



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS REITOR JOÃO DAVID FERREIRA LIMA
PROGRAMA DE GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Lívia Corazza Ferrão

**INTEGRAÇÃO DE TAXONOMIA HIERÁRQUICA DE PRODUTOS E APRENDIZADO
SEMISSUPERVISIONADO PARA DETECÇÃO DE FRAUDES EM LICITAÇÕES
PÚBLICAS**

Florianópolis, Santa Catarina – Brasil
2026

Livia Corazza Ferrão

**INTEGRAÇÃO DE TAXONOMIA HIERÁRQUICA DE PRODUTOS E APRENDIZADO
SEMISSUPERVISIONADO PARA DETECÇÃO DE FRAUDES EM LICITAÇÕES
PÚBLICAS**

Trabalho de Conclusão de Curso submetido
ao Programa de Graduação em Ciências da
Computação da Universidade Federal de Santa
Catarina para a obtenção do Grau de Bacharel em
Ciências da Computação.

Orientador: Prof. Jônata Tyska Carvalho, Dr.

Florianópolis, Santa Catarina – Brasil

2026

Catálogo na fonte pela Biblioteca Universitária da Universidade Federal de Santa Catarina.
Arquivo compilado às 18:48h do dia 11 de julho de 2026.

Lívia Corazza Ferrão

Integração de Taxonomia Hierárquica de Produtos e Aprendizado Semissupervisionado para Detecção de Fraudes em Licitações Públicas / Lívia Corazza Ferrão; Orientador, Prof. Jônata Tyska Carvalho, Dr. – Florianópolis, Santa Catarina – Brasil, 30 de junho de 2026.

82 p.

Trabalho de Conclusão de Curso – Universidade Federal de Santa Catarina, INE – Departamento de Informática e Estatística, CTC – Centro Tecnológico, Programa de Graduação em Ciências da Computação.

Inclui referências

1. Licitações, 2. Análise semântica, 3. BERTopic, 4. Modelos de linguagem, 5. Modelos de Mistura Gaussiana, I. Prof. Jônata Tyska Carvalho, Dr. II. Programa de Graduação em Ciências da Computação III. Integração de Taxonomia Hierárquica de Produtos e Aprendizado Semissupervisionado para Detecção de Fraudes em Licitações Públicas

CDU 02:141:005.7

Lívia Corazza Ferrão

**INTEGRAÇÃO DE TAXONOMIA HIERÁRQUICA DE PRODUTOS E APRENDIZADO
SEMISSUPERVISIONADO PARA DETECÇÃO DE FRAUDES EM LICITAÇÕES
PÚBLICAS**

Este(a) Trabalho de Conclusão de Curso foi julgado adequado(a) para obtenção do Título de Bacharel em Ciências da Computação, e foi aprovado em sua forma final pelo Programa de Graduação em Ciências da Computação do INE – Departamento de Informática e Estatística, CTC – Centro Tecnológico da Universidade Federal de Santa Catarina.

Florianópolis, Santa Catarina – Brasil, 30 de junho de 2026.

**Profa. Lúcia Helena Martins Pacheco,
Dra.**

Coordenadora do Programa de Graduação
em Ciências da Computação

Banca Examinadora:

Prof. Jônata Tyska Carvalho, Dr.

Orientador
Universidade Federal de Santa
Catarina – UFSC

Prof. Matheus Machado dos Santos, Dr.

Avaliador
Universidade de São Paulo - USP

Profa. Carina Friedrich Dorneles, Dra.

Avaliador
Universidade Federal de Santa Catarina –
UFSC

AGRADECIMENTOS

Os agradecimentos principais são direcionados aos meus pais, pelo apoio incondicional ao longo de toda a minha trajetória acadêmica. Agradeço aos meus amigos, que estiveram presentes durante todo o curso, oferecendo companhia, incentivo nos períodos mais intensos.

Agradeço também meu orientador Prof. Jônata Tyska, pela orientação e pelas discussões construtivas ao longo do desenvolvimento deste trabalho. Estendo esse agradecimento aos professores do curso de Ciência da Computação da UFSC, que contribuíram para minha formação.

Agradeço ainda aos colegas e colaboradores envolvidos neste projeto, especialmente aqueles que trabalhei de forma conjunta durante o projeto do CEOS.

RESUMO

Fraudes em processos licitatórios representam um dos problemas mais graves para a administração pública, gerando desperdício de recursos, enfraquecimento da concorrência e comprometimento da confiança nas instituições, no Brasil, onde bilhões de reais são movimentados anualmente em contratações públicas, o direcionamento de licitações, prática em que editais são elaborados com especificações técnicas restritivas para favorecer determinados fornecedores, constitui uma modalidade especialmente difícil de detectar, pois os indícios estão implicitamente embutidos na linguagem técnica dos documentos. A fiscalização tradicional, baseada em análise manual, é incapaz de acompanhar o volume crescente de processos, tornando a automação uma necessidade urgente e o aprendizado de máquina uma oportunidade concreta para detecção automatizada de irregularidades. Nesse contexto, este trabalho propõe uma abordagem integrada que conecta a construção automatizada de uma taxonomia hierárquica de atributos técnicos com o modelo semissupervisionado SubFraudGMM, baseado em Modelos de Mistura Gaussiana, incorporando quatro novas variáveis derivadas dos termos de referência. Os experimentos foram conduzidos sobre 100 termos de referência de licitações de escavadeiras hidráulicas, dos quais 30 são casos confirmados de fraude provenientes da Operação Patrola (MPSC) e 70 são documentos não rotulados, resultando em uma taxonomia com 176 caminhos hierárquicos distintos. A metodologia é aplicável a outros objetos de licitação, exigindo a disponibilização dos respectivos termos de referência e fraudes confirmadas do objeto analisado. Na validação, cada uma das 30 fraudes conhecidas foi, por sua vez, mascarada e tratada como não rotulada, sendo o modelo executado uma vez para cada fraude removida. Em cada execução, foi obtido o Indicador de Risco atribuído à fraude mascarada. A partir dos 30 valores resultantes, o modelo aprimorado apresentou mediana de 78,8% no Indicador de Risco dessas fraudes, contra 77,2% da versão base, indicando que a incorporação de características hierárquicas contribuiu para refinar a detecção de padrões de direcionamento em compras públicas, com potencial de aprimoramento adicional pela inclusão de novas características.

Palavras-chave: Licitações públicas. Detecção de fraudes. Taxonomia hierárquica. Aprendizado semissupervisionado. Modelos de Mistura Gaussiana.

ABSTRACT

Fraud in public procurement processes represents one of the most serious challenges for public administration, causing resource waste, weakened competition, and reduced trust in institutions. In Brazil, where billions of reais are spent annually on public contracts, bid steering, a practice in which public notices are designed with restrictive technical specifications to favor specific suppliers, constitutes a particularly difficult type of fraud to detect, as its indicators are implicitly embedded in the technical language of the documents. Traditional auditing approaches, based on manual analysis, are unable to keep pace with the growing volume of procurement processes, making automation an urgent necessity and machine learning a promising opportunity for the automated detection of irregularities. In this context, this work proposes an integrated approach that combines the automated construction of a hierarchical taxonomy of technical attributes with the semi-supervised SubFraudGMM model, based on Gaussian Mixture Models, incorporating four new variables derived from terms of reference. The experiments were conducted on 100 terms of reference from hydraulic excavator procurement processes, of which 30 are confirmed fraud cases from Operation Patrola (MPSC) and 70 are unlabeled documents, resulting in a taxonomy with 176 distinct hierarchical paths. The methodology is applicable to other procurement objects, requiring the availability of the respective terms of reference and confirmed fraud cases for the analyzed object. During validation, each of the 30 known fraud cases was individually masked and treated as unlabeled, with the model executed once for each removed fraud case. In each execution, the Risk Indicator assigned to the masked fraud was obtained. Based on the 30 resulting values, the enhanced model achieved a median Risk Indicator of 78.8% for these fraud cases, compared to 77.2% for the original version, indicating that the incorporation of hierarchical features contributed to refining the detection of bid steering patterns in public procurement, with additional improvement potential through the inclusion of new features.

Keywords: Public procurement. Fraud detection. Hierarchical taxonomy. Semi-supervised learning. Gaussian Mixture Models.

LISTA DE FIGURAS

Figura 1	– Exemplo simplificado da hierarquia de atributos.	31
Figura 2	– Fluxograma geral do TCC.	36
Figura 3	– Top 20 caminhos por número de documentos.	38
Figura 4	– Visão Geral das Frequências dos 176 Atributos.	38
Figura 5	– Análise de sobreposição de caminhos entre documentos de fraude e não rotulados	41
Figura 6	– Indicador de Risco para os modelos construídos.	47
Figura 7	– Curva de Recuperação de Fraudes em função do Top-N registros ranqueados pelo Indicador de Risco	47
Figura 8	– Distribuição do Indicador de Risco por grupo (RT vs Demais) no painel CEOS (2020–2026)	49
Figura 9	– Visualização do alerta de favoritismo na licitação.	52
Figura 10	– Listagem dos alertas de favoritismo na aba de alertas.	53
Figura 11	– Alerta de favoritismo da 1ª licitação com maior risco (Parte 1). . .	53
Figura 12	– Alerta de favoritismo da 1ª licitação com maior risco (Parte 2). . .	54
Figura 13	– Alerta de favoritismo da 2ª licitação com maior risco (Parte 1). . .	54
Figura 14	– Alerta de favoritismo da 2ª licitação com maior risco (Parte 2). . .	55
Figura 15	– Alerta de favoritismo da 3ª licitação com maior risco (Parte 1). . .	55
Figura 16	– Alerta de favoritismo da 3ª licitação com maior risco (Parte 2). . .	56
Figura 17	– Alerta de favoritismo da 4ª licitação com maior risco (Parte 1). . .	56
Figura 18	– Alerta de favoritismo da 4ª licitação com maior risco (Parte 2). . .	57
Figura 19	– Alerta de favoritismo da 5ª licitação com maior risco (Parte 1). . .	57
Figura 20	– Alerta de favoritismo da 5ª licitação com maior risco (Parte 2). . .	58
Figura 21	– Alerta de favoritismo da 1ª licitação com maior risco (Parte 1). . .	71
Figura 22	– Alerta de favoritismo da 1ª licitação com maior risco (Parte 2). . .	71

LISTA DE TABELAS

Tabela 1	–	Composição da base de dados de Termos de Referência	28
Tabela 2	–	Características utilizadas no SubFraudGMM	33
Tabela 4	–	Características selecionadas para o modelo de agrupamento . .	34
Tabela 5	–	Top 5 caminhos por categoria de frequência.	39
Tabela 6	–	Top 5 licitações não rotuladas com maior indicador de risco - modelo original	44
Tabela 7	–	Top 5 licitações não rotuladas com maior indicador de risco - novo modelo	45
Tabela 8	–	Comparação da mediana do Indicador de Risco (Grupo RT) . . .	46
Tabela 9	–	Impacto das novas características no Indicador de Risco médio do grupo RT	46
Tabela 10	–	Top 5 licitações não rotuladas com maior indicador de risco - painel CEOS (2020–2026)	49
Tabela 11	–	Composição da base de dados de Termos de Referência	67
Tabela 12	–	Novas características propostas	69
Tabela 13	–	Mediana do Indicador de Risco no grupo RT	70
Tabela 14	–	Top 5 licitações não rotuladas com maior indicador de risco - novo modelo	70

LISTA DE ABREVIATURAS E SIGLAS

BERT	Representações Bidirecionais Codificadas por Transformadores
BERTopic	Modelagem de Tópicos Baseada em BERT
GMM	Modelo de Mistura Gaussiana
LLM	Modelo de Linguagem de Larga Escala
PLN	Processamento de Linguagem Natural
LOO	Validação por Exclusão de Uma Observação
LOFO	Validação com Remoção de Fraudes
BIC	Critério de Informação Bayesiano
TCC	Trabalho de Conclusão de Curso
UFSC	Universidade Federal de Santa Catarina

SUMÁRIO

1	INTRODUÇÃO	12
1.1	MOTIVAÇÃO	13
1.2	OBJETIVOS	14
1.2.1	Objetivos Específicos	14
1.3	ORGANIZAÇÃO DO TEXTO	15
2	REVISÃO BIBLIOGRÁFICA	16
2.1	FUNDAMENTAÇÃO TEÓRICA	16
2.1.1	Licitações Públicas	16
2.1.2	Modelos de Linguagem de Larga Escala (LLMs)	17
2.1.3	Embeddings	17
2.1.4	BERTopic	18
2.1.5	Taxonomias Hierárquicas	19
2.1.6	Aprendizado Supervisionado	19
2.1.7	Aprendizado Não Supervisionado	20
2.1.8	Aprendizado Semissupervisionado	21
2.1.9	Modelos de Mistura Gaussiana	22
2.1.10	Leave-One-Out (LOO)	23
2.2	TRABALHOS CORRELATOS	24
3	MATERIAIS E MÉTODOS	27
3.1	ENTENDIMENTO E PREPARAÇÃO DOS DADOS	27
3.1.1	Operação Patrola	27
3.1.2	Preparação da base de dados	27
3.2	MODELAGEM	28
3.2.1	Pipeline de construção da taxonomia	28
3.2.2	Tratamento manual e agregação por documento	29
3.2.3	Construção da hierarquia de atributos	30
3.2.4	Modelo semissupervisionado baseado em subespaços	32
3.2.4.1	Características Utilizadas e Extensões Propostas	33
3.2.5	Novas características	34
3.2.6	Visão geral da pipeline de processamento	35
4	AVALIAÇÃO	37
4.1	ABORDAGEM METODOLÓGICA DA AVALIAÇÃO	37
4.2	ANÁLISE DOS DADOS	37
4.3	MÉTRICAS	42
4.4	VALIDAÇÃO	43

4.5	ANÁLISE COMPARATIVA DO INDICADOR DE RISCO	44
4.6	VALIDAÇÃO DO INDICADOR DE RISCO	45
4.7	ANÁLISE DO INDICADOR DE RISCO COM DADOS RECENTES DO PAINEL	48
4.8	DISCUSSÃO DOS RESULTADOS	50
5	APLICAÇÃO	52
5.1	INTEGRAÇÃO DO ALERTA DE FAVORITISMO AO PAINEL CEOS	52
6	CONCLUSÃO	59
6.1	TRABALHOS FUTUROS	60
	REFERÊNCIAS	61
	APÊNDICE A – ARTIGO EM FORMATO SBC	65
1	INTRODUÇÃO	66
2	TRABALHOS RELACIONADOS	67
3	MATERIAIS E MÉTODOS	67
3.1	Base de Dados	67
3.2	Construção da Taxonomia	67
3.3	Modelo Semissupervisionado SubFraudGMM	68
3.4	Novas Características	69
4	AVALIAÇÃO E RESULTADOS	69
4.1	Validação por Remoção de Fraudes (LOFO)	69
4.2	Resultados Comparativos	69
4.3	Aplicação Prática	71
5	CONCLUSÃO	72
6	REFERÊNCIAS	72
	APÊNDICE B – CÓDIGO-FONTE	73

1 INTRODUÇÃO

A administração pública brasileira movimenta anualmente bilhões de reais em contratações por meio de licitações ([Portal da Transparência do Governo Federal, 2025](#)). Embora exista um aparato legal robusto para garantir a transparência e a competitividade nesses processos, fraudes continuam a ocorrer, comprometendo não apenas a eficiência da gestão pública, mas também a confiança da população nas instituições ([Lyra et al., 2022](#)). Entre as diversas modalidades de irregularidade, este trabalho foca no direcionamento de licitações, prática em que editais são elaborados para favorecer um fornecedor específico, restringindo a concorrência e viabilizando o desvio de recursos públicos ([Schneider dos Santos et al., 2025](#)).

Dentre os elementos centrais de um processo licitatório está o termo de referência, documento que descreve detalhadamente o objeto da contratação, estabelecendo especificações técnicas, requisitos, critérios de execução e demais condições que orientam a elaboração do edital e a futura contratação. Esse documento tem como objetivo garantir que a necessidade da administração pública seja claramente definida e que os fornecedores possam formular propostas em condições de igualdade. Entretanto, por concentrar as especificações técnicas do objeto, é também no termo de referência que o direcionamento costuma se materializar, por meio da inclusão de requisitos excessivamente restritivos, linguagem tendenciosa ou reutilização de especificações entre diferentes processos, de forma a favorecer determinados fornecedores ([Singh; Singh, 2025](#)). Em investigações como a Operação Patrola, conduzida pelo Ministério Público de Santa Catarina¹, esse tipo de irregularidade foi identificado em licitações de máquinas pesadas.

A detecção desse tipo de irregularidade enfrenta desafios consideráveis. O volume de licitações realizadas anualmente no Brasil é imenso ([Lyra et al., 2022](#)), e a análise manual de documentos técnicos como os termos de referência exige conhecimento especializado e tempo elevado, tornando inviável uma cobertura abrangente por parte dos órgãos de controle. Além disso, os indícios de direcionamento estão frequentemente implícitos na linguagem técnica dos documentos, exigindo não apenas leitura atenta, mas também comparação entre múltiplos editais para identificar padrões suspeitos de repetição ou restrição. Esse cenário torna a fiscalização tradicional, baseada em denúncias ou auditorias pontuais, insuficiente para acompanhar a escala do problema ([Lyra et al., 2022](#)).

A crescente disponibilidade de dados públicos digitais e o avanço do aprendizado de máquina abrem oportunidades concretas para a detecção automatizada de irregu-

¹ Operação desenvolvida pelo MPSC que desarticulou um esquema de fraudes em licitações de máquinas pesadas em 39 municípios catarinenses. Disponível em: <https://www.mpsc.mp.br/w/noticias/operacao-patrola-ex-prefeito-de-tangara-e-condenado-a-13-anos-de-prisao>.

laridades. Este trabalho realiza a integração entre a construção automatizada da uma taxonomia com apoio de modelos de linguagem de larga escala e uma abordagem semissupervisionada baseada em Modelos de Mistura Gaussiana, aplicada à priorização de licitações suspeitas.

A abordagem utiliza modelos de linguagem de larga escala para realizar a análise semântica de termos de referência, em conjunto com métodos de agrupamento de termos e expressões (Schiavon, 2024). Essa estratégia permite transformar textos não estruturados em representações organizadas de atributos e características. A partir dessa estrutura hierárquica, cada licitação é representada como um vetor em um espaço multidimensional com doze características, que alimenta o modelo semissupervisionado SubFraudGMM. Um desafio dessa integração consistiu em converter a representação hierárquica gerada pela taxonomia em um conjunto de variáveis compatíveis com a entrada do modelo. Esse modelo gera um indicador de risco baseado na similaridade entre licitações analisadas e casos suspeitos, além da consistência dos agrupamentos formados (Schmitz, 2025).

Os experimentos foram conduzidos sobre 100 termos de referência de licitações de escavadeiras hidráulicas, selecionadas por terem casos de fraude previamente identificados, dos quais 30 correspondem a casos de fraude apontados pelo Ministério Público de Santa Catarina (MPSC) no âmbito da Operação Patrola e 70 são documentos não rotulados, resultando em uma taxonomia com 176 atributos técnicos distintos. Cada uma das 30 fraudes conhecidas foi mascarada e tratada como não rotulada, sendo o modelo executado uma vez para cada fraude removida; a partir dos 30 valores de Indicador de Risco assim obtidos, o modelo aprimorado apresentou mediana de 78,8% nas fraudes mascaradas, contra 77,2% da versão original, demonstrando que a incorporação de características dos termos de referência contribuiu para refinar a detecção de padrões de direcionamento em compras públicas.

1.1 MOTIVAÇÃO

A motivação deste trabalho surge da necessidade de melhorar os mecanismos de fiscalização e de detecção de irregularidades em contratações públicas. A avaliação dos processos licitatórios exige leitura atenta dos documentos técnicos e ferramentas tradicionais de auditoria costumam depender de denúncias ou análises manuais extensas, o que torna difícil acompanhar o grande volume de licitações realizadas no país. Essa demanda foi evidenciada no contexto do Projeto CEOS, voltado ao desenvolvimento de soluções computacionais para apoio à fiscalização de compras públicas, motivando a investigação de técnicas capazes de automatizar parte desse processo. Com o uso de métodos baseados em aprendizado de máquina, procura-se ampliar a capacidade dos órgãos de controle, identificando indícios de direcionamento ou inconsistências.

1.2 OBJETIVOS

Estender uma abordagem híbrida para identificação de indícios de fraude em processos licitatórios, unificando métodos já existentes: a análise semântica de termos de referência, realizada por meio da construção de taxonomias com Modelos de Linguagem de Larga Escala (LLMs) ([Schiavon, 2024](#)), e as técnicas de aprendizado de máquina baseadas em Modelos de Mistura Gaussiana (GMM) ([Schmitz, 2025](#)), utilizadas para priorização de licitações potencialmente suspeitas. O trabalho concentra-se em conectar essas duas etapas por meio da proposição de novas características derivadas dos termos de referência, permitindo que a representação semântica gerada pela taxonomia seja incorporada ao modelo de detecção de risco.

1.2.1 Objetivos Específicos

A fim de alcançar o objetivo geral deste trabalho, foram definidos os seguintes objetivos específicos:

- Levantar e organizar um conjunto representativo de termos de referência oriundos de processos licitatórios de escavadeiras hidráulicas;
- Aplicar a pipeline de [Schiavon \(2024\)](#) para obter uma base semântica hierárquica de atributos e características técnicas extraídas dos termos de referência, utilizando BERTopic e modelos de linguagem de larga escala (LLMs);
- Padronizar e consolidar os caminhos semânticos obtidos, isto é, as estruturas hierárquicas que representam os atributos técnicos extraídos dos termos de referência, de modo a eliminar redundâncias, alinhar termos semanticamente equivalentes e garantir consistência terminológica entre documentos;
- Possibilitar que a pipeline de [Schiavon \(2024\)](#) integre a taxonomia hierárquica construída ao processo de geração automática de caminhos semânticos, visando reduzir inconsistências decorrentes da geração por modelos de linguagem de larga escala;
- Reaplicar o algoritmo de aprendizado semissupervisionado do trabalho de [Schmitz \(2025\)](#), incorporando as novas características semânticas obtidas, a fim de verificar se elas contribuem para uma melhor diferenciação entre documentos legítimos e fraudulentos;
- Avaliar o impacto da inclusão dessas novas características sobre os resultados previamente alcançados.

1.3 ORGANIZAÇÃO DO TEXTO

O restante deste documento está organizado da seguinte forma. O Capítulo 2 é destinado a apresentar a fundamentação teórica necessária para a realização deste trabalho, bem como desenvolver uma análise aprofundada dos trabalhos correlatos. No Capítulo 3, são vistos os materiais e métodos utilizados para a realização deste trabalho, como a construção do dataset. Já no Capítulo 4, os resultados alcançados são apresentados, bem como uma análise a respeito dos experimentos realizados. Por fim, no Capítulo 5 são apresentadas as conclusões e as sugestões de trabalhos futuros. Após o capítulo de Conclusão, estão listadas as referências.

Durante a elaboração deste trabalho, foram utilizados os assistentes de inteligência artificial ChatGPT e Claude como apoio à revisão textual e à melhoria da clareza e fluidez da escrita. Essas ferramentas contribuíram principalmente para ajustes de forma e linguagem, sem prejuízo do conteúdo técnico apresentado. Todo o conteúdo metodológico e analítico foi cuidadosamente revisado ao longo do processo de elaboração do trabalho.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo está organizado em duas seções. A Seção 2.1 apresenta os conceitos fundamentais utilizados neste trabalho, incluindo os métodos de análise semântica, tipos de aprendizados e a construção de taxonomias.

A Seção 2.2 discute os principais trabalhos correlatos, com ênfase em pesquisas que empregam LLMs na construção de taxonomias e Modelos de Mistura Gaussiana, os quais serviram de referência para o desenvolvimento deste estudo.

2.1 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os principais conceitos, técnicas e métodos que fundamentam esta pesquisa. Inicialmente, são apresentadas as licitações públicas, contextualizando o domínio de aplicação do trabalho. Em seguida, são introduzidos os Modelos de Linguagem de Larga Escala (LLMs) e os embeddings, destacando seu papel na representação e análise semântica de textos. Posteriormente, discute-se o BERTopic, utilizado para agrupamento semântico e construção de taxonomias hierárquicas. Na sequência, são apresentados os métodos de aprendizado supervisionado, não supervisionado e semissupervisionado, bem como o Modelo de Mistura Gaussiana (GMM). Por fim, aborda-se o método Leave-One-Out para validação de modelos.

2.1.1 Licitações Públicas

As licitações públicas constituem o procedimento administrativo por meio do qual a administração pública seleciona a proposta mais vantajosa para a contratação de obras, serviços, compras e alienações, assegurando igualdade de condições entre os participantes e observando os princípios da legalidade, impessoalidade, moralidade, publicidade, eficiência e transparência. No Brasil, esse processo é regulamentado pela Lei nº 14.133/2021, conhecida como Nova Lei de Licitações e Contratos Administrativos, que unificou e modernizou as normas anteriormente vigentes, estabelecendo novas diretrizes para o planejamento, a condução e a fiscalização das contratações públicas (Brasil, 2021).

O processo licitatório é composto por diferentes etapas, iniciando-se com o planejamento da contratação, no qual são definidos a necessidade da administração pública, os requisitos técnicos do objeto e os critérios de julgamento. Entre os documentos produzidos nessa fase destaca-se o termo de referência, responsável por descrever detalhadamente as características técnicas, quantitativas e funcionais do bem ou serviço a ser contratado. Esse documento orienta tanto a elaboração do edital quanto a formulação das propostas pelos licitantes, sendo fundamental para garantir a adequada execução do contrato.

A Lei nº 14.133/2021 prevê diferentes modalidades de licitação, como concorrência, pregão, concurso, leilão e diálogo competitivo, cuja utilização depende das características do objeto e da natureza da contratação. Além disso, estabelece diferentes critérios de julgamento, incluindo menor preço, maior desconto, melhor técnica, técnica e preço, maior lance e maior retorno econômico, permitindo que o procedimento seja adequado às especificidades de cada contratação ([Brasil, 2021](#)).

2.1.2 Modelos de Linguagem de Larga Escala (LLMs)

Modelos de Linguagem de Larga Escala (LLMs) são sistemas de inteligência artificial baseados em redes neurais profundas, especialmente arquiteturas do tipo transformer, treinados com grandes volumes de dados textuais. Esses modelos são capazes de capturar padrões complexos de linguagem, incluindo sintaxe, semântica e contexto, permitindo a compreensão e geração de texto em níveis sofisticados ([Naveed et al., 2024](#)).

Uma das capacidades centrais dos LLMs é a representação semântica de textos por meio de vetores que codificam o significado de palavras, frases ou documentos em espaços multidimensionais. Essas representações facilitam a comparação, agrupamento e análise de informações textuais, sendo especialmente úteis para tarefas que envolvem grandes volumes de dados não estruturados ([Nie et al., 2025](#)).

LLMs apresentam habilidade de raciocínio em múltiplas etapas, permitindo a realização de inferências sequenciais e a síntese de informações a partir de contextos extensos. Esses modelos demonstram capacidade relevante em compreensão de textos complexos, resumo automático e adaptação de estilo de linguagem ([Naveed et al., 2024](#)). Essas propriedades tornam os LLMs adequados para aplicações em análise de documentos técnicos e extração de informações.

No presente trabalho, o LLM é empregado para gerar uma taxonomia hierárquica de atributos técnicos, recebendo como entrada os agrupamentos produzidos pelo BERTopic. O modelo, Qwen 2.5 (32B), processa esses dados, identifica relações de similaridade entre atributos e organiza-os em caminhos hierárquicos estruturados. O uso do LLM neste contexto permite automatizar a transformação de dados textuais em representações semânticas hierárquicas.

2.1.3 Embeddings

Embeddings são representações vetoriais de palavras, sentenças ou documentos em um espaço de alta dimensionalidade, projetadas para capturar relações semânticas e contextuais entre elementos textuais. Nessas representações, textos com significados semelhantes tendem a ocupar posições próximas no espaço vetorial, permitindo que algoritmos computacionais quantifiquem similaridade semântica por meio de medidas de distância ou similaridade, como a similaridade de cosseno ([Mikolov et al.,](#)

2013).

Os primeiros modelos de embeddings, como Word2Vec (Mikolov et al., 2013) e GloVe (Pennington; Socher; Manning, 2014), geravam representações para palavras individuais. Entretanto, esses métodos atribuíam um único vetor para cada palavra, independentemente do contexto em que ela aparecia. Com o avanço dos modelos baseados em transformadores, surgiram os embeddings contextuais, nos quais a representação de uma palavra ou sentença depende do contexto em que está inserida.

No contexto deste trabalho, são utilizados embeddings de sentenças gerados pelo modelo *paraphrase-multilingual-MiniLM-L12-v2*, baseado na arquitetura Sentence-BERT (Reimers; Gurevych, 2019). Esse tipo de representação permite capturar similaridades semânticas entre descrições técnicas escritas de formas distintas, servindo como base para o agrupamento realizado pelo BERTopic.

2.1.4 BERTopic

O BERTopic é um modelo de extração de tópicos que combina técnicas modernas de representação semântica e agrupamento para identificar assuntos recorrentes nos textos (Grootendorst, 2022). Diferentemente de abordagens tradicionais como a *Latent Dirichlet Allocation* (LDA), o BERTopic utiliza representações densas obtidas por meio de modelos de linguagem baseados em transformadores, como o *Sentence-BERT* (Reimers; Gurevych, 2019), que capturam relações semânticas contextuais entre palavras e sentenças. Essa característica permite uma detecção de tópicos mais coerente, especialmente em textos curtos ou com vocabulário variado.

A metodologia do BERTopic segue uma sequência de etapas. Inicialmente, cada documento é convertido em um vetor denso de alta dimensionalidade, utilizando um modelo pré-treinado de *embeddings* linguísticos. Em seguida, aplica-se a redução de dimensionalidade por meio do algoritmo UMAP (*Uniform Manifold Approximation and Projection*) (McInnes; Healy; Melville, 2018), que preserva a estrutura local dos dados enquanto reduz o custo computacional. Os vetores reduzidos são então agrupados utilizando o algoritmo HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) (Campello; Moulavi; Sander, 2013), que identifica agrupamentos de densidade variável e é robusto à presença de ruído.

Após a formação dos grupos, o BERTopic utiliza a técnica *class-based TF-IDF* (c-TF-IDF) para identificar as palavras mais representativas de cada tópico. Essa variação do TF-IDF tradicional calcula a relevância de termos considerando a frequência das palavras dentro de cada grupo em relação ao conjunto de documentos agregados, permitindo uma melhor interpretação semântica dos tópicos resultantes.

No contexto deste trabalho, o BERTopic é utilizado como etapa inicial de agrupamento semântico, responsável por identificar conjuntos de atributos técnicos semelhantes extraídos dos termos de referência das licitações públicas. O BERTopic utiliza

representações vetoriais geradas pelo modelo *paraphrase-multilingual-MiniLM-L12-v2*, uma variante leve e eficiente da arquitetura BERT, capaz de capturar similaridades contextuais entre expressões técnicas distintas. Os grupos de atributos formados pelo BERTopic são então submetidos a um modelo de linguagem de larga escala, responsável por gerar automaticamente uma estrutura hierárquica de taxonomia. Essa estrutura organiza os agrupamentos em caminhos semânticos, isto é, ramos hierárquicos que representam os atributos técnicos extraídos.

2.1.5 Taxonomias Hierárquicas

Taxonomias são estruturas utilizadas para organizar, classificar e representar conceitos de um determinado domínio de conhecimento. Seu principal objetivo é promover a padronização da informação, facilitando a navegação, a recuperação e a reutilização de dados em diferentes contextos de aplicação (Dias; Moreiro González, 2024).

De acordo com Dias e Moreiro González (2024), as taxonomias constituem um tipo de Sistema de Organização do Conhecimento (SOC), sendo amplamente empregadas em ambientes digitais para estruturar informações por meio de categorias e subcategorias organizadas de forma hierárquica. Essa organização permite estabelecer relações entre conceitos e reduzir ambiguidades decorrentes da utilização de diferentes termos para representar uma mesma informação (Dias; Moreiro González, 2024).

A construção de taxonomias envolve a identificação dos conceitos relevantes de um domínio, a definição de categorias e a organização desses elementos em uma estrutura coerente. Tradicionalmente, esse processo é realizado por especialistas, que utilizam conhecimento do domínio para definir os componentes da estrutura taxonômica. Entretanto, avanços recentes em Processamento de Linguagem Natural (PLN) e Modelos de Linguagem de Larga Escala (LLMs) têm possibilitado a automatização parcial desse processo por meio da extração e organização automática de conceitos a partir de documentos textuais.

No contexto deste trabalho, a taxonomia é utilizada para organizar e padronizar os atributos técnicos extraídos dos termos de referência de licitações de escavadeiras hidráulicas. A estrutura hierárquica desenvolvida permite agrupar características semanticamente relacionadas e uniformizar diferentes formas de representação de um mesmo conceito, contribuindo para a construção de um conjunto de dados mais consistente para as etapas posteriores de análise.

2.1.6 Aprendizado Supervisionado

O aprendizado supervisionado é um dos paradigmas mais difundidos em aprendizado de máquina. Nesse contexto, o treinamento do modelo é realizado a partir de um conjunto de dados previamente rotulados, no qual cada instância possui um conjunto

de atributos de entrada e uma saída esperada associada. O objetivo do algoritmo é aprender uma função capaz de mapear as entradas para as saídas corretas, de forma que possa generalizar esse conhecimento para novos dados não observados durante o treinamento ([Russell; Norvig, 2021](#)).

Os problemas de aprendizado supervisionado são geralmente divididos em duas categorias principais: classificação e regressão. Em tarefas de classificação, o objetivo é atribuir uma observação a uma ou mais categorias discretas, como a identificação de e-mails de spam ou a detecção de transações fraudulentas. Já em problemas de regressão, busca-se estimar valores numéricos contínuos, como preços de imóveis ou previsões de demanda.

Entre os algoritmos mais utilizados nesse paradigma destacam-se árvores de decisão, máquinas de vetores de suporte (SVM), redes neurais artificiais, regressão logística e métodos baseados em ensembles, como Random Forest e Gradient Boosting. Esses métodos têm sido amplamente empregados em diferentes domínios devido à sua capacidade de aprender padrões complexos a partir de dados históricos rotulados ([Russell; Norvig, 2021](#)).

Apesar de seu elevado desempenho em diversos cenários, o aprendizado supervisionado apresenta uma limitação importante: a necessidade de grandes volumes de dados rotulados. Em muitos problemas reais, especialmente aqueles relacionados à detecção de fraudes, a obtenção de rótulos confiáveis pode ser custosa, demorada ou até mesmo inviável, uma vez que a confirmação de irregularidades frequentemente depende de investigações conduzidas por especialistas ou órgãos de controle. Essa limitação motiva o uso de abordagens não supervisionadas e semissupervisionadas em contextos onde os dados rotulados são escassos.

2.1.7 Aprendizado Não Supervisionado

O aprendizado não supervisionado compreende técnicas capazes de identificar estruturas, padrões ou relacionamentos presentes em conjuntos de dados sem a utilização de rótulos previamente definidos. Diferentemente do aprendizado supervisionado, o algoritmo recebe apenas os atributos de entrada e busca descobrir regularidades inerentes aos dados analisados ([Russell; Norvig, 2021](#)).

Uma das aplicações mais comuns desse paradigma é a clusterização, que consiste em agrupar observações semelhantes em conjuntos denominados clusters. O objetivo é maximizar a similaridade entre elementos pertencentes ao mesmo grupo e minimizar a similaridade entre elementos de grupos distintos. Entre os algoritmos mais conhecidos estão o k-means, o DBSCAN, o HDBSCAN e os Modelos de Mistura Gaussiana (GMM) ([Jain; Murty; Flynn, 1999](#)).

Outra aplicação relevante envolve técnicas de redução de dimensionalidade, que permitem representar dados de alta dimensionalidade em espaços reduzidos, pre-

servando características relevantes para análise e visualização. Esses métodos são amplamente utilizados como etapa de pré-processamento para tarefas de agrupamento, visualização e exploração de dados.

No contexto de dados textuais, o aprendizado não supervisionado desempenha papel fundamental em tarefas de modelagem de tópicos, descoberta de temas recorrentes e agrupamento semântico. Tais técnicas permitem identificar automaticamente padrões de coocorrência e relações entre documentos, contribuindo para a organização e interpretação de grandes volumes de informação não estruturada.

A principal vantagem do aprendizado não supervisionado é sua capacidade de explorar grandes volumes de dados sem a necessidade de rotulação manual. Entretanto, a ausência de rótulos também torna a avaliação dos resultados mais desafiadora, uma vez que não existe uma referência explícita para validar os agrupamentos ou padrões encontrados. Ainda assim, essas técnicas são amplamente empregadas em mineração de dados, recuperação de informação, segmentação de usuários e detecção de anomalias (Jain; Murty; Flynn, 1999).

2.1.8 Aprendizado Semissupervisionado

O aprendizado semissupervisionado representa uma abordagem intermediária entre os paradigmas supervisionado e não supervisionado. Nesse cenário, o treinamento utiliza simultaneamente uma pequena quantidade de dados rotulados e uma quantidade significativamente maior de dados não rotulados. O objetivo é aproveitar a informação contida em ambos os conjuntos para melhorar a capacidade de generalização dos modelos (Chapelle; Schölkopf; Zien, 2009).

Essa abordagem tornou-se particularmente relevante em aplicações onde a obtenção de rótulos é cara ou limitada. Em diversos domínios, como processamento de linguagem natural, visão computacional e detecção de fraudes, a coleta de grandes volumes de dados é relativamente simples, enquanto a rotulação exige conhecimento especializado e elevado custo operacional.

Os métodos semissupervisionados baseiam-se na premissa de que os dados não rotulados contêm informações úteis sobre a estrutura subjacente do problema. Dessa forma, mesmo sem conhecer explicitamente suas classes, é possível utilizar essas observações para auxiliar na definição de fronteiras de decisão, agrupamentos ou distribuições probabilísticas mais representativas do conjunto analisado (Chapelle; Schölkopf; Zien, 2009).

Entre as estratégias mais utilizadas encontram-se a propagação de rótulos (label propagation), o treinamento auto supervisionado (self-training), métodos baseados em grafos e abordagens probabilísticas. Essas técnicas permitem incorporar conhecimento proveniente dos dados não rotulados, reduzindo a dependência exclusiva de exemplos previamente classificados.

No contexto da detecção de fraudes em licitações públicas, a aplicação de métodos semissupervisionados apresenta especial relevância. Em geral, apenas uma pequena parcela dos processos possui fraude confirmada por investigações conduzidas por órgãos de controle, enquanto a maioria das licitações permanece sem classificação formal. Consequentemente, abordagens semissupervisionadas permitem explorar simultaneamente os casos conhecidos e os registros não rotulados, aproveitando melhor a informação disponível.

Neste trabalho, adota-se uma abordagem semissupervisionada baseada em Modelos de Mistura Gaussiana, proposta por [Schmitz \(2025\)](#), na qual os agrupamentos obtidos a partir dos dados são posteriormente avaliados considerando a distribuição dos casos de fraude conhecidos.

2.1.9 Modelos de Mistura Gaussiana

Os Modelos de Mistura Gaussiana (GMM) constituem uma abordagem probabilística amplamente utilizada para modelar conjuntos de dados compostos por múltiplas distribuições. A principal premissa do método é que os dados observados podem ser representados por uma combinação ponderada de distribuições Gaussianas, sendo cada componente responsável por descrever um agrupamento ou subpopulação presente no conjunto analisado ([McLachlan; Peel, 2004](#); [Bishop, 2006](#)).

Diferentemente de métodos de agrupamento que realizam atribuições rígidas dos elementos aos grupos, os GMMs adotam uma abordagem probabilística. Dessa forma, cada observação possui uma probabilidade associada de pertencer a cada componente do modelo, permitindo representar situações em que os limites entre grupos não são claramente definidos. Essa característica torna os GMMs particularmente adequados para problemas em que existem comportamentos intermediários ou sobreposição entre agrupamentos.

Outra vantagem dos Modelos de Mistura Gaussiana está na capacidade de representar agrupamentos com diferentes formatos e dispersões. Enquanto algoritmos como o k-means assumem grupos aproximadamente esféricos e de variância semelhante, os GMMs utilizam matrizes de covariância para modelar a distribuição dos dados em cada componente, permitindo capturar estruturas mais complexas presentes no conjunto analisado ([Bishop, 2006](#)).

A função densidade de probabilidade pode ser expressa como mostrado na Equação 1.

$$f(x) = \sum_{k=1}^K \pi_k \cdot \mathcal{N}(x \mid \mu_k, \Sigma_k) \quad (1)$$

em que:

- K representa o número de componentes do modelo;

- π_k é o peso associado à componente k , de modo que $\sum_{k=1}^K \pi_k = 1$;
- $\mathcal{N}(x|\mu_k, \Sigma_k)$ corresponde à distribuição Gaussiana parametrizada pela média μ_k e pela matriz de covariância Σ_k .

Neste trabalho, os Modelos de Mistura Gaussiana são utilizados conforme a metodologia proposta por [Schmitz \(2025\)](#), com o objetivo de identificar agrupamentos e possíveis anomalias em processos licitatórios. Diferentemente do estudo original, o modelo é aplicado sobre um conjunto de dados ampliado, que inclui não apenas variáveis como preço unitário e número de participantes, mas também novas características derivadas dos termos de referência, extraídas por meio de análise semântica.

A estimativa dos parâmetros do GMM é realizada por meio do algoritmo *Expectation-Maximization* (EM), que se alterna entre dois passos:

- **E-step (Expectation)**: calcula a probabilidade de cada observação pertencer a um componente gaussiano;
- **M-step (Maximization)**: atualiza os parâmetros μ_k , Σ_k e π_k para maximizar a verossimilhança dos dados.

A escolha do número de componentes K constitui um aspecto central na aplicação dos GMMs. Dois critérios de informação são amplamente utilizados para orientar essa seleção: o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC). Ambos penalizam modelos mais complexos para evitar sobreajuste, sendo o BIC mais conservador por apresentar penalidade proporcional ao logaritmo do tamanho da amostra. O valor de K adotado é aquele que minimiza o critério escolhido. Neste trabalho, a seleção do número de componentes segue o procedimento definido por [Schmitz \(2025\)](#), que emprega o BIC como critério de referência.

2.1.10 Leave-One-Out (LOO)

O Leave-One-Out (LOO) é uma técnica de validação cruzada utilizada para avaliar a capacidade de generalização de modelos de aprendizado de máquina. O método consiste em remover temporariamente uma observação do conjunto de dados, utilizar as demais observações para construção do modelo e, posteriormente, avaliar o comportamento do método em relação à instância excluída. Esse procedimento é repetido sucessivamente até que todas as observações tenham sido utilizadas uma vez como elemento de validação ([Stone, 1974](#)).

O LOO pode ser interpretado como um caso particular da validação cruzada *k-fold*, no qual o número de partições é igual ao número total de observações ($k = n$). Como consequência, cada execução utiliza praticamente todo o conjunto de dados disponível para treinamento, o que reduz o viés da estimativa obtida e maximiza o aproveitamento das informações disponíveis. Por outro lado, essa estratégia apresenta

maior custo computacional, uma vez que o modelo precisa ser reconstruído repetidamente ao longo das iterações (Stone, 1974).

Uma das principais vantagens do Leave-One-Out está em sua aplicação em cenários com quantidade limitada de dados rotulados. Nesses casos, a utilização de praticamente todas as observações em cada etapa de treinamento permite reduzir a perda de informação causada pela separação entre conjuntos de treinamento e teste. Por esse motivo, o método é amplamente empregado em aplicações nas quais exemplos rotulados são escassos, como problemas de diagnóstico médico, detecção de anomalias e identificação de fraudes.

Além de sua utilização em tarefas supervisionadas tradicionais, o LOO também pode ser empregado para avaliar modelos de detecção de anomalias e abordagens semissupervisionadas. Nesses cenários, uma observação conhecida pode ser temporariamente removida do conjunto de referência e utilizada para verificar se o método é capaz de identificá-la novamente com base nos padrões aprendidos a partir das demais observações. É com base nesse princípio que o presente trabalho adota uma variação do LOO aplicada especificamente aos casos confirmados de fraude, descrita na Seção 4.4.

2.2 TRABALHOS CORRELATOS

A revisão dos trabalhos relacionados foi conduzida com o objetivo de identificar abordagens que empregam LLMs e métodos de aprendizado semissupervisionado aplicados à análise de documentos técnicos, à construção de taxonomias e à detecção de fraudes. O intuito foi compreender como diferentes estudos abordam a construção de taxonomias, o agrupamento de informações textuais, a similaridade semântica e a utilização de Modelos de Mistura Gaussiana em cenários com dados rotulados limitados.

Schmitz (2025) propôs uma metodologia semissupervisionada para a detecção de fraudes em licitações públicas, utilizando Modelos de Mistura Gaussiana para identificar padrões atípicos em grandes conjuntos de dados de cotações. O estudo avalia a similaridade entre licitações não rotuladas e casos confirmados de fraude, criando um indicador de risco composto, que considera: (i) a proximidade com fraudes conhecidas; (ii) a quantidade de subespaços em que uma licitação é considerada suspeita; e (iii) o desempenho do modelo na reorganização de fraudes dentro de cada subespaço. Os experimentos demonstraram que o modelo superou técnicas comparativas tradicionais, evidenciando a eficácia de abordagens probabilísticas em cenários semissupervisionados.

O estudo de Schmitz (2025) é utilizado como base metodológica, mas será re-aplicado com a inclusão de novas características extraídas dos termos de referência das licitações. O objetivo é avaliar se a adição dessas informações permite melhorar o

desempenho do modelo na detecção de fraudes.

O trabalho de Schiavon (2024) propõe uma pipeline automatizada para construção de taxonomias aplicadas à análise de termos de referência em licitações públicas. A abordagem combina técnicas de processamento de linguagem natural, agrupamento semântico e modelos de linguagem de larga escala (LLMs) para transformar descrições textuais em estruturas hierárquicas de atributos técnicos. Inicialmente, os atributos extraídos dos documentos são convertidos em embeddings semânticos e agrupados por similaridade. Em seguida, um LLM identifica hierarquias de atributos, valores e unidades de medida, permitindo a construção de uma taxonomia estruturada utilizada na análise de padrões técnicos em licitações.

Marchenko e Dvoichenkov (2024) propuseram o TaxoRankConstruct, um método iterativo baseado em ranks para construção de taxonomias com Large Language Models, priorizando as propriedades dos objetos em vez de apenas relações hierárquicas. O método demonstra como LLMs podem gerar subconceitos e organizar hierarquias complexas de forma flexível, sendo avaliado com comparações a WordNet e avaliação humana. Essa abordagem ilustra o potencial do uso de LLMs em aprendizado não-supervisionado para criação de estruturas conceituais.

Diferentemente de métodos tradicionais que utilizam clustering ou taxonomias preexistentes, como TaxoGen e CoRel, o TaxoRankConstruct adota uma abordagem iterativa rank-based que combina geração automática com refinamento humano, permitindo a construção de multitaxonomias e exploração de múltiplos critérios de classificação. Similar ao TaxoRankConstruct, este trabalho constrói taxonomias baseadas em propriedades dos objetos.

Zhang et al. (2024) propõem o TELEClass, um método de classificação hierárquica de textos com supervisão mínima, que combina Modelos de Linguagem de Larga Escala (LLMs) e enriquecimento automático de taxonomias. O sistema utiliza apenas os nomes das classes como fonte inicial de supervisão, explorando o conhecimento prévio dos LLMs e a extração de termos indicativos a partir de grandes volumes de textos não rotulados. Essa abordagem permite construir representações hierárquicas mais ricas e precisas, reduzindo a dependência de anotações manuais e custos de inferência típicos de modelos supervisionados.

Enquanto o TELEClass utiliza LLMs para enriquecer taxonomias existentes com termos-chave indicativos de cada classe, permitindo posterior classificação hierárquica de documentos, a abordagem deste trabalho emprega LLMs para gerar a estrutura hierárquica completa a partir dos agrupamentos semânticos produzidos pelo BERTopic.

Ibrahim et al. (2025) aplicaram Modelos de Mistura Gaussiana (GMM) na detecção de fraudes em transações de cartão de crédito. A abordagem utiliza distribuições gaussianas para modelar o comportamento das transações e estimar a probabilidade de cada observação pertencer a padrões legítimos ou fraudulentos. O estudo foi conduzido em uma base de dados altamente desbalanceada, cenário comum em aplicações

de detecção de fraude, onde os casos fraudulentos representam apenas uma pequena parcela do conjunto analisado. Os autores avaliaram diferentes configurações do modelo e compararam seus resultados com classificadores tradicionais, observando desempenho competitivo na identificação de transações suspeitas.

Os resultados demonstram que os GMMs são capazes de representar distribuições complexas e identificar comportamentos atípicos por meio de medidas probabilísticas, mesmo em contextos com escassez de exemplos fraudulentos. Embora o domínio de aplicação seja distinto, o estudo reforça a adequação de modelos probabilísticos para problemas de detecção de fraude. De forma semelhante ao presente trabalho, a abordagem busca identificar observações que se afastam dos padrões predominantes dos dados, utilizando informações estatísticas para estimar o grau de suspeição de cada registro analisado.

3 MATERIAIS E MÉTODOS

Este capítulo detalha a metodologia do trabalho, organizada em duas frentes principais: a seção 3.1 Entendimento e Preparação dos Dados, que aborda a contextualização do conjunto de dados utilizado; e a seção 3.2 de Modelagem, que descreve o desenvolvimento das variáveis, a arquitetura do modelo estatístico e o fluxo integrado para a detecção de risco.

3.1 ENTENDIMENTO E PREPARAÇÃO DOS DADOS

Nesta etapa, são descritos os procedimentos adotados para reunir os documentos de entrada do modelo, bem como as decisões metodológicas envolvidas na sua organização e categorização. A seção apresenta inicialmente o contexto da Operação Patrola, cujos casos confirmados de fraude fornecem o conjunto supervisionado essencial para o estudo. Em seguida, detalham-se os métodos de coleta e organização da base completa, formada pela combinação de documentos rotulados e não rotulados.

3.1.1 Operação Patrola

A Operação Patrola, iniciada pelo Ministério Público de Santa Catarina (MPSC), investigou fraudes e corrupção em licitações públicas de máquinas pesadas. As primeiras ações legais ocorreram contra autoridades locais envolvidas na manipulação de editais para desvio de recursos públicos, com valores superiores a R\$ 569.000,00 ([Ministério Público de Santa Catarina, 2019](#)). Durante a investigação, foram identificados 32 casos de fraude relacionados à aquisição de escavadeiras hidráulicas. Destes, foi possível coletar os 30 termos de referência correspondentes, que compõem parte da base de dados deste estudo.

3.1.2 Preparação da base de dados

A preparação dos dados consistiu na coleta e organização de termos de referência provenientes de editais de licitação voltados à aquisição de escavadeiras hidráulicas. Esses documentos podem conter cláusulas potencialmente restritivas em suas especificações técnicas, representando indícios de direcionamento a determinados fornecedores.

A base utilizada neste estudo é composta por 100 termos de referência, divididos conforme a Tabela 1. Dentre eles, 30 documentos foram obtidos a partir da Operação Patrola, enquanto os outros 70 foram coletados diretamente de portais de transparência de municípios de Santa Catarina e do governo federal, abrangendo processos licitatórios realizados no período de 2006 a 2021.

Essa divisão permite combinar informações supervisionadas (casos confirma-

Tabela 1 – Composição da base de dados de Termos de Referência

Origem	Quantidade	Status
Documentos associados à Operação Patrola	30	Fraudes
Documentos de portais de transparência	70	Não rotulados
Total	100	

dos de fraude) com dados não rotulados, essenciais para a aplicação de técnicas de aprendizado semissupervisionado.

É crucial observar que as licitações não categorizadas como fraudulentas são, na verdade, não rotuladas, e seu status permanece desconhecido, potencialmente ocultando casos adicionais de fraude.

3.2 MODELAGEM

O objetivo da modelagem é transformar os documentos textuais em uma representação estruturada, permitindo a aplicação de técnicas de aprendizado semissupervisionado para detecção de fraudes.

3.2.1 Pipeline de construção da taxonomia

Essa etapa da modelagem não foi desenvolvida neste trabalho, sendo baseada diretamente na metodologia proposta por [Schiavon \(2024\)](#). Os termos de referências foram inicialmente processados por uma pipeline de construção de taxonomia, que segue as etapas principais:

1. **Coleta de características (Feature Collection):** os atributos relevantes foram extraídos das descrições dos objetos da licitação, gerando listas de tuplas (`item_name`, `characteristic`).
2. **Clustering semântico (Semantic Clustering):** os atributos extraídos foram convertidos em vetores semânticos (`embeddings`) utilizando modelos como MiniLM, agrupados por similaridade via UMAP e HDBSCAN, e foram extraídas palavras-chave representativas de cada cluster.
3. **Processamento em lotes (Batch Processing):** cada cluster foi dividido em `batches` de tamanho fixo para fornecer contexto semântico consistente ao modelo de linguagem.
4. **Análise por LLM (LLM Analysis):** para cada `batch`, o LLM extraiu a hierarquia de atributos (`attribute_path`), o atributo final (`attribute`), o valor (`value`) e a unidade de medida (`unit`).

5. **Criação da estrutura de taxonomia (Taxonomy Structure Creation):** as respostas do LLM foram usadas para construir a hierarquia de atributos e vincular características aos atributos correspondentes.
6. **Persistência e indexação (Persistence & Indexing):** todos os dados foram armazenados em tabelas de banco de dados, incluindo atributos, links de características e respostas do LLM.

3.2.2 Tratamento manual e agregação por documento

Após a execução da pipeline, cada característica técnica extraída foi representada por um caminho na taxonomia gerada pela LLM. Um caminho corresponde a um ramo da hierarquia que descreve um atributo técnico, desde conceitos mais gerais até os mais específicos. Assim, a hierarquia organiza os atributos de forma estruturada. Por exemplo, atributos como "movimento_esteira" e "som_buzina" correspondem a caminhos distintos dentro dessa estrutura hierárquica.

Durante a análise dos resultados, verificou-se que alguns caminhos extraídos apresentavam inconsistências ou erros, como diferentes representações para um mesmo atributo. Esses casos foram revisados e ajustados manualmente para assegurar a qualidade dos dados. Ao final desse processo, obteve-se um total de 176 caminhos distintos, representando todas as características técnicas identificadas nos termos de referência. Em seguida, esses atributos foram consolidados por documento e convertidos em variáveis binárias, indicando a presença (1) ou ausência (0) de cada característica.

Durante a consolidação das características, foi necessário tomar decisões de projeto sobre como cada uma deveria ser representada no conjunto final de dados. Como muitas especificações apareciam com diferentes formas e níveis de granularidade, cada característica exigiu uma análise individual. Em alguns casos, optou-se por manter caminhos distintos e em outros casos decidiu-se unificá-los em uma única característica.

Essas decisões foram orientadas pelas sugestões de caminhos geradas pela LLM e foram embasadas em manuais técnicos e guias de especificação de escavadeiras, utilizados como referência para identificar quando duas expressões representavam o mesmo conceito ou quando descreviam propriedades distintas que deveriam ser preservadas separadamente. Esse processo assegurou que as características mantivessem coerência técnica e relevância para o domínio analisado. O trecho a seguir represente um exemplo de termo de referência, extraído de um documento presente na base de dados:

“Escavadeira Hidráulica sob esteiras, de fabricação nacional, nova, equipada com motor a diesel, de 04 (quatro) cilindros, turboalimentado, com

potência líquida mínima de 90 HP à 2200 RPM e que atenda aos padrões internacionais de controle de nível de emissão de poluentes (EPA TIER 3 e EU ESTÁGIO 3A), peso operacional de no máximo 14.000 Kg, com sapatas de no mínimo 700mm, caçamba com capacidade mínima de 0,60 m³ e sistema de monitoramento e gerenciamento de dados de operação e manutenção via satélite Standard do fabricante, com cabine fechada, com ar condicionado, câmara de vídeo na traseira, 05 (cinco) modos de operação selecionados manualmente pelo operador.”

O atributo “potência”, por exemplo, aparece em outros documentos da base sob formas distintas: às vezes registrado simplesmente como “potência”, outras vezes como “potência bruta” ou “potência líquida”, conceitos que, embora relacionados, possuem distinções técnicas relevantes. De forma semelhante, a exigência de conformidade com normas de emissão de poluentes apresenta variações consideráveis entre os documentos. Em alguns casos, o modelo extrai o nome específico da norma, como “EPA TIER 3” ou “EU ESTÁGIO 3A”; em outros, o atributo extraído é apenas “normas de emissão” ou “padrão de emissão de poluentes”, sem qualquer referência ao nome técnico. Sem um mecanismo de padronização, esses casos seriam tratados como atributos distintos, quando na prática descrevem o mesmo requisito com diferentes níveis de especificidade.

A opção pela binarização mostrou-se adequada diante da grande diversidade e granularidade dos atributos. Ela simplifica a análise e, no contexto da detecção de direcionamento, o aspecto mais relevante é a existência de determinadas especificações técnicas, e não seus valores numéricos. Por exemplo, a mera inclusão do atributo “peso operacional mínimo” já pode indicar um requisito potencialmente direcionador, independentemente do valor especificado.

Essa escolha garantiu padronização entre os documentos e permitiu construir um dataset, no qual cada linha representa um documento e cada coluna corresponde a uma das características extraídas.

3.2.3 Construção da hierarquia de atributos

Como contribuição adicional à pipeline proposta por [Schiavon \(2024\)](#), este trabalho desenvolveu uma hierarquia manual específica para o domínio de escavadeiras hidráulicas. Durante a análise dos caminhos extraídos pela LLM, observou-se que alguns atributos semanticamente equivalentes eram representados por diferentes caminhos hierárquicos, conforme discutido na Seção 3.2.2. Essas variações dificultavam a padronização e a comparação entre documentos, tornando necessária a definição de uma estrutura hierárquica capaz de organizar atributos relacionados.

Para reduzir essas inconsistências, foi construída uma estrutura hierárquica de atributos baseada em literatura técnica, manuais de fabricantes e conhecimento do

domínio. Essa estrutura organiza características relacionadas em categorias comuns, permitindo uniformizar diferentes representações de um mesmo conceito. As relações hierárquicas foram definidas considerando a organização semântica dos atributos extraídos pela LLM.

A Figura 1 apresenta um exemplo simplificado da hierarquia desenvolvida para representar atributos de escavadeiras hidráulicas.



Figura 1 – Exemplo simplificado da hierarquia de atributos.

A adoção dessa hierarquia permitiu reduzir a fragmentação dos atributos extraídos e aumentar a consistência das características utilizadas nas etapas posteriores da análise. Além disso, a estrutura construída poderá ser incorporada à pipeline proposta por [Schiavon \(2024\)](#), servindo como conhecimento prévio para orientar a geração automática de caminhos hierárquicos pela LLM. Dessa forma, as etapas de padronização e consolidação realizadas manualmente neste trabalho podem ser incorporadas ao processo automatizado, reduzindo inconsistências na geração dos caminhos e aumentando a escalabilidade da abordagem.

3.2.4 Modelo semissupervisionado baseado em subespaços

O modelo de detecção de risco utilizado neste trabalho, denominada Sub-FraudGMM, é uma abordagem semissupervisionada fundamentada em Modelos de Mistura Gaussiana (GMM) (Schmitz, 2025). A premissa central do método é que, em espaços de alta dimensionalidade, fraudes conhecidas tendem a apresentar similaridade consistente apenas em determinados subconjuntos de características, e não necessariamente em todas as dimensões simultaneamente.

Dessa forma, em vez de modelar o espaço completo de variáveis, o método explora múltiplos subespaços de características. Cada subespaço é avaliado individualmente por meio da aplicação de GMM com diferentes combinações de hiperparâmetros, variando-se o número de componentes e o tipo de matriz de covariância (esférica, amarrada, diagonal ou completa). A seleção do modelo ótimo em cada subespaço é realizada com base no Critério de Informação Bayesiano (BIC), privilegiando configurações que conciliem bom ajuste e menor complexidade.

Os subespaços são mantidos apenas quando conseguem concentrar uma proporção mínima de fraudes conhecidas em um mesmo agrupamento, indicando comportamento consistente. Adotou-se um limiar de 80%, de modo a garantir que apenas subespaços com forte capacidade de agrupamento fossem considerados. Esse limiar foi definido buscando um equilíbrio entre a retenção de informações relevantes e a exclusão de padrões pouco específicos. Valores menores teriam maior risco de incorporar padrões menos específicos, enquanto valores maiores poderiam tornar o modelo mais restritivo, descartando informações potencialmente relevantes. A escolha pelo GMM decorre de sua flexibilidade para modelar agrupamentos com diferentes formas e tamanhos.

A partir dos subespaços considerados relevantes, define-se ainda um indicador contínuo de risco, utilizado para ranqueamento das instâncias não rotuladas. Em vez de produzir uma classificação binária (fraude ou não fraude), adota-se uma métrica contínua de risco, capaz de refletir diferentes níveis de indícios de irregularidades. Sua formulação matemática e os procedimentos de validação são descritos posteriormente na seção de avaliação.

Além disso, para cada licitação, são identificadas as três características mais relevantes que mais contribuíram para seu agrupamento com casos de fraude. Essas características são obtidas a partir da frequência com que cada variável aparece nos subespaços considerados relevantes. Isso permite interpretar o resultado do modelo, indicando quais fatores, como número de participantes ou preço unitário, foram mais determinantes para aproximar aquela licitação de comportamentos suspeitos.

3.2.4.1 Características Utilizadas e Extensões Propostas

O conjunto original de características empregado no SubFraudGMM foi construído a partir de indicadores estruturais relacionados ao comportamento das empresas e das Unidades Gestoras ao longo dos processos licitatórios, abrangendo o período de 2006 a 2021. As principais variáveis utilizadas são descritas a seguir:

Tabela 2 – Características utilizadas no SubFraudGMM

Variável	Descrição
Win	Proporção de lances vencedores de uma empresa dentro da mesma Unidade Gestora. Frequências elevadas de vitória concentradas em um único órgão podem indicar possível favorecimento ou assimetria competitiva.
Met	Variável binária que indica se a empresa teve reuniões anteriores com a Unidade Gestora em licitações passadas. Funciona como indicador de interação prévia entre as partes.
Num	Número total de licitações das quais a empresa participou. Empresas com baixa participação podem indicar atuação pontual ou limitada experiência operacional.
Period	Quantidade total de dias em que a empresa esteve ativa no conjunto de dados, auxiliando na distinção entre empresas consolidadas e empresas com presença temporária.
Duration	Intervalo de tempo entre a abertura e a homologação da licitação. Durações excepcionalmente curtas podem indicar restrição à concorrência ou processos acelerados.
Unique	Proporção de vencedores únicos em licitações de uma mesma Unidade Gestora. Baixa diversidade de vencedores pode sinalizar concentração recorrente de contratos.
Num of Participants	Número total de participantes em cada licitação, representando o nível de concorrência do processo.
Unit Price	Preço unitário do item adquirido. Desvios significativos em relação ao padrão observado podem indicar sobrepreço ou possível coordenação entre licitantes.

Essas variáveis capturam aspectos estruturais de concorrência, relacionamento institucional, histórico de participação e padrões de precificação, compondo a base do modelo probabilístico original.

Neste trabalho são incorporadas novas características derivadas da análise estrutural dos documentos licitatórios, organizadas segundo uma taxonomia hierárquica. Diferentemente das variáveis originais, que refletem predominantemente padrões

comportamentais, as novas características capturam padrões presentes no termo de referência dos documentos.

A introdução dessas novas variáveis permite avaliar se a incorporação de características hierárquicas contribui para melhorar a separação dos agrupamentos e a priorização de instâncias suspeitas, ampliando a capacidade discriminativa do modelo semissupervisionado.

3.2.5 Novas características

A construção das variáveis teve como objetivo transformar a representação binária dos documentos em um conjunto reduzido de características. As características selecionadas para o modelo final estão apresentadas na Tabela 4. De forma geral, essas variáveis concentram-se em três aspectos principais: similaridade estrutural entre documentos, proximidade em relação a padrões previamente identificados como fraude e comportamento coletivo dos atributos no conjunto de dados.

Tabela 4 – Características selecionadas para o modelo de agrupamento

Característica	Descrição	Justificativa
mean_similarity	Similaridade média do documento em relação aos demais, calculada a partir da distância de Hamming.	Captura padrões de repetição estrutural entre documentos, sendo indicativa da possível utilização de templates.
fraud_similarity	Similaridade do documento em relação ao perfil médio dos documentos fraudulentos.	Permite identificar documentos não rotulados com características semelhantes às fraudes conhecidas.
score_fraud_like	Soma ponderada dos atributos com base na frequência observada no conjunto de fraudes.	Mede o grau de aderência do documento aos padrões típicos de fraude.
mean_attr_freq	Frequência média dos atributos presentes em cada documento no conjunto de dados.	Diferencia documentos com atributos comuns (padronizados) daqueles com atributos mais específicos.

Embora diversas variáveis tenham sido inicialmente construídas, optou-se por selecionar apenas aquelas que apresentaram maior relevância e menor redundância no contexto dos dados analisados. A seleção reduzida também foi motivada por limitações inerentes do SubFraudGMM. Como o modelo avalia diferentes combinações de variáveis para construção dos agrupamentos, o aumento no número de variáveis provoca crescimento exponencial do espaço de combinações possíveis. Esse comportamento

eleva significativamente o custo computacional.

Entre as variáveis inicialmente consideradas, estavam aquelas relacionadas à quantidade de atributos presentes em cada documento, isto é, ao número de requisitos técnicos especificados. Também foram avaliadas medidas de densidade, que indicam quão completo ou detalhado é um documento em relação ao conjunto total de possíveis atributos. No entanto, essas informações não foram utilizadas no modelo final, pois o nível de detalhamento das especificações não apresentou uma relação consistente com a ocorrência de fraude. Em muitos casos, documentos mais completos refletem apenas maior cuidado na elaboração, e não necessariamente um comportamento suspeito.

Optou-se por utilizar uma medida baseada na frequência média dos atributos presentes em cada documento, que captura, de forma mais estável, o grau de especificidade das características utilizadas. Em termos práticos, essa variável representa o quão comuns são, em média, os atributos presentes em um documento, permitindo verificar se ele utiliza características frequentemente observadas no conjunto de dados ou se apresenta atributos mais específicos e menos recorrentes.

Além disso, foi considerada uma variável que mede o quão parecido um documento é em relação aos demais documentos do conjunto. Essa medida avalia se diferentes documentos apresentam estruturas semelhantes, isto é, se tendem a incluir e omitir os mesmos atributos. Esse tipo de padrão pode indicar a reutilização de modelos ou templates.

Outra variável mede a similaridade de cada documento em relação a um perfil médio de fraude. Essa variável indica se um documento apresenta uma estrutura semelhante àquela observada nas fraudes conhecidas. E a última segue uma lógica complementar, avaliando o quanto os atributos presentes em um documento são característicos do conjunto de fraudes. Dessa forma, documentos que utilizam com maior frequência atributos comuns em fraudes tendem a obter valores mais elevados nessa métrica, refletindo maior aderência aos padrões já observados.

3.2.6 Visão geral da pipeline de processamento

A Figura 2 apresenta uma visão de toda a pipeline desenvolvida neste trabalho, desde a obtenção dos documentos até a geração do indicador de risco. A estrutura foi organizada em duas macroetapas, correspondentes à criação da taxonomia (Etapa 1) e à modelagem semissupervisionada (Etapa 2).

Essas macroetapas correspondem a componentes previamente desenvolvidos de forma independente: a Etapa 1 corresponde à pipeline de construção de taxonomia proposta por [Schiavon \(2024\)](#), enquanto a Etapa 2 se baseia no modelo SubFraudGMM aplicado para identificação de padrões suspeitos de acordo com a metodologia proposta por [Schmitz \(2025\)](#).

O presente estudo se posiciona entre esses dois blocos, realizando a integração completa entre eles e estabelecendo a ponte necessária para que a representação semântica dos documentos possa alimentar, de maneira estruturada e padronizada, o modelo semissupervisionado de risco.

Originalmente, o SubFraudGMM utilizava como entrada um conjunto de oito características previamente definidas para representar os documentos. Neste trabalho, as novas características, extraídas a partir dos caminhos hierárquicos identificados na etapa de construção da taxonomia são integradas às características originais do modelo. A etapa de criação do novo dataset consiste na consolidação dessas informações, resultando em um conjunto de dados enriquecido, que é utilizado como entrada para a aplicação do modelo semissupervisionado.

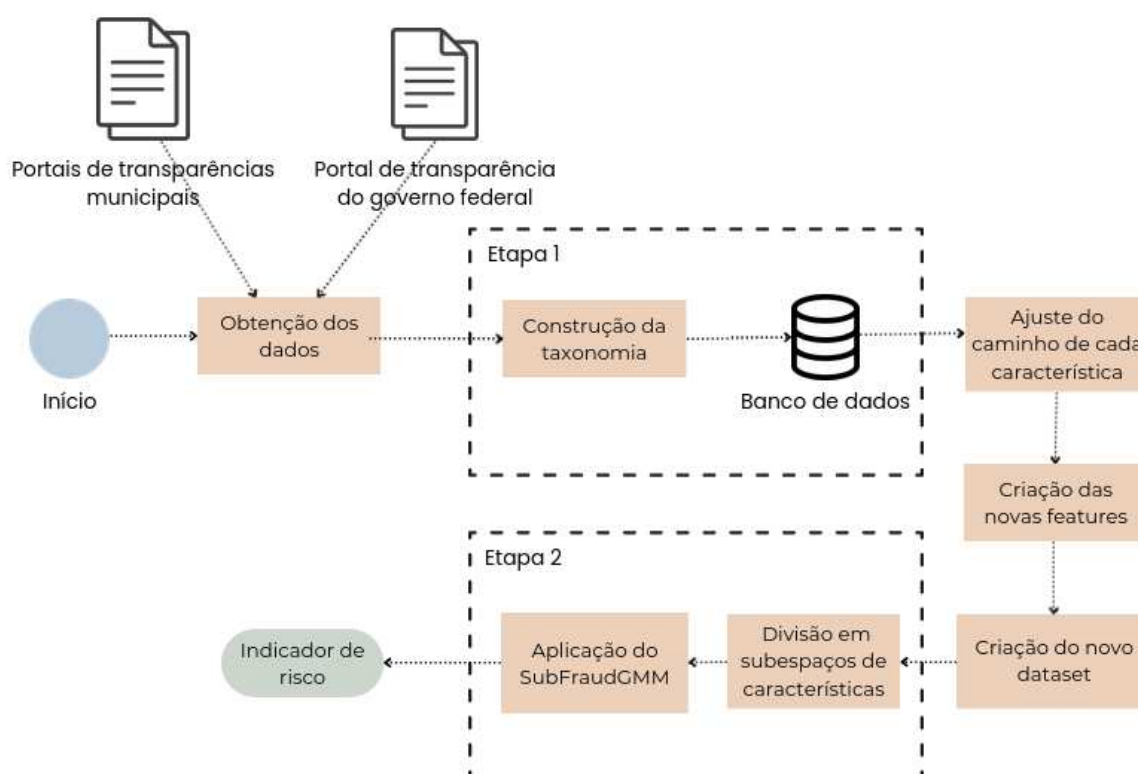


Figura 2 – Fluxograma geral do TCC.

Fonte: Elaborado pela autora (2025).

4 AVALIAÇÃO

A avaliação está organizada em sete seções: a Seção 4.1 descreve a abordagem metodológica adotada para a análise dos dados; a Seção 4.2 apresenta os resultados quantitativos da análise de frequência e distribuição dos atributos; a Seção 4.3 apresenta as métricas utilizadas; a Seção 4.4 descreve o procedimento de validação; a Seção 4.5 apresenta os resultados do indicador de risco; a Seção 4.6 aborda a validação do indicador de risco; a Seção 4.7 avalia o modelo para os dados do painel e a Seção 4.8 discute os principais resultados e as limitações.

4.1 ABORDAGEM METODOLÓGICA DA AVALIAÇÃO

O processo de avaliação contemplou tanto a construção do novo dataset quanto a aplicação dos modelos de detecção de fraude. Inicialmente, foi realizada uma análise exploratório de frequência dos caminhos dos atributos identificados em cada documento. Cada caminho representa uma sequência hierárquica extraída das especificações técnicas, sendo tratado como um possível indicativo de direcionamento.

Para a construção do dataset, além da representação binária dos caminhos extraídos dos termos de referência, foram incorporadas ao conjunto de dados duas colunas adicionais: uma indicando a classificação de cada documento como fraude ou não fraude, conforme mapeamento previamente estabelecido, e outra contendo os identificadores dos itens de licitação na base de dados da Esfinge.

Posteriormente, foram adicionadas quatro novas colunas derivadas, com o objetivo de capturar informações dos termos de referências dos documentos. O dataset final contém 100 linhas (documentos) e 182 colunas (176 caminhos) no total. Essa combinação permitiu analisar a distribuição geral dos caminhos e comparar como esses mesmos atributos se manifestam entre documentos suspeitos e não suspeitos. Com o dataset estruturado, a avaliação concentrou-se na análise do Indicador de Risco.

4.2 ANÁLISE DOS DADOS

A análise mostrou que os documentos apresentam uma estrutura hierárquica bem definida, na qual cada caminho representa os atributos técnicos específicos dos termos de referência. Alguns caminhos são recorrentes em grande parte dos documentos, indicando elementos comuns e amplamente padronizados nos editais, enquanto outros aparecem apenas em poucos casos, o que indica variações e especificidades nos requisitos.

A frequência dos caminhos permitiu identificar os padrões mais representativos da base. Esse comportamento é ilustrado na Figura 3, que apresenta os 20 caminhos mais frequentes no conjunto analisado, segmentados entre documentos rotulados co-

mo fraude e não rotulados. A comparação mostra que certos caminhos predominam entre documentos não rotulados, sugerindo maior padronização em editais considerados regulares, enquanto outros aparecem de forma proporcionalmente mais intensa nos documentos rotulados como fraude, como por exemplo "monitoramento por satélite" e "origem nacional", o que possivelmente pode indicar um direcionamento.

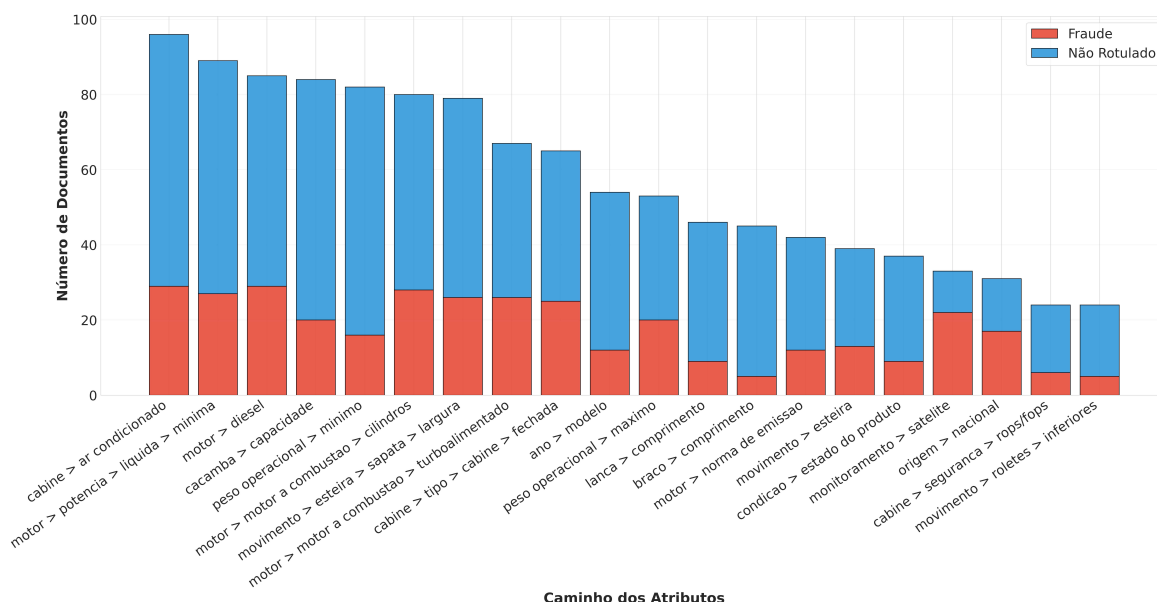


Figura 3 – Top 20 caminhos por número de documentos.

Além disso, com a finalidade de mostrar uma visão geral de todo o dataset, a Figura 4 apresenta as frequências ordenadas dos 176 caminhos identificados na base de dados, sem a inclusão de rótulos individuais.

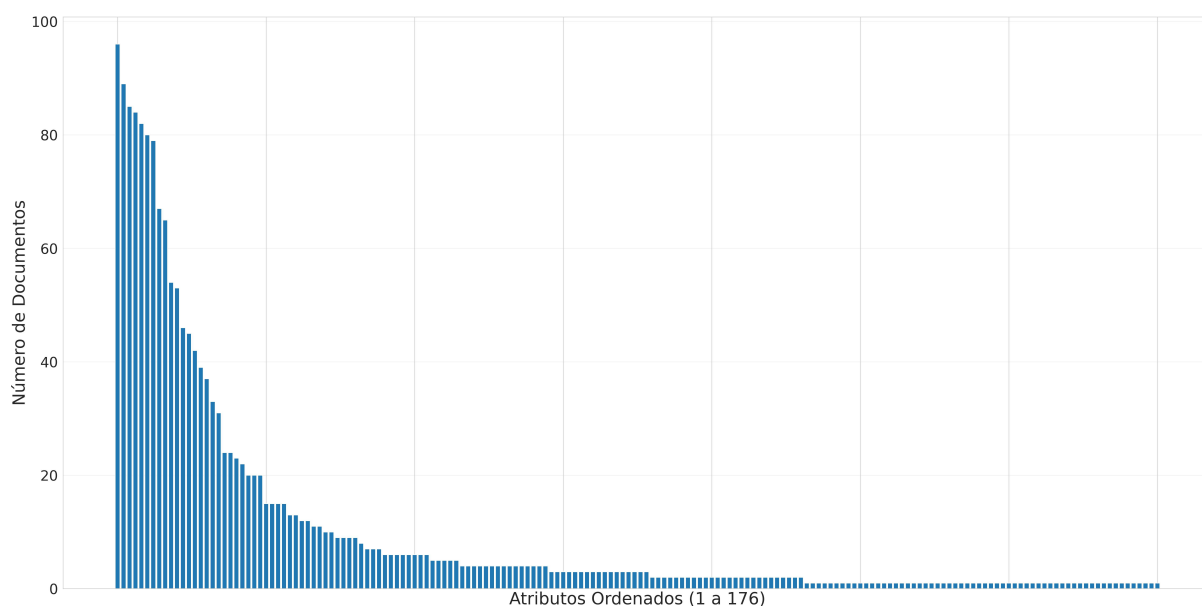


Figura 4 – Visão Geral das Frequências dos 176 Atributos.

Os caminhos foram avaliados separadamente entre documentos rotulados como fraude e documentos não rotulados. Para isso, calculou-se a frequência média de cada caminho em ambos os grupos, permitindo identificar três categorias relevantes de recorrência estrutural. Caminhos cuja presença média ultrapassa 40% dos documentos foram classificados como frequentes; aqueles com frequência igual ou superior a 5% foram considerados intermediários; e os demais, abaixo desse limiar, foram categorizados como raros.

A Tabela 5 apresenta, de forma ilustrativa, os cinco caminhos mais representativos de cada categoria de frequência em ambos os conjuntos. Embora não represente a totalidade dos caminhos em cada categoria, essa amostra permite visualizar exemplos concretos dos padrões estruturais identificados.

Tabela 5 – Top 5 caminhos por categoria de frequência.

Categoria	Conjunto	Rank	Caminho	Frequência
Frequentes	Fraude	1	Ar condicionado	0.966667
		2	Motor diesel	0.966667
		3	Número de cilindros	0.933333
		4	Potência líquida mínima	0.900000
		5	Turboalimentado	0.866667
	Não rotulados	1	Ar condicionado	0.957143
		2	Peso operacional mínimo	0.942857
		3	Capacidade da caçamba	0.914286
		4	Potência líquida mínima	0.885714
		5	Motor diesel	0.800000
Intermediários	Fraude	1	Força escavação caçamba	0.366667
		2	Modos de operação	0.333333
		3	Comprimento da lança	0.300000
		4	Condição do estado	0.300000
		5	Escavação alcance	0.300000
	Não rotulados	1	Movimento esteiras	0.371429
		2	Áudio rádio	0.285714
		3	Rolos inferiores	0.271429
		4	Rolos superiores	0.257143
		5	ROPS/FOPS	0.257143
Raros	Fraude	1	Vidros temperados	0.033333
		2	Vazão de bomba	0.033333
		3	Joystick	0.033333
		4	Painel de monitoramento	0.033333
		5	Aquecedor da cabine	0.033333
	Não rotulados	1	Extintor de incêndio	0.028571
		2	Buzina	0.028571
		3	Jogo de chaves	0.028571
		4	Porta copos	0.028571
		5	Altura da cabine	0.028571

Caminhos nos documentos

- **Caminhos muito frequentes:** tendem a compor a estrutura padrão dos termos de referência, indicando elementos comuns à maioria dos editais da base. Exemplos típicos incluem caminhos como "ar condicionado", "motor diesel" ou "potência líquida mínima", que aparecem de forma recorrente em praticamente todos os documentos.
- **Caminhos intermediários:** correspondem a atributos presentes em uma parcela moderada dos documentos. Essa categoria apresentou maior quantidade de caminhos entre os documentos não rotulados, sugerindo uma maior diversidade de combinações de especificações técnicas nesse grupo. Exemplos incluem atributos como "ROPS/FOPS" e "movimento esteiras".
- **Caminhos raros:** desempenham um papel importante na distinção entre ruído e especificação. Esses caminhos podem corresponder a descrições muito particulares, como itens incomuns ou serviços pouco frequentes, ou alternativamente, podem indicar ruído gerado por variações não estruturadas do texto ou erros de extração que não seguem o padrão típico dos documentos.

Essa análise exploratória permitiu compreender como os caminhos se distribuem estruturalmente entre os dois grupos. A Figura 5 apresenta a sobreposição desses caminhos e evidencia a existência de 107 atributos cuja frequência é zero em todos os documentos rotulados como fraude, mas que aparecem apenas nos não rotulados.

Entre os exemplos desses atributos exclusivos estão especificações muito particulares, como acessórios jogo de chaves; requisitos específicos de cabine, como cabine ventilação positiva com filtragem; e itens administrativos, como origem código FINAME. Esses caminhos foram mantidos na base de dados, pois sua presença pode fornecer informações discriminativas relevantes, contribuindo para diferenciar documentos não rotulados dos padrões observados nos casos de fraude conhecidos.

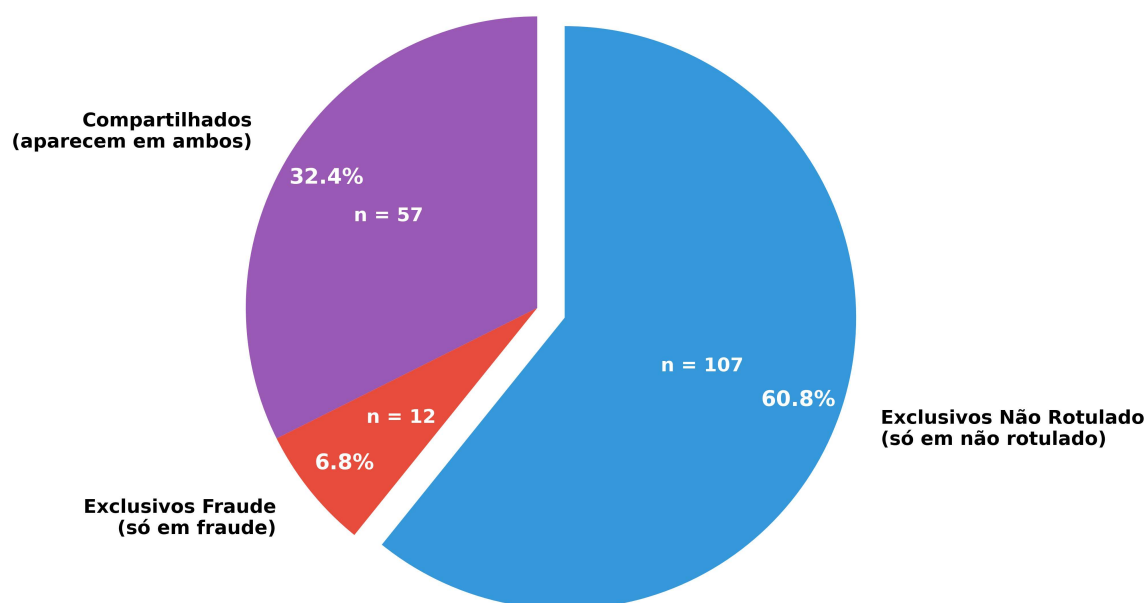


Figura 5 – Análise de sobreposição de caminhos entre documentos de fraude e não rotulados

Para ilustrar concretamente os padrões identificados, serão apresentados dois casos representativos que exemplificam diferentes perfis de especificação técnica presentes no dataset.

1. Caso 1: Especificação Genérica (Documento Não Rotulado)

- **Documento ID:** cunhatai_2016_29
- **Total de atributos:** 8 caminhos
- **Análise:** O documento apresenta um conjunto de atributos predominantemente frequentes entre os casos não rotulados, incluindo potência líquida mínima do motor, peso operacional mínimo, cabine fechada, ar condicionado, largura da sapata e capacidade da caçamba. As especificações seguem padrões técnicos amplamente encontrados em processos regulares. Apenas o caminho relacionado à origem nacional apresenta frequência mais baixa entre os documentos não rotulados, sugerindo um pequeno desvio do padrão dominante.

2. Caso 2: Documento Suspeito (Fraude)

- **Documento ID:** guatambu_2012_500
- **Total de atributos:** 18 caminhos
- **Análise:** O documento apresenta forte presença de caminhos frequentes associados a atributos estruturais comuns, como potência líquida mínima,

cabine fechada, ar condicionado e peso operacional máximo. Há também presença de atributos intermediários que ocorrem com maior frequência em documentos de fraude do que em documentos não rotulados, como força de escavação da caçamba, modos de operação, rotação do motor e monitoramento por satélite. Alguns caminhos apresentam frequências altas em documentos fraudulentos e que são menos comuns no grupo não fraude, sugerindo um perfil característico e restritivo.

4.3 MÉTRICAS

Com o objetivo de avaliar o desempenho do SubFraudGMM, as métricas descritas a seguir foram utilizadas.

1. Indicador de Risco do SubFraudGMM

A metodologia identifica agrupamentos que concentram fraudes conhecidas e utiliza essa estrutura para avaliar a similaridade de instâncias não rotuladas.

O risco atribuído a cada instância não rotulada baseia-se na hipótese central do modelo: quanto maior a recorrência de similaridade com fraudes conhecidas em múltiplos subespaços relevantes, maior deve ser o grau de suspeição. Para capturar essa ideia, o indicador combina três dimensões complementares:

- **Proximidade estrutural:** complemento da distância euclidiana média em relação às fraudes conhecidas nos subespaços selecionados. Quanto menor a distância, maior a similaridade.
- **Consistência intersubespaços:** número de subespaços relevantes nos quais a instância foi associada a clusters que concentram fraudes.
- **Capacidade discriminativa:** recall relativo médio obtido na validação leave-one-out (LOFO), refletindo a habilidade do GMM em reagrupar fraudes conhecidas.

O indicador final de risco é definido a partir da combinação dessas três componentes, considerando suas contribuições de forma equilibrada:

$$Risco = \frac{(1 - Dist_{euclid}) + NumClusters + RecallRel}{3} \quad (2)$$

Para assegurar comparabilidade entre as variáveis e evitar que métricas com escalas distintas dominem o cálculo do indicador, todas as componentes são previamente normalizadas no intervalo $[0, 1]$ por meio da transformação Min-Max:

$$x'_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (3)$$

Essa padronização preserva a estrutura relativa dos dados, garantindo que o indicador final permaneça limitado ao intervalo $[0, 1]$, no qual valores mais próximos de 1 indicam maior grau de suspeição.

O indicador resultante não representa uma probabilidade direta de fraude, mas sim um índice relativo de priorização, permitindo ordenar as instâncias não rotuladas conforme sua proximidade estrutural com padrões fraudulentos previamente identificados.

2. Ranking Simples (Baseline)

Método heurístico baseado na combinação de variáveis do conjunto de dados interpretadas como sinais de risco, por meio de uma média simples, na qual todas contribuem com o mesmo peso para o cálculo final, conforme definido na Equação 4. Na versão estendida do modelo, as novas variáveis propostas também foram incorporadas ao cálculo do risco. Variáveis cuja interpretação original está associada a menor risco para valores elevados são invertidas, de modo a manter a coerência na direção do score. O Ranking Simples não incorpora mecanismos de agrupamento, sendo utilizado apenas como referência.

$$\text{Risco} = \frac{\text{win} + \text{met} + \text{unit_price} + (1 - \text{unique}) + (1 - \text{period}) + (1 - \text{num_partic}) + (1 - \text{num}) + (1 - \text{duration})}{8} \quad (4)$$

4.4 VALIDAÇÃO

A avaliação dos modelos foi conduzida por meio do protocolo *Leave-One-Fraud-Out* (LOFO). O conjunto de dados utilizado contém 30 licitações previamente identificadas como fraudulentas. No protocolo LOFO, realiza-se uma iteração para cada fraude disponível. Em cada iteração, uma das 30 fraudes é temporariamente removida do conjunto de instâncias rotuladas como fraude, sendo tratada como não rotulada durante o processo de treinamento. As demais 29 fraudes permanecem disponíveis para a construção do modelo.

Esse procedimento resulta na geração de 30 conjuntos intermediários de dados, cada um correspondente a uma execução independente do modelo. Em cada execução:

- 29 fraudes são utilizadas como referência para aprendizado estrutural;

- 1 fraude é retirada do conjunto supervisionado e inserida no grupo de instâncias não rotuladas.

O subconjunto formado pelas fraudes temporariamente removidas é denominado grupo RT (Registro de Teste). Esse grupo contém, em cada iteração, a instância fraudulenta que o modelo nunca observou como fraude durante o treinamento. Após o ajuste do modelo em cada iteração, calcula-se o Indicador de Risco para todas as instâncias não rotuladas, incluindo aquela fraude removida. O objetivo é verificar se o modelo atribui a essa instância um valor elevado de risco, mesmo sem tê-la utilizado como referência explícita.

Esse procedimento permite avaliar a capacidade de generalização do modelo em cenários realistas, nos quais novas fraudes, ainda não previamente catalogadas, precisam ser identificadas. Ao final das 30 execuções, obtém-se um conjunto de 30 valores de risco para o grupo RT, cada um correspondente à fraude retirada em determinada iteração. Esses valores são então utilizados para análise estatística comparativa entre os modelos.

4.5 ANÁLISE COMPARATIVA DO INDICADOR DE RISCO

A avaliação do modelo foi realizada comparando-se o desempenho da versão original com a versão enriquecida por novas variáveis, considerando o ranking de risco gerado. No modelo original, observou-se que os registros com maior indicador de risco concentraram-se em casos com alta recorrência em subespaços, baixa distância em relação a fraudes conhecidas e consistência moderada do agrupamento.

A Tabela 6 apresenta os cinco casos não rotulados com maior indicador de risco. Observa-se que ESC38 e ESC94 combinam baixa distância e alta recorrência em subespaços relevantes, configurando os casos mais críticos. O registro ESC93 mantém alta recorrência, embora com maior distância, indicando um perfil ainda fortemente suspeito. Já ESC3 e ESC49 apresentam valores mais moderados e um conjunto distinto de características (*met*, *unique*, *period*), sugerindo comportamentos menos centrais, porém ainda relevantes para investigação.

Tabela 6 – Top 5 licitações não rotuladas com maior indicador de risco - modelo original

ID	Indicador Risco	Distância	Subsets	LOO	Principais Features
ESC38	0.8085	0.251	0.93	0.678	period, met, num
ESC94	0.8041	0.245	0.91	0.676	period, met, num
ESC93	0.7813	0.332	0.93	0.678	period, met, num
ESC3	0.7325	0.282	0.72	0.695	met, unique, period
ESC49	0.7265	0.302	0.72	0.695	met, unique, period

Com a introdução das novas características, verificou-se que a ordem dos principais registros no ranking permaneceu inalterada, embora os valores absolutos do

indicador de risco tenham apresentado leve aumento. Esse comportamento sugere que o modelo original já capturava de forma adequada os principais padrões de anomalia presentes nos dados, enquanto as novas variáveis contribuíram principalmente para refinar a separação entre os grupos, sem alterar significativamente a estrutura global do ranking.

Adicionalmente, observou-se uma mudança nas variáveis mais relevantes que explicam os agrupamentos. No modelo original, predominavam características estruturais como *period*, *met* e *num*, enquanto na versão atual passaram a aparecer com maior frequência variáveis como *fraud_similarity* e *mean_attr_freq*. Isso indica que o modelo passou a incorporar não apenas padrões estruturais, mas também informações provenientes dos termos de referências.

A Tabela 7 apresenta os principais registros não rotulados com maior indicador de risco após a inclusão das novas variáveis, juntamente com suas respectivas métricas e principais características associadas.

Tabela 7 – Top 5 licitações não rotuladas com maior indicador de risco - novo modelo

ID	Indicador Risco	Distância	Subsets	LOO	Principais Features
ESC38	0.8418	0.2512	0.9317	0.7631	num, period, fraud_similarity
ESC94	0.8410	0.2542	0.9313	0.7625	num, period, fraud_similarity
ESC93	0.7638	0.4232	0.8638	0.7716	num, period, fraud_similarity
ESC3	0.7415	0.2856	0.6860	0.7513	mean_attr_freq, unit_price, score_fraude_like
ESC49	0.7335	0.2969	0.6722	0.7513	mean_attr_freq, fraud_similarity, unit_price

4.6 VALIDAÇÃO DO INDICADOR DE RISCO

Esta seção apresenta a comparação entre os resultados obtidos no trabalho de [Schmitz \(2025\)](#) e aqueles verificados no dataset atual, considerando a mediana do Indicador de Risco atribuída às licitações fraudulentas no grupo RT.

A principal diferença entre os experimentos não está na metodologia ou nos modelos utilizados, mas sim na composição e no tamanho do conjunto de dados. No trabalho original, havia um número maior de licitações do tipo escavadeira, o que proporcionava maior diversidade de padrões estruturais e maior estabilidade na formação dos subespaços. Já no presente trabalho, o conjunto reduzido de aproximadamente 100 instâncias resulta em menor variabilidade.

A Tabela 8 apresenta a comparação direta entre a mediana geral do trabalho ori-

ginal (SubFraudGMM) e a mediana obtida no dataset atual de escavadeiras hidráulicas (SubFraudGMM-Base).

Tabela 8 – Comparação da mediana do Indicador de Risco (Grupo RT)

Modelo	SubFraudGMM	SubFraudGMM-Base
SubFraudGMM	82.6%	77.2%
Ranking Simples	65.0%	61.0%

Observa-se que, tanto no trabalho original quanto no dataset atual, o SubFraudGMM apresentou mediana superior ao Ranking Simples, indicando maior capacidade de priorização das licitações fraudulentas no grupo RT. Apesar da redução observada no dataset atual (77.2% em comparação a 82.6%), o modelo manteve desempenho consistente mesmo em um cenário com menor quantidade de dados e menor variabilidade estrutural entre os documentos analisados.

Em relação às novas características adicionadas na versão estendida do SubFraudGMM-Base, observa-se que a mediana do grupo RT apresentou um leve aumento no Indicador de Risco médio, passando de 77.2% para 78.8%, conforme apresentado na Tabela 9, representando um ganho aproximado de 2%. Esse incremento indica que a inclusão das novas variáveis contribuiu para elevar o indicador de risco das fraudes removidas e avaliadas como casos não rotulados.

Tabela 9 – Impacto das novas características no Indicador de Risco médio do grupo RT

Versão do Modelo	Mediana RT	Diferença
SubFraudGMM	77.2%	—
SubFraudGMM-Estendido	78.8%	+1.6%

A Figuras 6 apresenta a comparação entre o SubFraudGMM-Base, a versão estendida com as novas características e o ranking simples base e o ranking simples com as novas características. Observa-se que o SubFraudGMM-Estendido tende a apresentar distribuição deslocada para valores mais altos do Indicador de Risco no grupo RT. Esse comportamento reforça que as novas características não alteram drasticamente o comportamento global do modelo, mas atuam no sentido distinguir melhor fraudes e casos não rotulados. Já os modelos baseados em Ranking Simples exibem distribuições bastante semelhantes entre RT e Demais, mesmo após a inclusão das novas características, indicando menor capacidade discriminativa

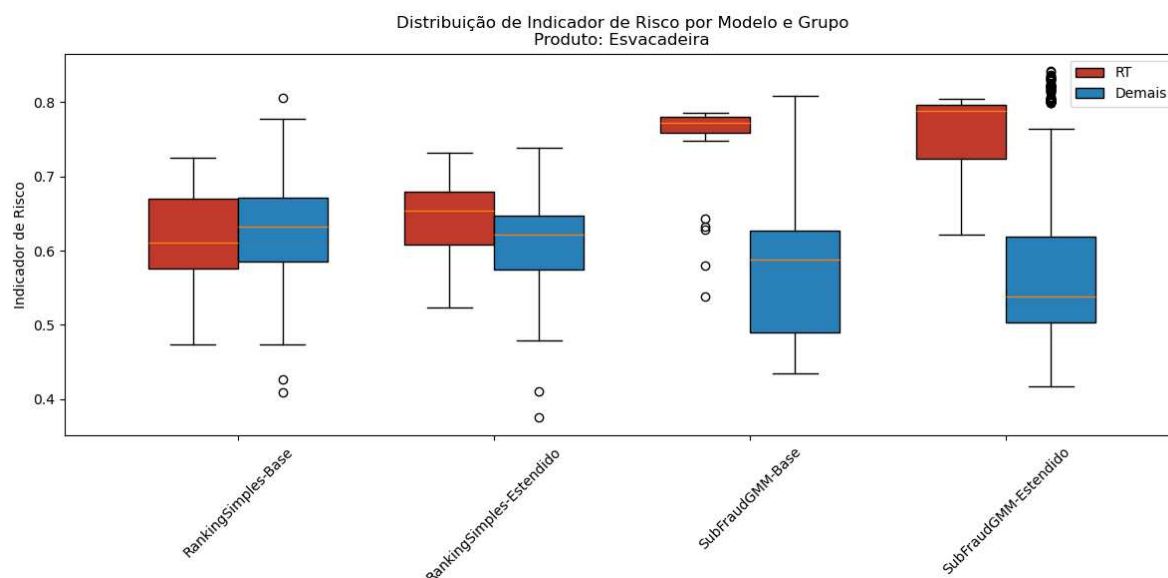


Figura 6 – Indicador de Risco para os modelos construídos.

Complementarmente, para avaliar a efetividade do modelo sob a perspectiva de ranqueamento, foi analisada a Curva de Recuperação de Fraudes em função do Top-N registros mais bem classificados pelo Indicador de Risco, conforme ilustrado na Figura 7. Essa análise permite verificar a proporção acumulada de fraudes conhecidas que são recuperadas à medida que se ampliam as posições consideradas no ranking.

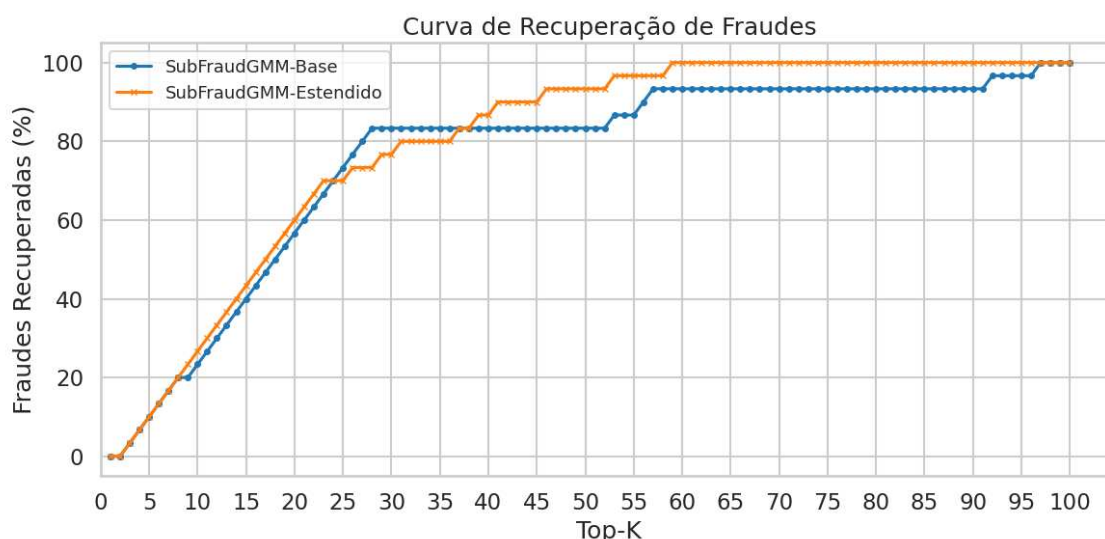


Figura 7 – Curva de Recuperação de Fraudes em função do Top-N registros ranqueados pelo Indicador de Risco

A análise da curva de recuperação de fraudes evidencia um comportamento inicial muito semelhante entre os dois modelos até aproximadamente o Top-8, indicando que ambos priorizam de forma equivalente os casos mais críticos no início da ordenação. A partir desse ponto, o modelo estendido apresenta uma leve vantagem em relação ao modelo base, recuperando uma fraude adicional no mesmo intervalo

de análise, sugerindo um ganho inicial de capacidade de ranqueamento dos casos fraudulentos.

Entretanto, essa superioridade não se mantém ao longo de toda a curva. Em etapas intermediárias, observa-se que o modelo base passa a superar o estendido em termos de recuperação acumulada. Apesar dessa inversão, o modelo estendido retoma o desempenho e consegue recuperar 100% das fraudes já por volta do Top-60, enquanto o modelo base atinge a cobertura completa apenas próximo ao Top-95.

Esses resultados indicam que, embora haja oscilações na ordenação intermediária, o modelo estendido demonstra maior eficiência global na concentração de fraudes em posições mais altas do ranking,

4.7 ANÁLISE DO INDICADOR DE RISCO COM DADOS RECENTES DO PAINEL

Com o objetivo de avaliar a abordagem proposta em um cenário mais recente, o Indicador de Risco base foi aplicado a dados extraídos do painel CEOS referentes ao período de 2020 a 2026. Nesse experimento, as licitações foram ranqueadas de acordo com o indicador calculado a partir das oito características base utilizadas pelo SubFraudGMM, reconstruídas especificamente para esse novo conjunto de dados.

Foram utilizadas as 29 licitações de fraudes associadas à Operação Patrola e um conjunto de 53 registros não rotulados de escavadeiras hidráulicas. Em relação às variáveis utilizadas, a única adaptação necessária ocorreu na característica *period*. Diferentemente da base histórica empregada nos experimentos anteriores, os registros mais recentes apresentavam elevada quantidade de valores ausentes para a data de homologação. Dessa forma, a data limite para envio das propostas foi utilizada como substituta no cálculo dessa característica.

A Tabela 10 apresenta os cinco registros não rotulados com maior Indicador de Risco, juntamente com as principais características que agruparam essas licitações junto com as fraudes previamente conhecidas. Esse resultado sugere que tais licitações não rotuladas apresentam padrões mais semelhantes aos de casos previamente identificados como fraudes.

Tabela 10 – Top 5 licitações não rotuladas com maior indicador de risco - painel CEOS (2020–2026)

ID	Indicador Risco	Distância	Subsets	LOO	Principais Features
ESC60	0.838438	0.256983	0.944444	0.788793	duration, unit_price, met
ESC90	0.826294	0.296329	0.944444	0.792850	unit_price, duration, met
ESC68	0.744802	0.342330	0.736111	0.804582	duration, unit_price, met
ESC67	0.715475	0.426641	0.722222	0.812666	duration, met, unit_price
ESC69	0.709895	0.443089	0.722222	0.812666	duration, met, unit_price

A variação observada nos valores do Indicador de Risco para os dados mais recentes do painel CEOS pode estar relacionada às diferenças na distribuição dos dados em relação à base antiga, uma vez que as características foram reconstruídas para o novo conjunto e podem sofrer influência de mudanças nos períodos analisados, no preenchimento dos documentos e nos padrões técnicos das contratações. Entretanto, como o Indicador de Risco é baseado em uma medida de similaridade, valores elevados continuam indicando que essas licitações apresentam maior proximidade com os padrões associados aos casos de fraude previamente identificados.

A Figura 8 apresenta o boxplot do Indicador de Risco para o grupo RT e para os demais registros no painel CEOS. Observa-se uma mediana mais elevada para o grupo RT, indicando que o modelo consegue atribuir valores mais altos a instâncias efetivamente associadas as fraudes quando tratadas como licitações desconhecidas.

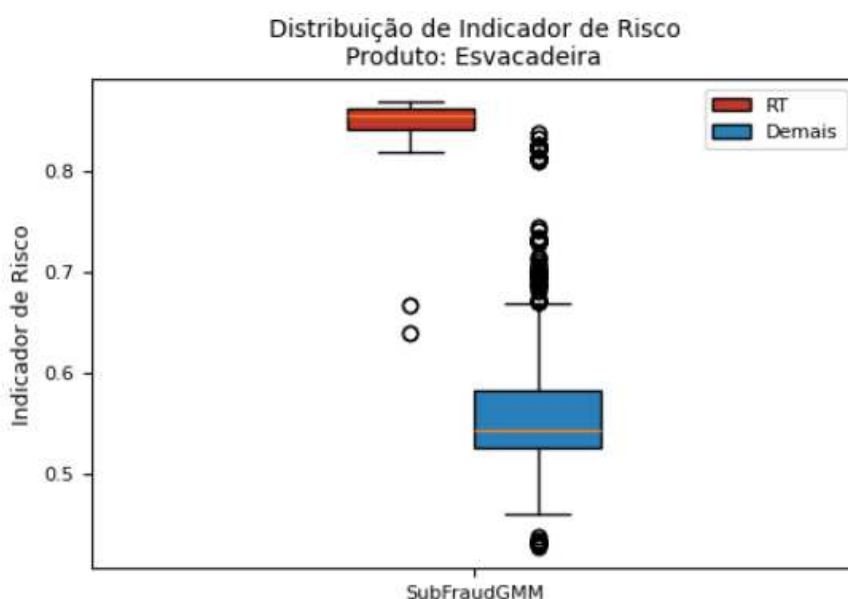


Figura 8 – Distribuição do Indicador de Risco por grupo (RT vs Demais) no painel CEOS (2020–2026)

De forma geral, os resultados obtidos reforçam a capacidade do SubFraudGMM em identificar o grupo RT mesmo sob condições de mudança completa de base de dados e reconstrução do espaço de características.

4.8 DISCUSSÃO DOS RESULTADOS

A análise exploratória dos caminhos mostra que os termos de referência podem ser organizados em uma estrutura hierárquica, na qual certos caminhos aparecem de forma recorrente em muitos documentos, enquanto outros são mais específicos e menos frequentes. Esse padrão indica que, embora os editais sigam algumas normas e padrões comuns estabelecidos pela legislação e boas práticas de contratação pública, ainda apresentam variações significativas que mostram as necessidades específicas de cada órgão contratante.

Percebe-se que a construção das quatro novas características foi diretamente motivada pelos padrões identificados nessa análise exploratória. Com base na classificação dos caminhos em frequentes, intermediários e raros foi definida a característica *mean_attr_freq*, que calcula a frequência média dos atributos presentes em cada documento.

A característica *mean_similarity* foi criada para quantificar o grau de similaridade estrutural entre um documento e os demais termos de referência da base. Adicionalmente, a observação de que determinados atributos, como "monitoramento por satélite" e "origem nacional", ocorrem proporcionalmente mais nos documentos fraudulentos motivou duas features complementares: *fraud_similarity*, que mede a proximidade de um documento em relação ao perfil médio das fraudes conhecidas, e *score_fraud_like*, que atribui maior peso aos atributos mais recorrentes entre os casos fraudulentos.

É importante reconhecer algumas limitações da análise atual. Primeiro, o conjunto de dados é relativamente pequeno, com 100 documentos, o que pode limitar a generalização dos padrões identificados. Segundo, a categorização binária entre fraude e não fraude, embora útil para análise exploratória, não captura gradações de direcionamento que podem existir na prática. Além disso, os dados de fraudes são provenientes de apenas uma operação, o ideal seria ter maior diversidade de fraudes.

Apesar de este estudo tenha sido conduzido com licitações de motoniveladoras, a metodologia proposta pode ser aplicada a outros objetos de contratação pública. Para isso, é necessário dispor de um conjunto de termos de referência referentes ao equipamento ou serviço de interesse, bem como de exemplos de licitações fraudulentas relacionados ao mesmo objeto.

Os resultados comparativos indicam que a inclusão das novas características não alterou significativamente a ordem dos casos mais críticos no ranking de risco, mas promoveu um refinamento na separação entre documentos suspeitos e não suspeitos. Esse comportamento sugere que o modelo original já capturava os principais padrões

estruturais associados às fraudes, enquanto as novas variáveis atuaram como um mecanismo de ajuste fino, aumentando a capacidade de discriminação em casos mais ambíguos.

Embora o modelo estendido tenha apresentado ganho na mediana do Indicador de Risco, esse incremento foi relativamente modesto quando comparado ao aumento de complexidade introduzido pela construção e manutenção da taxonomia hierárquica. Sob a perspectiva de aplicação prática por órgãos de controle, a necessidade de manter uma pipeline de extração, padronização e atualização contínua da taxonomia representa um custo adicional que, neste estudo, não se mostrou proporcional ao ganho observado. Ainda assim, a taxonomia permanece como uma contribuição relevante, por demonstrar que características semânticas extraídas dos termos de referência podem complementar o modelo original e servir de base para futuras extensões da abordagem.

5 APLICAÇÃO

5.1 INTEGRAÇÃO DO ALERTA DE FAVORITISMO AO PAINEL CEOS

Como forma de aplicação prática da abordagem proposta, o indicador de risco gerado pelo SubFraudGMM-Estendido foi integrado ao painel CEOS por meio da criação de um alerta de favoritismo em licitações públicas. Os alertas foram implementados no nível de item de licitação e incluem informações descritivas sobre os principais fatores que contribuíram para o risco identificado, como recorrência de participações, interações anteriores com o órgão, tempo de atuação das empresas e similaridade comportamental com fraudes conhecidas.

Ressalta-se que essa integração ainda não foi realizada para os dados mais recentes do painel CEOS, sendo aplicada apenas sobre a primeira análise do trabalho. Como demonstrado na Figura 9, a integração permite visualizar os alertas diretamente no painel da licitação.

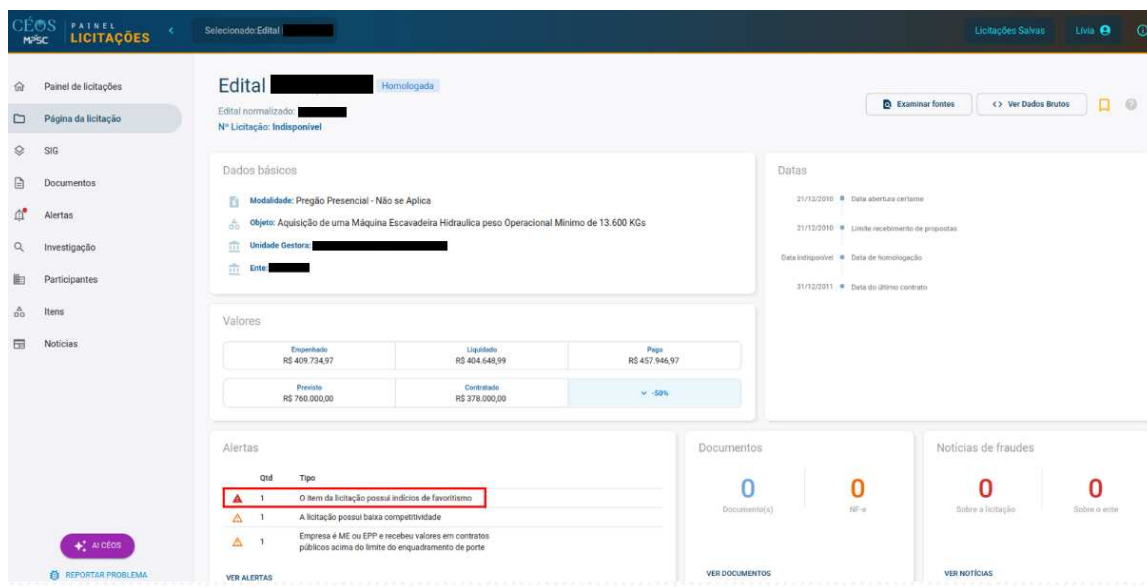


Figura 9 – Visualização do alerta de favoritismo na licitação.

Além da visualização na licitação, na figura 10 os alertas são exibidos na aba específica de alertas do painel, permitindo a consulta resumida das ocorrências identificadas pelo modelo.

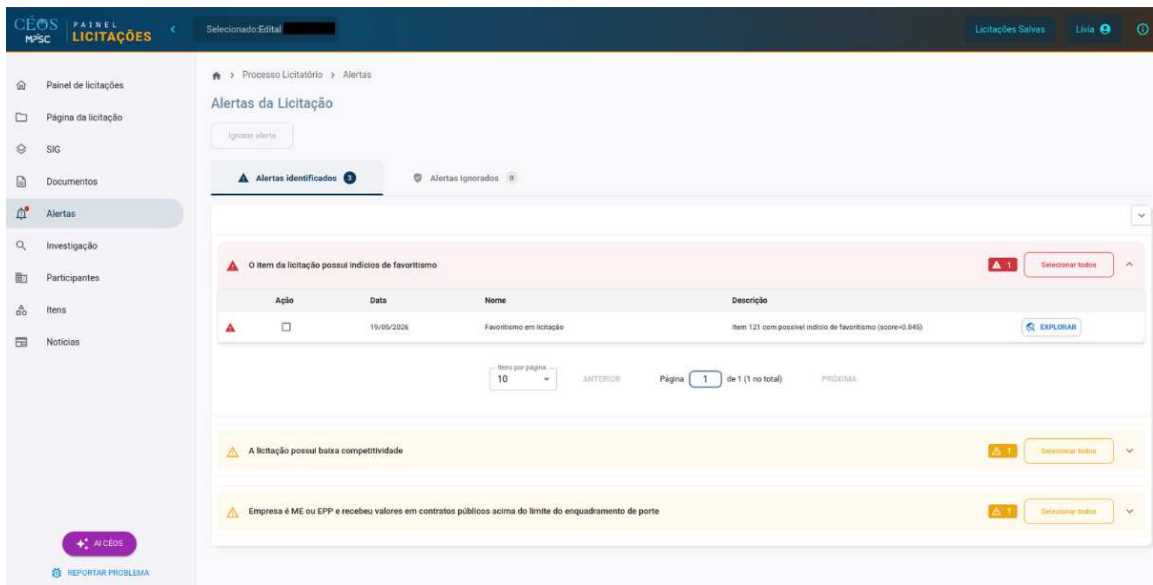


Figura 10 – Listagem dos alertas de favoritismo na aba de alertas.

Ao selecionar a opção de explorar, o sistema apresenta informações mais detalhadas sobre os fatores que contribuíram para o indicador de risco atribuído ao item analisado. As figuras 11 e 12 mostram o alerta da licitação com maior indicador de risco.

De forma análoga, as figuras 13 e 14 apresentam o detalhamento da licitação classificada com o segundo maior indicador de risco; as figuras 15 e 16, da terceira; as figuras 17 e 18, da quarta; enquanto as figuras 19 e 20 exibem o detalhamento da quinta licitação com maior indicador de risco.

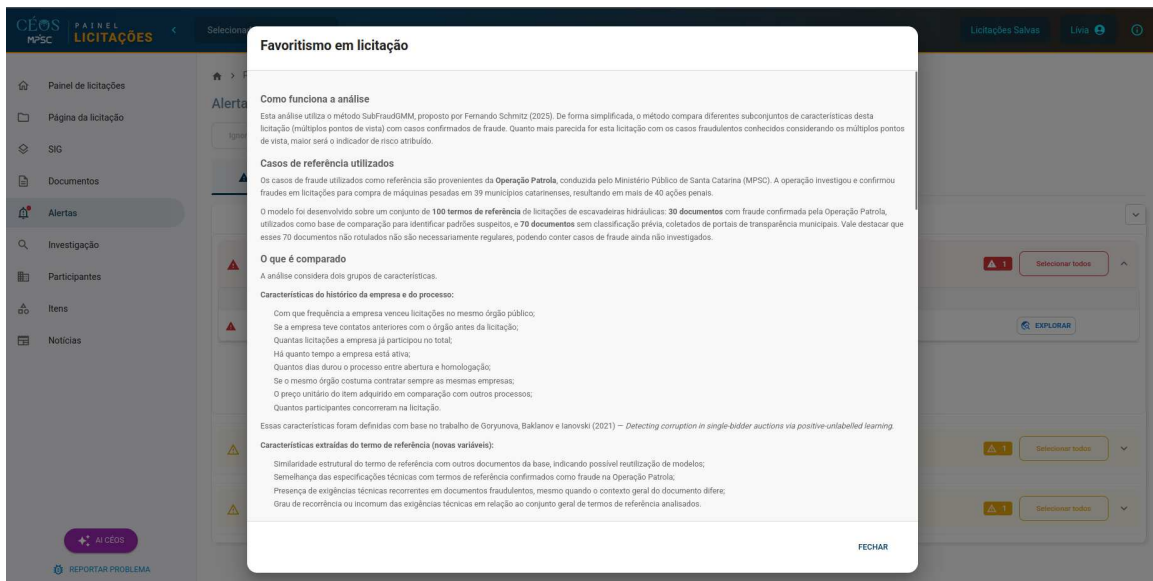


Figura 11 – Alerta de favoritismo da 1ª licitação com maior risco (Parte 1).

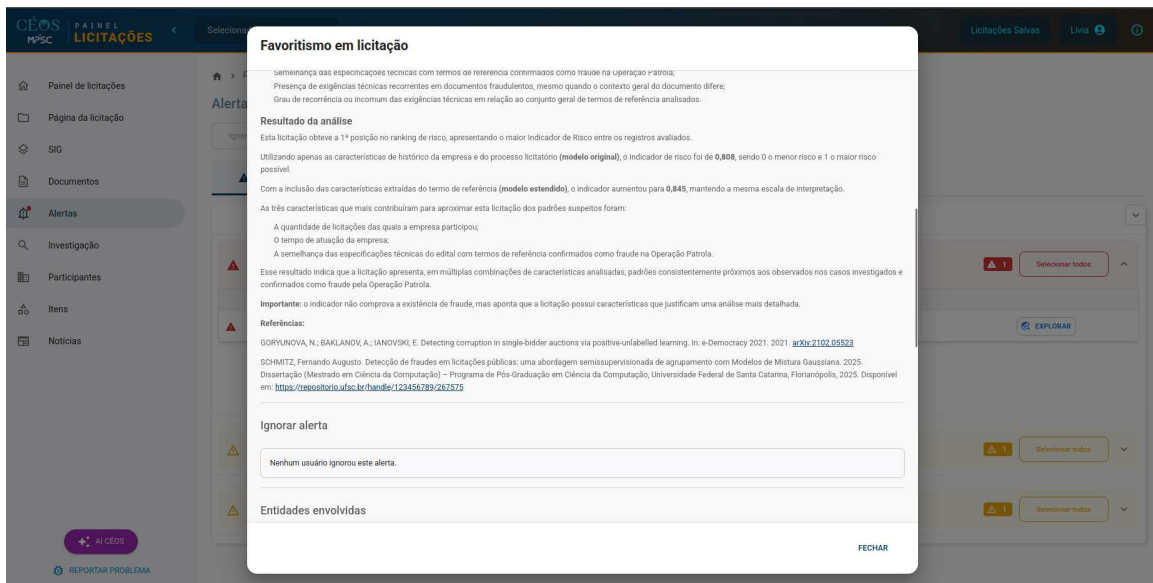


Figura 12 – Alerta de favoritismo da 1ª licitação com maior risco (Parte 2).

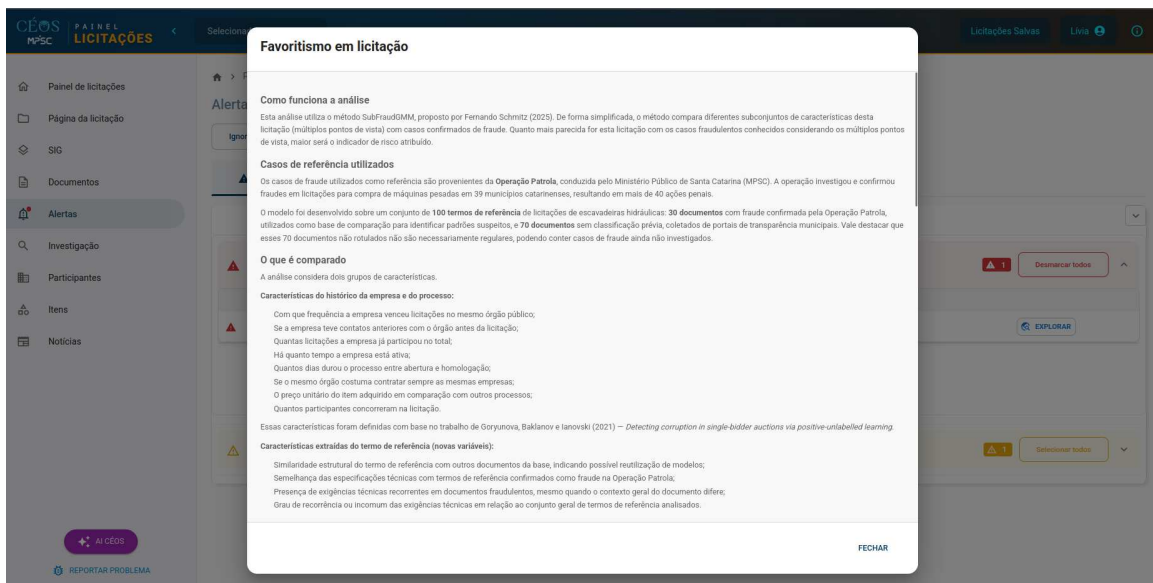


Figura 13 – Alerta de favoritismo da 2ª licitação com maior risco (Parte 1).

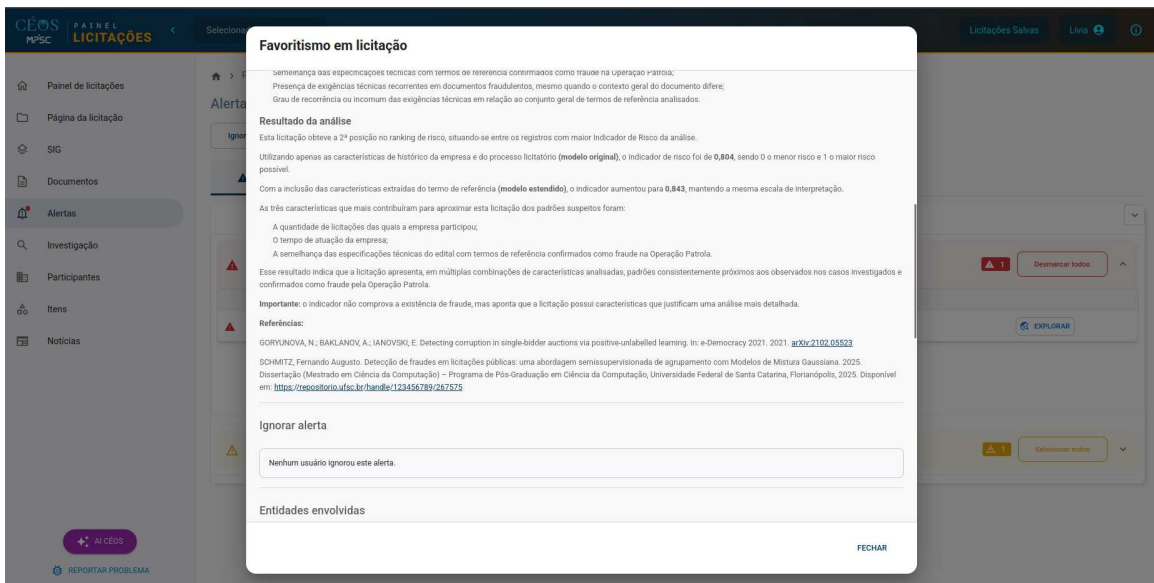


Figura 14 – Alerta de favoritismo da 2ª licitação com maior risco (Parte 2).

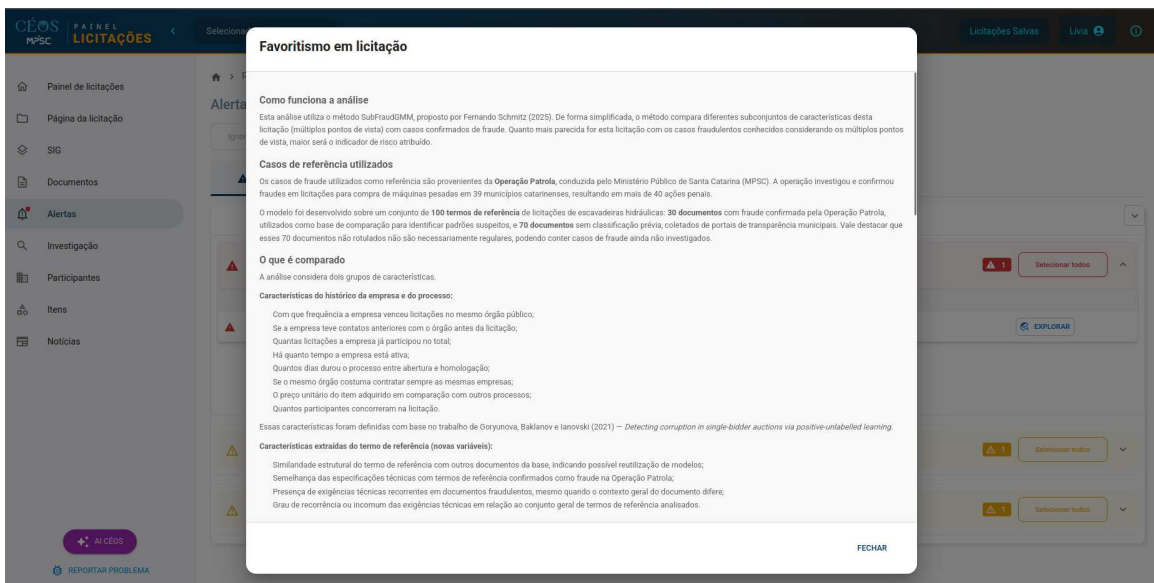


Figura 15 – Alerta de favoritismo da 3ª licitação com maior risco (Parte 1).

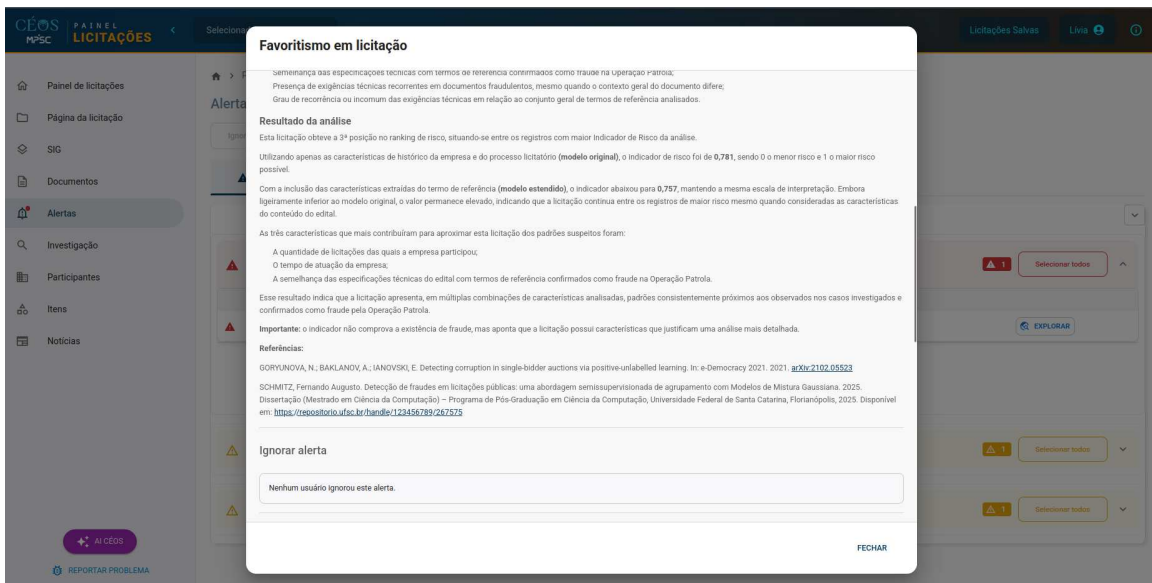


Figura 16 – Alerta de favoritismo da 3ª licitação com maior risco (Parte 2).

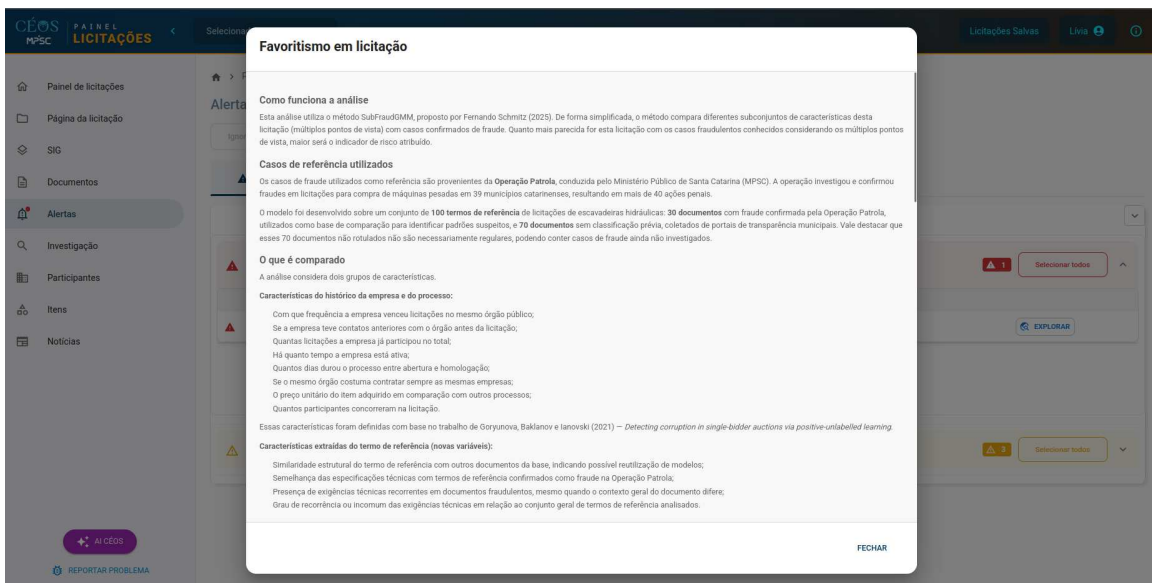


Figura 17 – Alerta de favoritismo da 4ª licitação com maior risco (Parte 1).

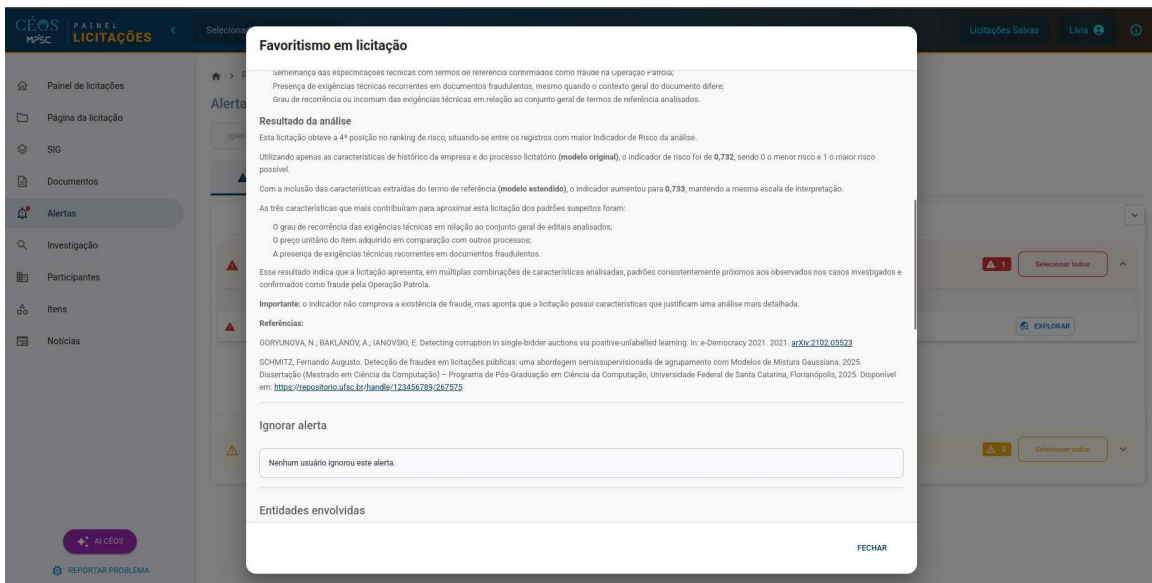


Figura 18 – Alerta de favoritismo da 4ª licitação com maior risco (Parte 2).

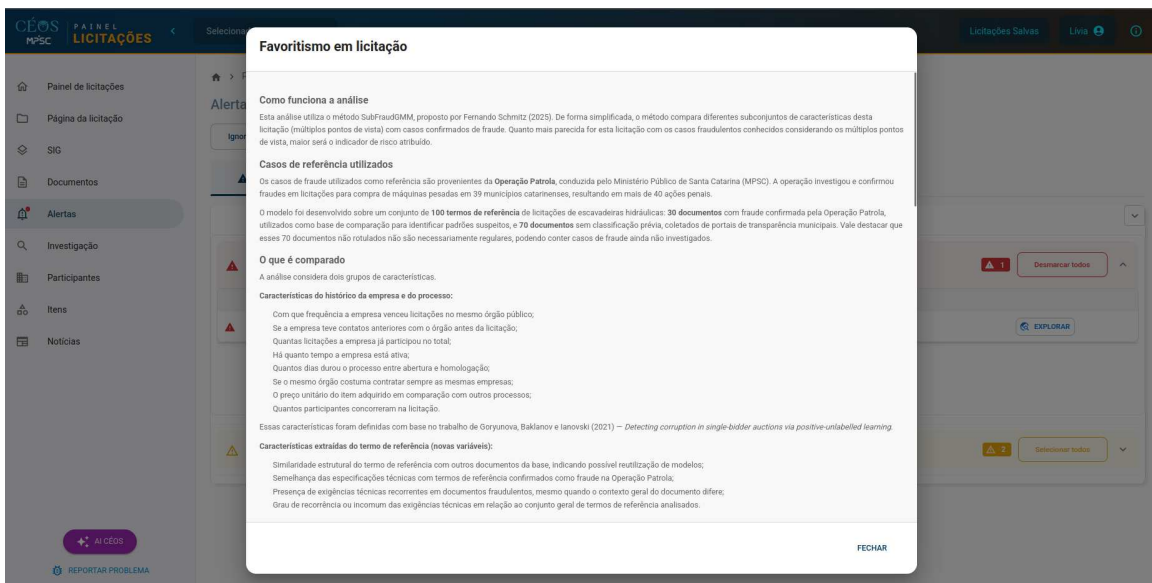


Figura 19 – Alerta de favoritismo da 5ª licitação com maior risco (Parte 1).

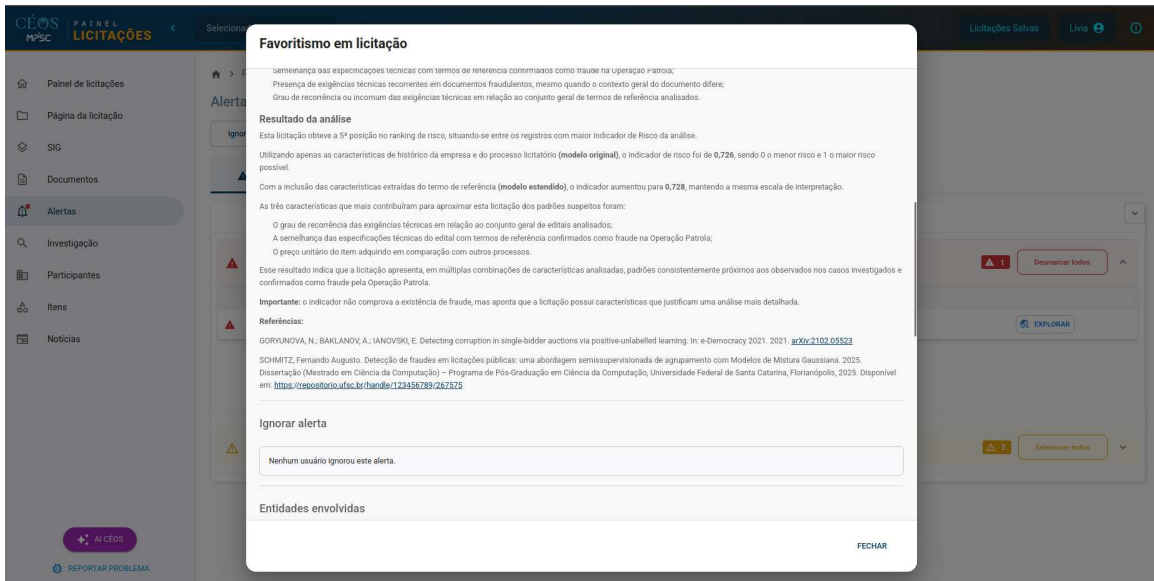


Figura 20 – Alerta de favoritismo da 5ª licitação com maior risco (Parte 2).

6 CONCLUSÃO

Este trabalho apresentou uma abordagem integrada para detecção de direcionamento em licitações públicas, fundamentada na análise de termos de referência. A metodologia conecta duas frentes previamente desenvolvidas de forma independente: a construção automatizada de taxonomias hierárquicas via BERTopic e LLM, e a aplicação do modelo semissupervisionado SubFraudGMM para identificação de padrões suspeitos.

A taxonomia construída revelou 176 caminhos distintos de atributos técnicos a partir de 100 termos de referência de escavadeiras hidráulicas. A análise exploratória evidenciou diferenças na distribuição dos atributos entre os grupos analisados. Enquanto alguns caminhos apresentaram frequências semelhantes em ambos os grupos, determinados atributos ocorreram proporcionalmente mais nos documentos rotulados como fraude, indicando a existência de padrões característicos desse conjunto.

Os resultados quantitativos confirmam que o SubFraudGMM-Estendido atribuiu valores significativamente mais elevados de risco às instâncias fraudulentas com mediana de 78.8% no grupo RT, apesar do conjunto de dados reduzido.

A incorporação de novas variáveis introduziram uma camada de informação contextual que o modelo original não contemplava, passando a capturar não apenas padrões comportamentais, mas também padrões de conteúdo dos documentos. Vale destacar que a limitação no número de características incorporadas não é arbitrária, ela decorre diretamente da natureza combinatória do SubFraudGMM, em que o crescimento do espaço de combinações possíveis é exponencial.

Por fim, a integração do indicador de risco ao painel CEOS como alerta de favoritismo representa a materialização prática da pesquisa, demonstrando que a abordagem vai além do experimento acadêmico e pode ser efetivamente utilizada por órgãos de controle na fiscalização automatizada de compras públicas.

6.1 TRABALHOS FUTUROS

Como continuidade deste trabalho, destacam-se algumas direções para pesquisas futuras:

- **Ampliação do conjunto de dados:** Durante o desenvolvimento deste trabalho, a análise ficou limitada a 100 documentos, em função da disponibilidade de editais. Com a ampliação recente do acesso a editais públicos no painel do CEOS, torna-se possível aplicar o método em um conjunto de dados mais extenso e diversificado.
- **Exploração de novas variáveis:** Investigar a construção de novas variáveis representativas dos caminhos extraídos dos termos de referência, considerando aspectos como frequência, especificidade e relações estruturais entre os atributos.
- **Avaliação em outros domínios de licitação:** Aplicar a abordagem proposta a diferentes tipos de objetos licitatórios, incluindo aqueles já explorados no trabalho base, como pá carregadeira, rolo compactador, trator de esteira e motoniveladora, além das escavadeiras hidráulicas analisadas neste estudo.
- **Aplicação em outros tipos de documentos:** Além dos termos de referência, a abordagem pode ser estendida para outros documentos que contêm especificações técnicas ou justificativas da contratação, como estudos técnicos preliminares.
- **Automatização da descoberta de alertas:** Atualmente, os alertas considerados pelo modelo são cadastrados manualmente. Pretende-se estudar como realizar a integração da abordagem proposta ao painel do CEOS, permitindo a geração automática de alertas.

REFERÊNCIAS

BISHOP, Christopher M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer, 2006. Citado na p. [22](#).

BRASIL. **Lei nº 14.133, de 1º de abril de 2021: Lei de Licitações e Contratos Administrativos**. [S.l.: s.n.], 2021.
https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm.
Citado nas pp. [16](#), [17](#).

CAMPELLO, Ricardo JGB; MOULAVI, Davoud; SANDER, Jörg. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **ADVANCES in Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2013. Citado na p. [18](#).

CHAPELLE, Olivier; SCHÖLKOPF, Bernhard; ZIEN, Alexander. **Semi-Supervised Learning**. [S.l.]: MIT Press, 2009. Citado na p. [21](#).

DIAS, Celia da Consolação; MOREIRO GONZÁLEZ, José Antonio. Metodologia para construção de taxonomias: uma revisão sistemática de literatura. **Ciência da Informação em Revista**, v. 11, 2024. Citado na p. [19](#).

GROOTENDORST, Maarten. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. **arXiv preprint arXiv:2203.05794**, 2022. Citado na p. [18](#).

JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. Data clustering: a review. **ACM Computing Surveys (CSUR)**, ACM, v. 31, n. 3, p. 264–323, 1999. Citado nas pp. [20](#), [21](#).

LYRA, Marcos S et al. Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods. **Applied Network Science**, Springer, v. 7, n. 1, p. 83, 2022. Citado na p. [12](#).

MCINNES, Leland; HEALY, John; MELVILLE, James. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **arXiv preprint arXiv:1802.03426**, 2018. Citado na p. [18](#).

MCLACHLAN, G.J.; PEEL, D. **Finite Mixture Models**. [S.l.]: Wiley, 2004. (Wiley Series in Probability and Statistics). ISBN 9780471654063. Disponível em:
<https://books.google.com.br/books?id=c2_fAox0DQoC>. Citado na p. [22](#).

MIKOLOV, Tomas et al. Efficient Estimation of Word Representations in Vector Space. In: **INTERNATIONAL Conference on Learning Representations**. [S.l.: s.n.], 2013. Citado nas pp. [17](#), [18](#).

MINISTÉRIO PÚBLICO DE SANTA CATARINA. **Operação Patrola: ex-Prefeito de Tangará é condenado a 13 anos de prisão**. 2019. Disponível em:

<<https://www.mpssc.mp.br/w/noticias/operacao-patrola-ex-prefeito-de-tangara-e-condenado-a-13-anos-de-prisao>>. Citado na p. 27.

NAVEED, Humza et al. **A Comprehensive Overview of Large Language Models**. [S.l.: s.n.], 2024. arXiv: 2307.06435 [cs.CL]. Disponível em: <<https://arxiv.org/abs/2307.06435>>. Citado na p. 17.

NIE, Zhijie et al. **When Text Embedding Meets Large Language Model: A Comprehensive Survey**. [S.l.: s.n.], 2025. arXiv: 2412.09165 [cs.CL]. Disponível em: <<https://arxiv.org/abs/2412.09165>>. Citado na p. 17.

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. GloVe: Global Vectors for Word Representation. In: PROCEEDINGS of the 2014 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2014. p. 1532–1543. Citado na p. 18.

PORTAL DA TRANSPARÊNCIA DO GOVERNO FEDERAL. **Licitações**. [S.l.: s.n.], 2025. <https://portaldatransparencia.gov.br/licitacoes>. Brasília: Controladoria-Geral da União. Acesso em: 22 jun. 2025. Citado nas pp. 12, 66.

REIMERS, Nils; GUREVYCH, Iryna. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: PROCEEDINGS of the 2019 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2019. Citado na p. 18.

RUSSELL, Stuart; NORVIG, Peter. **Artificial Intelligence: A Modern Approach**. 4. ed. [S.l.]: Pearson, 2021. Citado na p. 20.

SCHIAVON, João Pedro. **Feature Extractor Taxonomy**. [S.l.: s.n.], 2024. <https://github.com/SchiavonJP/Feature-Extractor-Taxonomy>. Acesso em: 6 maio 2026. Citado nas pp. 13, 14, 28, 30, 31, 35, 66, 68.

SCHMITZ, Fernando Augusto. **Deteção de fraudes em licitações públicas: uma abordagem semissupervisionada de agrupamento com Modelos de Mistura Gaussiana**. 2025. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Santa Catarina, Florianópolis. Disponível em: <<https://repositorio.ufsc.br/handle/123456789/267575>>. Citado nas pp. 13, 14, 22, 23, 32, 35, 45, 66, 68.

SCHNEIDER DOS SANTOS, Everton et al. Detection of fraud in public procurement using data-driven methods: a systematic mapping study. **EPJ Data Science**, v. 14, n. 52, 2025. DOI: 10.1140/epjds/s13688-025-00569-3. Citado nas pp. 12, 66.

SINGH, Rajdutt Shekhar; SINGH, Rounak. Tendering With Intent: Dissecting Tailor-made Bid Conditions in Public Procurement vis-à-vis the Judicial Precedents. **Mondaq**, 2025. Acesso em: 11 nov. 2025. Disponível em: <<https://www.mondaq.com/india/government-contracts-procurement>>

[ppp/1682144/tendering-with-intent-dissecting-tailor-made-bid-conditions-in-public-procurement-vis-%C3%A0-vis-the-judicial-precedents>](#). Citado na p. 12.

STONE, M. Cross-validators choice and assessment of statistical predictions. **Journal of the Royal Statistical Society. Series B**, v. 36, n. 2, p. 111–147, 1974. Citado nas pp. 23, 24.

Apêndices

Integração de Taxonomia Hierárquica de Produtos e Aprendizado Semissupervisionado para Detecção de Fraudes em Licitações Públicas

Livia Corazza Ferrão, Jônata Tyska Carvalho

Programa de Graduação em Ciências da Computação – Universidade Federal de Santa Catarina (UFSC)

livia.ferrao@grad.ufsc.br, jonata.carvalho@ufsc.br

Resumo. *Fraudes em licitações públicas, em especial o direcionamento por meio de especificações técnicas restritivas, são difíceis de detectar porque seus indícios estão implícitos na linguagem técnica dos termos de referência. Este trabalho integra a construção automatizada de uma taxonomia hierárquica de atributos técnicos, apoiada por BERTopic e por um modelo de linguagem de larga escala, ao modelo semissupervisionado SubFraudGMM, baseado em Modelos de Mistura Gaussiana, incorporando quatro novas variáveis derivadas dos termos de referência. Os experimentos foram conduzidos sobre 100 termos de referência de licitações de escavadeiras hidráulicas (30 fraudes confirmadas pela Operação Patrola e 70 documentos não rotulados), resultando em uma taxonomia com 176 caminhos hierárquicos. Na validação por remoção de fraudes (LOFO), o modelo estendido apresentou mediana de 78,8% no Indicador de Risco das fraudes mascaradas, contra 77,2% da versão base, indicando que as características hierárquicas contribuem para refinar a detecção de direcionamento em compras públicas.*

Abstract. *Fraud in public procurement, particularly bid steering through restrictive technical specifications, is hard to detect because its indicators are implicit in the technical language of terms of reference. This work integrates the automated construction of a hierarchical taxonomy of technical attributes, supported by BERTopic and a large language model, with the semi-supervised SubFraudGMM model, based on Gaussian Mixture Models, incorporating four new variables derived from terms of reference. Experiments were conducted on 100 terms of reference for hydraulic excavator procurements (30 confirmed fraud cases from Operação Patrola and 70 unlabeled documents), producing a taxonomy with 176 hierarchical paths. In leave-one-fraud-out validation, the extended model achieved a median Risk Indicator of 78.8% for masked fraud*

cases, versus 77.2% for the base version, indicating that hierarchical features help refine the detection of bid-steering patterns in public procurement.

1 INTRODUÇÃO

A administração pública brasileira movimenta anualmente bilhões de reais em contratações por meio de licitações ([Portal da Transparência do Governo Federal, 2025](#)), e mesmo diante de um aparato legal robusto, fraudes continuam comprometendo a eficiência da gestão pública e a confiança da população nas instituições. Entre as modalidades de irregularidade, este trabalho concentra-se no direcionamento de licitações, prática em que editais são elaborados com especificações técnicas restritivas para favorecer um fornecedor específico, restringindo a concorrência e viabilizando o desvio de recursos público ([Schneider dos Santos et al., 2025](#)).

O termo de referência, documento que detalha as especificações técnicas do objeto contratado, é justamente onde esse tipo de direcionamento costuma se materializar, por meio de requisitos excessivamente restritivos, linguagem tendenciosa ou reutilização de especificações entre editais. A análise manual desses documentos exige conhecimento especializado e tempo elevado, o que torna inviável o acompanhamento do volume de licitações realizadas no país, motivando o uso de aprendizado de máquina para a detecção automatizada de irregularidades.

Este artigo integra duas linhas de pesquisa desenvolvidas de forma independente: (i) uma pipeline de construção automatizada de taxonomias hierárquicas de atributos técnicos a partir de termos de referência, apoiada por BERTopic e por um modelo de linguagem de larga escala (LLM) ([Schiavon, 2024](#)); e (ii) o modelo semissupervisionado SubFraudGMM, baseado em Modelos de Mistura Gaussiana (GMM), utilizado para priorizar licitações potencialmente suspeitas por meio de um Indicador de Risco ([Schmitz, 2025](#)). A contribuição central deste trabalho é conectar essas duas etapas, propondo quatro novas características derivadas da representação semântica hierárquica dos documentos e avaliando seu impacto sobre a capacidade de detecção do modelo.

Os experimentos foram conduzidos sobre 100 termos de referência de licitações de escavadeiras hidráulicas, das quais 30 são casos de fraude confirmados pelo Ministério Público de Santa Catarina (MPSC) no âmbito da Operação Patrola e 70 são documentos não rotulados coletados em portais de transparência. A validação, realizada por meio do protocolo Leave-One-Fraud-Out (LOFO), mostrou que a incorporação das novas variáveis elevou a mediana do Indicador de Risco atribuído às fraudes mascaradas de 77,2% para 78,8%.

2 TRABALHOS RELACIONADOS

Schmitz (2025) propôs o SubFraudGMM, uma metodologia semissupervisionada baseada em Modelos de Mistura Gaussiana para identificar padrões atípicos em conjuntos de dados de licitações, combinando proximidade com fraudes conhecidas, consistência entre subespaços de características e capacidade discriminativa via validação *leave-one-out*. Este trabalho utiliza o SubFraudGMM como base metodológica, reaplicando-o com a inclusão de características derivadas dos termos de referência.

Schiavon (2024) propôs uma pipeline automatizada de construção de taxonomias para termos de referência de licitações, combinando extração de atributos, agrupamento semântico com BERTopic e geração de hierarquias por meio de um LLM. Essa pipeline é reutilizada neste trabalho como etapa inicial de processamento textual, complementada por uma hierarquia manual de atributos específica para o domínio de escavadeiras hidráulicas.

Outros trabalhos correlatos incluem o TaxoRankConstruct, método iterativo baseado em ranks para construção de taxonomias com LLMs priorizando propriedades de objetos, o TELEClass, que combina LLMs com enriquecimento automático de taxonomias sob supervisão mínima, e trabalhos que aplicam GMMs à detecção de fraude em transações financeiras, reforçando a adequação de modelos probabilísticos para problemas de detecção de anomalias em cenários com poucos exemplos rotulados.

3 MATERIAIS E MÉTODOS

3.1 Base de Dados

A base de dados é composta por 100 termos de referência de licitações de escavadeiras hidráulicas, conforme a Tabela 11. Trinta documentos correspondem a casos de fraude confirmados pela Operação Patrola, conduzida pelo MPSC, e os demais 70 foram coletados em portais de transparência municipais e federais, permanecendo não rotulados.

Tabela 11 – Composição da base de dados de Termos de Referência

Origem	Quantidade	Status
Operação Patrola	30	Fraudes
Portais de transparência	70	Não rotulados
Total	100	—

3.2 Construção da Taxonomia

A construção da taxonomia segue a pipeline de Schiavon (2024), composta por seis etapas: coleta de características a partir das descrições técnicas; agrupamento

semântico dos atributos via *embeddings* (paraphrase-multilingual-MiniLM-L12-v2), U-MAP e HDBSCAN; processamento em lotes; análise por um LLM (Qwen 2.5, 32B) que extrai a hierarquia de atributos, valor e unidade de medida; construção da estrutura de taxonomia; e persistência dos dados (Schiavon, 2024).

Após tratamento manual das inconsistências geradas pelo LLM, obteve-se um total de 176 caminhos hierárquicos distintos (por exemplo, *motor > potência > líquida > mínima*), convertidos em variáveis binárias de presença/ausência por documento. Como contribuição adicional, foi corrigida a hierarquia manual de atributos específicos para escavadeiras hidráulicas gerada pela LLM, baseada em manuais técnicos, com o objetivo de padronizar atributos semanticamente equivalentes representados por caminhos distintos (por exemplo, “potência”, “potência bruta” e “potência líquida”). Essa estrutura poderá futuramente ser incorporada à própria pipeline de geração automática, servindo como conhecimento prévio para o LLM.

3.3 Modelo Semissupervisionado SubFraudGMM

O SubFraudGMM (Schmitz, 2025) parte da premissa de que, em espaços de alta dimensionalidade, fraudes conhecidas tendem a apresentar similaridade consistente apenas em subconjuntos de características, não necessariamente em todas as dimensões simultaneamente. O método avalia múltiplos subespaços de características por meio de GMMs com diferentes números de componentes e tipos de matriz de covariância, selecionados via Critério de Informação Bayesiano (BIC). Um subespaço é mantido apenas quando concentra ao menos 80% das fraudes conhecidas em um mesmo agrupamento (Schmitz, 2025).

A partir dos subespaços relevantes, define-se um Indicador de Risco contínuo, combinando três dimensões:

$$\text{Risco} = \frac{(1 - \text{Dist}_{\text{euclid}}) + \text{NumClusters} + \text{RecallRel}}{3} \quad (5)$$

em que $\text{Dist}_{\text{euclid}}$ é a distância euclidiana média em relação às fraudes conhecidas, NumClusters é a proporção de subespaços relevantes em que a instância foi agrupada com fraudes, e RecallRel é o *recall* relativo médio obtido na validação *leave-one-out*. Todas as componentes são normalizadas por Min-Max antes da combinação.

O conjunto original de características do SubFraudGMM inclui oito variáveis estruturais e comportamentais: proporção de vitórias por Unidade Gestora (*win*), histórico de reuniões prévias (*met*), número de licitações da empresa (*num*), tempo de atuação (*period*), duração do processo licitatório (*duration*), proporção de vencedores únicos (*unique*), número de participantes e preço unitário do item.

3.4 Novas Características

Este trabalho propõe quatro novas variáveis derivadas da representação binária dos caminhos da taxonomia, descritas na Tabela 12.

Tabela 12 – Novas características propostas

Característica	Descrição
mean_similarity	Similaridade média do documento em relação aos demais, via distância de Hamming
fraud_similarity	Similaridade do documento em relação ao perfil médio dos documentos fraudulentos
score_fraud_like	Soma ponderada dos atributos conforme sua frequência no conjunto de fraudes
mean_attr_freq	Frequência média dos atributos presentes no documento, no conjunto de dados

Essas variáveis foram selecionadas por concentrarem-se em três aspectos: similaridade estrutural entre documentos, proximidade a padrões de fraude e comportamento coletivo dos atributos. Variáveis relacionadas à contagem bruta ou à densidade de atributos foram descartadas por não apresentarem relação consistente com a ocorrência de fraude, além de aumentarem exponencialmente o espaço de combinações de subespaços avaliado pelo SubFraudGMM.

4 AVALIAÇÃO E RESULTADOS

4.1 Validação por Remoção de Fraudes (LOFO)

A avaliação seguiu o protocolo *Leave-One-Fraud-Out* (LOFO): a cada iteração, uma das 30 fraudes conhecidas é removida do conjunto rotulado e tratada como instância não rotulada, enquanto as demais 29 permanecem como referência. O modelo é então executado e o Indicador de Risco é calculado para a fraude mascarada, denominada grupo RT (Registro de Teste). O procedimento é repetido para cada uma das 30 fraudes, gerando 30 valores de risco comparáveis entre a versão base e a versão estendida do modelo.

4.2 Resultados Comparativos

A Tabela 13 apresenta a mediana do Indicador de Risco obtida para o grupo RT, comparando a versão base do SubFraudGMM com a versão estendida (incluindo as quatro novas características) e com um *Ranking Simples*, heurística de referência baseada em média das variáveis originais.

O SubFraudGMM superou o Ranking Simples em ambas as versões, confirmando a vantagem da modelagem por subespaços via GMM sobre uma heurística linear. A versão estendida apresentou ganho de aproximadamente 1,6 ponto percentual em

Tabela 13 – Mediana do Indicador de Risco no grupo RT

Modelo	Mediana RT
Ranking Simples (base)	61,0%
SubFraudGMM (base)	77,2%
SubFraudGMM (estendido)	78,8%

relação à base, sem alterar drasticamente a ordem dos casos mais críticos no ranking, indicando que o modelo original já capturava os principais padrões estruturais, enquanto as novas variáveis atuaram como mecanismo de ajuste fino. Nas análises de importância de características, variáveis originais como *period* e *met* continuaram entre as mais relevantes, mas passaram a ser acompanhadas por *fraud_similarity* e *mean_attr_freq*, evidenciando que o modelo passou a considerar tanto padrões comportamentais quanto padrões de conteúdo dos documentos.

A Tabela 14 apresenta os principais registros não rotulados com maior indicador de risco após a inclusão das novas variáveis, juntamente com suas respectivas métricas e principais características associadas.

Tabela 14 – Top 5 licitações não rotuladas com maior indicador de risco - novo modelo

ID	Indicador Risco	Distância	Subsets	LOO	Principais Features
ESC38	0.8418	0.2512	0.9317	0.7631	num, period, fraud_similarity
ESC94	0.8410	0.2542	0.9313	0.7625	num, period, fraud_similarity
ESC93	0.7638	0.4232	0.8638	0.7716	num, period, fraud_similarity
ESC3	0.7415	0.2856	0.6860	0.7513	mean_attr_freq, unit_price, score_fraude_like
ESC49	0.7335	0.2969	0.6722	0.7513	mean_attr_freq, fraud_similarity, unit_price

A análise da curva de recuperação de fraudes em função do Top- N registros ranqueados mostrou comportamento equivalente entre os modelos até aproximadamente o Top-8, com o modelo estendido atingindo 100% de recuperação das fraudes por volta do Top-60, contra aproximadamente o Top-95 do modelo base, um ganho relevante de eficiência global no ranqueamento, apesar de oscilações em posições intermediárias.

O modelo base também foi aplicado a um conjunto mais recente de dados do painel CEOS (2020–2026), com 29 fraudes da Operação Patrola e 53 registros não rotulados de escavadeiras hidráulicas, apresentando mediana de risco mais elevada para o grupo RT do que para os demais registros, o que reforça a capacidade de generalização do modelo mesmo sob mudança completa da base de dados.

4.3 Aplicação Prática

Como forma de validar a aplicabilidade da abordagem, o Indicador de Risco gerado pelo SubFraudGMM estendido foi integrado ao painel CEOS por meio de um alerta de favoritismo em licitações públicas, implementado no nível de item de licitação. O alerta apresenta os principais fatores que contribuíram para o risco atribuído, como recorrência de participação, interações prévias com o órgão contratante e similaridade estrutural com fraudes conhecidas, tornando o resultado do modelo interpretável para uso por órgãos de controle.

As figuras 21 e 22 mostram o alerta da licitação com maior indicador de risco.

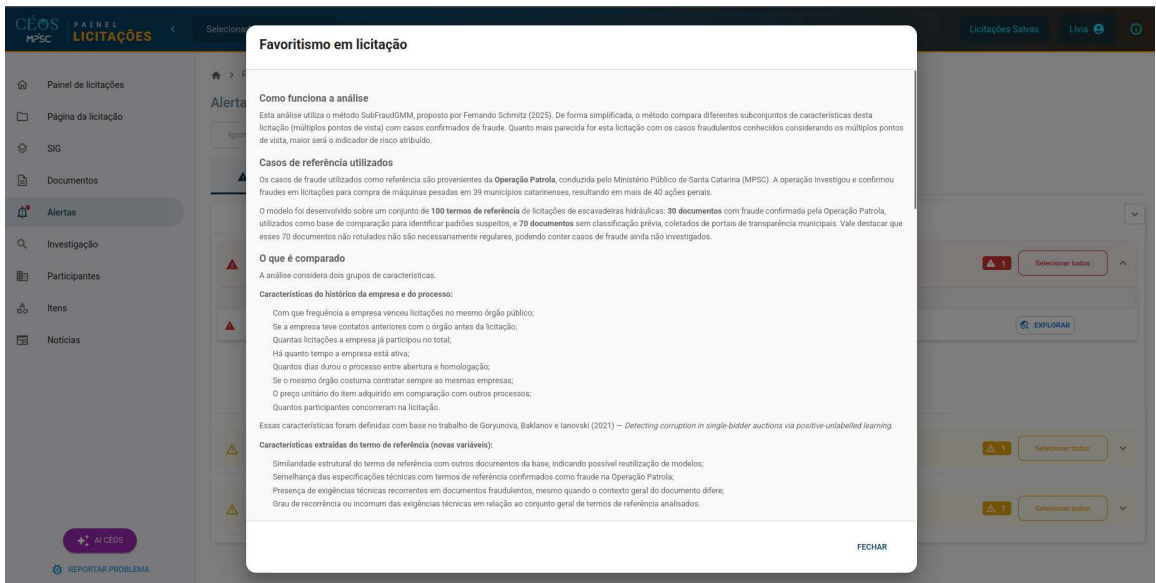


Figura 21 – Alerta de favoritismo da 1ª licitação com maior risco (Parte 1).

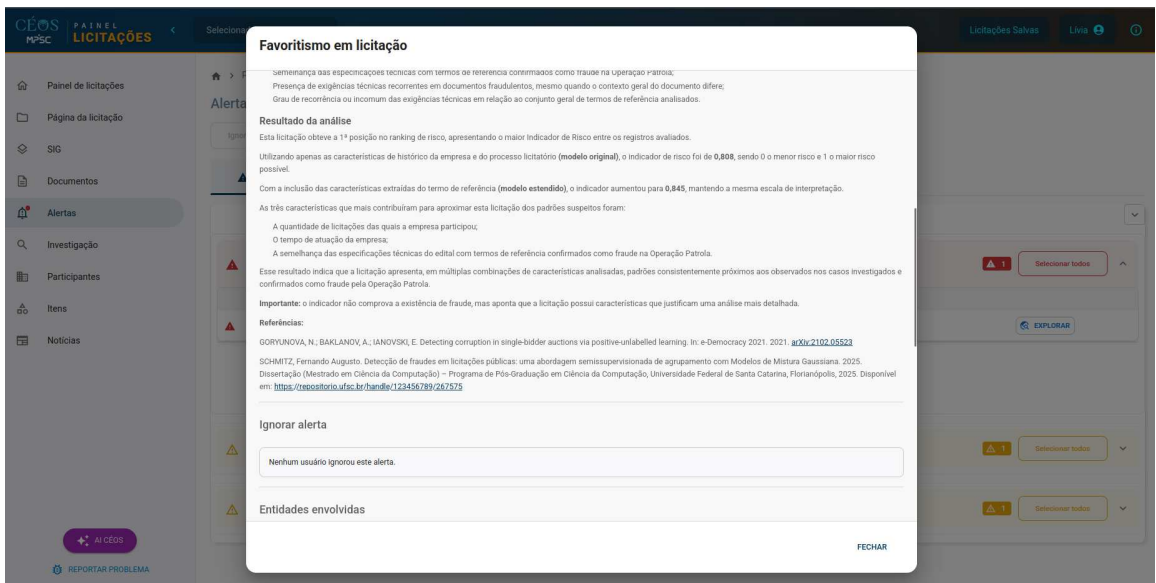


Figura 22 – Alerta de favoritismo da 1ª licitação com maior risco (Parte 2).

5 CONCLUSÃO

Este trabalho integrou a construção automatizada de uma taxonomia hierárquica de atributos técnicos ao modelo semissupervisionado SubFraudGMM, propondo quatro novas características derivadas dos termos de referência de licitações de escavadeiras hidráulicas. A validação por remoção de fraudes mostrou que a incorporação dessas características elevou a mediana do Indicador de Risco das fraudes mascaradas de 77,2% para 78,8%, e melhorou a eficiência da curva de recuperação de fraudes no ranqueamento, sem alterar significativamente a priorização dos casos mais críticos.

Como limitações, destacam-se o tamanho reduzido da base de dados (100 documentos), a categorização binária, e a origem das fraudes confirmadas restrita a uma única operação. Como trabalhos futuros, pretende-se ampliar a base de dados, explorar novas variáveis derivadas dos caminhos da taxonomia, avaliar a abordagem em outros objetos de licitação (pá carregadeira, rolo compactador, trator de esteira, motoniveladora) e outros tipos de documentos técnicos, além de automatizar a geração dos alertas de favoritismo diretamente no painel CEOS.

6 REFERÊNCIAS

PORTAL DA TRANSPARÊNCIA DO GOVERNO FEDERAL. *Licitações*. [S.l.: s.n.], 2025. Disponível em: <https://portaldatransparencia.gov.br/licitacoes>. Brasília: Controladoria-Geral da União. Acesso em: 22 jun. 2025.

Schiavon, J. P. (2024) *Feature Extractor Taxonomy*. Disponível em: <https://github.com/SchiavonJP/Feature-Extractor-Taxonomy>. Acesso em: 6 maio 2026.

Schmitz, F. A. (2025) *Detecção de fraudes em licitações públicas: uma abordagem semissupervisionada de agrupamento com Modelos de Mistura Gaussiana*. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Santa Catarina, Florianópolis. Disponível em: <https://repositorio.ufsc.br/handle/123456789/267575>.

Schneider dos Santos, E. et al. (2025) “Detection of fraud in public procurement using data-driven methods: a systematic mapping study”. *EPJ Data Science*, v. 14, n. 52. DOI: 10.1140/epjds/s13688-025-00569-3.

APÊNDICE B – CÓDIGO-FONTE

Este apêndice apresenta os principais trechos de código utilizados.

```
1 # CONFIGURAÇÕES E CONSTANTES
2 SQLITE_PATH = "analysis.sqlite3"
3 OUTPUT_CSV_PATH = "dataframe_binario_100.csv"
4
5 # Valores reais não são publicados no código-fonte por conterem dados
  sensíveis/identificáveis
6 DOCUMENT_IDS = {
7     # Fraudes
8     "DOC_01": numero_licitacao01,
9     "DOC_02": numero_licitacao02,
10    ...
11 }
12
13 FRAUDE_DOCS = {
14     "DOC_01": numero_licitacao01,
15     ...
16 }
17
18 # FUNÇÕES AUXILIARES DE TEXTO
19 def extract_doc_id(text):
20     """Extraí o identificador de documento do texto baseado em padrão
      Regex."""
21     match = re.search(r"\(([a-zA-ZÀ- 0 -9]+\d{4}_\d+)\)\s*$",
22                       text.strip())
23     return match.group(1) if match else text.strip()
24
25 def normalize_text(text):
26     """Remove acentos e padroniza o texto em minúsculas."""
27     text = unicodedata.normalize("NFKD", text).encode("ASCII",
28               "ignore").decode("utf-8")
29     return text.lower().strip()
30
31 def parse_json_safely(raw_json):
32     """Realiza o parse de strings JSON garantindo o retorno estruturado em
      lista."""
33     try:
34         parsed = json.loads(raw_json)
35     except (json.JSONDecodeError, TypeError):
36         return []
37
38     if isinstance(parsed, dict):
39         return [parsed]
40     return parsed if isinstance(parsed, list) else []
41
```

```
40 # ETAPAS DO PIPELINE
41 def fetch_data_from_sqlite(sqlite_path):
42     """Busca as linhas necessárias do banco SQLite (sem salvar arquivos
43         locais)."""
44     conn = sqlite3.connect(sqlite_path)
45     cur = conn.cursor()
46     query = """
47         SELECT item_name, expansion_json
48         FROM feature_semantic_expansions
49         WHERE taxonomy_run_id IN (1, 3)
50         AND LOWER(item_name) LIKE '%escavadeira%'
51     """
52     cur.execute(query)
53     rows = cur.fetchall()
54     conn.close()
55
56     return rows
57
58 def build_binary_dataframe(rows):
59     """Processa caminhos taxonômicos e cria a matriz binária de
60         features."""
61     all_paths = set()
62
63     for _, raw_json in rows:
64         parsed = parse_json_safely(raw_json)
65         for exp in parsed:
66             path = exp.get("attribute_path") or []
67             path_norm = tuple(normalize_text(p) for p in path)
68             if path_norm:
69                 all_paths.add(path_norm)
70
71     print(f"Total de caminhos distintos: {len(all_paths)}")
72
73     docs = sorted({extract_doc_id(item_name) for item_name, _ in rows})
74     df = pd.DataFrame({"documento": docs})
75
76     col_names = ["path_" + "_".join(p) for p in all_paths]
77     df_paths = pd.DataFrame(0, index=df.index, columns=col_names)
78     doc_index_map = {doc: i for i, doc in enumerate(df["documento"])}
79
80     for item_name, raw_json in rows:
81         doc_id = extract_doc_id(item_name)
82         row_idx = doc_index_map[doc_id]
83         parsed = parse_json_safely(raw_json)
84
85         for exp in parsed:
86             path = exp.get("attribute_path") or []
```

```

86         path_norm = tuple(normalize_text(p) for p in path)
87         col_name = "path_" + "_".join(path_norm)
88         if col_name in df_paths.columns:
89             df_paths.iat[row_idx, df_paths.columns.get_loc(col_name)]
90                 = 1
91
92     return pd.concat([df, df_paths], axis=1).copy()
93
94 def merge_metadata_and_labels(df, document_ids, fraude_docs):
95     """Mapeia IDs de documentos e rotula as classes de fraude
96     conhecidas."""
97     df_ids = pd.DataFrame({
98         "documento": list(document_ids.keys()),
99         "documento_id": list(document_ids.values())
100     })
101
102     df_final = df.merge(df_ids, on="documento", how="left")
103     df_final["documento_id"] =
104         df_final["documento_id"].fillna(0).astype(int)
105     df_final["fraude"] = df_final["documento"].apply(lambda d: 1 if d in
106         fraude_docs else 0)
107
108     return df_final
109
110 def compute_statistical_features(df, path_cols):
111     """Calcula frequência de atributos e similaridade Hamming média geral
112     usando colunas ordenadas."""
113     X = df[path_cols]
114     attr_freq = X.mean(axis=0)
115
116     df["mean_attr_freq"] = (
117         X.multiply(attr_freq, axis=1).sum(axis=1) /
118         X.sum(axis=1).replace(0, 1)
119     )
120
121     dist_matrix = cdist(X.values, X.values, metric='hamming')
122     np.fill_diagonal(dist_matrix, np.nan)
123     df["mean_similarity"] = 1 - np.nanmean(dist_matrix, axis=1)
124
125     return df
126
127 def compute_fraud_profile_scores(df, path_cols):
128     """Calcula scores do perfil de fraude respeitando rigorosamente a
129     ordem geométrica de path_cols."""
130     df["similaridade_fraude"] = 0.0
131     df["score_fraude_like"] = 0.0
132
133     fraudes_idx = df[df["fraude"] == 1].index

```

```
128
129     for idx in df.index:
130         if idx in fraudes_idx:
131             fraudes_temp = df.loc[fraudes_idx.drop(idx)]
132         else:
133             fraudes_temp = df.loc[fraudes_idx]
134
135         if len(fraudes_temp) == 0:
136             continue
137
138         perfil_fraude = fraudes_temp[path_cols].mean()
139         perfil_vetor = perfil_fraude[path_cols].values.reshape(1, -1)
140
141         x = df.loc[idx, path_cols].values.reshape(1, -1)
142
143         df.loc[idx, "similaridade_fraude"] = cosine_similarity(x,
144             perfil_vetor)[0][0]
145         df.loc[idx, "score_fraude_like"] = (df.loc[idx, path_cols] *
146             perfil_fraude).sum()
147
148     return df
149
150 def scale_features(df, cols_to_scale):
151     """Normaliza as colunas numéricas especificadas no intervalo [0, 1]."""
152     scaler = MinMaxScaler()
153     df[cols_to_scale] = scaler.fit_transform(df[cols_to_scale])
154     return df
155
156 # CONTROLADOR PRINCIPAL (EXECUÇÃO DO PIPELINE)
157 def main():
158     # 1. Extração de Dados (Direto da memória do SQLite)
159     rows = fetch_data_from_sqlite(SQLITE_PATH)
160
161     # 2. Construção do DataFrame Base das matrizes
162     df_base = build_binary_dataframe(rows)
163
164     # 3. Integração de metadados e Rótulos (Labels)
165     df_processed = merge_metadata_and_labels(df_base, DOCUMENT_IDS,
166         FRAUDE_DOCS)
167
168     # 4. Engenharia de Features Estatísticas e de Similaridade
169     path_cols = sorted([c for c in df_processed.columns if
170         c.startswith("path_")])
171
172     df_processed = compute_statistical_features(df_processed, path_cols)
173     df_processed = compute_fraud_profile_scores(df_processed, path_cols)
174
175     # 5. Normalização das features geradas
```

```

172     cols_to_scale = ["mean_attr_freq", "mean_similarity",
173                     "similaridade_fraude", "score_fraude_like"]
174     df_final = scale_features(df_processed, cols_to_scale)
175
176     # 6. Salvamento do Resultado Final em CSV
177     df_final.to_csv(OUTPUT_CSV_PATH, index=False, encoding="utf-8")
178     print(f"Pipeline concluído com sucesso! Features finais gravadas em
179           '{OUTPUT_CSV_PATH}'")
180
181     return df_final
182
183 if __name__ == "__main__":
184     df_finalizado = main()

```

Código 1 – Construção de caminhos hierárquicos e novas características.

```

1 def executar_merge():
2     print("Carregando dados locais para o cruzamento...")
3
4     # Ler os arquivos CSV utilizando os seus caminhos relativos específicos
5     df_escavadeira_teste = pd.read_csv('escavadeira.csv')
6     df_escavadeira_teste['Produto'] = 'escavadeira'
7
8     df_novas_colunas = pd.read_csv('dataframe_binario_100.csv')
9
10    df_merge = df_escavadeira_teste.merge(
11        df_novas_colunas,
12        left_on='ID_ItemLicitacao',
13        right_on='documento_id',
14        how='inner'
15    )
16
17    # Tratamento e alinhamento de colunas de fraude clonadas
18    if 'fraude_y' in df_merge.columns:
19        df_merge = df_merge.drop(columns=['fraude_y'])
20    if 'fraude_x' in df_merge.columns:
21        df_merge = df_merge.rename(columns={'fraude_x': 'fraude'})
22
23    # Salva o arquivo de saída unificado
24    df_merge.to_csv("df_merge.csv", index=False, encoding="utf-8")
25    print("Cruzamento concluído com sucesso! Dataset final gerado em:
26          'df_merge.csv'")
27
28 if __name__ == "__main__":
29     executar_merge()
30
31 columns = ["duration", "unique", "win", "period", "num", "met",

```

```
    "num_partic", "unit_price", "mean_similarity", "similaridade_fraude",
    "score_fraude_like", "mean_attr_freq"]
32 dataframes = {'esvacadeira_final': df_merge}
33 thresholds = [80]
34
35 tasks = []
36
37 for df_name, df in dataframes.items():
38     for threshold in thresholds:
39         tasks.append((df_name, df, threshold, None))
40         fraud_ids = df[df['fraude'] == 1]['ID']
41         for fraud_id in fraud_ids:
42             df_temp = df.copy()
43             df_temp.loc[df_temp['ID'] == fraud_id, 'fraude'] = 0
44             tasks.append((df_name, df_temp, threshold, fraud_id))
45
46 def process_main_task(args):
47     df_name, df, threshold, fraud_id = args
48
49     if fraud_id is None:
50         print(f"[PROCESSANDO] Dataset principal - DF: {df_name},
51             threshold: {threshold}")
52         task_type = "dataset principal"
53         main(df, columns, threshold, df_name)
54         result_msg = f"Concluído dataset principal: {df_name} com
55             threshold {threshold}"
56     else:
57         print(f"[PROCESSANDO] Fraude {fraud_id} - DF: {df_name},
58             threshold: {threshold}")
59         task_type = "fraude"
60         main(df, columns, threshold, df_name, fraud_id)
61         result_msg = f"Concluído fraude {fraud_id} para {df_name} com
62             threshold {threshold}"
63
64     return {
65         "df_name": df_name,
66         "threshold": threshold,
67         "fraud_id": fraud_id,
68         "task_type": task_type,
69         "elapsed_seconds": elapsed,
70         "result_message": result_msg
71     }
72
73 results = []
74
75 for args in tasks:
76     df_name, df, threshold, fraud_id = args
```

```

74     print("\n-----")
75     if fraud_id is None:
76         print(f"[PROCESSANDO SEQUENCIAL] Dataset principal - DF:
              {df_name}, threshold: {threshold}")
77     else:
78         print(f"[PROCESSANDO SEQUENCIAL] Fraude {fraud_id} - DF:
              {df_name}, threshold: {threshold}")
79     print("-----")
80
81     result = process_main_task(args)
82
83     results.append(result)

```

Código 2 – Integração entre dataframe binário e SubFraudGMM.

```

1  DIRETORIO = 'DIRETORIO_DADOS_SAIDA_SUBFRAUDGMM'
2  PADRAO = re.compile(r'^.+_80_ESC.*\.csv$')
3
4  def calculate_f1_score(precision, recall):
5      return np.where(
6          (precision + recall) > 0,
7          2 * (precision * recall) / (precision + recall),
8          0,
9      )
10
11 def normalize_column(column):
12     return (column - column.min()) / (column.max() - column.min())
13
14 def count_features(feature_series):
15     all_features = []
16     for row in feature_series:
17         all_features.extend(f.strip() for f in row.split(','))
18     return dict(Counter(all_features))
19
20 def merge_dicts(series):
21     result = {}
22     for d in series:
23         if isinstance(d, dict):
24             for k, v in d.items():
25                 result[k] = result.get(k, 0) + v
26     return result
27
28 def processar_arquivo(nome_arquivo):
29     df = pd.read_csv(
30         os.path.join(DIRETORIO, nome_arquivo),
31         delimiter=',',
32         encoding='utf-8',
33     )

```

```

34
35     nome_sem_ext = nome_arquivo.replace('.csv', '')
36     parametros = nome_sem_ext.split('_')
37
38     produto = parametros[0].replace('-', ' ')
39     df['Produto'] = produto.capitalize()
40     df['Threshold'] = pd.to_numeric(parametros[2], errors='coerce')
41     df['FraudCV'] = parametros[3] if len(parametros) > 3 else ''
42
43     df['fraude'] = df['fraude'].apply(lambda x: 'Sim' if x == 1 else 'Não')
44     df['num_features'] = df['features'].apply(
45         lambda x: x.count(',') + 1 if pd.notna(x) else 0
46     )
47     df['Tamanho_Cluster'] =
48         df.groupby('features')['features'].transform('count')
49
50     return df
51
52 def carregar_arquivos(padrao=PADRAO):
53     arquivos_csv = [f for f in os.listdir(DIRETORIO) if padrao.match(f)]
54     df = pd.concat([processar_arquivo(a) for a in arquivos_csv],
55                     ignore_index=True)
56
57     df['Subespaço'] = (
58         df['features'].astype(str) + '|'
59         + df['Produto'].astype(str) + '|'
60         + df['Threshold'].astype(str)
61     )
62     df['Subespaço_ID'] = df['Subespaço'].astype('category').cat.codes + 1
63     df['Métrica_Desempenho'] = df['num_fraud_reclustered'] /
64         df['Tamanho_Cluster']
65     df['f1_score'] = calculate_f1_score(df['Métrica_Desempenho'],
66         df['loo'])
67
68     df['Normalized Distance'] =
69         normalize_column(df['euclidean_distance_to_fraud'])
70     df['Normalized Subsets'] = normalize_column(df['ocorrencias'])
71     df['Normalized L00'] = normalize_column(df['loo'])
72     df['Risk Indicator'] = (
73         (1 - df['Normalized Distance']) + df['Normalized Subsets'] +
74         df['Normalized L00']
75     ) / 3
76
77     return df
78
79 def agregar_por_id(df):
80     df_mean = df.groupby(
81         ['ID', 'Threshold', 'FraudCV', 'Produto', 'numeroEditais',

```

```

76         'Ente', 'ID_ProcedimentoLictatorio', 'fraude']
77     ).agg({
78         'Normalized Distance': 'mean',
79         'Normalized Subsets': 'max',
80         'Normalized L00': 'mean',
81         'win': 'mean',
82         'unique': 'mean',
83         'period': 'mean',
84         'num_partic': 'mean',
85         'num': 'mean',
86         'unit_price': 'mean',
87         'duration': 'mean',
88         'met': 'mean',
89         'features': count_features,
90     }).reset_index()
91
92     df_mean['Risk Indicator'] = (
93         (1 - df_mean['Normalized Distance'])
94         + df_mean['Normalized Subsets']
95         + df_mean['Normalized L00']
96     ) / 3
97
98     df_mean['Rank'] = df_mean.groupby(
99         ['Produto', 'FraudCV', 'Threshold']
100     )['Risk Indicator'].rank(method='min', ascending=False)
101
102     df_mean['Grupo'] = df_mean.apply(
103         lambda row: 'RT' if row['ID'] == row['FraudCV'] else 'Demais',
104         axis=1
105     )
106
107     return df_mean
108
109 def extract_top_features(row):
110     feature_dict = row['features']
111     if not isinstance(feature_dict, dict):
112         return None
113
114     top = sorted(feature_dict.items(), key=lambda x: x[1],
115                 reverse=True)[:3]
116     top_features = [f[0] for f in top]
117     top_values = [feature_dict.get(f) for f in top_features]
118
119     return pd.Series({
120         'ID': row['ID'],
121         'Produto': row['Produto'],
122         'Risk Indicator': row['Risk Indicator'],
123         'Fraude': row['fraude'],

```

```

122     'Normalized Distance': row['Normalized Distance'],
123     'Normalized Subsets': row['Normalized Subsets'],
124     'Normalized L00': row['Normalized L00'],
125     '3 Principais Características': ', '.join(top_features),
126     'Característica 1': top_values[0] if len(top_values) > 0 else None,
127     'Característica 2': top_values[1] if len(top_values) > 1 else None,
128     'Característica 3': top_values[2] if len(top_values) > 2 else None,
129 })
130
131 def construir_resultado_final(df_mean):
132     df_id = df_mean.groupby(['ID', 'Produto']).agg({
133         'Risk Indicator': 'max',
134         'fraude': 'max',
135         'features': merge_dicts,
136         'Normalized Distance': 'mean',
137         'Normalized Subsets': 'mean',
138         'Normalized L00': 'mean',
139     }).reset_index()
140
141     resultados = []
142     for _, group in df_id.groupby('Produto'):
143         top100 = group.sort_values(by='Risk Indicator',
144                                   ascending=False).head(100)
145         enriched = top100.apply(extract_top_features, axis=1)
146         resultados.append(enriched)
147
148     return pd.concat(resultados, ignore_index=True)
149
150 def main():
151     df_total = carregar_arquivos()
152     df_mean = agregar_por_id(df_total)
153     final_df = construir_resultado_final(df_mean)
154
155     return final_df
156
157 if __name__ == '__main__':
158     resultado = main()
159     print(resultado)

```

Código 3 – Análise dos dados após SubFraudGMM.