaço de características.

KNN também tem algumas limitações, que incluem:

- Pode não funcionar bem com conjuntos de dados que possuem recursos irrelevantes ou redundantes, pois pode levar a superajuste ou subajuste.
- 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 sensível à escolha da métrica de distância, o que pode afetar o desempenho do algoritmo.
- Pode ser computacionalmente caro para grandes conjuntos de dados, pois requer o cálculo da distância entre cada instância e todas as outras instâncias no conjunto de dados.



Foram utilizados na pesquisa de hiper-parâmetros os seguintes parâmetros: *contamination*, *leaf\_size*, *method* e *n\_neighbors*.

Para explicar o parâmetro *Leaf\_size* ou tamanho da folha, é preciso entender as estruturas de árvore que são utilizadas para organizar e pesquisar dados em espaços multidimensionais. As principais estruturas são a *KDTree* e a *BallTree*.

A *KDTree* é uma árvore binária onde cada nó representa um ponto em um espaço k-dimensional. Ela utiliza planos de divisão alinhados com os eixos coordenados para particionar o espaço, cada nível da árvore divide o espaço com base em um dos k eixos coordenados, alternando entre eles à medida que se desce na árvore. Esta estrutura é utilizada nos algoritmos de IA por ser eficaz para buscas de vizinhos mais próximos.

A *BallTree* organiza os pontos em uma estrutura de árvore agrupando-os em hiperesferas. Cada nó da árvore representa uma hiperesfera contendo um conjunto de pontos. Ao contrário da *KDTree*, não se baseia em planos de divisão alinhados com os eixos, permitindo uma melhor adaptação à distribuição dos dados. É particularmente útil para conjuntos de dados onde a métrica de distância não é bem adaptada aos cortes alinhados com os eixos, como em espaços com muitas dimensões.

O parâmetro *Leaf\_size* é utilizado quando o algoritmo de busca dos vizinhos mais próximos é baseado em uma estrutura de árvore, como *KDTree* ou *BallTree*. O *Leaf\_size* determina o tamanho das folhas na estrutura da árvore utilizada para a busca dos vizinhos mais próximos. Um tamanho de folha maior significa que cada folha da árvore conterá mais pontos que o algoritmo deve considerar.

o parâmetro *N\_neighbors* é o número de vizinhos a serem utilizados por padrão para as pesquisas de vizinhança.

Já o parâmetro *method* é o método a ser utilizado para definir a pontuação que define o que é um *outlier*. Os métodos são *largest*, *mean* e *median*. No método *largest* é utilizada a distância para o k-ésimo vizinho como a pontuação; no método *mean* é utilizada a média da distância de todos os k vizinhos e no método *median* é utilizada a mediana de todos os k vizinhos.

#### 3.5.4.3 Local Outlier Factor (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 dimensão<sup>4</sup>.

A fórmula do Local Outlier Factor (LOF) é usada para medir a anomalia local de uma amostra em relação às suas vizinhas. O LOF de uma amostra é calculado como o quociente da

<sup>4</sup> Dados de alta dimensão são dados com grande quantidade de variáveis.

densidade média de alcance local das amostras vizinhas e a própria densidade de alcance local da amostra. Matematicamente, o LOF é expresso pela Equação 3.5.

$$LOF(k) = \frac{\sum_{p \in N_k(x)} LRD_k(p)}{|N_k(x)|} / LRD_k(x) \quad (3.5)$$

Onde:

$N_k(x)$  é o conjunto dos k-vizinhos mais próximos de x,

$LRD_k(x)$  é a densidade de alcance local de x, que é o inverso da média das distâncias de alcance dos k-vizinhos mais próximos de x,

$LRD_k(p)$  é a densidade de alcance local de um ponto p dentro dos k-vizinhos mais próximos de x,

$|N_k(x)|$  é o número de vizinhos em  $N_k(x)$ .

Um valor de LOF maior que 1 sugere que a amostra é um outlier, pois tem uma densidade significativamente menor do que seus vizinhos

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.

Para o algoritmo LOF foram utilizados os parâmetros *contamination*, *leaf\_size* e *n\_neighbors* já explicados em outros algoritmos.

#### 3.5.4.4 Principal Component Analysis (PCA)

Segundo Elkhadir, Chougali e Benattou (2015), o algoritmo *Principal Component Analysis* (PCA) 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 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.

Para o PCA, além da contaminação, foram utilizados os parâmetros *n\_components* e *n\_selected\_components*, onde são definidos a quantidade de atributos que devem ser mantidos e o número de componentes utilizados para determinar a pontuação de *outlier*, respectivamente.

#### 3.5.4.5 *Rapid Distance-Based Outlier Detection via Sampling (Sampling)*

Sugiyama e Borgwardt (2013) apresentaram O *Rapid Distance-Based Outlier Detection via Sampling* como um algoritmo eficiente para detecção de anomalias que não depende da modelagem da distribuição de probabilidade subjacente, o que é particularmente desafiador em dados de alta dimensão.

O algoritmo funciona seguindo os seguintes passos:

1. Amostragem: O algoritmo começa realizando uma amostragem única dos dados.
2. Cálculo da Distância: Para cada ponto de dado, o algoritmo mede a distância até o vizinho mais próximo dentro do conjunto de amostra.
3. Pontuação de Anomalia: Cada ponto de dado recebe uma pontuação de anomalia baseada nessa distância.

4. Detecção de Outliers: Os pontos com as maiores pontuações de anomalia são considerados outliers, ou seja, desvios significativos em relação aos outros pontos de dados.

Este algoritmo tem como vantagem ser escalável com complexidade linear em relação ao número de pontos de dados, eficaz em média entre os métodos existentes e fácil de usar com apenas um parâmetro necessário, o número de amostras.

Para o algoritmo Sampling foram testados apenas dois parâmetros, a contaminação e o `subset_size`, que é justamente o número de amostras.

#### 3.5.4.6 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.

Para entender o ECOD, é necessário entender o que é a função de distribuição cumulativa empírica (ECDF). A ECDF é uma função que representa a proporção ou percentual de observações de um conjunto de dados que são menores ou iguais a um determinado valor. Ela é uma forma de estimar a distribuição de probabilidade de uma variável aleatória baseada em uma amostra de dados.

A ECDF é definida matematicamente para um conjunto de dados  $(x_1, x_2, \dots, x_n)$  pela Equação 3.6.

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq t) \quad (3.6)$$

Onde:

$F_n(t)$  é o valor da ECDF no ponto  $t$ ,

$n$  é o número total de observações no conjunto de dados,

$I$  é a função indicadora, que é igual a 1 se a condição  $(x_i \leq t)$  for verdadeira e 0 caso contrário.

A ECDF é uma função degrau que aumenta em  $(\frac{1}{n})$  em cada ponto de dado  $(x_i)$ . Ela é útil para visualizar e comparar a distribuição de diferentes conjuntos de dados, pois não faz suposições sobre a forma da distribuição subjacente dos dados.

A fórmula do ECOD é baseada na função de distribuição cumulativa empírica (ECDF) e é utilizada para calcular a pontuação de anomalia de cada ponto de dados. O ECOD funciona da seguinte maneira:

1. Estimativa da Distribuição: primeiro, o ECOD estima a distribuição subjacente dos dados de entrada de forma não paramétrica, calculando a distribuição cumulativa empírica por dimensão dos dados.

2. Probabilidades de Cauda: em seguida, o ECOD usa essas distribuições empíricas para estimar as probabilidades de cauda por dimensão para cada ponto de dados.
3. Pontuação de Anomalia: finalmente, o ECOD calcula uma pontuação de anomalia para cada ponto de dados agregando as probabilidades de cauda estimadas em todas 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.

Para o ECOD, apenas a contaminação foi utilizada para o ajuste, já que seu outro parâmetro é o *n\_jobs* que informa quantos jobs paralelos podem ser executados para o treinamento, como forma de acelerar o processo. Utilizou-se o valor -1, que informa para utilizar como valor do parâmetro o número de núcleos de processamento existentes.

#### 3.5.4.7 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 copula. Copula é uma função de distribuição que descreve a estrutura de dependência entre variáveis aleatórias.

A ideia básica por trás das copulas é que elas podem separar a modelagem das margens (distribuições univariadas de cada variável) da modelagem da dependência entre essas variáveis. Isso permite que os pesquisadores escolham as margens que melhor se ajustam aos dados individuais e, em seguida, escolham uma copula que melhor descreva como essas margens estão relacionadas.

Matematicamente, uma copula ( $C$ ) é uma função com a seguinte propriedade:  $C : [0, 1]^n \rightarrow [0, 1]$ , onde  $(C(u_1, u_2, \dots, u_n))$  é a copula para as variáveis aleatórias  $(U_1, U_2, \dots, U_n)$  com distribuições uniformes no intervalo  $[0, 1]$ .

A copula ( $C$ ) então une as distribuições marginais para formar a distribuição conjunta das variáveis aleatórias originais. O Teorema de Sklar afirma que qualquer distribuição multivariada pode ser expressa em termos de suas margens e uma copula que descreve a dependência entre elas. Este teorema é a base para a aplicação de copulas em estatística multivariada.

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.

Assim como no ECOD, no COPOD apenas a contaminação foi utilizada para o ajuste, já que os parâmetros em ambos os casos eram os mesmos.

#### 3.5.4.8 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 selecionados, permitindo que, com um único treinamento, fosse gerado um modelo que utilizasse todos eles e que gerasse um único resultado final a ser avaliado. O funcionamento do SUOD é baseado no princípio de que a combinação de vários algoritmos pode melhorar a robustez e a precisão da detecção de outliers. Existem, porém, situações onde o SUOD pode ter um desempenho pior do que o dos algoritmos individuais que ele agrega, pois um desempenho ruim de um determinado algoritmo pode afetar negativamente o desempenho geral agregado.

Partindo então dos algoritmos escolhidos, decidiu-se utilizar os cinco materiais com maior número de registros para selecionar a melhor combinação de algoritmos a ser utilizada no SUOD. Neste caso específico, os dados a serem utilizados deveriam ser os mesmos para todos os algoritmos que compõem a agregação. Sendo assim, como a maioria dos algoritmos performou melhor sem escalonamento, utilizou-se os dados sem escalonamento no modelo do SUOD. Por este motivo, foram adotados no modelo iNNE os melhores parâmetros encontrados quando não se utilizava escalonamento de dados.

Para a definição do modelo final a ser incorporado à ferramenta de detecção de sobrepreço, foram realizados testes combinando os algoritmos selecionados em todas as formas possíveis, num total de 127 combinações, incluindo cada algoritmo isoladamente. Após os testes, foram selecionadas duas opções com resultados próximos onde decidiu-se testar qual apresentava o melhor tempo de resposta. Como o resultado deste trabalho é um aplicativo, e o tempo de resposta contribui para uma experiência de usuário mais ágil e satisfatória, decidimos avaliar este tempo, sabendo que um tempo de resposta eficiente é fundamental para o sucesso de um aplicativo.

Para a avaliação do tempo de resposta das duas opções foi realizado um teste com os oito materiais com maior número de registros e apurado o tempo necessário para treinar a ferramenta em cada uma das avaliações. O resultado do teste é apresentado no Capítulo 4, na Tabela 1. Com esse teste foi possível chegar à melhor solução.

### 3.5.5 Avaliação final do modelo

Após a definição do modelo final, foi realizado um teste e uma análise da performance em cada um dos 13 materiais com maior número de registros listados no Quadro 6. A avaliação foi realizada por meio do método de validação cruzada chamado de *LeaveOneOut* (LOOCV) que, segundo Berrar (2018), é uma abordagem exaustiva de divisão de validação que aprimora *k-fold*. Nesta técnica se define o valor de *k* como sendo a quantidade de registros, então para cada iteração teremos um modelo construído com todos os registros menos um, que é utilizado para o teste.

A técnica de validação cruzada *LeaveOneOut* possui várias vantagens quando utilizada para testar modelos de algoritmos de aprendizado de máquina:

- **Máxima Utilização dos Dados:** Como cada amostra é usada uma vez como teste e o restante para treinamento, essa técnica aproveita ao máximo o conjunto de dados disponível, o que é particularmente útil em conjuntos de dados pequenos.
- **Redução do Viés:** Ao testar cada amostra individualmente, o *LeaveOneOut* pode reduzir o viés associado à escolha aleatória de conjuntos de treino e teste.
- **Estimativa de Desempenho:** Fornece uma estimativa detalhada do desempenho do modelo, pois cada amostra é validada independentemente.
- **Menos Variação:** Como cada amostra é testada individualmente, a variação entre as diferentes iterações de treino/teste é minimizada.
- **Sensibilidade a *Outliers*:** Pode identificar se o modelo é sensível a *outliers*, já que cada amostra tem a chance de ser um conjunto de teste por si só.
- **Generalização:** Ajuda a entender como o modelo se comportará com novos dados, fornecendo uma visão sobre sua capacidade de generalização.

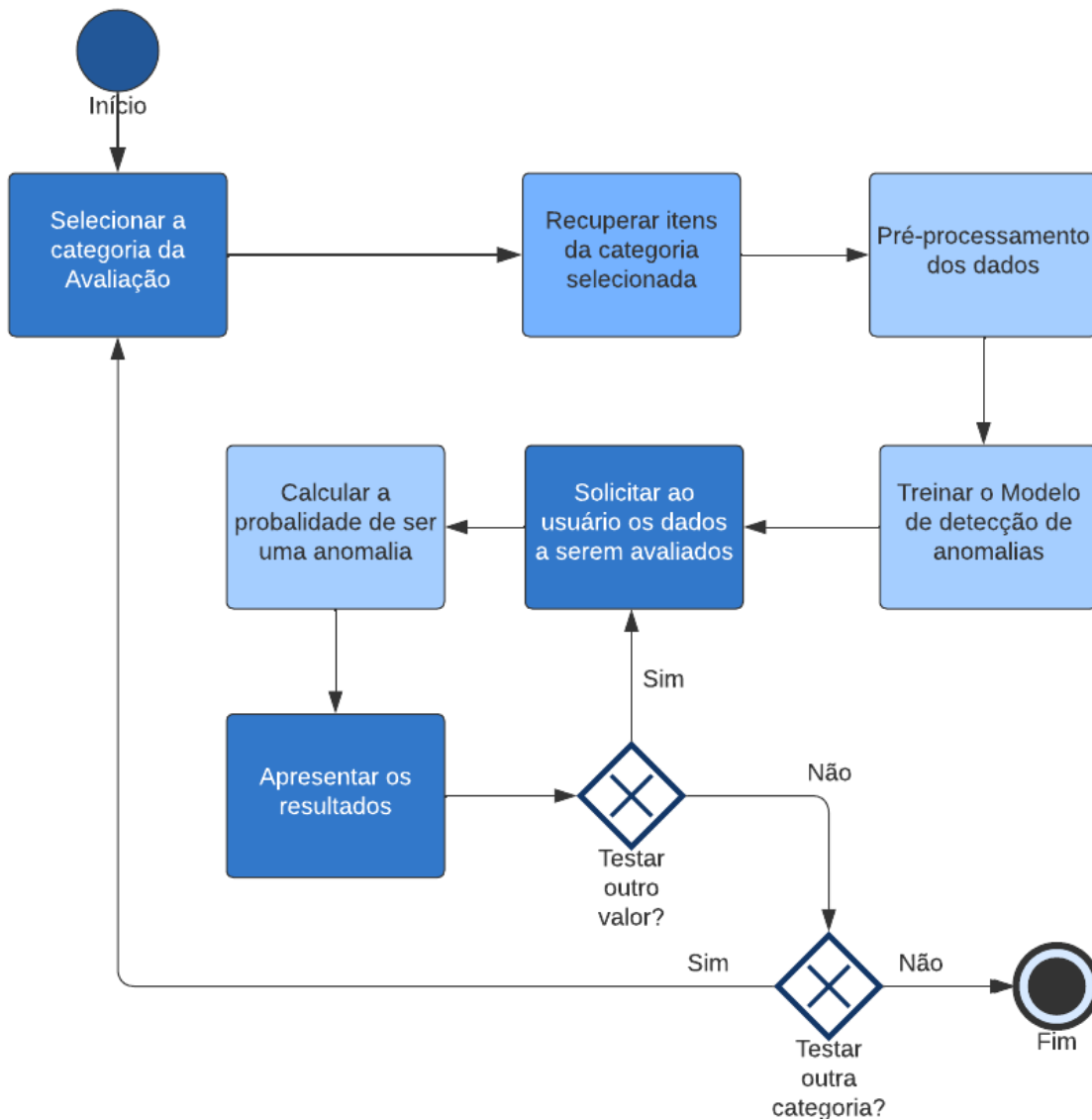
No entanto, é importante notar que o *LeaveOneOut* pode ser computacionalmente intensivo, especialmente para conjuntos de dados grandes, pois requer que o modelo seja treinado tantas vezes quanto o número de amostras, motivo pelo qual não foi utilizado nas demais fases da pesquisa.

Os resultados obtidos são apresentados no Capítulo 4.

### 3.6 Modelo do Aplicativo de detecção de sobrepreço

Escolhidos os algoritmos, partiu-se para a definição do modelo a ser utilizado na ferramenta de detecção de sobrepreço. A Figura 8 ilustra o diagrama dos processos da ferramenta. Nas próximas seções, cada um dos processos será detalhado.

**Figura 8** – Modelo da ferramenta de detecção de sobrepreço



Fonte: Elaboração própria

#### 3.6.1 Processo Selecionar a categoria da avaliação

Neste ponto do modelo, foi criada uma tela de interação com o usuário, onde se pode escolher a categoria (atributo Catmat) do item que se deseja avaliar.

As opções de categorias são recuperadas dos dados importados do portal Dados Abertos de Compras Governamentais.

### 3.6.2 Processo Recuperar itens da categoria selecionada

Neste processo são recuperados todos os itens da categoria selecionada de contratos e licitações da data especificada até o dia de hoje. O código deste processo está listado abaixo no Código-Fonte 1

**Código-Fonte 1** – Função recuperar\_itens\_catmat

```

1 #####
2 # Função recuperar_itens_catmat
3 # Objetivo: Retornar todos os registros de um determinado material
4 #           desde uma data especificada até o dia de hoje
5 # Parâmetros: catmat - código do material a ser recuperado
6 #           data - Data a partir da qual os registros serão
7 #           selecionados
8 # Retorno: dataframe pandas com todos os registros selecionados
9 #####
10 def recuperar_itens_catmat(catmat, data):
11
12     # Cria a conexão 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     # Lê o banco de dados para buscar os registros que atendem os
18     # parâmetros informados
19     df = pd.read_sql(
20         "SELECT quantidade, valor_unitario, distancia_uasg_fornecedor
21         FROM siasg.itens where catmat_id=" + str(catmat) + " and data > '" +
22         data
23         + "'", dbConnection);
24     return df

```

Todos os itens recuperados são repassados à função seguinte de pré-processamento dos dados.

### 3.6.3 Processo Pré-processamento dos dados

Para os seis primeiros algoritmos selecionados no Quadro 7, apenas duas formas de pré-processamento estão presentes: sem escalonamento e com escalonamento Robust.

Portanto, a função de pré-processamento considera apenas o método *Robust* conforme apresentado no Código-Fonte 2 abaixo.

**Código-Fonte 2** – Função preprocessar\_dados

```

1 #####
2 # Função preprocessar_dados
3 # Objetivo: Realizar o pré-processamento dos dados
4 # Parametros: df - dataframe pandas com os dados a serem escalonados
5 # Retorno: dataframe pandas com todos os registros ajustados utilizando
6 #           o método Robust de escalonamento
7 #####
8 from sklearn.preprocessing import RobustScaler
9
10 def preprocessar_dados(df):
11
12     # Cria uma instância do RobustScaler
13     robust_scaler = RobustScaler()
14
15     # Ajusta e transforma os dados com o RobustScaler
16     df_ajustado = robust_scaler.fit_transform(df)
17
18     # retorna o dataframe ajustado
19     return df_ajustado

```

O resultado deste processo de pré-processamento é utilizado apenas no treinamento do algoritmo INNE, que teve uma melhor performance com sua utilização.

### 3.6.4 Processo Treinar o modelo de detecção de anomalias

Como citado anteriormente, foi utilizado o SUOD para agrupar os algoritmos INNE, KNN, LOF, PCA, Sampling, ECOD e COPOD, fazendo com que um único treinamento gerasse um modelo combinando 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 detecção de anomalias
4 # Parâmetros: df - dataframe pandas com os dados a
5 #             serem utilizados no treinamento
6 # Retorno: clf - modelo treinado utilizando o SUOD
7 #####
8 # Importa bibliotecas do PyOD com os algoritmos de detecção de anomalias
9 from pyod.models.inne import INNE
10 from pyod.models.knn import KNN
11 from pyod.models.lof import LOF
12 from pyod.models.pca import PCA
13 from pyod.models.sampling import Sampling
14 from pyod.models.ecod import ECOD
15 from pyod.models.copod import COPOD
16 from pyod.models.suod import SUOD

```



```

17
18 def treina_modelo(df):
19     # Lista de detectores
20     detector_list = [
21         INNE(contamination=0.05, n_estimators=50, random_state= 69),
22         KNN(contamination=0.05, leaf_size=10, method='largest',
23             n_neighbors=10),
24         LOF(contamination=0.05, leaf_size=1, n_neighbors=34),
25         PCA(contamination=0.05, n_components=3, n_selected_components=1),
26         Sampling(contamination=0.05, subset_size=10),
27         ECOD(contamination=0.09),
28         COPOD(contamination=0.12)
29     ]
30
31     # configurar o SUOD com a lista de parâmetros e detectores
32     clf = SUOD(base_estimators=detector_list, n_jobs=2, combination='
33     maximization', contamination=0.05,
34     verbose=False)
35
36     # treinar o modelo
37     clf.fit(df)
38     return clf

```

Uma função alternativa utilizando apenas um dos algoritmos foi desenvolvida e encontra-se detalhada no Código-Fonte 4.

#### Código-Fonte 4 – Função treina\_modelo

```

1 #####
2 # Funcao treina_modelo
3 # Objetivo: Treina o modelo de detecção de anomalias
4 # Parâmetros: df - dataframe pandas com os dados a
5 #             serem utilizados no treinamento
6 # Retorno: clf - modelo treinado utilizando o SUOD
7 #####
8 # Importa bibliotecas do PyOD com os algoritmos de detecção de anomalias
9 from pyod.models.copod import COPOD
10
11 def treina_modelo(df):
12
13     # configurar o algoritmo com a lista de parâmetros
14     clf = COPOD(contamination=0.12)
15
16     # treinar o modelo
17     clf.fit(df)
18     return clf

```

### 3.6.5 Processo Solicitar ao usuário os dados a serem avaliados

Nesta parte do processo é solicitado ao usuário que entre com os dados que deseja avaliar, podendo informar a quantidade e valor unitário obtidos na sua pesquisa de preços e, opcionalmente, o município do fornecedor. Caso não seja apresentado o município do fornecedor, será considerado que é o mesmo da UASG, ou seja, que a distância da UASG ao fornecedor é igual a 0.

### 3.6.6 Processo Calcular a possibilidade de ser uma anomalia

Tendo sido informados a quantidade, o valor obtido na pesquisa e a distância do município do fornecedor, os mesmos são passados para a função de avaliação detalhada no Código-Fonte 5.

**Código-Fonte 5** – Função avalia\_dados

```

1 #####
2 # Função avalia_dados
3 # Objetivo: Avaliar os dados da pesquisa de preços no modelo treinado
4 # Parâmetros: clf - modelo treinado
5 #               quantidade - quantidade utilizada na pesquisa de preços
6 #               valor - valor obtido na pesquisa de preços
7 #               distancia - distância do município do fornecedor ao
8 #                       município da UASG
9 # Retorno: predicacao - resultado da avaliação
10 #####
11 def avalia_dados(clf, quantidade, valor, distancia):
12     # cria um dataframe de apenas 1 linha com os valores recebidos
13     dfteste=pd.DataFrame({'quantidade':quantidade,'valor_unitario':valor,'
14     distancia':distancia},index=[0])
15     # Calcula a predição
16     predicacao=clf.predict(dfteste)
17     # Retorna o resultado
18     return predicacao

```

### 3.6.7 Processo Apresentar os resultados

Neste processo são mostradas na tela, apresentada na Figura 10, as seguintes informações:

- indício de sobrepreço ou não;
- o valor mínimo, máximo, média, mediana e quartil 97,5% para os atributos do item selecionado valor unitário e quantidade;
- 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, são apresentadas as partes que compõem a ferramenta de avaliação de indícios de sobrepreços: a tela de seleção de categoria e a tela de avaliação e resultados.

### 3.7.1 Tela de seleção da categoria da avaliação

A tela do protótipo segue o modelo apresentado na Figura 9.

**Figura 9** – Tela de seleção de material

A imagem mostra a interface de usuário da ferramenta. No topo, há uma barra de navegação com o logotipo do Senado Federal e o título 'IA para Pesquisa de Preços'. Abaixo, uma barra de menu contém links para 'IAPP', 'Home', 'Listagens', 'Preparação do Repositório', 'Licitações', 'Preparação de Dados' e 'Avaliação da Pesquisa de Preços'. O conteúdo principal da tela é dividido em duas colunas. A coluna da esquerda, intitulada 'Selecione o Material', contém um menu suspenso com o texto '462729 - GELATINA ALIMENTÍCIA, APRESENTAÇÃO PÔ S', um campo para a 'Data inicial de preços para validação' com o valor '09/06/2022' e dois botões, 'Avaliar' e 'Voltar'. A coluna da direita, intitulada 'Descrição', exibe o texto 'GELATINA ALIMENTÍCIA, APRESENTAÇÃO PÔ SABOR VARIADO ORIGEM ANIMAL'. Na base da tela, uma barra de rodapé fornece o endereço: 'Senado Federal - Praça dos Três Poderes - Brasília DF - CEP 70165-900 - Fone: (61)3303-4141'.

Fonte: elaboração própria

O controle (*combobox*) de seleção do material é carregado com todas as categorias disponíveis na base de itens do repositório.

Ao selecionar um material, é apresentada a descrição conforme o que consta no catálogo de materiais da API de Compras Governamentais.

### 3.7.2 Tela de avaliação de resultados

Após a seleção do material, a tela de avaliação de resultados, conforme ilustrada na Figura 10, apresenta as medidas estatísticas sobre a categoria selecionada e irá receber o valor e quantidade da pesquisa.

Diversos valores e quantidades podem ser testados a critério do usuário.

**Figura 10** – Tela de avaliação da pesquisa

SENADO FEDERAL IA para Pesquisa de Preços

IAPP Obtenção de dados ▾ Preparação do Repositório ▾ Avaliação da Pesquisa de Preços ▾

### Avaliação de Pesquisa de Preços

Material selecionado = 464381 - FRUTA, TIPO BANANA PRATA / BANANA BRANCA APRESENTAÇÃO NATURAL

Quantidade a ser adquirida Valor encontrado na pesquisa

200 6,00

**Avaliar** **Voltar**

Dados avaliados desde de 01/01/2022 - Número de registros do item = 205

**Valor unitário:**

Mínimo	Máximo	Média	Mediana	Quartil 97,5%
1,47	31,00	4,88	4,66	8,57

**Quantidade:**

Mínimo	Máximo	Média	Mediana	Quartil 97,5%
30	83000	2452,27	750,00	14242,00

**Resultado da avaliação**

Valor aceitável

**Preços analisados**

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

Fonte: elaboração própria

### 3.7.3 Tela de preços analisados

Na tela de avaliação de resultados pode ser selecionado o botão "Preços analisados", que irá apresentar os contratos ou licitações utilizados na avaliação. Os registros apresentados serão os que tem o menor preço, e no número máximo de quinze, conforme ilustrada na Figura 11.

**Figura 11** – Tela de preços analisados

SENADO FEDERAL IA para Pesquisa de Preços								
IAPP Obtenção de dados ▾ Preparação do Repositório ▾ Avaliação da Pesquisa de Preços ▾								
Preços Analisados								
Tipo	Número	Orgão	CPF/CNPJ Fornecedor	Nome Fornecedor	Data	Valor unitário	Quantidade	
Contrato	00059/2022	COMANDO DO EXERCITO	73.373.243/0001-49	REFISERVI REFEICOES INDUSTRIAIS LTDA	14/10/2022	R\$ 2,00	100	
Licitação	6320/22	STM - SUPERIOR TRIBUNAL MILITAR/DF	28634818000185	N.S.S. COMERCIAL & CONSTRUTORA LTDA	20/10/2022	R\$ 2,52	264	
Licitação	320/22	IBAMA - SUPERINTENDENCIA ESTADUAL/AP	37584954000107	M E PINTO DE OLIVEIRA	28/12/2022	R\$ 3,30	1508	
Licitação	1720/22	CENTRO NACIONAL DE PRIMATAS	20290559000100	E A ALCANTARA & CIA LTDA	05/12/2022	R\$ 3,60	5824	
Contrato	00049/2022	UNIVERSIDADE FEDERAL DE JUIZ DE FORA	29.455.568/0001-89	DISTRIBUIDORA VIB LTDA	18/10/2022	R\$ 3,96	4500	
Contrato	00009/2022	INST.FED.DE EDUC.,CIENC.E TEC.DE PERNAMBUCO	04.323.064/0001-84	ASSOCIACAO DOS PRODUTORES RURAIS DO ENGENHO BOM JARDIM/	14/10/2022	R\$ 4,05	1482	
Contrato	00015/2022	INST.FED.DE EDUC.,CIENC.E TEC.DE SERGIPE	45.340.904/0001-02	ASSOCIACAO SERGIPANA DA AGRICULTURA FAMILIAR	07/10/2022	R\$ 4,16	480	
Contrato	00007/2022	INST.FED.DE EDUC.,CIENC.E TEC.DA PARAIBA	31.860.198/0001-07	THIAGO GOMES BARBOSA COMERCIO	05/10/2022	R\$ 4,50	3087	
Contrato	00015/2022	INST.FED.DE EDUC.,CIENC.E TEC.DA BAHIA	28.716.605/0001-00	COOPERATIVA AGRICOLA DE DESENVOLVIMENTO SUSTENTAVEL DO	25/10/2022	R\$ 4,66	90	
Contrato	00047/2022	INST.FED.DE EDUC.,CIENC.E TEC.DO PIAUI	044.427.963-60	JOSE AUGUSTO DE SOUSA MAGALHAES	05/10/2022	R\$ 4,99	400	
Licitação	620/23	TRIBUNAL REGIONAL FEDERAL-SEC.1A.REG./DF	46099785000100	BMOT DISTRIBUIDORA DE HORTIFRUTIGRANJEIROS LTDA	24/02/2023	R\$ 5,00	400	
Contrato	00004/2022	INST.FED.DE EDUC.,CIENC.E TEC.DO TRIA.MINEIRO	11.465.646/0001-60	COOPERATIVA MISTA DOS ASSENTADOS E AGRICULTORES FAMILIA	18/10/2022	R\$ 5,35	7321	

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

Fonte: elaboração própria

Apesar de a ferramenta no geral não conseguir responder diretamente as tarefas presentes no checklist apresentado no Apêndice C, a tela de preços analisados pode ser utilizada pelo gestor para procurar outras contratações públicas para servir de amostras necessárias para atender ao item **1. CONTRATAÇÕES PÚBLICAS** do checklist.

### 3.8 Questionário

Para avaliar a proposta do sistema e poder aprimorá-lo foram feitas entrevistas com gestores, auditores e responsáveis pela avaliação das pesquisas de preços da área de compras do Senado Federal. Os entrevistados foram convidados a responder três perguntas, avaliando-as em uma escala Likert de 1 a 5:

Pergunta 1: A avaliação apresentada neste sistema pode servir na validação da pesquisa de preços?

1. Não pode ser utilizada
2. Quase não há utilidade
3. Há alguma utilidade, mas limitada
4. É útil na maioria das situações
5. É totalmente útil

Pergunta 2: A consulta dos contratos e licitações utilizados na avaliação dos preços é útil para aprimorar a pesquisa de preços.

1. Discordo totalmente
2. Discordo
3. Nem concordo nem discordo
4. Concordo
5. Concordo totalmente

Pergunta 3: Que percentual aproximado de esforço na validação de uma pesquisa de preços é reduzido com a utilização do sistema?

1. 0%
2. 25%
3. 50%
4. 75%
5. 100%

O questionário também dispunha de um campo textual para que os respondentes pudessem apresentar outras considerações de forma livre.

As questões foram formuladas com o objetivo avaliar o sistema. As perguntas 1 e 3 buscam avaliar a utilidade da validação. A pergunta 2 está relacionada ao nível de utilidade das informações apresentadas.

### 3.9 Considerações finais do capítulo

Neste 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 e do 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

Neste capítulo são apresentados os resultados obtidos na avaliação qualitativa da rotulagem realizada pelo Snorkel, nos testes de seleção dos algoritmos para composição do modelo final e na aplicação do modelo final em cada um dos 13 materiais com maior número de registros. Após a apresentação dos resultados é efetuada uma análise do desempenho obtido pela ferramenta nos testes e nas entrevistas realizadas.

### 4.1 Avaliação qualitativa da rotulagem do Snorkel

Para a avaliação da qualidade da rotulagem realizada pelo Snorkel optou-se por utilizar o material de Catmat 464381, cuja descrição é **FRUTA, TIPO BANANA PRATA / BANANA BRANCA APRESENTAÇÃO NATURAL**. Este material foi escolhido, por ser um dos 10 materiais com maior número de registros, mas que não tem uma quantidade muito grande de registros, e por ser um produto conhecido e de fácil entendimento quanto ao preço.

O item selecionado apresenta 205 registros, nestes registros avaliamos para rotulagem os atributos: preço unitário, quantidade e distância do fornecedor. No Quadro 8 apresentamos as medidas de estatística descritiva como mediana, média, valor mínimo, valor máximo, valor do percentil 97,5% e o desvio padrão, de cada um dos atributos.

**Quadro 8** – Medidas de Estatística Descritiva por Atributo

Atributo	Mediana	Média	Mínimo	Máximo	P. 97,5%	$\sigma$
Valor unitário	4,66	4,88	1,47	31,00	8,57	2,39
Quantidade	750,0	2452,3	30	83000	14242	6926,7
Distância do fornecedor	69,33	640,76	0,0	1627,52	1454,23	711,05

Fonte: Elaboração própria.

Quatro funções de rotulagem<sup>1</sup> foram criadas para que o Snorkel pudesse avaliar os dados. A avaliação do Snorkel pode produzir 3 resultados: 0, quando considerado normal; 1, quando considerada uma anomalia; e -1 quando o registro não se encaixou em nenhuma das regras e o Snorkel se abstém de definir.

A primeira função de rotulagem está relacionada ao preço unitário. Ela define como anomalia um preço maior que o percentil 97,5% do valor unitário ou um preço menor que o valor unitário mínimo onde a quantidade fornecida não seja maior que a do percentil 97,5% da quantidade. No Quadro 9 podemos ver todos os registros que esta primeira função de rotulagem avaliou como anomalias e o resultado final da avaliação do Snorkel após a avaliação por todas as funções.

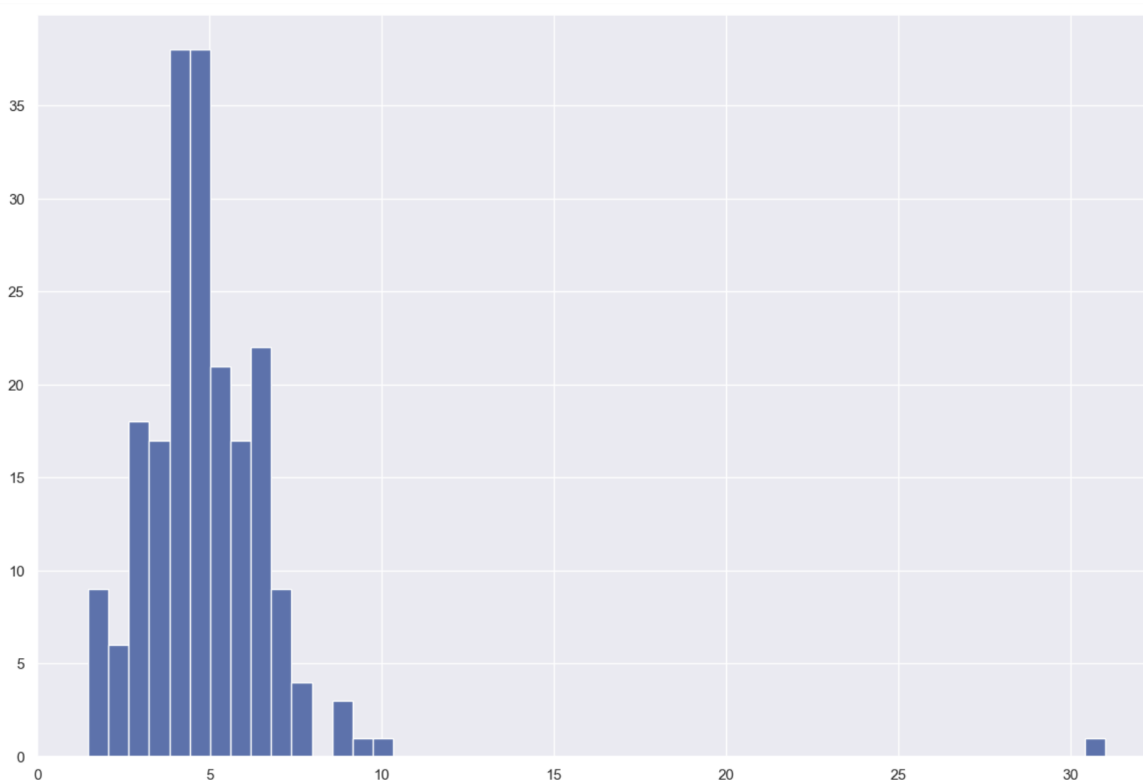
<sup>1</sup> Disponíveis no Apêndice A

**Quadro 9** – Registros avaliados como anomalias pela função de rotulagem **preco\_anomalo**

Registro	Valor unitário	Quantidade	Distância do Fornecedor	Rótulo final
11	31,00	30	0,00	1
49	8,65	600	0,00	1
101	9,75	700	0,00	1
106	9,30	475	1454,23	1
154	8,80	748	1454,23	1
199	8,92	1100	278,25	1

Fonte: Elaboração própria.

Na Figura 12 pode-se verificar que os resultados apontados pela função de rotulagem são aqueles que estão mais a direita e apartados da maioria da distribuição dos valores, validando a avaliação feita pela função de rotulagem.

**Figura 12** – Histograma do Valor Unitário

Fonte: elaboração própria

A segunda função de rotulagem está relacionada à quantidade. Ela define como anomalia uma quantidade maior que o percentil 97,5% da quantidade. No Quadro 10 podemos ver todos os registros que esta função de rotulagem avaliou como anomalias e o resultado final da avaliação do Snorkel após a avaliação por todas as funções.

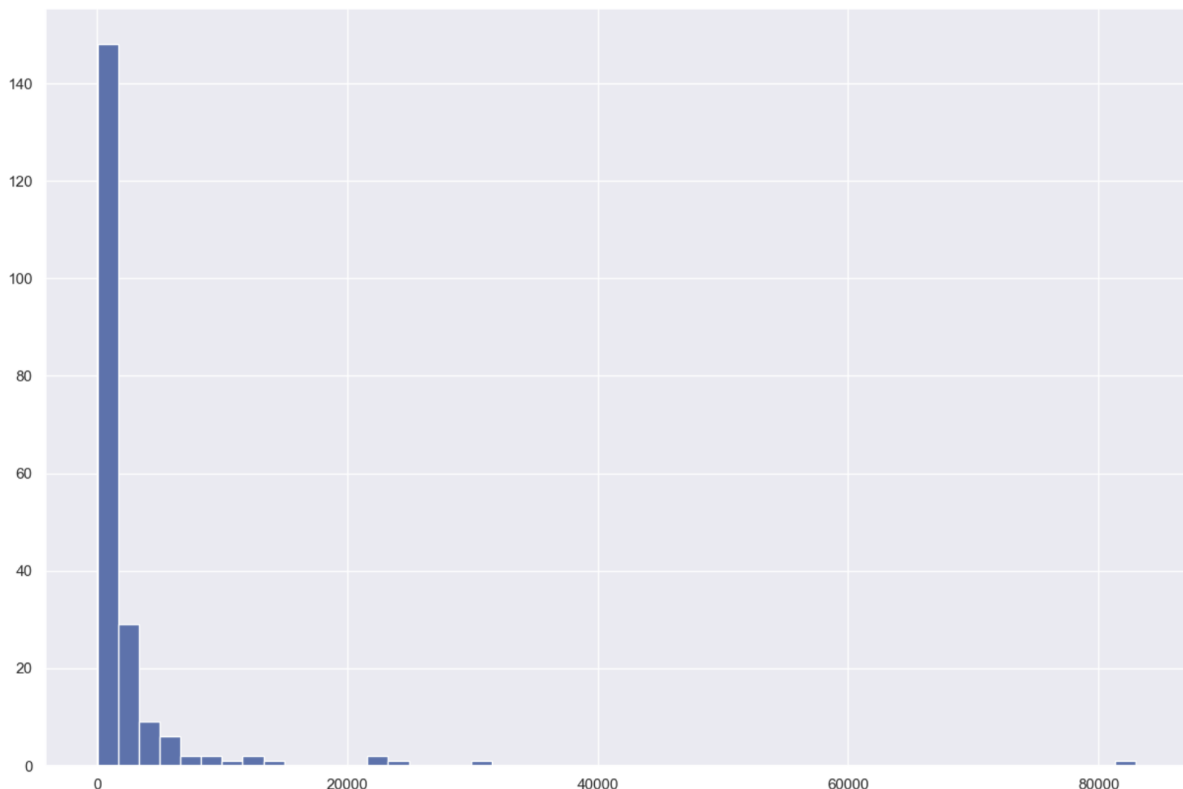


**Quadro 10** – Registros avaliados como anomalias pela função de rotulagem **quantidade\_alta**

Registro	Valor unitário	Quantidade	Distância do Fornecedor	Rótulo final
34	3,99	24.400	69,33	0
42	4,50	30.939	0,00	0
76	3,10	83.000	50,06	0
79	6,84	14.380	1454,23	0
160	4,28	22.594	0,00	0
203	4,27	22.594	0,00	0

Fonte: Elaboração própria.

Na Figura 13 pode-se verificar que os resultados apontados pela função de rotulagem são aqueles que estão mais a direita e apartados da maioria da distribuição dos valores, validando a avaliação feita pela função de rotulagem. Pode-se perceber que apesar da quantidade ter sido considerada anômala, o resultado final não foi. O motivo será analisado quando da avaliação da quarta função de rotulagem.

**Figura 13** – Histograma da quantidade

Fonte: elaboração própria

A terceira função de rotulagem está relacionada à distância da UASG ao fornecedor. Ela define como anomalia uma distância maior que o percentil 97,5% das distâncias. No Quadro 11

podemos ver todos os registros que esta função de rotulagem avaliou como anomalias e o resultado final da avaliação do Snorkel após a avaliação por todas as funções.

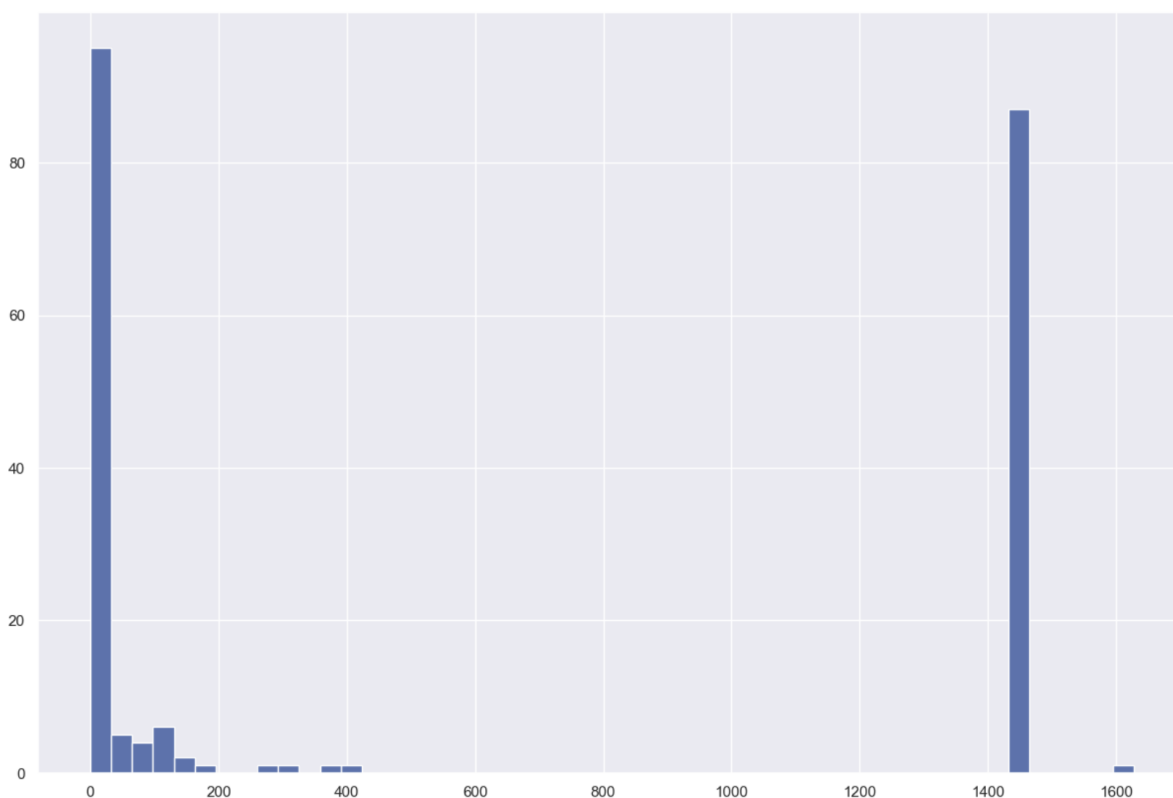
**Quadro 11** – Registros avaliados como anomalias pela função de rotulagem **distancia\_alta**

Registro	Valor unitário	Quantidade	Distância do Fornecedor	Rótulo final
62	3,54	12.000	1627,52	0

Fonte: Elaboração própria.

Na Figura 14 pode-se verificar que o resultado apontado pela função de rotulagem é aquele que está mais a direita e apartado da maioria da distribuição dos valores, validando a avaliação feita pela função de rotulagem. Pode-se perceber que apesar da distância ter sido considerada anômala, o resultado final não foi. O motivo será analisado quando da avaliação da quarta função de rotulagem.

**Figura 14** – Histograma da distância



Fonte: elaboração própria

A quarta função de rotulagem foi especificada para definir o que é considerado um registro normal. A definição é que se o preço é menor ou igual a mediana e maior que o menor preço, ou se o preço sendo maior que a mediana não é maior que o quantil 97,5% é um valor normal, caso contrário, esta função irá se abster. No Quadro 12 podemos ver todos os registros

considerados como anomalias pelas funções anteriores, a avaliação feita por esta função e o resultado final da avaliação do Snorkel após a avaliação por todas as funções.

**Quadro 12** – Avaliação da normalidade dos Registros avaliados como anomalias pelas funções anteriores

Registro	Valor unitário	Quantidade	Distância do Fornecedor	Rótulo de Normalidade	Rótulo final
11	31,00	30	0,00	-1	1
34	3,99	24.400	69,33	0	0
42	4,50	30.939	0,00	0	0
49	8,65	600	0,00	-1	1
62	3,54	12.000	1627,52	0	0
76	3,10	83.000	50,06	0	0
79	6,84	14.380	1454,23	0	0
101	9,75	700	0,00	-1	1
106	9,30	475	1454,23	-1	1
154	8,80	748	1454,23	-1	1
160	4,28	22.594	0,00	0	0
199	8,92	1100	278,25	-1	1
203	4,27	22.594	0,00	0	0

Fonte: Elaboração própria.

Pode-se perceber que todos os registros que foram considerados anomalias pelas funções anteriores e não foram considerados normais tiveram seu resultado final apontado como anomalia. Acredita-se, que pelos resultados aqui apontados, que as funções de rotulagem estão funcionando a contento. E a rotulagem por elas realizado foi a utilizada para a avaliação dos resultados das seções seguintes.

## 4.2 Resultados da seleção dos algoritmos

Para a escolha dos algoritmos componentes do modelo, foram testadas todas as 127 combinações possíveis entre os 7 (sete) algoritmos selecionados, incluindo-se o teste de cada algoritmo separadamente. O resultado da avaliação dos 15 modelos com maior *recall* se encontram no Quadro 13.

**Quadro 13** – Os 15 modelos com maior *Recall*

Algoritmos do Modelo	<i>Recall</i> (%)	Acurácia (%)	F <sub>1</sub> -Score (%)
KNN+Sampling	98,00	94,03	54,09

*Continua na próxima página*

**Quadro 13** – Continuação

<b>Algoritmos do Modelo</b>	<b>Recall (%)</b>	<b>Acurácia (%)</b>	<b>F1-score (%)</b>
COPOD	97,97	90,22	41,91
INNE+COPOD	97,14	90,94	44,06
INNE+KNN+COPOD	97,14	90,94	44,06
INNE+Sampling+COPOD	97,14	90,94	44,06
INNE+KNN+Sampling+COPOD	97,14	90,94	44,06
INNE+PCA+COPOD	97,08	90,93	44,03
INNE+KNN+PCA+COPOD	97,08	90,93	44,03
INNE+PCA+Sampling+COPOD	97,08	90,93	44,03
INNE+KNN+PCA+Sampling+COPOD	97,08	90,93	44,03
INNE+LOF+COPOD	96,95	91,02	44,46
INNE+KNN+LOF+COPOD	96,95	91,02	44,46
INNE+LOF+Sampling+COPOD	96,95	91,02	44,46
INNE+KNN+LOF+Sampling+COPOD	96,95	91,02	44,46
INNE+LOF+PCA+COPOD	96,88	91,01	44,43

Fonte: Elaboração própria.

No Quadro 13 o modelo com o melhor *recall* combina os algoritmos KNN e Sampling com 98% de sensibilidade na identificação das anomalias, seguido pelo modelo com apenas o algoritmo COPOD, com *recall* de 97,97%. Ambos os resultados podem ser considerados muito bons e muito próximos. Como o algoritmo COPOD fazia parte de 14 dos 15 melhores modelos, foi suscitada a dúvida de qual modelo utilizar. Decidiu-se avaliar qual modelo possuiria o melhor tempo de resposta para o usuário final, já que o treinamento é realizado após a seleção pelo usuário dos dados desejados. Optou-se, então, por realizar o teste das duas soluções nos 8 materiais com maior número de registros, apurando-se o tempo necessário para realizar o treinamento. Os resultados obtidos são apresentados na Tabela 1.

**Tabela 1** – Comparativo dos modelos COPOD e KNN+Sampling

<b>Material</b>	<b>COPOD</b>			<b>KNN+Sampling</b>		
	<b>Recall (%)</b>	<b>Acurácia (%)</b>	<b>Tempo (s)</b>	<b>Recall (%)</b>	<b>Acurácia (%)</b>	<b>Tempo (s)</b>
104671	99,83	90,49	0,94	99,83	92,53	26,93
150658	100,00	90,49	0,26	100,00	92,49	21,83
445485	100,00	90,69	0,04	100,00	92,17	20,64
402920	100,00	89,30	0,03	100,00	99,77	21,10
150877	100,00	89,84	0,03	100,00	92,20	18,47

*Continua na próxima página*

**Tabela 1** – Continuação

Material	COPOD			KNN+Sampling		
	<i>Recall</i> (%)	Acurácia (%)	Tempo (s)	<i>Recall</i> (%)	Acurácia (%)	Tempo (s)
481567	100,00	91,05	0,05	100,00	93,29	20,91
460872	100,00	93,35	0,04	75,00	92,24	21,85
469793	100,00	62,50	0,04	100,00	100,00	21,93
Médias	99,98	87,21	0,18	96,85	94,34	21,71

Fonte: Elaboração própria.

Pode-se perceber na Tabela 1 que, com a utilização de mais materiais, o recall médio do modelo COPOD superou o do modelo KNN+Sampling. Em parte a explicação para esta mudança está no fato de que o algoritmo COPOD possui apenas um hiper-parâmetro, que é a contaminação, enquanto o KNN e o Sampling possuem outros hiper-parâmetros a serem ajustados. Poder-se-ia pensar em incluir o ajuste de hiper-parâmetros dentro da fase de treinamento realizada após a seleção dos dados pelo usuário. Porém, como fica claro na Tabela 1, o tempo necessário para treinar o modelo com KNN e Sampling é em média mais de 120 vezes maior que para treinar o COPOD. Além disso, realizar o ajuste dos hiper-parâmetros aumentaria ainda mais esse tempo. Sendo assim, optou-se por utilizar o algoritmo COPOD.

### 4.3 Resultados do modelo final

Tendo então sido definido o algoritmo a ser utilizado na ferramenta de detecção de sobrepreço, restava realizar a avaliação do modelo final. Para esta avaliação, foi utilizada a técnica de validação cruzada *LeaveOneOut* explicada na subseção 3.5.5.

No caso desta pesquisa, a utilização desta técnica para realizar a avaliação final se aproxima muito da utilização real da ferramenta, pois quando a ferramenta for utilizada, todos os registros da base serão utilizados no treinamento e apenas o resultado inserido pelo usuário da aplicação será testado.

**Tabela 2** – Avaliação Final do Modelo

Material	TN	FP	FN	TP	<i>Recall</i>	Acurácia	F <sub>1</sub> -Score
104671	41704	4505	2	1183	0,9983	0,9049	0,3442
150658	15319	1654	0	436	1,0000	0,9050	0,3452
445485	651	69	0	21	1,0000	0,9069	0,3784
402920	421	0	0	10	1,0000	1,0000	1,0000
150877	337	38	0	9	1,0000	0,9010	0,3214

*Continua na próxima página*

**Tabela 2** – Continuação

<b>Material</b>	<b>TN</b>	<b>FP</b>	<b>FN</b>	<b>TP</b>	<b>Recall</b>	<b>Acurácia</b>	<b>F<sub>1</sub>-Score</b>
481567	313	33	0	11	1,0000	0,9076	0,4000
460872	305	22	0	20	1,0000	0,9366	0,6452
469793	264	0	0	13	1,0000	1,0000	1,0000
461652	183	21	0	5	1,0000	0,8995	0,3226
464381	180	19	0	6	1,0000	0,9073	0,3871
461506	161	17	5	7	0,5833	0,8842	0,3889
463795	140	15	0	4	1,0000	0,9057	0,3478
150515	137	17	1	3	0,7500	0,8861	0,2500
<b>Total</b>	60115	6410	8	1728	0,9954	0,9060	0,3500

Fonte: Elaboração própria.

Na Tabela 2 <sup>2</sup> encontram-se os resultados da avaliação final do modelo obtidos com o teste realizado nos 13 (treze) materiais com maior número de registros. Pode-se perceber que o *recall* total ficou em 99,54%, e como pode-se ver, apenas 8 registros não foram identificados. A acurácia ficou em 90,60%, apontando como anomalias, 6.410 sem realmente serem. Em um caso real significaria que em aproximadamente 10% das possíveis avaliações seria necessário que o gestor fizesse uma avaliação mais aprofundada da pesquisa para garantir sua lisura.

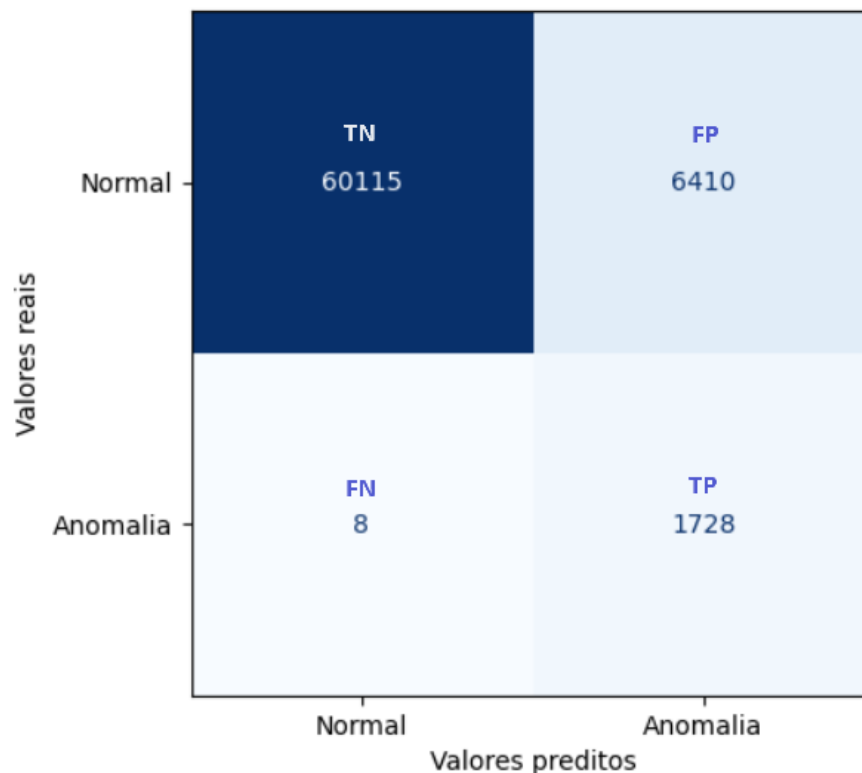
A Figura 15 apresenta a matriz de confusão da avaliação final gerada com a avaliação dos 13 materiais com maior número de registros. Nela fica demonstrada que o algoritmo identificou corretamente 61.843 dos 68.261 registros avaliados, tendo 8 anomalias não identificadas e 1.728 anomalias identificadas corretamente.

Observa-se, ainda, que os valores preditos como anomalia são aproximadamente 12% do total dos valores avaliados. Este número está relacionado com o hiper-parâmetro contaminação utilizado no algoritmo COPOD, já que este hiper-parâmetro define a linha de corte que o algoritmo utiliza para determinar o que é e o que não é uma anomalia. Pode-se dizer que este parâmetro define a sensibilidade que a ferramenta terá na detecção de sobrepreços. Por este motivo, optou-se por criar uma possibilidade de ajuste deste hiper-parâmetro pelos gestores, que poderiam aprender a configurá-lo com o tempo.

A Tabela 2 mostra que a acurácia do modelo ficou em média em 90,6%, mas tivemos materiais com acurácia de 88,42% e outros com 100%. Quanto ao recall, que no caso desta pesquisa é a métrica mais importante, pois indica o percentual de sobrepreços que foram identificados, uma média excelente de 99,54%, mas com uma variação um pouco maior do que a acurácia, variando entre 58,33% a 100% dos casos. O que significa que, no pior caso, foi possível identificar sobrepreços existentes em 58,33% dos casos.

Esses resultados podem ser considerados satisfatórios, uma vez que os gestores, atual-

<sup>2</sup> As métricas utilizadas na tabela estão explicadas na subseção 2.3.3

**Figura 15** – Matriz de confusão (13 materiais com mais registros)

Fonte: elaboração própria

mente, não dispõem de qualquer ferramental de auxílio na elaboração da pesquisa de preços. Logo, um indicativo que aponta para uma probabilidade de identificação de sobrepreço em torno de 99,54% com uma acurácia de 90,6% é um importante balizamento para a avaliação.

#### 4.4 Resumo do capítulo

Neste capítulo foram apresentados os resultados obtidos durante os testes de seleção dos algoritmos para composição do modelo final, onde decidiu-se por utilizar o Algoritmo COPOD que apresentou um *recall* de 97,97% na análise com os 8 materiais com maior número de registros, e o tempo necessário para seu treinamento foi substancialmente menor que seu melhor concorrente. Os resultados obtidos na aplicação do modelo final em cada um dos 13 materiais com maior número de registros apontou para um *recall* médio de 99,54% com uma acurácia de 90,6%, que foi um excelente desempenho obtido pela ferramenta.





## 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. Foi possível elaborar um modelo inicial com capacidade de apontar indícios de sobrepreço nas contratações do Senado Federal, que pode servir 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 as 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, foi o escolhido para servir de subsídio para a pesquisa, por ser a fonte de dados mais relevante para o propósito deste trabalho.

O desenvolvimento da ferramenta de detecção de sobrepreço incluiu a obtenção e pré-processamento dos dados necessários ao treinamento de algoritmos de aprendizagem. Foram levantados dados de licitações e contratos que datam de 01/01/2022 a 14/09/2023. Neste ponto foram encontradas as maiores dificuldades, já que a API apresentou diversos problemas e erros que exigiram a criação de novos métodos para contornar suas deficiências.

Uma vez organizados os dados e disponibilizados em repositório público ([TERRA NETO, 2023](#)), passou-se à tarefa de levantar 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, escrita em *Python* e especializada em detecção de anomalias. A PyOD também 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 na área de detecção de anomalias, como por exemplo o COPOD, de 2020, que apresentou melhor performance nos testes realizados.

Com a utilização da biblioteca PyOD e do SUOD foi possível treinar e utilizar um modelo de AM na detecção de sobrepreço nas contratações públicas do banco de dados preparado. Realizaram-se diversos testes. Primeiramente, testaram-se diversos algoritmos não supervisionados existentes na biblioteca para selecionar aqueles que tinham melhor performance sobre os dados disponíveis. Após a seleção, diversas simulações foram feitas para escolher a melhor combinação de algoritmos e determinar os parâmetros a serem utilizados. Chegou-se à conclusão de que o algoritmo COPOD era a solução mais adequada para o problema em questão considerando os dados levantados.

O protótipo da ferramenta de avaliação de indícios de sobrepreços para uso efetivo no Senado Federal, da perspectiva de sua interface, foi desenvolvido de forma que o usuário possa

selecionar o material desejado e o período de tempo das amostras, possibilitando que os preços estejam dentro da validade ao serem considerados na elaboração de um documento de pesquisa de preços. Mesmo que preliminares, os resultados apresentados pela ferramenta podem ser utilizados de forma efetiva pelos gestores.

Finalmente, ao avaliar o grau de confiabilidade da ferramenta na predição de sobrepreço, foi encontrada uma detecção de anomalias na ordem de 99,54%, com acurácia de 90,6% nos testes realizados. Tendo em vista que no presente momento não existe ferramenta semelhante, há um ganho potencial na sua utilização. Porém, apesar do alto fator de detecção na predição realizada, o número de falsos positivos é significativo, o que indica que muitos itens que não eram sobrepreços foram encarados como suspeitos, fazendo com que o gestor tenha que examiná-los com maior cuidado, o que deixa espaço para um aperfeiçoamento da ferramenta.

## 5.1 Limitações

Nos testes efetuados escolheram-se os materiais com maior quantidade de registros no repositório, o que possibilitou a execução de 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 na medida em 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.

Cabe destacar que a utilização do Snorkel para a determinação dos rótulos utilizados para ajuste e teste dos algoritmos é um fator limitante, já que a qualidade dos rótulos está diretamente ligada a qualidade das funções de rotulagem, que podem não representar toda a complexidade e nuance dos dados. Precisa-se ter em mente, também, que podem acontecer anomalias raras que não estão previstas nestas funções de rotulagem, fazendo com que a determinação não seja correta, ou que as funções possam introduzir algum viés baseado em suposições incorretas ou incompletas sobre os dados. Apesar disso, a contribuição da ferramenta oferece vantagem significativa, já que é possível rotular grande quantidade de dados de treinamento em situações onde os rótulos manuais seriam escassos ou caros de se obterem.

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.

Foram realizadas duas sugestões de aprimoramentos durante a banca de qualificação que não puderam ser implementadas. A primeira é a sugestão de comparar a especificação/descrição dos itens licitados, o que demandaria o desenvolvimento de uma IA para ler e interpretar os requisitos do edital de forma a poder compará-los, este projeto pode ser desenvolvido no futuro, mas não foi possível no momento em face do tempo necessário para sua realização. A segunda sugestão é a verificação da possibilidade de apresentar a informação do desconto da licitação, que seria a diferença entre o valor da pesquisa e o valor licitado, neste caso, os dados referentes aos valores da pesquisa não estão disponíveis, mas caso a ferramenta evolua, poderiam ser armazenados para que esta informação possa vir a fazer parte.

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

ALHAMID, Mohammed. **What is Cross-Validation?** Estados Unidos da América: Medium, 2020. Disponível em: <https://towardsdatascience.com/what-is-cross-validation-60c01f9d9e75>. Acesso em: 9 fev. 2024. Citada 6 vezes nas páginas 61(4x) e 62(2x).

ARAUJO, 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 27.

BANDARAGODA, Tharindu *et al.* Isolation-based anomaly detection using nearest-neighbor ensembles: inne. **Computational Intelligence**, v. 34, Jan. 2018. Disponível em: [https://www.researchgate.net/publication/322359651\\_Isolation-based\\_anomaly\\_detection\\_using\\_nearest-neighbor\\_ensembles\\_iNNE](https://www.researchgate.net/publication/322359651_Isolation-based_anomaly_detection_using_nearest-neighbor_ensembles_iNNE). Acesso em: 29 abr. 2024. Citada 3 vezes nas páginas 65, 66 e 67.

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](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122020000501382&tlng=pt). Acesso em: 12 maio 2023. Citada 3 vezes na página 39(3x).

BERRAR, Daniel. Cross-validation. *In: ENCYCLOPEDIA of Bioinformatics and Computational Biology*. Tokyo, Japan: Elsevier, 2018. v. 1, p. 542–545. ISBN 9780128096338. Disponível em: [https://www.researchgate.net/publication/324701535\\_Cross-Validation](https://www.researchgate.net/publication/324701535_Cross-Validation). Acesso em: 16 abr. 2024. Citada 1 vez na página 75.

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 28 e 33.

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](http://www.planalto.gov.br/ccivil_03/leis/L8666compilado.htm). Acesso em: 7 maio 2019. Citada 2 vezes nas páginas 33 e 35.

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](http://www.planalto.gov.br/ccivil_03/leis/2002/110520.htm). Acesso em: 25 ago. 2022. Citada 1 vez na página 33.

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/l12527.htm](https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm). Acesso em: 25 mar. 2023. Citada 1 vez na página 28.

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 34(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](http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14133.htm). Acesso em: 21 mar. 2022. Citada 5 vezes nas páginas 27, 29, 33 e 34(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 28, 37, 52, 117, 118, 119(2x), 120, 121 e 122.

CALANCA, Paulo; MATHEUS, Yuri; RAPHAELL, Bruno. **Quais são os 4 tipos de aprendizagem na IA, algoritmos e usos no dia a dia**. [S. l.]: Alura, 2023. Disponível em: <https://www.alura.com.br/artigos/quais-sao-tipos-aprendizagem-ia-inteligencia-artificial>. Acesso em: 17 out. 2023. Citada 2 vezes na página 41(2x).

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 1 vez na página 42.

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 45(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&lng=pt](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122020000501382&lng=pt). Acesso em: 22 mar. 2022. Citada 1 vez na página 43.

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 2 vezes nas páginas 37 e 52.

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](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3904483). Acesso em: 22 mar. 2022. Citada 1 vez na página 42.

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](https://link.springer.com/chapter/10.1007/978-3-319-30298-0_50). Acesso em: 20 abr. 2023. Citada 2 vezes nas páginas 70 e 71.
- 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 34.
- 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 29(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 7 vezes nas páginas 40, 46, 47, 59, 60, 62 e 71.
- 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](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 56 e 57(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 1 vez na página 37.
- 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 29.
- 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 73.
- 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 72.
- 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 35.



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 29.

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/repositorio/article/view/2484>. Acesso em: 14 out. 2022. Citada 1 vez na página 35.

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 nas páginas 38 e 39.

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 35.

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 1 vez na página 37.

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 27.

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 29.

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 27 e 29.

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 43.

OPEN KNOWLEDGE BRASIL. **Painel Jarbas**. São Paulo/SP: Open Knowledge Brasil, 20—. Disponível em: [https://jarbas.serenata.ai/dashboard/chamber\\_of\\_deputies/reimbursement](https://jarbas.serenata.ai/dashboard/chamber_of_deputies/reimbursement). Acesso em: 3 abr. 2023. Citada 1 vez na página 44.



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 43 e 44(2x).

PANDEY, Pallavi. **Feature Scaling**. [S. l.]: Machine Learning Geek, 2020. Disponível em: <https://machinelearninggeek.com/feature-scaling-minmax-standard-and-robust-scaler/>. Acesso em: 9 fev. 2024. Citada 3 vezes nas páginas 59(2x) e 60.

PANG, Guansong *et al.* Deep learning for anomaly detection: A review. **ACM Computing Surveys**, ACM PUB27 New 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 4 vezes nas páginas 42, 67, 69 e 70.

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 27, 29 e 45.

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§Ã£oecapacidadesassimÃªtricasdagestÃ£oedocontrole-estadao.pdf?sequence=1&isAllowed=y](https://repositorio.ufsc.br/bitstream/handle/123456789/222445/[corrupÃ§Ã£o] corrupÃ§Ã£oecapacidadesassimÃªtricasdagestÃ£oedocontrole-estadao.pdf?sequence=1&isAllowed=y). Acesso em: 21 mar. 2022. Citada 3 vezes na página 28(3x).

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citada 3 vezes nas páginas 60 e 62(2x).

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 39.

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 36.

RAPHAELL, Bruno. Desmistificando termos em machine learning: tipos de aprendizado. v. 16, 2021. Disponível em: <https://www.alura.com.br/artigos/desmistificando-termos-machine-learning-tipos-aprendizado>. Acesso em: 27 mar. 2023. Citada 1 vez na página 41.

RASCHKA, Sebastian; PATTERSON, Joshua; NOLET, Corey. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. **Information**, MDPI, v. 11, n. 7, p. 345, 2020. Disponível em: <https://arxiv.org/abs/2002.04803>. Acesso em: 5 abr. 2024. Citada 1 vez na página 40.

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 35.

- 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 36.
- SENADO FEDERAL. **Dados Abertos**. [S. l.: s. n.], 2022. Disponível em: <https://www12.senado.leg.br/dados-abertos>. Acesso em: 29 abr. 2022. Citada 1 vez na página 38.
- SHARMA, Prashant. **Different Types of Cross-Validations in Machine Learning**. Indore, Índia: Analytics Vidhya, 2022. Disponível em: <https://www.analyticsvidhya.com/blog/2022/02/different-types-of-cross-validations-in-machine-learning/>. Acesso em: 16 fev. 2024. Citada 1 vez na página 62.
- 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 29.
- SILVA, Fernando da. **REAMOSTRAGEM EM MODELOS PREDITIVOS: SEPARAÇÃO TREINO E TESTE**. Rio de Janeiro/RJ: Análise Macro, 2023. Disponível em: <https://analisemacro.com.br/econometria-e-machine-learning/reamostragem-em-modelos-preditivos-separacao-treino-e-teste/>. Acesso em: 14 fev. 2024. Citada 2 vezes nas páginas 60 e 62.
- 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 27, 34(2x) e 35.
- SNORKEL AI. **Get Started**. [S. l.: s. n.], 2020. Disponível em: <https://www.snorkel.org/get-started/>. Acesso em: 6 fev. 2024. Citada 1 vez na página 63.
- 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 42(2x).
- SUGIYAMA, Mahito; BORGWARDT, Karsten. Rapid distance-based outlier detection via sampling. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 27., 2013, Lake Tahoe, Nevada, USA. **Proceedings** [...]. USA: Curran Associates, Inc., 2013. v. 26, p. 467-475. ISBN 9781632660244. Disponível em: [https://proceedings.neurips.cc/paper\\_files/paper/2013/file/d296c101daa88a51f6ca8cfc1ac79b50-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2013/file/d296c101daa88a51f6ca8cfc1ac79b50-Paper.pdf). Acesso em: 29 fev. 2024. Citada 1 vez na página 71.
- 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 5 vezes nas páginas 51(2x), 55, 63 e 95.

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 36.

VAZ, Arthur Lamblet. **Otimizando os hiperparâmetros**. Estados Unidos da América: Medium, 2019. Disponível em: <https://medium.com/data-hackers/otimizando-os-hiperpar%C3%A2metros-621de5e9be37>. Acesso em: 17 fev. 2024. Citada 2 vezes na página 62(2x).

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 35.

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 40.

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 42 e 115.

ZHAO, Yue. **pyod 1.1.4 documentation**. [S. l.: s. n.], 2023. Disponível em: <https://pyod.readthedocs.io/en/latest/>. Acesso em: 16 abr. 2024. Citada 2 vezes nas páginas 42 e 66.

ZHAO, Yue. **pyod 1.1.4 documentation - API Reference - INNE**. [S. l.: s. n.], 2023. Disponível em: <https://pyod.readthedocs.io/en/latest/pyod.models.html#pyod.models.inne.INNE>. Acesso em: 16 abr. 2024. Citada 1 vez na página 65.

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 74(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 42.



## APÊNDICES



## APÊNDICE A – FUNÇÕES DE ROTULAGEM (LFS) UTILIZADAS NO SOFTWARE SNORKEL

### Código-Fonte 6 – Funções de rotulagem dos dados

```

1 #####
2 # Funções de rotulagem de dados
3 # Objetivo: Rotular os dados
4 # Parâmetros: catmat - código do material a ser recuperado
5 #               data - Data a partir da qual os registros serão
6 #               selecionados
7 # Retorno: dataframe pandas com todos os registros selecionados
8 #####
9 @labeling_function()
10 def preco_anomalo(v_df):
11     preco=v_df['valor_unitario']
12     quantidade=v_df['quantidade']
13     return ANOMALY if ((preco > preco_maior) or ((preco < preco_menor) and
14         ((quantidade<quantidade_maior)))) else ABSTAIN
15
16 @labeling_function()
17 def quantidade_alta(v_df):
18     quantidade=v_df['quantidade']
19     #Retorna um label de anomalia se o valor é maior que 97,5% dos valores
20     #se não se ABSTAIN
21     return ANOMALY if quantidade > quantidade_maior else ABSTAIN
22
23 @labeling_function()
24 def distancia_alta(v_df):
25     distancia=v_df['distancia_uasg_fornecedor']
26     #Retorna um label de anomalia se o valor é
27     #maior que 97,5% dos valores se não se ABSTAIN
28     return ANOMALY if distancia > distancia_maior else ABSTAIN
29
30 @labeling_function()
31 def normal(v_df):
32     preco=v_df['valor_unitario']
33     quantidade=v_df['quantidade']
34     distancia=v_df['distancia_uasg_fornecedor']
35     return NORMAL if (((preco<=preco_mediana) and (preco>preco_menor)) or
36         ((preco>preco_mediana) and (preco<preco_maior))) else ABSTAIN

```





## ANEXOS



## ANEXO A – ALGORITMOS DO PYOD

**Quadro 14** – Algoritmos do PyOD

<b>Tipo</b>	<b>Sigla</b>	<b>Algoritmo</b>	<b>Ano</b>
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 14 – 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 14** – Continuação

<b>Tipo</b>	<b>Sigla</b>	<b>Algoritmo</b>	<b>Ano</b>
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: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 15** – Métodos de Consultas Básicas de Contratos

<b>Método</b>	<b>Descrição</b>
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 16** – Métodos de Informações Detalhadas de Contratos

<b>Método</b>	<b>Descrição</b>
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 16 – Continuação**

<b>Método</b>	<b>Descrição</b>
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 17 – Métodos de Consultas Básicas de Fornecedores**

<b>Método</b>	<b>Descrição</b>
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 17 – Continuação**

<b>Método</b>	<b>Descrição</b>
tipos_ocorrendia	Retorna lista com tipos de ocorrências.

Fonte: API de Compras Governamentais (BRASIL, 2021)

**Quadro 18 – Métodos de Informações Detalhadas de Fornecedores**

<b>Método</b>	<b>Descrição</b>
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 19 – Métodos de Consultas Básicas de Licitações**

<b>Método</b>	<b>Descrição</b>
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.
rdcs	Fornece uma lista com informações relacionadas às RDCs.

Fonte: API de Compras Governamentais (BRASIL, 2021)

**Quadro 20 – Métodos de Informações Detalhadas das Licitações**

<b>Método</b>	<b>Descrição</b>
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 21 – 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 22** – Métodos de Informações Detalhadas de Materiais

<b>Método</b>	<b>Descrição</b>
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 da 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.					