Identificação de Irregularidades em Contratos Públicos por Meio de Aprendizado de Máquina: Um Estudo com Dados do e-ContratosDF

Sara Borges L. Sousa (252107073)
Programa de Pós-Graduação em Computação APicada - PPCA
Departamento de Computação, Universidade de Brasília (UnB), Brasília, Brasil
Prof. Dr. Marcelo Ladeira (mladeira@unb.br) and Pr. MSc. Gustavo Cordeiro G. V. Erven (gvanerven@gmail.com)
Email: sara.blsousa@gmail.com

I. Contextualização do Problema

A forma como o Brasil gerencia seus contratos públicos afeta diretamente a qualidade de vida da população, já que centenas de bilhões de reais são destinados anualmente a áreas essenciais, como saúde, educação e infraestrutura. Para se ter uma ideia do impacto financeiro atrelado à má gestão governamental, estimativas do Banco Interamericano de Desenvolvimento apontam que a simples otimização dessas compras poderia impulsionar o Produto Interno Bruto (PIB) de países em desenvolvimento em até 4%. Contudo, mesmo com o notório avanço de leis focadas em transparência e responsabilidade fiscal, a exemplo da Lei de Acesso à Informação e da histórica Nova Lei de Licitações, a fiscalização prática na ponta da linha ainda sofre com barreiras estruturais severas.

O principal obstáculo é que o controle tradicional se mostra engessado: atua de forma reativa, quase sempre depois que o desvio já ocorreu, e limita-se a analisar pequenas amostras de processos devido à sobrecarga operacional dos auditores. Um exemplo nítido desse gargalo ocorre no Governo do Distrito Federal (GDF). A plataforma e-ContratosDF, utilizada por 107 unidades administrativas, acumula um acervo que ultrapassa a marca de 17 mil registros. Nesse cenário, o fiscal humano precisa garimpar indícios de fraude em um verdadeiro labirinto de formatos desconexos, lidando com códigos estruturados (JSON e bancos SQL), planilhas eletrônicas corporativas (XLSX), sistemas restritos de business intelligence (QVD) e milhares de páginas de editais em PDF.

É humanamente impossível monitorar um volume de informações tão massivo e heterogêneo de forma manual. Para contornar essa barreira, o texto propõe a adoção de uma esteira analítica automatizada, capaz de ingerir e auditar o universo completo de contratações do GDF de uma só vez. Essa arquitetura tecnológica orquestra ferramentas avançadas de forma inteligente: utiliza sistemas de computação distribuída, como Apache Spark e Dask, para viabilizar o processamento em larga escala; emprega a inteligência artificial do BERTimbau para varrer e interpretar a linguagem oculta em textos não estruturados; e aplica modelos matemáticos preditivos (gradient boosting) para calcular a pontuação de risco de cada transação. Toda essa estrutura foi arquitetada em um ambiente de nuvem, garantindo que o rastreamento permaneça veloz e perfeitamente escalável à medida que o governo firme novos contratos com o passar dos anos.

II. Objetivos

O grande objetivo deste projeto é criar e validar um sistema inteligente capaz de identificar, de forma automática, suspeitas de irregularidades nos contratos públicos do Governo do Distrito Federal (GDF). Para tornar isso realidade, a proposta é construir uma esteira de processamento de dados que conecte e organize informações de diversas origens. Essa estrutura vai buscar dados diretamente da plataforma e-ContratosDF e cruzá-los com outras fontes de informação do governo. O desafio aqui é lidar com registros organizados de maneiras muito diferentes: desde bancos de dados e planilhas tradicionais (como SQL e XLSX), passando por formatos mais flexíveis (como o JSON), até chegar a documentos complexos e soltos, como editais em PDF e arquivos QVD. A grande inovação está na forma como esses dados serão analisados, indo muito além do cruzamento básico de planilhas. O sistema combinará inteligência artificial, utilizando técnicas de aprendizado supervisionado e não supervisionado, com o mapeamento das redes de relacionamentos contratuais. Além disso, a ferramenta empregará modelos avançados de Processamento de Linguagem Natural (baseados na tecnologia BERT) para literalmente ler e interpretar o contexto dos textos nos documentos. A meta central de todo esse esforço técnico é dar um salto de qualidade no trabalho investigativo da Controladoria-Geral do Distrito Federal (CGDF). Com essa tecnologia, o órgão poderá fiscalizar todo o ciclo de vida dos contratos da administração pública de forma muito mais rápida, com maior alcance e precisão, protegendo de maneira mais eficaz os recursos da sociedade

A. Objetivos Específicos

Na prática, a pesquisa funciona como uma linha de montagem inteligente que começa pela coleta e unificação automática das informações do e-ContratosDF e de outras bases

do governo, passando sempre por rigorosos testes de qualidade. Como o volume de dados é gigantesco, a solução orquestra tecnologias de processamento em larga escala, como Apache Spark e Dask, para conseguir calcular mais de cem indicadores de risco de forma simultânea. É nessa etapa que a análise ganha profundidade: enquanto a inteligência artificial do BERTimbau literalmente "lê" os editais para identificar cláusulas que restrinjam a concorrência, um sistema de grafos mapeia as relações entre as empresas para descobrir conexões ocultas e denunciar possíveis conluios. Todo esse rico volume de informações deságua no treinamento de algoritmos preditivos, que utilizam dados sintéticos (por meio da técnica SMOTE) para aprender a identificar anomalias com precisão, contornando o desafio de as fraudes serem estatisticamente raras. Para garantir que a tecnologia não seja uma "caixa preta" incompressível, a arquitetura é guiada por diretrizes éticas e utiliza frameworks de interpretabilidade, como o SHAP, garantindo que cada alerta gerado pela máquina seja explicado de forma clara, operando como um radar transparente e auditável sob a supervisão constante do fiscal humano.

- **OE1: Ingestão e Coleta Sistemática de Dados via API.** Para viabilizar a auditoria em larga escala, o primeiro desafio metodológico consiste em construir uma esteira automatizada de coleta de informações conectada diretamente à interface de programação (API) do e-ContratosDF. Como o acervo distribuído entre as cento e sete unidades gestoras do governo é massivo e sofre alterações constantes, a ferramenta de extração foi projetada com navegação iterativa e mecanismos de proteção robustos. Para evitar interrupções no fluxo causadas por instabilidades no servidor, implementou-se uma estratégia de tentativas progressivas, conhecida como exponential backoff, garantindo que o sistema gerencie falhas de rede de forma autônoma e resiliente. Além de assegurar a estabilidade da conexão, a arquitetura evita o retrabalho e o acúmulo de dados redundantes ao utilizar códigos de verificação (hashes de integridade), os quais permitem que o sistema reconheça e incorpore apenas as informações inéditas ou modificadas. Por fim, todo esse volume bruto capturado é preservado no formato Parquet dentro de um repositório organizado em estágios evolutivos de maturidade (camadas bronze, silver e gold), consolidando uma fundação de dados íntegra, segura e permanentemente atualizada para nutrir os algoritmos preditivos nas fases seguintes.
- **OE2: Integração Multiformato e Construção do Dataset Unificado.** Avançando na esteira computacional, a etapa seguinte concentra-se na captura e padronização de cinco vertentes informacionais que complementam os registros originais da plataforma de contratos. Na prática, a arquitetura precisa traduzir um autêntico mosaico de formatos para uma base coesa, absorvendo desde o fluxo de pagamentos do sistema relacional SIGGO e planilhas gerenciais descentralizadas, até o texto de milhares de editais em PDF, painéis de inteligência de negócios (arquivos QVD) e listagens de sanções empresariais. Para amarrar todo esse quebra-cabeça em um banco de dados unificado e sem duplicidades, a pesquisa adota uma estratégia de cruzamento alicerçada em uma chave tripla exata, conectando o número do instrumento, o CNPJ da instituição contratante e o respectivo exercício financeiro. Por fim, como a mescla de origens tão diversas eleva naturalmente o risco de inconsistências, o processo assegura a confiabilidade do acervo valendo-se da biblioteca Great Expectations, que atua como um inspetor automatizado para validar a saúde e a integridade de cada novo lote de dados antes que eles sigam para a análise matemática
- **OE3: Processamento Distribuído e Escalável com Spark e Dask.** Avançando para a fase de transformação e criação de variáveis analíticas, a pesquisa estrutura uma engenharia de dados capaz de processar registros estruturados e semiestruturados sem criar gargalos de performance. Para dar conta da volumetria do acervo, a arquitetura divide as tarefas computacionais de forma inteligente. O trabalho mais pesado recai sobre o Apache Spark, que assume as operações em larga escala, como o cruzamento massivo entre as tabelas de contratos, empenhos financeiros e fornecedores, além de calcular métricas de risco cruciais, a exemplo do grau de concentração de mercado (índice HHI) e da variação financeira causada por aditivos. Paralelamente, o sistema delega as demandas de porte intermediário para a tecnologia Dask, que se encarrega de ler planilhas eletrônicas de forma simultânea e de aplicar em lote a inteligência artificial do modelo BERT sobre os textos extraídos. Essa divisão cirúrgica de esforços assegura que a plataforma seja horizontalmente escalável, mantendo a estabilidade e a rapidez analítica mesmo à medida que novos registros continuem inflando a base de compras do governo nos próximos anos
- **OE4: Extração de Entidades e Análise Semântica com BERTimbau.** Para lidar com a complexidade inerente aos dados textuais não estruturados, a pesquisa estrutura a criação de um módulo avançado de Processamento de Linguagem Natural. O motor dessa análise é o BERTimbau, um modelo de inteligência artificial robusto que já absorveu mais de dois bilhões de palavras do vocabulário brasileiro. Na prática, essa tecnologia atua como um inspetor incansável, capaz de vasculhar as páginas de contratos e editais em formato PDF para pinçar automaticamente informações cruciais, como as empresas envolvidas, os valores financeiros, os prazos e as penalidades previstas. Para ir muito além da simples extração de dados, os pesquisadores calibraram o algoritmo utilizando o próprio acervo do governo, ensinando a máquina a reconhecer sutilezas e intenções nas entrelinhas. Com isso, o sistema ganha a habilidade de classificar se a descrição de uma compra é propositalmente vaga ou se o texto impõe exigências que restringem a livre concorrência. Fechando esse ciclo de auditoria inteligente, a arquitetura compara o sentido das frases das novas publicações com licitações antigas, gerando um índice

matemático de similaridade que funciona como um radar para detectar se um edital foi "reciclado" com o objetivo de direcionar a contratação para um fornecedor específico

- **OE5: Engenharia de Features e Construção dos Indicadores de Risco.** O núcleo analítico do projeto consiste em calcular e validar uma série de indicadores estratégicos que apontam para o risco de irregularidades. Para que o sistema consiga avaliar a lisura de cada acordo governamental, ele monitora tanto fatores quantitativos quanto qualitativos, avaliando indícios de sobrepreço, o excesso ou o impacto financeiro das renovações por aditivos, e a proporção exata entre os recursos empenhados e os que foram efetivamente liquidados. A ferramenta também investiga o comportamento das unidades gestoras, detectando se um órgão concentra seus contratos em um número suspeitamente pequeno de fornecedores ou se fragmenta as compras para burlar as exigências tradicionais de licitação. Somado a isso, o algoritmo acende alertas para trâmites anormais, como demoras injustificadas na publicação no diário oficial, empresas recém-criadas assumindo serviços públicos, ou até mesmo o grave conflito de interesses verificado quando um servidor público atua como sócio da empresa contratada. Ao fim desse intenso mapeamento de variáveis, consolida-se uma base de dados riquíssima e detalhada, na qual cada contrato passa a ser radiografado por mais de 120 atributos distintos
- **OE6: Análise de Redes Contratuais por Grafos.** Para desvendar conexões suspeitas que muitas vezes passam despercebidas, o projeto constrói um verdadeiro mapa de relacionamentos interligando os órgãos do governo, as empresas fornecedoras, os servidores responsáveis pela fiscalização e o próprio objeto de cada contrato. A partir dessa grande teia de informações, a ferramenta consegue calcular o nível de influência e a posição estratégica de cada fornecedor, aplicando algoritmos inteligentes de detecção de comunidades para identificar grupos que atuam de forma excessivamente próxima. O grande trunfo dessa tecnologia de análise de grafos é a sua capacidade de cruzar dados rapidamente para emitir alertas automáticos sempre que empresas que deveriam estar competindo em uma mesma licitação compartilham o mesmo endereço, sócio ou representante legal com a organização que venceu a disputa. Ao expor essas sobreposições, o sistema joga luz sobre padrões clássicos de conluio, desmascarando fraudes e falsas competições que, de outra forma, ficariam escondidas na burocracia do processo
- **OE7: Modelagem Preditiva e Detecção de Anomalias.** Treinar, ajustar e comparar modelos de aprendizado de máquina supervisionado, XGBoost, Random Forest e redes neurais densas, e não supervisionado de detecção de anomalias, Isolation Forest, DBSCAN e Autoencoder, avaliando os modelos supervisionados por métricas de *precision*, *recall*, F1-Score e AUC-ROC com validação cruzada estratificada, e adotando a estratégia combinada de SMOTE e ponderação de classes para mitigação do desbalanceamento severo característico de bases de dados de irregularidades, onde registros positivos tipicamente representam menos de 5% do total.
- **OE8: Interpretabilidade e Auditabilidade dos Modelos.** A etapa de modelagem preditiva desta pesquisa apoia-se no treinamento, na otimização de parâmetros e na comparação analítica rigorosa entre diferentes algoritmos de aprendizado de máquina. Sob a ótica do aprendizado supervisionado, o estudo selecionou arquiteturas consagradas na literatura, especificamente o XGBoost, o Random Forest e as redes neurais densas, com o propósito de mapear os padrões históricos de risco já existentes na base de dados. Simultaneamente, com o objetivo de rastrear comportamentos atípicos e desvios que não possuam rotulagem prévia, a metodologia incorpora métodos não supervisionados desenhados para a detecção de anomalias, fundamentando-se nas abordagens de Isolation Forest, DBSCAN e Autoencoder. Para assegurar a confiabilidade estatística e a real capacidade de generalização dos resultados, os classificadores supervisionados são submetidos ao método de validação cruzada estratificada. A aferição do desempenho dessas ferramentas ocorre por meio de indicadores de avaliação robustos, valendo-se das métricas de precisão (precision), sensibilidade (recall), média harmônica (F1-Score) e a área sob a curva ROC (AUC-ROC). Um obstáculo metodológico intrínseco à auditoria de informações governamentais é o desbalanceamento severo das classes estatísticas, um cenário no qual os registros de irregularidades representam, na maior parte dos casos, uma fração inferior a 5% da totalidade analisada. Vale pontuar que, no escopo específico dos dados deste estudo, a taxa de irregularidades confirmadas atingiu a marca de 12,71%. Ainda assim, para mitigar esse forte viés e impedir que os algoritmos negligenciem a classe minoritária em favor do padrão de normalidade, a investigação aplica uma estratégia corretiva híbrida. Essa solução combina a técnica de sobreamostragem sintética (SMOTE) com a ponderação matemática de classes durante a fase de treinamento da máquina.
- **OE9: Avaliação Ética, Vieses e Governança Algorítmica.** Para garantir a transparência e a plena auditabilidade das predições elaboradas pelos modelos supervisionados, a arquitetura metodológica fundamenta sua análise na ferramenta SHAP (SHapley Additive exPlanations). Essa estrutura analítica viabiliza a construção de explicações nas dimensões local e global, o que permite quantificar com precisão o peso de cada variável na composição do risco atribuído a um contrato individual. No contexto da Administração Pública, a explicabilidade dos algoritmos supera a simples exigência técnica, estabelecendo a base fundamental para a aplicação ética e responsável da inteligência artificial. Ao traduzir formulações matemáticas complexas em evidências claras, a solução instrumentaliza os auditores da Controladoria-Geral do Distrito Federal (CGDF), oferecendo o suporte adequado para que a equipe possa avaliar criticamente,

questionar e validar as indicações de risco geradas pelo sistema computacional. Além dessa camada de interpretabilidade, a pesquisa estabelece protocolos rigorosos de avaliação ética e governança algorítmica. O estudo prevê a análise sistemática de potenciais vieses e o monitoramento contínuo da distribuição de erros empíricos, com atenção especial aos falsos positivos, segmentando essas ocorrências por órgão contratante, porte empresarial e categoria do objeto. Esse cuidado metodológico assegura que as predições da máquina operem estritamente como um subsídio investigativo qualificado, jamais como uma evidência definitiva de irregularidade. Com essa abordagem, o projeto alinha a inovação tecnológica às diretrizes de uso responsável exigidas pelos órgãos de controle governamental

III. Descrição dos Dados

A. Fonte Primária: API REST e-ContratosDF

A base de dados principal desta pesquisa provém do e-ContratosDF, o ambiente oficial de gestão contratual do Governo do Distrito Federal, cuja utilização foi regulamentada pela Portaria nº 314/2018 e estabelecida como obrigatória a partir do Decreto nº 40.447 de 2020. A coleta dessas informações ocorreu mediante acesso autenticado à interface de programação da plataforma, a qual fornece os metadados estruturados no formato JSON e abrange o acervo de 107 unidades gestoras atualmente em operação. Durante a etapa de extração, o repositório contava com um montante de 17.885 acordos firmados, categorizados em sete diferentes estágios de tramitação. Observou-se uma predominância expressiva de contratos já vencidos, representando mais da metade da amostra (56,09%), seguidos pelas categorias de termos quitados (16,50%) e publicados (16,11%). A parcela remanescente do conjunto de dados dividia-se entre registros apenas cadastrados (6,13%) e aqueles efetivamente em execução (4,94%), restando uma fração estatisticamente marginal de ocorrências rescindidas (0,23%) ou suspensas (0,01%)

TABLE I
Distribuição de Contratos por Status no e-ContratosDF

Status	Quantidade	Percentual
Cadastrados	1.096	6,13%
Em Execução	883	4,94%
Publicado	2.880	16,11%
Quitado	2.950	16,50%
Rescindido	42	0,23%
Suspenso	1	0,01%
Vencidos	10.033	56,09%
Total	**17.885**	**100,00%**

B. Fontes Complementares

Para enriquecer e aprofundar a análise da base de dados primária, a pesquisa incorpora informações estratégicas provenientes de cinco fontes secundárias distintas. O detalhamento financeiro, por exemplo, é extraído do Sistema Integrado de Gestão Governamental (SIGGO), cujos registros de execução são acessados por meio de consultas SQL parametrizadas com o uso de um conector SQLAlchemy. Todo esse volume financeiro é posteriormente processado na plataforma Apache Spark, o que permite agregar variáveis cruciais de empenho, liquidação e pagamento que não constam na interface de programação original do governo. Em paralelo, a metodologia abrange o tratamento rigoroso de arquivos corporativos descentralizados. Planilhas de controle interno mantidas pelas unidades gestoras são absorvidas pelo sistema utilizando a biblioteca Pandas e padronizadas para o formato otimizado Parquet. Adicionalmente, arquivos no formato QVD, extraídos do ambiente de inteligência de negócios QlikSense e QlikView, são processados para fornecer um histórico consolidado de indicadores de desempenho governamental. A extração de informações textuais ganha contornos técnicos específicos no tratamento de documentos em formato PDF. Para realizar a leitura de editais e contratos digitais nativos, aplica-se a ferramenta pdfplumber, ao passo que os arquivos digitalizados como imagem passam pelo processo de reconhecimento óptico de caracteres com o suporte do Tesseract OCR. Por fim, o arcabouço de dados é fortalecido pela integração com os cadastros de empresas punidas mantidos pela Controladoria-Geral da União, especificamente o Cadastro de Empresas Inidôneas e Suspensas (CEIS) e o Cadastro Nacional de Empresas Punidas (CNEP). O cruzamento minucioso dessas bases federais com o CNPJ dos fornecedores permite gerar marcadores binários essenciais, os quais alertam o sistema sobre a inidoneidade prévia das empresas envolvidas nos processos de contratação

C. Grupos de Variáveis e Engenharia de Características

A base de dados analítica consolidada ao final de todo o processo de extração e tratamento resulta em uma matriz estruturada e altamente detalhada, contemplando cento e vinte atributos distintos para cada instrumento contratual avaliado. Para otimizar a etapa de modelagem preditiva e garantir uma trilha lógica de interpretação para os algoritmos de inteligência artificial, esse amplo volume de características foi metodologicamente organizado em cinco grandes eixos funcionais. Essa arquitetura de categorização estratégica é fundamental para agrupar as variáveis de acordo com a sua natureza e o seu propósito analítico, pavimentando o caminho para uma auditoria governamental precisa, fluida e cientificamente embasada. Agrupamento por naturea:

1) Grupo I: cobre a identificação e estrutura do contrato.

2) Grupo II: descreve as partes contratantes, incluindo CNPJ do fornecedor, porte empresarial, data de registro, sinalizações CEIS/CNEP e a flag fl_socio_servidor, que indica sobreposição entre sócios do fornecedor e servidores públicos da agência.

3) Grupo III: contém valores financeiros.

4) Grupo IV: agrega oito indicadores de risco engenheirados: o índice de sobrepreço (idx_sobrepreco), índice de variação de aditivos (idx_variacao_aditivo), contagem de aditivos (qt_aditivos), razão liquidação-empenho (idx_liquidacao_empenho), índice de concentração Herfindahl-Hirschman (idx_concentracao_hhi), índice de

fragmentação de contrato (idx_fragmentacao), atraso de publicação em dias (dias_publicacao_apos_assinatura) e flag de empresa nova (fl_empresa_nova).

5) Grupo V: contém características derivadas de PLN do BERTimbau:

- Transformar a descrição do objeto contratual em um vetor numérico de 768 dimensões (embedding), representando o significado do texto. Contratos semanticamente equivalentes, mesmo com redações diferentes, ficam próximos no espaço vetorial, permitindo detectar possível fracionamento artificial de compras.
- Medir o grau de similaridade entre um edital novo e editais anteriores do mesmo órgão. Similaridade muito alta (próxima de 1,0) pode indicar reutilização do edital e possível direcionamento para o mesmo fornecedor.
- Classificar se a descrição do objeto é excessivamente vaga (flag `fl_objeto_vago`) ou se o edital contém cláusulas excessivamente restritivas (`fl_clausula_restritiva`), o que pode indicar irregularidades.

D. Variável Alvo

Para dar sustentação ao modelo de aprendizado supervisionado, a variável principal de predição da pesquisa, identificada tecnicamente como fl_irregularidade_confirmada, foi construída por meio do cruzamento rigoroso de três fontes oficiais de rotulagem. O mapeamento metodológico extraiu evidências de relatórios públicos de auditoria elaborados pela Corregedoria-Geral e pelo Tribunal de Contas do Distrito Federal, além de incorporar decisões de processos administrativos disciplinares já finalizados e registros punitivos das bases federais CEIS e CNEP associados diretamente aos certames avaliados. Ao término dessa consolidação, a matriz de dados revelou que 12,71% dos casos analisados apresentavam inconsistências ou fraudes comprovadas. Visto que esse percentual configura um cenário de agudo desbalanceamento estatístico entre os contratos regulares e os problemáticos, a estratégia do estudo demandou a aplicação da técnica de sobreamostragem sintética (SMOTE). Essa medida corretiva foi implementada especificamente na fase de treinamento algorítmico, nivelando a proporção das classes para impedir que a inteligência artificial desenvolvesse vieses preditivos e garantindo a confiabilidade da ferramenta de auditoria

E. Pipeline de Integração e Qualidade de Dados

A consolidação de todas as fontes de informação fundamenta-se na utilização de uma chave de integração composta. Esse elemento central unifica o número do contrato, o Cadastro Nacional da Pessoa Jurídica (CNPJ) do órgão vinculante e o respectivo exercício fiscal, servindo como o eixo definitivo de cruzamento de dados no ambiente de processamento Apache Spark. Para atestar a integridade contínua dessa massa de informações, a arquitetura implementa a biblioteca computacional Great Expectations, responsável por executar rigorosas validações automáticas a cada lote ingerido no sistema. Esse mecanismo avalia diversos critérios de qualidade, garantindo o preenchimento dos campos obrigatórios, a coerência das tipagens, a conformidade matemática das faixas de valores monetários, a exclusividade dos identificadores e a autenticidade jurídica dos CNPJs por meio da conferência de dígitos verificadores. Após o término de toda a fase de engenharia de características, o conjunto analítico ganha corpo e atinge um escopo superior a 120 atributos individuais por contrato. Toda essa pluralidade de dados distribui-se de forma organizada e estratégica, compreendendo 32 variáveis numéricas contínuas, 18 categóricas nominais, 14 lógicas (booleanas) e oito marcadores temporais. Para adequar as informações mais complexas ao aprendizado da máquina, a matriz acolhe as seis métricas resultantes do mapeamento de grafos e as representações semânticas textuais, as quais passam por uma redução de dimensionalidade utilizando a Análise de Componentes Principais (PCA), enxugando os 768 vetores originais para 50 componentes perfeitamente consumíveis pelos algoritmos preditivos tabulares. Por fim, o escopo metodológico da pesquisa é sintetizado em um quadro referencial que detalha a origem de cada base, os formatos de arquivos abordados, as ferramentas de ingestão empregadas e a contribuição exata de cada fonte para a consolidação do repositório final.

TABLE II
SÍNTESE DAS FONTES DE DADOS, FORMATOS E CONTRIBUIÇÕES AO DATASET ANALÍTICO

Fonte	Formato	Ingestão	Processamento	Contribuição Principal
API e-ContratosDF	JSON	requests / aiohttp	Dask / Spark	Metadados contratuais (Grupos I–II)
SIGGO (execução financeira)	SQL relacional	SQLAlchemy	Apache Spark	Valores de empenho, liquidação e pagamento (Grupo III)
Editais e Contratos	PDF	pdfplumber / Tesseract	BERTimbau / SBERT	Embeddings e entidades NER (Grupo V)
Planilhas corporativas	XLSX / CSV	pandas → Parquet	Dask	Complementação de metadados
BI Corporativo GDF	QVD (proprietário)	Biblioteca qvd	Dask	Histórico de indicadores de desempenho
CEIS / CNEP (CGU)	CSV	pandas	Dask	Flags de inidoneidade (Grupo II)