Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Score de Risco para Priorização de Auditoria em Contratos Públicos: Uma abordagem com Inteligência Artificial Explicável (XAI).

Nilson Romero Michiles Junior

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. João Gabriel de Moraes Souza

Brasília
2025

Ficha Catalográfica de Teses e Dissertações

Está página existe apenas para indicar onde a ficha catalográfica gerada para dissertações de mestrado e teses de doutorado defendidas na UnB. A Biblioteca Central é responsável pela ficha, mais informações nos sítios:

http://www.bce.unb.br
http://www.bce.unb.br/elaboracao-de-fichas-catalograficas-de-teses-e-dissertacoes

Esta página não deve ser inclusa na versão final do texto.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Score de Risco para Priorização de Auditoria em Contratos Públicos: Uma abordagem com Inteligência Artificial Explicável (XAI).

Nilson Romero Michiles Junior

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. João Gabriel de Moraes Souza (Orientador)
CIC/UnB

Prof. Dr. Peng Yaohao
Universidade de Brasília

Prof. Dr. João Carlos Félix Souza
Universidade de Brasília

Prof. Dr. Ticiana Linhares Coelho da Silva
Universidade Federal do Ceará

Prof. Dr. Edna Dias Canedo
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 18 de dezembro de 2025

Dedicatória

Dedico este trabalho, em primeiro lugar, aos meus filhos, cuja curiosidade, entusiasmo e presença constante renovam diariamente em mim o compromisso com um futuro mais justo e promissor. São fontes de inspiração e lembram-me da importância de agir com responsabilidade, integridade e propósito.

A minha esposa, pela parceria incondicional ao longo desta jornada. Sua paciência nos momentos de ausência, sua compreensão diante dos desafios e sua presença constante foram fundamentais para a realização deste percurso.

A minha família, especialmente a minha mãe, é um exemplo de dedicação, coragem e confiança. Mesmo à distância, ela sempre acreditou nas minhas escolhas e me incentivou a seguir com determinação. Sua força e apoio foram essenciais em todas as etapas desta trajetória.

Estendo esta dedicatória a todos que, de diferentes formas, contribuíram para a concretização deste trabalho. Cada gesto de apoio, cada palavra de incentivo e cada momento de compreensão desempenharam um papel importante na superação dos desafios enfrentados.

Agradecimentos

A construção desta dissertação foi marcada por uma trajetória de múltiplos aprendizados e colaborações valiosas. Cada pequena contribuição integrou um percurso maior — uma jornada coletiva que desejo reconhecer com gratidão.

Agradeço ao meu orientador, Professor Dr. João Gabriel de Moraes Souza, pelo apoio constante, pela disponibilidade em todos os momentos em que precisei e pela confiança no desenvolvimento deste trabalho. A orientação técnica e humana recebida foi fundamental.

Aos demais professores do Programa de Pós-Graduação em Computação Aplicada (PPCA), especialmente àqueles que atuam na linha de pesquisa em Gestão de Riscos e nos módulos comuns a todas as linhas de pesquisa, expresso minha sincera gratidão. Seus ensinamentos foram decisivos para consolidar meu conhecimento técnico e ampliar minha visão crítica, contribuindo diretamente para as diversas entregas realizadas ao longo da jornada acadêmica.

Aos colegas de pesquisa Marco Schyeres (St. Gallen University) e Qing (Rutgers University), agradeço pelo rico intercâmbio de ideias e pelo aprendizado compartilhado na fronteira entre Inteligência Artificial e Auditoria — uma experiência que elevou significativamente o rigor e a relevância deste trabalho.

Aos amigos e colegas de trabalho, Carlos Jesus e Gutemberg Assunção, agradeço pelo incentivo constante e pelo suporte prático em diversas etapas. Agradeço também aos colegas de trabalho da coordenação de empresarial e pessoas pela compreensão quando atuavam de forma autônoma e proativa durante os períodos mais intensos.

O propósito de inovar no setor público não é uma tarefa simples — é uma missão para aqueles que não estão satisfeito com o "sempre foi assim". Ao combinar dados, conhecimento e coragem, é possível superar opiniões e construir soluções concretas. Que esta pesquisa reflita minha escolha de empreender com propósito dentro do Estado, acreditando que decisões públicas podem — e devem — ser mais justas, eficientes e baseadas em evidências. Como bem disse W. Edwards Deming: “Sem dados, você é apenas mais uma pessoa com uma opinião.”

Resumo

A priorização eficiente de auditorias em contratos públicos é fundamental para combater a corrupção, melhorar a governança e otimizar o uso dos recursos públicos. Este trabalho apresenta uma metodologia inovadora, organizada em um framework que integra aprendizado de máquina e explicabilidade de inteligência artificial (XAI) para classificar contratos e fornecedores com maior risco de irregularidades. Foram utilizadas bases públicas para a construção do dataset, seguidas por etapas de pré-processamento, balanceamento(SMOTE, ADASYN, TOMEK Link e variações), seleção de atributos via LASSO e ajuste de Hiperparâmetros. Entre os modelos testados, com 177 variações de modelo e balanceamento treinados, o modelo Ensemble (Tabular Prior-Data Fitted Network (TabPFN) + XGBoost +LightGBM) obteve o melhor desempenho, com AUC-ROC de 0,86, Recall de 0,767 e F2-Score de 0,631, métrica que mede a eficácia na detecção de empresas com padrão de fraude, porém com maior penalização de falsos negativos — visando reduzir o risco da auditoria. Para garantir a robustez da escolha, aplicou-se o Model Confidence Set (MCS) com nível de confiança de 95%, permitindo a comparação e seleção dos modelos com menor variabilidade estatística (desvio padrão reduzido na métrica Weight LogLoss ao longo das validações cruzadas). Adicionalmente, foram aplicadas técnicas de interpretabilidade com valores de Shapley (SHAP), permitindo compreender os fatores determinantes no cálculo do risco para cada contrato ou fornecedor. Os resultados foram apresentados em um painel analítico de governança e priorização de auditorias, demonstrando que a abordagem proposta pode transformar a auditoria pública ao torná-la mais estratégica, eficiente, transparente e orientada por dados.

Palavras-chave: contratações públicas, auditoria preditiva, aprendizado de máquina, explicabilidade (XAI), SHAP, TabPFN, Model Confidence Set (MCS), detecção de risco.

Abstract

Efficient prioritization of audits in public procurement is crucial for combating corruption, enhancing governance, and optimizing the use of public resources. This work presents an innovative methodology organized as a framework that integrates Machine Learning (ML) and Explainable Artificial Intelligence (XAI) to classify contracts and suppliers with a higher risk of irregularities. Public databases were used to construct the dataset, followed by preprocessing, balancing (SMOTE, ADASYN, Tomek Links, and variations), feature selection via LASSO, and hyperparameter tuning steps. Among the 177 trained models and balancing variations tested, the Ensemble Model (Tabular Prior-Data Fitted Network (TabPFN) + XGBoost + LightGBM) achieved the best performance, with an AUC-ROC of 0.86, a recall of 0.767, and an F2-Score of 0.631. The latter metric measures the effectiveness of detecting companies with fraud patterns while imposing a higher penalty on false negatives, aiming to reduce audit risk. To ensure the robustness of the selection, the Model Confidence Set (MCS) procedure was applied with a 95% confidence level, allowing for the comparison and selection of models with lower statistical variability (reduced standard deviation in the Weighted LogLoss metric) across cross-validations. Additionally, interpretability techniques using Shapley values (SHAP) were applied, enabling the understanding of the determining factors in the risk calculation for each contract or supplier. The results were presented in an analytical governance and audit prioritization dashboard, demonstrating that the proposed approach can transform public auditing by making it more strategic, efficient, transparent, and data-driven.

Keywords: public procurement, predictive auditing, machine learning, explainable AI (XAI), SHAP, TabPFN, Model Confidence Set (MCS), risk detection.

Sumário

Referências **118**

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O processo de compras e aquisições públicas no Brasil é caracterizado por elevados valores e alta complexidade, exigindo conformidade com rigorosos ritos legais e envolvendo múltiplos atores e interesses. De acordo com o Portal de Compras do Governo Federal Brasil (2023), em 2022, o valor total de compras homologadas em nível federal atingiu R$ 167 bilhões.

Nesse contexto, as compras e aquisições públicas são suscetíveis a fraudes, que podem ser definidas como "ato intencional, realizado por um ou mais indivíduos da administração, responsáveis pela governança, empregados ou terceiros, que envolvem dolo para a obtenção de vantagem injusta ou ilegal" (CFC, 2009).

Entre abril de 2020 e março de 2021, a Controladoria-Geral da União (CGU) analisou contratações relacionadas ao enfrentamento da pandemia de COVID-19, totalizando R$ 1,38 bilhão. As fraudes identificadas nesses contratos resultaram em um prejuízo efetivo de R$ 39,06 milhões (CGU, 2022). Diante disso, há consenso entre as organizações de que o fortalecimento dos controles internos é essencial para a prevenção de fraudes.

Ampliando a perspectiva, a corrupção em processos de compras públicas constitui um problema relevante não apenas no Brasil, mas em diversas partes do mundo. Segundo relatório da OECD (2015), a vulnerabilidade dos processos de compras públicas à corrupção pode ocorrer em todas as fases, desde a avaliação das necessidades até a execução contratual. Estimativas indicam que a corrupção pode gerar perdas significativas, atingindo até 30% do valor de projetos financiados publicamente. Nesse sentido, um estudo da consultoria PricewaterhouseCoopers PwC (2013), focado na União Europeia, estimou os custos diretos da corrupção em compras públicas e identificou diversos sinais de alerta que indicam uma maior probabilidade de ocorrência de corrupção, como alterações significativas nos termos dos contratos após a adjudicação e contatos indevidos entre licitantes.

No contexto brasileiro, destaca-se a obra "Avaliação da Qualidade do Gasto Público e Mensuração da Eficiência", publicada pelo Ministério da Fazenda (Boueri, Rocha, & Ro-

dopoulos, 2015), que evidencia a necessidade de melhorar a qualidade dos gastos governamentais, visando promover maior eficiência na prestação de serviços públicos e otimizar a utilização dos recursos disponíveis. Os fatores que comprometem essa qualidade incluem a ineficiência nos processos de compras, a inadimplência nos procedimentos de aquisição, dispositivos de prestação de serviços ineficazes, além da corrupção e da fraude.

Diante desse panorama, a implementação de um processo sistemático para a priorização de auditorias em contratos públicos torna-se necessária. A definição dos critérios a serem considerados na avaliação de contratos e fornecedores constitui uma etapa fundamental para identificar e priorizar aqueles com maior probabilidade de apresentar irregularidades, especialmente devido à escassez de pessoal nas unidades de controle interno.

Considerando o grande volume de dados e valores envolvidos, a aplicação de técnicas analíticas apresenta-se como uma alternativa eficaz para refinar a seleção de processos licitatórios a serem auditados, reduzindo custos e o tempo de fiscalização ao concentrar esforços nos contratos com maior potencial de risco. Os recursos necessários estão acessíveis por meio de dados abertos governamentais e bancos de dados relacionados a processos licitatórios. A solução proposta fundamenta-se na utilização de técnicas de mineração de dados e aprendizado de máquina supervisionado.

Foram utilizados dados abertos disponíveis em portais de transparência ou obtidos por meio da Lei de Acesso à Informação, incluindo bases de dados de licitações, cadastros de servidores do Poder Executivo, registros de pessoas físicas e jurídicas e Relação Anual de Informações Sociais (RAIS), conforme a disponibilidade temporal. Também foram integradas bases mantidas pela CGU, como o Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS) e o Cadastro Nacional de Empresas Punidas (CNEP).

A hipótese de pesquisa propõe o desenvolvimento de uma ferramenta de suporte à decisão, baseada em mineração de dados provenientes de fontes públicas do Governo Federal, com o objetivo de gerar uma métrica de mensuração de riscos associada a indícios de irregularidades em licitações e fornecedores. A metodologia a ser apresentada neste trabalho segue as fases do modelo de referência *Cross-Industry Standard Process for Data Mining* (CRISP-DM) (Chapman et al., 2000), o qual organiza as atividades de mineração de dados em seis etapas, em abordagem *top-down*, desde o entendimento do negócio até a implantação do produto.

Espera-se que essa ferramenta auxilie na priorização de contratos e fornecedores a serem avaliados por auditores internos. Como contribuição para a sociedade, a metodologia proposta visa aprimorar a qualidade do gasto público por meio da identificação antecipada de potenciais casos de fraude em compras governamentais.

1.1 Problema da Pesquisa

A identificação eficiente dos riscos que devem ser priorizados em auditorias de contratações públicas visa evitar prejuízos aos cofres públicos e assegurar o fornecimento tempestivo de bens e serviços de qualidade à população. Por outro lado, as fraudes em processos licitatórios têm se tornado cada vez mais sofisticadas, envolvendo atos como abuso de poder, favoritismo, suborno, corrupção, peculato, nepotismo, simulação e apropriação indébita (Padhi & Mohapatra, 2011), o que dificulta sua detecção por meio de avaliações tradicionais de conformidade (Wensink & Vet, 2006).

O escopo das irregularidades em licitações abrange diversas tipologias. Dentre elas, Wensink and Vet (2006), destaca-se o conluio, entendido como um acordo ilícito entre concorrentes para manipular e fraudar os resultados dos processos licitatórios, sem a participação de agentes da Administração Pública.

Já Carpanese, Velasco, Interian, Paulo Neto, and Ribeiro (2021) trata da corrupção como uma atividade relacional entre agentes corruptores e entidades corrompidas, caracterizada, neste caso, pela participação de servidores públicos. Outro exemplo é a fraude documental, conforme descrito por Mlondo (2013), que abrange a falsificação de documentos, propostas fraudulentas, uso de empresas de fachada, simulações indevidas para a obtenção de benefícios, fraudes de identidade e ações que visam comprometer a eficiência da administração pública. Dessa forma, as bases de punições mantidas pela CGU reúnem diversas irregularidades que podem ser utilizadas como insumo para o treinamento de modelos de aprendizado de máquina.

Diante disso, o uso de técnicas avançadas de mineração de dados e aprendizado de máquina apresenta-se como uma solução potencialmente eficaz ao permitir a análise de padrões e relações complexas. Destaca-se, ainda, a possibilidade de que essas ferramentas auxiliem na automatização, escalabilidade e aperfeiçoamento do processo de detecção de fraudes e avaliação de riscos, contribuindo para o aumento da transparência e da eficiência nas compras governamentais.

Surge, portanto, a seguinte questão de pesquisa: como desenvolver uma metodologia eficaz que permita avaliar e priorizar riscos nas auditorias de contratações públicas, utilizando técnicas de mineração de dados e aprendizado de máquina, de forma a reduzir a exposição a irregularidades e aumentar a eficiência no uso dos recursos públicos?

1.2 Objetivos

1.2.1 Objetivo Geral

Desenvolver e validar uma metodologia, integrando aprendizado de máquina e interpretabilidade de Inteligência Artificial, para a criação de um score de risco destinado à priorização de auditorias em contratos públicos. , com vistas à redução da exposição a irregularidades e ao aumento da eficiência na utilização dos recursos públicos.

1.2.2 Objetivos Específicos

Para atingir o objetivo geral, estabelecem-se os seguintes objetivos específicos:

1. Revisar a literatura sobre avaliação de riscos e auditorias em compras públicas, com destaque para as principais metodologias aplicadas na área.
2. Desenvolver um modelo de aprendizado de máquina que auxilie na priorização de riscos, gerando *scores* de risco para contratos e fornecedores envolvidos em contratações públicas.
3. Implementar uma ferramenta capaz de automatizar a análise de riscos e a priorização de auditorias, proporcionando maior escalabilidade, eficiência e explicabilidade.
4. Validar a metodologia proposta com dados reais, avaliando sua eficácia na identificação de riscos e no aprimoramento da governança nas auditorias de contratações públicas.

1.3 Estrutura de Pesquisa

A pesquisa foi estruturada em cinco capítulos assim como é referenciado na Figura 1.1.

Figura 1.1: Estrutura da pesquisa
Fonte: Produzido pelo autor

Capítulo 1 : *Introdução.* Apresenta o tema do estudo, contextualizando o problema de pesquisa, os objetivos e a justificativa para o desenvolvimento da investigação científica.

Capítulo 2 : *Embasamento Teórico.* Expõe os fundamentos conceituais e as principais técnicas utilizadas, com ênfase na relevância teórica para o entendimento dos métodos aplicados no trabalho

Capítulo 3 : *Metodologia.* Detalha os procedimentos metodológicos adotados, os materiais utilizados, bem como as técnicas de coleta, tratamento e análise dos dados aplicadas na pesquisa.

Capítulo 4 : *Resultados e Discussão.* Aplicação da metodologia proposta, comparando os resultados aos estudos relacionados e resultados esperados.

Capítulo ?? : *Considerações Finais.* Conclusão das análises realizadas, com os riscos e desafios encontrados na execução da pesquisa.

Capítulo 2

Embasamento Teórico

2.1 Referencial Teórico

2.1.1 Trabalhos Relacionados

Conforme recomendado por Kitchenham and Charters (2007), foi realizada uma revisão sistemática da literatura com o objetivo de identificar publicações relevantes ao tema, agregando conhecimento ao estudo em desenvolvimento. A seleção dos trabalhos considerou a capacidade de elucidar as questões centrais do problema investigado e fornecer subsídios para a formulação da proposta metodológica.

O processo de busca concentrou-se em pesquisas voltadas à avaliação de riscos e à priorização de auditorias em compras públicas, com foco no uso de mineração de dados e aprendizado de máquina. A estratégia de busca foi planejada de forma sistemática, organizada em dois grupos temáticos principais: *Avaliação de Riscos em Compras Públicas*, que aborda a gestão de riscos e suas irregularidades; e *Mineração de Dados e Aprendizado de Máquina*, que se centra na análise de padrões e na priorização de auditorias com base em dados.

A busca foi realizada em março de 2024 na base Web of Science, considerando artigos publicados na última década. Foram utilizadas combinações de termos-chave, como "*Risk Assessment*", "*Public Procurement*", "*Fraud Detection*", "*Machine Learning*", "*Graph Analysis*", "*Shapley Value*"e "*Audit Prioritization*". Operadores booleanos foram aplicados para maximizar a abrangência e a relevância dos resultados.

A Tabela 2.1 apresenta as combinações de termos utilizadas e a quantidade de resultados obtidos. Após a coleta inicial, os artigos foram filtrados com base em critérios de qualidade metodológica, clareza dos objetivos e relevância dos resultados para o escopo da pesquisa.

Tabela 2.1: Combinação dos termos de pesquisa

Palavras-Chave	**Base de Dados**
("public procurement"OR "public bidding"OR "public contracts")	3,887
("public procurement"OR "public bidding"OR "public contracts") AND ("fraud"OR "prioritization "OR "risk")	356
("machine learning"OR "artificial intelligence"OR "cluster*"OR "regression"OR "deep learning"OR "data mining"OR "graph"OR "shapley"OR "explainability") AND ("public procurement"OR "public bidding"OR "public contracts") AND ("fraud detection"OR "prioritization"OR "risk")	40

Fonte: Elaborado pelo autor

Em seguida, realizou-se uma análise bibliométrica com o auxílio do software VOSviewer para a extração dos termos mais recorrentes, conforme ilustrado na Figura 2.1. O mapeamento desses termos permitiu identificar tendências temáticas e refinar a base de conhecimento utilizada no desenvolvimento do modelo proposto.

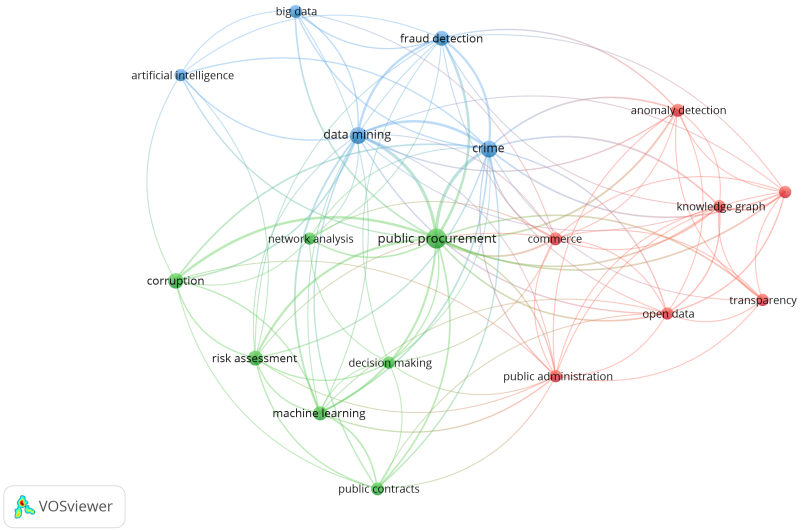

Figura 2.1: Rede Bibliométrica de termos mais recorrentes relacionados à riscos em compras públicas.
Fonte: (Vos viewer) Elaborado pelo autor

Estudos sobre a aplicação de ferramentas analíticas e estatísticas para a detecção de fraudes em compras públicas foram desenvolvidos tanto no Brasil quanto no exterior, conforme indicado pela revisão sistemática. A Tabela 2.2 apresenta publicações relevantes que abordam esse tema, fornecendo um panorama do estado atual da pesquisa na área.

Tabela 2.2: Trabalhos relacionados à prevenção de riscos em compras públicas com base em dados

Autores	**Síntese**
A decision support system for fraud detection in public procurement (Velasco, Carpanese, Interian, Paulo Neto, & Ribeiro, 2021)	O modelo do DSS para detecção de fraudes em contratações públicas utiliza mineração de dados e técnicas de pesquisa operacional para identificar padrões de risco, como conluio entre licitantes e conflitos de interesse. A abordagem envolve um processo ETL para consolidar dados de contratos públicos em um data lake, possibilitando análises com algoritmos baseados em teoria dos grafos e clusterização. O sistema gera relatórios de risco para priorizar investigações, ajudando na recuperação de milhões de reais em operações policiais como "Xeque-Mate"e "Calvário".
Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies. (Modrušan, Rabuzin, & Mršic, 2021)	Este artigo revisa técnicas de detecção de fraudes em compras públicas, utilizando tecnologias emergentes, como inteligência artificial e aprendizado de máquina. Os métodos incluem regras de associação e bancos de dados em grafos para identificar colusão entre operadores econômicos e autoridades contratantes. A eficiência da detecção depende da qualidade dos dados e dos indicadores de risco utilizados. O estudo foca em técnicas que permitam a detecção antecipada de comportamentos suspeitos, visando criar um sistema de alerta precoce.

From Corruption to State Capture: A New Analytical Framework with Empirical Applications from Hungary. (Fazekas, Tóth, & King, 2016)	O artigo utiliza técnicas de análise de redes sociais, aplicando um Índice de Risco de Corrupção (CRI) para contratos públicos. A análise identifica padrões de captura do Estado por meio de clusters de organizações de alto risco, detectando a concentração de empresas capturadas conectadas entre si. Métodos como o uso do pacote igraph no R e a biblioteca NetworkX no Python são usados para mapear essas conexões e identificar padrões de corrupção em diferentes períodos.
Corruption risk in contracting markets: a network science perspective. (Fazekas, Wachs, & Kertész, 2021)	O artigo aplica métodos de ciência de redes para analisar o risco de corrupção em mais de 4 milhões de contratos públicos da União Europeia. Usando redes bipartidas para mapear mercados de compras públicas, o estudo identifica padrões de risco em mercados centralizados e descentralizados. Para isso, foram utilizadas métricas como centralização, clustering modular e análise de núcleos (cores), o que permitiu medir a prevalência e distribuição do risco de corrupção, destacando variações entre os países analisados.
Prediction of Public Procurement Corruption Indices using Machine Learning Methods (Modrušan & Rabuzin, 2019)	Este artigo compara modelos de previsão usando técnicas de text mining e métodos de machine learning para detectar licitações suspeitas em compras públicas na Croácia. O modelo inclui algoritmos como Naive Bayes, regressão logística e máquinas de vetores de suporte (SVM) aplicados a documentos de licitação. A análise é feita utilizando o framework CRISP-DM, incluindo etapas como tokenização e stemming, seguido pela classificação dos documentos usando técnicas como tf-idf para representar os textos em forma de vetores para os modelos de aprendizado.

Controlling Corruption in Development Aid: New Evidence from Contract-Level Data.(Fazekas, Dávid-Barrett, Hellmann, Márk, & McCorley, 2020)	O artigo analisa contratos financiados por doações de bancos de desenvolvimento em mais de 100 países (1998–2008) para avaliar a eficácia dos regulamentos contra corrupção. Utiliza análise de regressão e matching por escore de propensão em nível de contrato, comparando contratos antes e depois da mudança nas regras de supervisão do Banco Mundial em 2003. A análise mostra que países com baixa capacidade estatal beneficiam-se mais do aumento de supervisão e ampliação de acesso aos editais, utilizando métodos como regressão logística e análise de variáveis latentes.
Characterization of the firm–firm public procurement co-bidding network from the State of Ceará (Brazil). (Lyra, Curado, Damásio, Bação, & Pinheiro, 2021)	O artigo analisa a rede de co-participação em licitações públicas no Estado do Ceará (Brasil) entre 2015 e 2019, usando métodos de ciência de redes. Utilizando a projeção de co-bidding para criar uma rede firm-firm, os autores empregam o coeficiente de similaridade de Jaccard centralizado para medir a similaridade entre as firmas e identificar comunidades de empresas com padrões de atividade semelhantes. Foram identificadas 22 comunidades de empresas, algumas das quais apresentaram um risco elevado de manipulação de mercado devido à baixa diversidade de contratos e regionalização. A modularidade da rede foi de 0,66, e o coeficiente de cluster médio foi de 0,52, indicando uma estrutura modular significativa.
Data Quality Barriers for Transparency in Public Procurement. (Soylu et al., 2022)	O artigo discute barreiras de qualidade dos dados em processos de compras públicas, utilizando dados de contratos, empresas e gastos na Eslovênia integrados em um Knowledge Graph (KG). Aplicou técnicas de detecção de anomalias sobre esses dados para identificar fraudes e melhorar a transparência, utilizando aprendizado de máquina e tecnologias da Web Semântica para integrar dados heterogêneos. Problemas comuns incluem dados ausentes, duplicados ou mal formados, o que impacta a análise.

Corruption red flags in public procurement: new evidence from Italian calls for tenders. (Decarolis & Giorgiantonio, 2022)	O artigo avalia o uso de "red flags"para prever o risco de corrupção em licitações públicas na Itália. A abordagem inclui técnicas de aprendizado de máquina, como LASSO, Ridge regression e random forest, aplicadas a dados detalhados de contratos de rodovias. O estudo destaca que o uso de critérios de adjudicação com múltiplos parâmetros (MEAT) está fortemente associado ao risco de corrupção. Além disso, indicadores menos previsíveis, como características específicas dos editais, aumentam a precisão dos modelos, especialmente no uso do random forest, que apresentou melhor desempenho na predição dos casos de corrupção.
O Controle Interno Preventivo à Luz do Sistema ALICE: Propostas de Trilhas para Detecção de Anomalias na Execução de Programas Sociais.(C. L. Nascimento, 2022)	O artigo discute o aprimoramento do sistema ALICE, que utiliza trilhas específicas para detectar anomalias em programas sociais como FUNDEB, PNATE e PNAE. A metodologia inclui abordagem mista quali-quanti, com coleta de dados de relatórios de fiscalização e análise de legislações pertinentes. Foram desenvolvidas 51 trilhas para auditoria que automatizam a identificação de irregularidades, como licitações de alta materialidade e favorecimento de licitantes. O uso do sistema ALICE melhora a eficácia das auditorias preventivas da CGU, gerando economia aos cofres públicos pela detecção precoce de fraudes.
Identificação Automática de Conluio em Pregões do Comprasnet com Aprendizado de Máquina. (Souza, 2023)	Este estudo apresenta a aplicação de algoritmos de aprendizado de máquina para identificar possíveis conluios em licitações realizadas no Comprasnet. Foram testados modelos de Ensemble Learning, como Extra Trees, Random Forest e Ada Boost, em quatro cenários distintos e em datasets provenientes do Comprasnet e outros países. Extra Trees se destacou, alcançando acurácia superior a 87%. O modelo gerou novas trilhas de auditoria para o sistema ALICE da CGU, aumentando a capacidade de detecção de conluios entre empresas participantes de pregões.

Proposta de modelo de classificação do risco de contratos públicos. (Sales, 2016)	O artigo propõe um modelo de classificação do risco de contratos públicos, utilizando Regressão Logística e Análise de Decisão Multicritério para auxiliar na seleção de contratos a serem auditados. Dois modelos de regressão foram desenvolvidos: o primeiro avalia o risco dos fornecedores com base em características como capacidade operacional e histórico de contratações; o segundo quantifica o risco dos contratos, levando em consideração as características do fornecedor, do contrato e da licitação. Ambos os modelos apresentaram acurácia acima de 80%. A técnica de Analytic Hierarchy Process (AHP) foi usada para desenvolver um modelo multicritério para priorização dos contratos a serem auditados.
Modelo Preditivo de Risco de Irregularidades em Compras Públicas no Estado de Goiás. (Jesus, 2020)	O artigo propõe um modelo preditivo para estimar o risco de irregularidades em licitações no Estado de Goiás. Utilizando o modelo CRISP-DM, o estudo aplica técnicas de mineração de dados e aprendizado supervisionado, como Regressão Logística, SVM e Gradient Boosting Machine, para identificar licitações com maior risco. A análise foi feita em duas fases da licitação: publicação do edital e disputa. Na fase de publicação, os modelos para pregão, dispensa e inexigibilidade apresentaram AUROC superior a 70%, enquanto a modalidade concorrência não teve resultados satisfatórios. Já na fase de disputa, todas as modalidades obtiveram AUROC acima de 70%, validando a eficácia do modelo para priorização de auditorias pelo TCE-GO.

A Machine Learning-Based Analysis on the Causality of Financial Stress in Banking Institutions. (de Moraes Souza, de Castro, Peng, et al., 2024)	O estudo aplica técnicas de aprendizado de máquina para analisar o risco de inadimplência e risco sistêmico em instituições financeiras usando um grande conjunto de dados de 2325 bancos ao longo de 17 anos. Utilizando uma abordagem metodológica com Random Forest, XGBoost e um framework de inteligência artificial explicável (XAI), os autores identificaram as variáveis mais importantes para prever o risco de inadimplência, como a probabilidade de resgate financeiro e a razão market-to-book. Para risco sistêmico, fatores como o número de bancos e os níveis de taxa de juros foram destacados. Este trabalho utilizará a mesma técnica aplicada ao contexto de risco em compras públicas, empregando valores de SHAP para análise da importância das variáveis, conciliando poder preditivo com interpretabilidade prática.
Explainable machine learning in credit risk management. (Bussmann, Giudici, Marinelli, & Papenbrock, 2021)	Propôs um modelo explicável de inteligência artificial que pode ser usado na gestão do risco de crédito. O modelo aplica redes de correlação aos valores de Shapley para que as previsões sejam agrupadas de acordo com as similaridades nas explicações subjacentes. Informa que tomadores de risco podem ser agrupados por conjuntos e características financeiras semelhantes que podem ser utilizadas para explicar sua pontuação de crédito e prever seu comportamento.

Fonte: Elaborado pelo autor

O presente estudo busca compreender de que forma a pesquisa acadêmica aborda os riscos relacionados a irregularidades em contratações públicas e como essas informações podem fortalecer o desempenho dos controles internos. A adoção de mecanismos de auditoria mais detectivos e preventivos, associada à identificação tempestiva de contratos de alto risco, tem ganhado relevância diante do aumento no volume de dados gerados e da complexidade dos processos licitatórios, em contraste à manutenção da força de trabalho.

As abordagens de avaliação de riscos em contratações públicas têm passado por transformações significativas, impulsionadas pelos avanços tecnológicos e pelas técnicas de *machine learning*. Entre 2010 e 2015, prevalecia o uso de métodos de auditoria tradicionais, baseados em análises documentais e em processos manuais. Em paralelo, Fazekas et al.

(2016) sinalizou a relevância da adoção de outros recursos, como a análise de redes sociais, para identificar capturas do Estado, detectando padrões de corrupção e possíveis aglomerados (clusters), reforçando as relações entre empresas e agentes políticos por meio de dados não supervisionados.

A partir de 2020, observou-se um direcionamento maior em direção à automação e ao uso de técnicas quantitativas, introduzindo o conceito de *red flags*, ou alertas fundamentados em padrões de risco detectados. Sistemas como "ALICE" (C. L. Nascimento, 2022), que aplica trilhas automatizadas para identificar informações atípicas em editais de licitação e "Malha Fina", voltado a regras em programas sociais, exemplificam essa evolução. De forma concomitante, Fazekas et al. (2020) aprimorou técnicas de ciência de redes para identificar corrupção em contratos públicos da União Europeia, utilizando redes bipartidas e métricas gráficas como suporte.

Nos últimos anos, técnicas de aprendizado de máquina e inteligência artificial foram adotadas de forma mais ampla em pesquisas relacionadas à auditoria e ao combate à fraude. Trabalhos como os de Modrušan et al. (2021), Souza (2023) e Lyra et al. (2021) empregaram modelos de aprendizado de Máquina (*Naive Bayes*, Regressão Logística, SVM) para prever índices de corrupção em contratos públicos, indicando a transição para enfoques preditivos. Em 2023, Souza (2023) utilizou o *ensemble learning* (*Random Forest, Extra Trees*) para detectar conluios em pregões eletrônicos do Comprasnet, resultando em uma acurácia superior a 87% em algumas avaliações.

Simultaneamente, a busca por maior interpretabilidade nos modelos de risco tem enfatizado técnicas de inteligência artificial explicável (XAI) com o objetivo de viabilizar a compreensão das saídas geradas pelos algoritmos. Em de Moraes Souza et al. (2024), voltado ao risco sistêmico financeiro em instituições bancárias, foi proposta uma abordagem utilizando XAI, possibilitando a interpretação prática dos modelos de *machine learning*. Técnicas como SHAP (*Shapley Additive Explanations*) oferecem explicações tanto locais quanto globais, contribuindo para a transparência das decisões automatizadas. Em contratações públicas, esses recursos podem auxiliar na priorização de auditorias de contratos com base em riscos preditivos, fornecendo visibilidade sobre os fatores que influenciam cada pontuação.

A evolução das técnicas de controle e auditoria em compras públicas apresenta uma transição gradual de processos manuais e revisões documentais para métodos mais automatizados e orientados por dados. As contribuições recentes, ancoradas em aprendizado de máquina e inteligência artificial explicável, possibilitam uma análise preditiva robusta e transparência no processo de priorização de contratos. Diante desse cenário, a presente pesquisa sugere a aplicação desses métodos modernos no contexto das compras públicas, ampliando a precisão e a fundamentação analítica da auditoria ao considerar critérios

objetivos e fatores de risco previamente identificados.

2.2 Compras Públicas, Riscos e Irregularidades

O processo de compras públicas no Brasil segue um conjunto normativo que visa garantir a legalidade, a transparência e a eficiência na utilização dos recursos (Carvalho, 2019). Em nível constitucional, o art. 37, inciso XXI, determina que contratos envolvendo obras, prestação de serviços, compras e alienações sejam, em regra, precedidos de processo licitatório, assegurando tratamento isonômico aos potenciais fornecedores e a seleção da proposta mais vantajosa para a Administração Pública.

Além disso, a Lei Federal nº 14.133/2021 (Brasil, 2021), em seu art. 3º, reforça o objetivo de selecionar a melhor proposta, em consonância com os princípios de legalidade, impessoalidade, publicidade, moralidade e eficiência.Conhecida como a Lei de Licitações e Contratos Administrativos, essa legislação moderniza o arcabouço normativo, propondo maior simplicidade e eficiência no processo licitatório. Além disso, a lei estabelece critérios para a seleção de fornecedores, procedimentos de licitação, exigências para a formalização de contratos administrativos e penalidades aplicáveis em caso de descumprimento contratual (art. 90; art. 156).

Destaca-se que o processo licitatório abrange diferentes fases. A primeira fase consiste no planejamento, durante o qual o órgão público identifica suas necessidades, define objetivos e estabelece os recursos disponíveis. Essa etapa é fundamental para direcionar aquisições de maneira eficaz, em conformidade com o art. 11 da Lei nº 14.133/2021. Em seguida, ocorre a elaboração do edital, documento que veicula especificações técnicas, critérios de julgamento e outros requisitos. A publicidade e a isonomia são pilares desse estágio, garantindo que todos os interessados tenham igual acesso às informações, conforme estabelecido no art. 5º da Lei nº 14.133/2021.

Em seguida, de acordo com Rodrigues and Lima Filho (2017), a fase externa inicia-se com a abertura do certame e o recebimento das propostas. As distintas modalidades de licitação — concorrência, tomada de preços, convite, pregão e leilão — regem o rito licitatório. A concorrência é considerada genérica e é utilizada para contratações de maior complexidade ou vulto financeiro. Adicionalmente, a Lei 14.133/2021 deixou de adotar o critério de valor da contratação como determinante para a escolha da modalidade concorrência. Assim, a escolha entre o pregão e a concorrência será determinada pela natureza do objeto: bens e serviços especiais e obras, bem como serviços comuns e especiais de engenharia no caso da concorrência; e bens e serviços comuns, incluindo os de engenharia, no caso do pregão.

Incluem-se entre os serviços a serem licitados por concorrência os serviços de natureza predominantemente intelectual, ressalvando-se a hipótese do uso do concurso ou da contratação desses serviços por meio de inexigibilidade de licitação, quando verificada a inviabilidade de competição. A tomada de preços aplica-se a valores intermediários e requer, em regra, o prévio cadastramento dos licitantes até o terceiro dia útil antes da licitação (Carvalho, 2019). O convite, por sua vez, é direcionado a contratações de menor valor, cabendo ao órgão selecionar pelo menos três licitantes devidamente qualificadas.

Outras modalidades incluem o leilão, que se aplica à alienação de bens inservíveis ou apreendidos, e o concurso, voltado à seleção de trabalhos técnicos, artísticos ou científicos. Por fim, o pregão, instituído pela Lei nº 10.520/02, é destinado à aquisição de bens e serviços comuns sem limite máximo de valor, podendo ser presencial ou eletrônico.

A legislação ainda prevê hipóteses de contratação direta, como a inexigibilidade e a dispensa de licitação (Lei nº 14.133/2021, art. 23). A inexigibilidade aplica-se quando não há possibilidade de competição (ex.: fornecedor exclusivo, objeto de natureza singular ou contratação de profissionais de notória especialização). A dispensa ocorre quando existe viabilidade competitiva, mas a realização do certame se mostra contraproducente em relação ao interesse público.

A Lei nº 14.133/2021 também introduz diversas penalidades para infrações contratuais (art. 156), incluindo advertência, multa, suspensão temporária de participação em licitações e declaração de inidoneidade por até dois anos. A advertência caracteriza-se como uma medida educativa aplicada a infrações menores; a multa é proporcional à gravidade da irregularidade; a suspensão temporária veda a participação do fornecedor em novas licitações; e a declaração de inidoneidade configura a sanção mais severa, direcionada a condutas como fraudes ou corrupção (Brasil, 2021).

As modalidades adotadas na execução de compras públicas podem influenciar a probabilidade de irregularidades. Pesquisas indicam que formas menos competitivas ou com rito menos formal, como convite e dispensa, podem apresentar uma maior incidência de fraudes, conforme identificado por Rodrigues and Lima Filho (2017). Em contrapartida, a concorrência e o pregão tendem a apresentar índices menores de irregularidades, conforme verificado em relatórios de controle interno da CGU (2004–2014). A definição de controles internos e a observância dos princípios constitucionais (art. 37, inciso XXI, CF/88) permanecem cruciais para mitigar riscos e assegurar a integridade nos procedimentos licitatórios.

2.2.1 Licitações e Compras Públicas

O processo de compras públicas no Brasil, embora amparado por um conjunto normativo amplo e por princípios constitucionais, ainda apresenta vulnerabilidades significativas.

Entre os principais desafios, encontram-se a complexidade burocrática, a insuficiência de qualificação técnica por parte dos servidores e as oportunidades para fraudes e corrupção (Lopes & de Jesus, 2024). Falhas no planejamento, na elaboração de editais e na fiscalização dos contratos podem resultar em atrasos, custos adicionais e contratações que não atendem às necessidades da administração.

A possibilidade de conluio entre fornecedores e a ocorrência de práticas irregulares refletem riscos que comprometem a integridade do processo licitatório (Lopes & de Jesus, 2024). Em cada fase do certame, desde a etapa inicial de orçamento até a adjudicação do objeto, podem ocorrer irregularidades, como a restrição indevida de competitividade, o direcionamento contratual e a ausência de comprovação da capacidade técnica. Tais práticas podem caracterizar simulações de concorrência ou manipulações voltadas a maximizar ganhos ilegítimos (Santos & Souza, 2023).

A fiscalização exercida por órgãos de controle e auditoria, como tribunais de contas e a Controladoria-Geral da União (CGU), tem o papel de identificar discrepâncias entre o cenário esperado, fundamentado em leis e normas vigentes, e o cenário efetivamente observado. A análise do auditor abrange a verificação de documentos comprobatórios, a coerência de preços em relação aos valores de mercado e a adequação do objeto licitado às finalidades da administração pública.

A administração pública dispõe de mecanismos de contratação direta (inexigibilidade e dispensa de licitação), previstos na Lei nº 14.133/2021, para situações em que a competição se revela inviável ou contrária ao interesse público. Em paralelo, estudos sugerem que modalidades com procedimentos menos formais, como carta convite e dispensa, apresentam maior probabilidade de ocorrência de irregularidades, enquanto concorrência e pregão costumam registrar percentuais menores de fraudes (Rodrigues & Lima Filho, 2017).

No que se refere aos controles internos, é fundamental a existência de procedimentos que identifiquem inconformidades e assegurem a conformidade do processo com a legislação aplicável (C. L. Nascimento, 2022; Souza, 2023). A pesquisa adequada de preços, a comprovação da capacidade técnica dos fornecedores e a realização de auditorias periódicas podem mitigar os riscos de superfaturamento e de uso inadequado de recursos públicos. Assim, a consolidação de uma cultura de transparência e responsabilidade no setor público depende da melhoria contínua desses controles e da competência técnica dos agentes envolvidos, contribuindo para a legalidade e a legitimidade das contratações.

Tabela 2.3: Exemplos de tipologias de irregularidades em licitações e contratos

Irregularidade	**Base Legal**
Ausência de comprovação de capacidade técnica	Art. 122 da Lei 14.133/22

Ausência ou deficiência de pesquisa de preços	Art. 17 e 40 da Lei 14.133/22
Autor do projeto e o licitante vinculados	Art. 9 da Lei 14.133/22
Conluio	Art. 9 da Lei 14.133/22
Direcionamento de contratação	Art. 9 da Lei 14.133/22
Dispensa ou inexigibilidade sem fundamentação legal	Art. 37 XXI da Constituição Federal
Fracionamento indevido de despesa	Art. 40 da Lei 14.133/22
Ausência de previsão de preferência para contratações de ME e EPP	Lei Federal 123/2006
Objeto impreciso, genérico, incompreensível ou incompleto	Art. 40 e art. 47 da Lei 14.133/22
Previsão de sub-contratação irregular	Art. 74 da Lei 14.133/22
Restrição de competitividade	Art. 40 da Lei 14.133/22; Acórdão 461/2014-TCU
Superfaturamento ou sobrepreço	Art. 337-L da Lei 14.133/22
Uso de modalidade indevida	Art. 5 da Lei 12.232/2010
Vínculos entre licitantes e servidores públicos	Art. 9 da Lei 14.133/22; Acórdão 1198/2007

Fonte: Adaptado de (Jesus, 2020)

Diante desses desafios, a adoção de tecnologias e metodologias avançadas, como técnicas de aprendizado de máquina e sistemas de apoio à decisão, pode contribuir para a melhoria da qualidade do gasto público e aumentar a confiança da sociedade nas contratações governamentais.

2.2.2 Critérios de Risco em Contratações Públicas

Conforme observado nos tópicos anteriores, a seleção de contratos para auditoria constitui uma etapa relevante para assegurar a gestão eficaz de riscos e a conformidade contratual. Conforme relatado na literatura, tem-se registrado uma crescente utilização de análises multicritério na priorização de contratos para auditoria. Diversos estudos têm explorado a aplicação desses métodos, destacando a importância de considerar múltiplos fatores de risco (Ferwerda, Deleanu, & Unger, 2017).

Nesta análise, são abordados os principais critérios discutidos na literatura: **Risco Empresa**, **Risco Sócio**, **Risco Contrato/Licitação** e **Alertas Relevantes**, oferecendo uma visão aprofundada sobre cada um deles.

O critério de **Risco relacionado à Empresa** contempla indicadores que avaliam a vulnerabilidade e o histórico das empresas contratadas. Jesus (2020) e J. L. R. Nascimento (2022) ressaltam fatores como a idade da empresa (Fazekas & Tóth, 2016), a diversidade de atividades cadastradas no CNAE e o valor total das compras vencidas. Empresas recém-criadas ou com múltiplas atividades podem indicar tentativa de burlar especificações contratuais ou diluir responsabilidades (Carpanese et al., 2021; Jesus, 2020; J. L. R. Nascimento, 2022), elevando o risco de fraude.

Adicionalmente, a realização de doações eleitorais por parte da empresa pode sugerir favorecimento indevido em processos licitatórios (Carpanese et al., 2021; J. L. R. Nascimento, 2022; Sales & Carvalho, 2016). Empresas com um número reduzido de funcionários ou com um histórico de punições configuram elementos de risco significativos (Jesus, 2020; Sales & Carvalho, 2016). Essas informações são obtidas a partir de bases como CNPJ, SIASG, TSE, RAIS, CEIS e CEPIM, que fornecem suporte factual para a análise de risco.

Os indicadores de **Risco Sócio** concentram-se no perfil e no comportamento dos sócios das empresas contratadas. Estudos de Jesus (2020), Sales and Carvalho (2016) e C. L. Nascimento (2022) indicam que fatores como participação em doações eleitorais, inscrição em programas sociais e filiação partidária são relevantes para avaliação de risco.

Sócios com baixa escolaridade ou com salários reduzidos, conforme dados da RAIS, tendem a apresentar maior vulnerabilidade econômica, o que pode aumentar a propensão a práticas fraudulentas (C. L. Nascimento, 2022).

A presença de sócios falecidos (Jesus, 2020; C. L. Nascimento, 2022; Sales & Carvalho, 2016) ou de servidores públicos federais entre os sócios pode indicar irregularidades na composição societária. Esses indicadores possibilitam identificar conflitos de interesse, utilizando bases como TSE, RAIS, BPC, Auxílio Emergencial, Defeso, Bolsa Família e CPF, além de registros de pagamentos públicos.

O critério de **Risco Contrato/Licitação** aborda os aspectos contratuais e procedimentais das licitações. Jesus (2020) e Sales and Carvalho (2016) enfatizam que a existência de múltiplos aditivos contratuais (de valor ou prazo), licitações de alta materialidade e o uso de modalidades menos competitivas, como dispensa ou inexigibilidade, são indicativos de risco elevado. Contratos que sofrem alterações sucessivas ao longo da execução podem evidenciar planejamento inadequado ou tentativa de manipulação contratual. A adoção de modalidades sem fundamentação técnica robusta pode também indicar favorecimento (Jesus, 2020; Sales & Carvalho, 2016).

Esses indicadores são extraídos de sistemas corporativos internos e regulamentos específicos, oferecendo um panorama detalhado dos riscos envolvidos. Em síntese, a análise sistemática dos critérios de risco nas contratações públicas é essencial para a integridade e a eficiência do processo licitatório. Os elementos destacados na literatura constituem

uma base analítica sólida para auditoria e gestão de riscos, contribuindo para a promoção da transparência e da responsabilização na administração pública.

2.3 Aprendizado de Máquina (AM) ou Machine Learning (ML)

Conceitos e Tipos de Aprendizado

O Aprendizado de Máquina (Machine Learning – ML) é uma subárea da Inteligência Artificial (IA) dedicada ao desenvolvimento de algoritmos capazes de identificar padrões em conjuntos de dados e tomar decisões com base nesses padrões. Diferentemente dos sistemas tradicionais, que requerem programação explícita para cada tarefa, os algoritmos de ML aprendem a partir de exemplos e ajustam seus parâmetros automaticamente para aprimorar o desempenho em novas instâncias (J. Han, Pei, & Kamber, 2011; Mitchell, 1997).

O campo de aprendizado de máquina (ML) pode ser classificado em diferentes paradigmas, conforme a forma como os algoritmos processam os dados disponíveis. Entre os principais, destacam-se o aprendizado supervisionado, o aprendizado não supervisionado, as abordagens mais recentes, como o aprendizado semi-supervisionado e o aprendizado por reforço. Cada método apresenta características específicas e aplicações distintas. Neste trabalho, a abordagem adotada será o aprendizado supervisionado.

Assim, o uso de aprendizado de máquina na detecção de riscos viabiliza o desenvolvimento de modelos capazes de automatizar a análise de contratos e fornecedores, possibilitando a priorização de auditorias com base em critérios objetivos. Este estudo explorará técnicas de aprendizado supervisionado voltadas à classificação de fornecedores e contratos, com o objetivo de aumentar a eficiência dos processos de fiscalização e mitigar os riscos associados às contratações públicas.

A combinação de métodos preditivos com abordagens explicáveis permitirá maior transparência na interpretação dos resultados, contribuindo para uma gestão mais eficaz e baseada em evidências.

Aprendizado Não Supervisionado

No aprendizado não supervisionado, os algoritmos operam sobre dados sem rótulos, ou seja, sem uma resposta previamente conhecida. O objetivo principal é identificar padrões subjacentes ou estruturas ocultas no conjunto de dados (Kotsiantis, 2007). Técnicas como clustering (agrupamento) e análise de regras de associação são amplamente utilizadas nesse contexto.

Entre os algoritmos mais empregados nessa abordagem, destacam-se o k-Means, um método iterativo responsável por agrupar instâncias com base na similaridade, minimizando a soma das distâncias entre os pontos e o centróide de cada grupo, e o DBSCAN (Density-Based Spatial Clustering of Applications with Noise), que identifica agrupamentos em regiões de alta densidade de dados. Este último é especialmente eficaz na detecção de ruídos e outliers, devido à sua abordagem baseada em densidade.

Apesar de sua ampla aplicação em diversos domínios, o aprendizado não supervisionado não será adotado neste estudo. A decisão justifica-se pelo foco na previsão de riscos com base em um conjunto de dados rotulados, composto por fornecedores e sócios que já sofreram punições anteriores, o que demanda a aplicação de técnicas de aprendizado supervisionado.

Aprendizado Supervisionado

No aprendizado supervisionado, os algoritmos são treinados com um conjunto de dados rotulados, no qual cada entrada está associada a uma saída esperada. Essa abordagem permite que o modelo aprenda as relações entre variáveis e seja capaz de realizar previsões quando novos dados são apresentados (James, Witten, Hastie, & Tibshirani, 2013).

Os problemas de aprendizado supervisionado podem ser categorizados em tarefas de classificação e de regressão. Os algoritmos de classificação têm como objetivo prever uma categoria ou classe para cada instância do conjunto de dados. Exemplos comuns incluem a detecção de fraudes em transações financeiras e a classificação de contratos públicos como regulares ou irregulares. Por outro lado, os algoritmos de regressão são utilizados quando a saída esperada é um valor numérico contínuo, como na previsão de preços de ativos ou na estimativa de riscos financeiros.

No contexto da detecção de riscos em compras públicas, os algoritmos de classificação são mais apropriados, pois permitem categorizar fornecedores e contratos com base em características previamente identificadas. Entre os modelos de classificação mais utilizados, destacam-se a Regressão Logística, as Árvores de Decisão, os métodos de ensemble, como Random Forest e Gradient Boosting, além das Redes Neurais Artificiais. A escolha do algoritmo mais adequado depende da natureza dos dados e da necessidade de interpretabilidade dos resultados. No presente estudo, priorizar-se-á a combinação de técnicas de classificação com métodos explicáveis, como os valores de SHAP (SHapley Additive exPlanations), a fim de assegurar maior transparência no processo decisório.

Outras Abordagens de Aprendizado de Máquina

Além do aprendizado supervisionado e não supervisionado, existem técnicas avançadas que exploram diferentes formas de aprendizado a partir dos dados. Entre esses méto-

dos, destacam-se o Aprendizado Semi-Supervisionado, o Aprendizado por Reforço e o Aprendizado Auto-Supervisionado.

O Aprendizado Semi-Supervisionado constitui uma abordagem híbrida que combina dados rotulados e não rotulados, com o objetivo de melhorar o desempenho dos modelos preditivos. Essa técnica é especialmente útil em contextos nos quais a rotulagem manual de dados é onerosa ou inviável em larga escala (Chapelle, Schölkopf, & Zien, 2006).

Em seguida, o Aprendizado por Reforço baseia-se na interação contínua entre um agente e um ambiente, por meio da qual o agente aprende a tomar decisões com base em recompensas e penalizações. Essa abordagem tem sido amplamente aplicada em áreas como o controle de sistemas, jogos e robótica, devido à sua capacidade de lidar com ambientes dinâmicos e sequenciais (Sutton & Barto, 2018).

Adicionalmente, o Aprendizado Auto-Supervisionado tem ganhado destaque recente por permitir que o próprio modelo gere rótulos a partir de relações intrínsecas nos dados. Com isso, torna-se possível construir representações úteis de forma não supervisionada, as quais são posteriormente utilizadas na fase de aprendizado supervisionado (LeCun, Misra, & Ba, 2021).

Portanto, a escolha do tipo de algoritmo de aprendizado de máquina está diretamente relacionada a diversos fatores, como a natureza da tarefa, a estrutura dos dados disponíveis e os objetivos específicos da aplicação. No presente trabalho, optou-se pela utilização de técnicas de Aprendizado Supervisionado focadas na tarefa de classificação, decisão motivada pela existência de um conjunto de dados previamente rotulado com informações sobre fornecedores e contratos sancionados.

Na seção seguinte, detalham-se algumas das principais abordagens dos algoritmos supervisionados, com ênfase nas que são aplicadas à construção dos modelos desenvolvidos nesta pesquisa.

Regressão Logística

A regressão logística, também denominada classificador de máxima entropia, pode ser organizada em duas modalidades principais: *logit binomial*, quando há apenas um par de categorias para classificação, e *logit multinomial*, quando o problema envolve múltiplas classes. Esse tipo de modelo estatístico encontra aplicação em diversos cenários de classificação, estimando a probabilidade de ocorrência de um evento específico a partir de variáveis explicativas.

Esse método modela a probabilidade da variável resposta (categórica), assumindo uma relação linear entre as variáveis explicativas (preditoras) e a transformação logarítmica das chances (o *logit*). Essa relação linear é, portanto, mapeada pela função *sigmoide*, que a transforma em uma probabilidade restrita ao intervalo entre 0 e 1 (Hosmer Jr, Lemeshow,

& Sturdivant, 2013). A Figura 2.2 ilustra de modo esquemático a curva resultante da regressão logística.

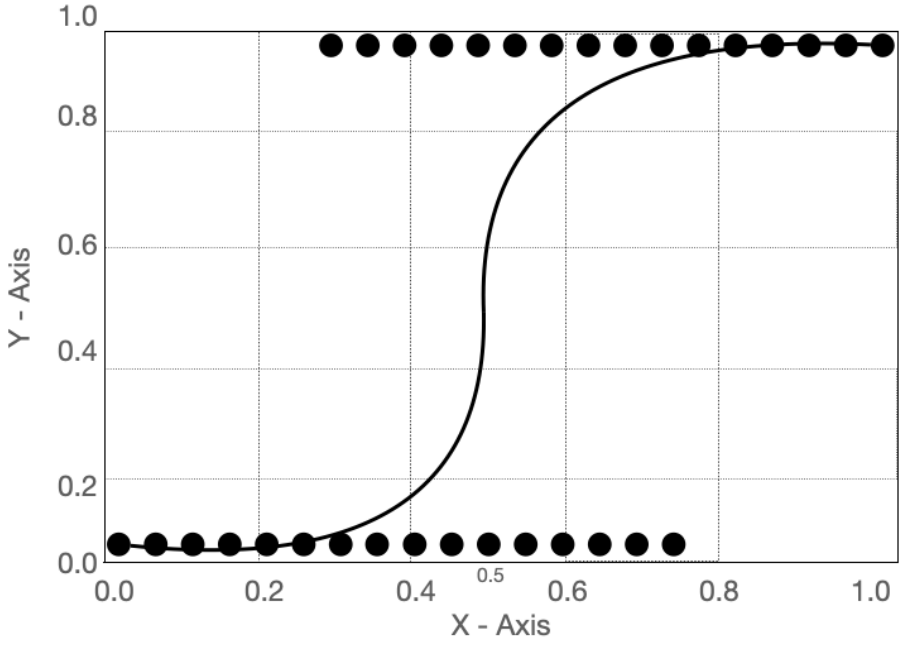

Figura 2.2: Representação gráfica da função *logit*
Fonte: Adaptado de (Hosmer Jr et al., 2013).

Segundo (Trueck & Rachev, 2009), a regressão logística é reconhecida como uma abordagem eficaz para problemas de classificação, especialmente na previsão de respostas binárias, como a existência ou a ausência de inadimplência (ou outro evento de interesse). Em geral, o foco está em investigar a relação entre a variável dependente, representada por Y, e um conjunto de variáveis explicativas, $X_1, X_2, \ldots, X_p$. A probabilidade de $Y = 1$ em função de X é expressa por:

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}}, \tag{2.1}$$

onde

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{2.2}$$

Assim, a probabilidade do evento de interesse pode ser reescrita da seguinte forma:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}. \tag{2.3}$$

Quando $g(x)$ excede o valor de 0, ou seja, quando $P(Y = 1)$ é superior a 0,5, a instância é classificada como $Y = 1$. Caso contrário, a predição recai em $Y = 0$.

Ao contrário das regressões lineares, que utilizam o método dos mínimos quadrados, os coeficientes $\beta_0, \beta_1, \ldots, \beta_p$ da regressão logística são estimados pelo método da máxima verossimilhança. Esse procedimento busca maximizar a probabilidade de observar as amostras, dadas as estimativas dos parâmetros.

A interpretação usual dos coeficientes baseia-se no conceito de *odds* (Hosmer Jr et al., 2013), representados por:

$$\text{odds} = \frac{p}{1-p} \tag{2.4}$$

em que p representa a probabilidade de ocorrência do evento. A razão de chances (*odds ratio*) compara duas probabilidades para quantificar a força da associação entre variáveis independentes e a variável dependente.

$$\text{odds ratio} = \frac{\text{odds}_1}{\text{odds}_2} \tag{2.5}$$

Para obter uma implementação consistente da regressão logística, as seguintes premissas devem ser validadas:

- Relação aproximadamente linear entre as variáveis explicativas (contínuas) e o *logit* da variável resposta;
- Ausência de multicolinearidade severa entre as variáveis independentes;
- Independência das observações (as amostras devem ser independentes entre si);
- É necessário contar com uma amostra representativa e suficientemente grande para reduzir o risco de viés e permitir a convergência do estimador de máxima verossimilhança.

Diferentemente da regressão linear, não é necessária a pressuposição de normalidade dos resíduos ou de homogeneidade das variâncias (homocedasticidade).

No contexto de processos licitatórios, a regressão logística pode auxiliar na identificação de casos com maior propensão a irregularidades, ao atribuir uma probabilidade de risco associada a cada instância.

Árvore de Decisão

As Árvores de Decisão, ou *Decision Tree* (DT), são modelos de aprendizado supervisionado conhecidos pela clareza interpretativa e pela versatilidade em diversas tarefas. Quando utilizadas em problemas de classificação, as variáveis-alvo são categóricas (como na Árvore de Classificação), enquanto os problemas que envolvem variáveis numéricas caracterizam

uma Árvore de Regressão. A Figura 2.3 ilustra um exemplo da estrutura de uma Árvore de Decisão.

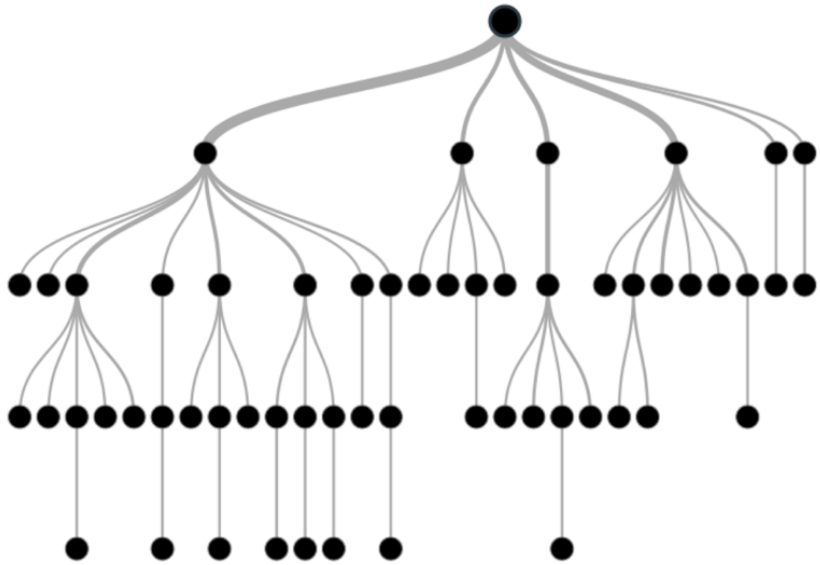

Figura 2.3: Exemplo de Árvore de Decisão
Fonte: Adaptado de (Charbuty & Abdulazeez, 2021)

Charbuty and Abdulazeez (2021) apresentam alguns conceitos fundamentais para compreender o funcionamento interno desses modelos:

- *Nó inicial*: corresponde à amostra total de dados que será segmentada de acordo com as decisões tomadas.
- *Nó intermediário*: representa subamostras geradas após a aplicação de um critério decisional. Cada nó intermediário pode continuar a ser particionado.
- *Nó final*: não sofre mais divisões, resultando em um caminho concluído no processo decisional.
- *Nó de probabilidade (intermediário ou folha)*: exibido graficamente como um círculo, indicando a probabilidade de um determinado resultado ocorrer.
- *Nó de decisão*: ilustrado por um quadrado, denota o ponto em que uma escolha é realizada para dividir os dados.

O processo de construção da árvore segue, de modo geral, os seguintes passos:

- Atribuição do conjunto completo de dados ao nó inicial.

- Aplicação de sucessivas condições de decisão, onde cada subdivisão cria duas novas partições, que são repassadas ao nó subsequente. O procedimento se repete até atingir os nós finais.
- Cada nó final representa um resultado definitivo, considerando os critérios de segmentação adotados.

A Figura 2.4 ilustra, de forma esquemática, a dinâmica das partições sucessivas em uma Árvore de Decisão:

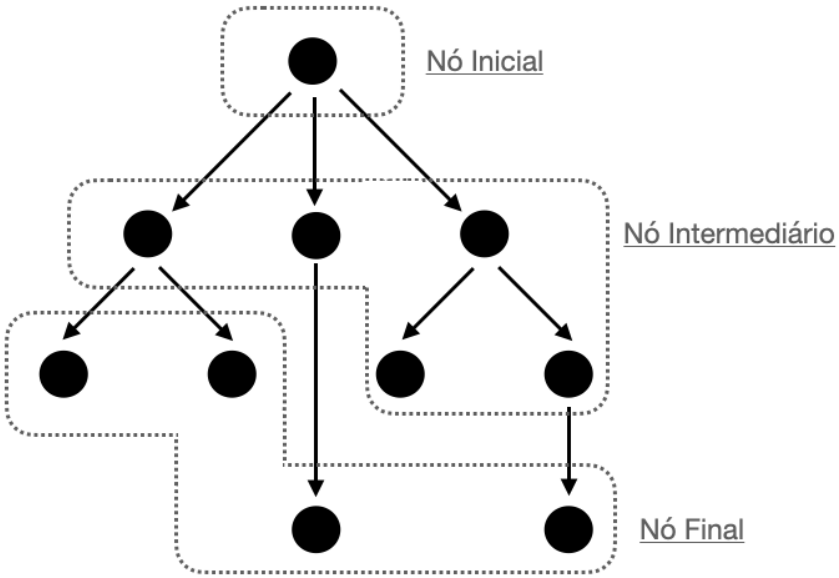

Figura 2.4: Funcionamento de uma Árvore de Decisão
Fonte: Adaptado de Charbuty and Abdulazeez (2021)

A formalização do modelo pode ser representada por uma função de mapeamento binário ou por um conjunto de variáveis de entrada e uma variável de saída Y:

$$f : \{0,1\}^n \rightarrow \{0,1\} \quad \text{ou} \quad (\mathbf{x}, Y) = (x_1, x_2, x_3, ..., x_k, Y). \tag{2.6}$$

Estudos empíricos indicam que as Árvores de Decisão geralmente oferecem bom desempenho e compreensibilidade em problemas de classificação. Em uma comparação de métodos conduzida por Caruana and Niculescu-Mizil (2006) em onze bases de dados distintas — avaliadas por oito métricas de desempenho —, as melhores posições foram ocupadas por *Boosted Trees*, *Random Forest*, *Bagged Trees* (variante de Árvores de Decisão) e *Support Vector Machines* (SVM), nessa ordem.

No mesmo sentido, Guégan and Hassani (2018) analisou diferentes estratégias de aprendizado de máquina para *credit scoring* em instituições financeiras, sugerindo que as *Random Forest* tendem a mostrar resultados competitivos, embora ressalte a necessidade de considerar a evolução dos dados ao longo do tempo. Já Munkhdalai, Namsrai,

Lee, and Ryu (2019) comparou abordagens humanas de previsão de crédito com algoritmos de aprendizado de máquina, concluindo que os modelos baseados em redes neurais e *Extreme Gradient Boosting* obtiveram elevada *AUC* e precisão, possibilitando uma menor perda de crédito esperada.

Em termos conceituais, uma Árvore de Decisão pode ser vista como uma sucessão de instruções condicionais (*if-else*), nas quais cada critério de divisão gera novas ramificações, resultando em partições de dados cada vez mais homogêneas. Embora métodos como SVM, regressão logística e redes neurais profundas abordem problemas semelhantes, as Árvores de Decisão oferecem maior interpretabilidade, pois permitem identificar claramente quais variáveis determinam a separação das classes. Contudo, quando é necessária alta performance em grandes volumes de dados, algoritmos mais complexos, como ensembles (*Boosted Trees* ou *Random Forests*), podem ser considerados em função de seus ganhos de acurácia, ainda que exijam um nível maior de ajustes de hiperparâmetros.

Floresta Aleatória (*Random Forest*)

A Floresta Aleatória é um método robusto de aprendizado supervisionado, baseado na técnica de *bagging* (*bootstrap aggregating*), cujo principal objetivo é reduzir a correlação entre Árvores de Decisão por meio da introdução de aleatoriedade tanto na seleção dos subconjuntos de dados quanto na escolha dos preditores. Diferentemente do *bagging* tradicional, esse algoritmo constrói múltiplas árvores a partir de subconjuntos aleatórios de amostras e variáveis, assegurando que cada árvore apresente características distintas em sua capacidade preditiva (Mishina, Murata, Yamauchi, Yamashita, & Fujiyoshi, 2015).

Segundo Breiman (2001), a Floresta Aleatória associa vetores aleatórios às amostras, gerando Árvores de Decisão independentes, apesar de serem oriundas da mesma distribuição estatística. À medida que o número de árvores aumenta, a previsão agregada do modelo tende a convergir para a média (no caso de regressão) ou para a moda (no caso de classificação) das previsões produzidas por todas as árvores individuais. A Figura 2.5 ilustra o funcionamento esquemático desse processo.

Para definir o número de variáveis consideradas em cada nó de decisão, utilizam-se critérios de amostragem aleatória. Uma fórmula frequentemente adotada é:

$$m = \sqrt{p} \quad \text{ou} \quad \log_2(p), \tag{2.7}$$

onde m representa a quantidade de preditores selecionados aleatoriamente do total disponível em cada nó, e p indica o número total de preditores disponíveis. A cada divisão, um novo subconjunto de variáveis é sorteado, proporcionando diversidade entre as árvores construídas. Essa estratégia evita que todas as árvores tenham a mesma estrutura, pro-

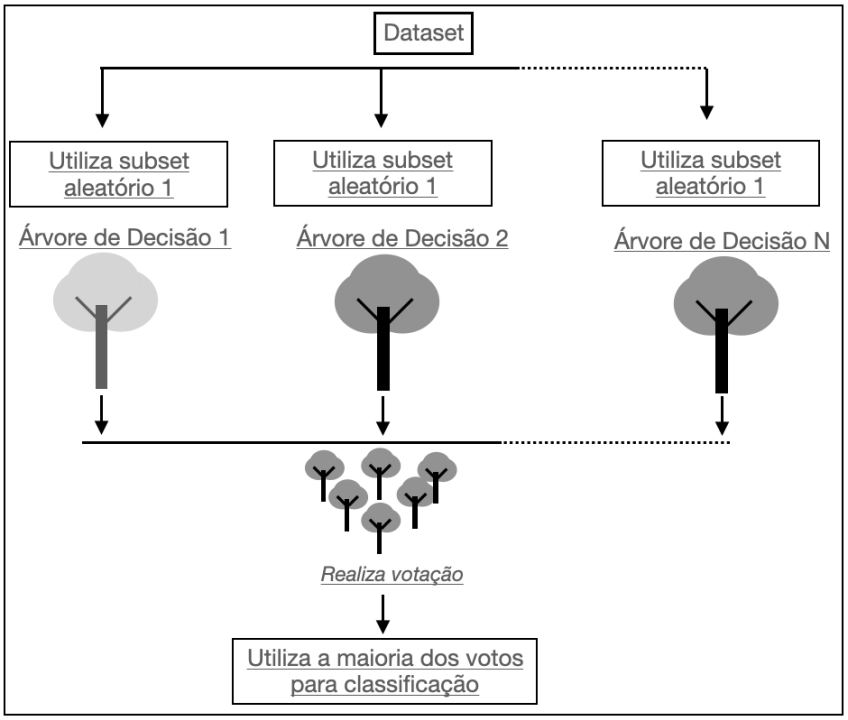

Figura 2.5: *Random Forest*
Fonte: Elaborado pelo autor

movendo a independência entre elas e mitigando o risco de sobreajuste (*overfitting*). Uma das principais vantagens da Floresta Aleatória reside em sua capacidade de manter um bom desempenho mesmo em contextos com dados ruidosos, desbalanceados ou contendo interações complexas entre variáveis. Além disso, o algoritmo fornece estimativas internas de erro por meio da técnica *Out-of-Bag* (OOB), na qual as amostras não utilizadas para treinar uma árvore específica são aproveitadas para avaliar sua acurácia. Esse recurso elimina a necessidade de um conjunto de validação separado, otimizando o uso dos dados disponíveis.

Outro aspecto relevante da Floresta Aleatória é a sua capacidade de mensurar a importância das variáveis preditoras. Essa análise é baseada na contribuição de cada variável para a redução da impureza nos nós das árvores ou na variação da acurácia do modelo ao permutar aleatoriamente os valores dessa variável. A mensuração da importância dos atributos auxilia na seleção de variáveis mais relevantes e fornece subsídios interpretativos para a compreensão dos resultados obtidos.

Apesar de sua eficácia e robustez, a Floresta Aleatória apresenta como limitação a redução da interpretabilidade em relação a modelos lineares, como a Regressão Logística. No entanto, essa desvantagem pode ser parcialmente contornada com o uso de técnicas de inteligência artificial explicável (*Explainable AI*), como os valores de SHAP (*Shapley Additive Explanations*), que permitem decompor a predição final em contribuições individuais

de cada variável, promovendo maior transparência nos resultados.

Gradient Boosting

O algoritmo *Gradient Boosting*, introduzido por Breiman (1997), representa uma evolução das técnicas de impulsionamento (*boosting*) no contexto do aprendizado supervisionado. Sua principal proposta é construir um modelo preditivo robusto por meio da combinação sequencial de modelos fracos, geralmente árvores de decisão de baixa profundidade. A cada iteração, o objetivo é reduzir o erro residual gerado pelas predições anteriores por meio da otimização de uma função de perda.

Diferentemente do *AdaBoost*, que ajusta os pesos das observações com base em erros anteriores, o *Gradient Boosting* realiza a atualização dos modelos com base no gradiente da função de perda, calculado em relação aos parâmetros do modelo atual. Esse gradiente fornece a direção mais promissora para minimizar o erro, orientando a construção de novos estimadores. A Figura 2.6 ilustra, de forma esquemática, o processo iterativo desse algoritmo.

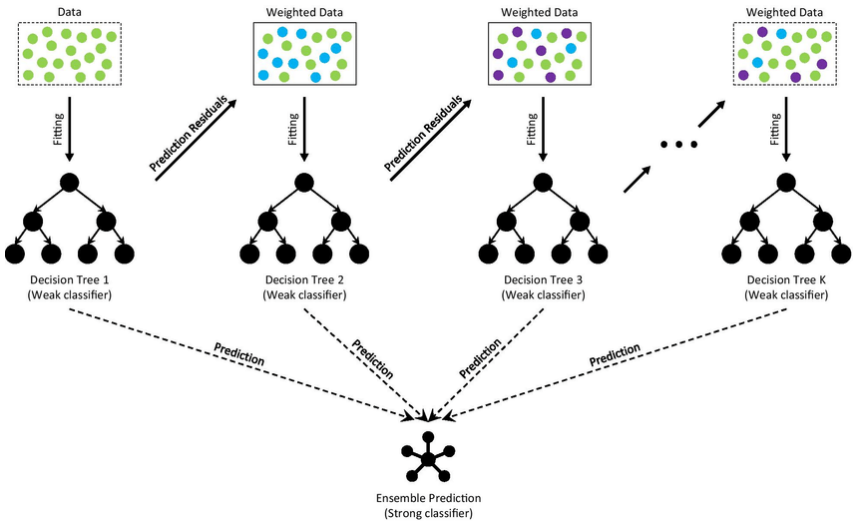

Figura 2.6: Esquema do *Gradient Boosting*
Fonte: Adaptado de Deng, Zhou, Wang, and Zhang (2021)

Em cada estágio, o algoritmo calcula o gradiente dos erros relativos aos parâmetros atuais do modelo, definindo a direção na qual esses parâmetros devem ser atualizados para reduzir a diferença entre os valores reais e previstos (C. Li, 2016). Matematicamente, em cada iteração m, o modelo preditivo $F_m(x)$ é atualizado com base no erro entre o valor verdadeiro y e a predição corrente. A atualização é definida pela equação:

$$h\_m(x) = y - F_m(x) \tag{2.8}$$

Na qual $h_m(x)$ representa a estimativa do resíduo. A função $F_m(x)$ é então ajustada ao longo das iterações, com a adição ponderada de novos estimadores $h_m(x)$, formando:

$$F_{m+1}(x) = F_m(x) + \alpha h_m(x) \tag{2.9}$$

em que α representa a taxa de aprendizado (*learning rate*), um parâmetro que regula a influência de cada nova árvore no modelo final. A escolha apropriada de α e do número total de iterações é fundamental para o desempenho do algoritmo, influenciando diretamente o equilíbrio entre viés e variância.

O *Gradient Boosting* é considerado altamente eficaz em tarefas de classificação e regressão, especialmente em problemas com interações complexas e dados heterogêneos. Sua flexibilidade permite o uso de diferentes funções de perda, como erro quadrático médio, entropia cruzada e funções robustas a outliers.

Apesar de sua elevada acurácia preditiva, o algoritmo pode ser sensível ao sobreajuste (*overfitting*), especialmente quando mal parametrizado. Por essa razão, técnicas complementares, como validação cruzada, regularização por penalidades (*shrinkage*) e *early stopping* , são frequentemente aplicadas.

Por fim, no contexto de contratações públicas e análise de risco, o *Gradient Boosting* se destaca por sua habilidade em capturar relações não lineares entre variáveis explicativas e desfechos binários, sendo frequentemente empregado em sistemas de detecção de fraudes e priorização de auditorias.

Extreme Gradient Boosting (XGBoost)

Conforme descrito por Chen and Guestrin (2016), o *Extreme Gradient Boosting* (*XGBoost*) representa uma implementação otimizada e escalável do algoritmo de *Gradient Boosting*, projetada para oferecer desempenho computacional e precisão preditiva superiores. Embora compartilhe os mesmos fundamentos teóricos — a adição sequencial de modelos fracos (árvores de decisão) para corrigir os resíduos de um modelo anterior — o *XGBoost* introduz otimizações substanciais em sua engenharia e, crucialmente, em sua formalização matemática (Chen & Guestrin, 2016).

A principal inovação do *XGBoost* reside em sua função objetivo regularizada. Enquanto o *Gradient Boosting* tradicional (descrito na Seção 2.3.5) tipicamente otimiza uma função de perda L, o *XGBoost* minimiza uma função objetivo $\mathcal{L}$ que combina a perda (treinamento) e a complexidade do modelo (regularização) em cada etapa t:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2.10}$$

Onde l é uma função de perda diferenciável (como *log-loss* ou erro quadrático), $\hat{y}_i^{(t-1)}$ representa a predição da iteração anterior, $f_t(x_i)$ denota a nova árvore a ser adicionada, e Ω refere-se ao termo de regularização que penaliza a complexidade do modelo (Chen & Guestrin, 2016).

O termo de regularização Ω é fundamental e impede o sobreajuste (*overfitting*). Ele é definido não apenas pela magnitude dos pesos, mas também pela estrutura da própria árvore:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + \alpha \sum_{j=1}^{T} |w_j| \tag{2.11}$$

onde T é o número de folhas na árvore, w_j é o escore (peso) da folha j, γ é a penalidade pela complexidade de adicionar uma nova folha, λ é o parâmetro de regularização L_2 (*Ridge*) e α é o parâmetro de regularização L_1 (*Lasso*) aplicado aos pesos das folhas (Chen & Guestrin, 2016, p. 786).

Para otimizar essa função objetivo de forma eficiente, o *XGBoost* utiliza uma expansão de Taylor de segunda ordem para aproximar a função de perda Justificando o uso da Hessiana, como mencionado em Chen and Guestrin (2016):

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{2.12}$$

Onde $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ é o gradiente (primeira derivada) e $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ é a Hessiana (segunda derivada) da função de perda. Esta abordagem, que utiliza derivadas de segunda ordem, converge mais rapidamente do que o *Gradient Boosting* tradicional, que geralmente se baseia apenas no gradiente (g_i).

Outra inovação significativa reside no tratamento eficiente de valores ausentes (*missing values*) (Chen & Guestrin, 2016). O XGBoost aprende uma "direção padrão"(*default direction*) em cada nó da árvore durante o treinamento. Ao encontrar um valor ausente, a instância é direcionada para essa direção padrão aprendida, eliminando a necessidade de imputação prévia dos dados (Chen & Guestrin, 2016).

Do ponto de vista computacional, o XGBoost foi projetado para escalabilidade. Ele utiliza uma estrutura de dados chamada *Column Block*, que armazena os dados de forma colunar e pré-organizada. Isso permite que o processo de busca pelo melhor ponto de corte (*split point*) em cada nó seja paralelizado, uma vez que a varredura das *features* é independente. O algoritmo também é *cache-aware*, otimizando o uso da hierarquia de memória da CPU (Chen & Guestrin, 2016).

Embora seu ajuste exija maior atenção à calibração de hiperparâmetros — como profundidade das árvores (*max_depth*), taxa de aprendizado (*learning_rate*), *subsample* e

colsample_bytree — o *XGBoost* oferece considerável flexibilidade e capacidade preditiva. Isso o torna especialmente eficaz em aplicações complexas de classificação binária e multiclasse, bem como em problemas de regressão e classificação por ranking.

Por fim, devido à sua robustez, escalabilidade e compatibilidade com métodos de interpretabilidade, como *SHAP values*, o *XGBoost* tem tido grande relevância em trabalhos de *Machine Learning*.

Light Gradient Boosting Machine (*LightGBM*)

O *Light Gradient Boosting Machine* (LightGBM), desenvolvido por Ke et al. (2017), é um algoritmo de aprendizado supervisionado que se baseia na técnica de *gradient boosting* e foi projetado para oferecer alto desempenho em tarefas de classificação e regressão, especialmente em contextos que envolvem grandes volumes de dados e alta dimensionalidade. Embora compartilhe princípios fundamentais com algoritmos como o *XGBoost*, o *LightGBM* introduz inovações estruturais que otimizam significativamente a eficiência computacional e o uso da memória.

Uma das principais diferenças em relação aos métodos tradicionais de construção de árvores está na estratégia de particionamento adotada. Enquanto os algoritmos convencionais seguem uma abordagem level-wise — realizando divisões simultâneas em todos os nós de um mesmo nível — o LightGBM adota a estratégia denominada *Leaf-wise with Depth Limitation*. Nessa abordagem, a árvore é expandida a partir do nó com o maior ganho na função objetivo, promovendo divisões mais seletivas e eficientes. Embora essa técnica possa resultar em árvores mais profundas, o uso de uma limitação explícita de profundidade permite mitigar os riscos de sobreajuste e manter a interpretabilidade.

Além disso, o *LightGBM* incorpora dois mecanismos que potencializam sua escalabilidade. O primeiro é o *binning*, que consiste na discretização de variáveis contínuas em intervalos (bins) fixos. Esse procedimento reduz drasticamente o número de comparações necessárias para encontrar os pontos de divisão ótimos, contribuindo para a agilidade no treinamento. O segundo é o método *Gradient-Based One-Side Sampling* (GOSS), responsável por selecionar preferencialmente as amostras com maiores gradientes residuais — ou seja, aquelas que mais contribuem para o erro atual do modelo — enquanto realiza amostragem aleatória das demais. Com isso, o GOSS acelera a convergência do modelo sem comprometer sua acurácia preditiva.

Um diferencial do *LightGBM* é seu suporte nativo a execução paralela e distribuída, o que o torna altamente eficiente em ambientes com múltiplos núcleos de processamento ou *clusters* computacionais. Essa característica é particularmente vantajosa em aplicações industriais e competições de ciência de dados, nas quais o tempo de treinamento e a escalabilidade são fatores críticos.

Em termos de desempenho, o *LightGBM* apresenta resultados competitivos em comparação com outras implementações de *boosting*, como o *XGBoost*, com a vantagem adicional de reduzir significativamente o tempo de processamento em grandes conjuntos de dados. No entanto, assim como ocorre com outros algoritmos baseados em árvores, seu desempenho pode ser sensível a hiperparâmetros, como profundidade máxima, número de folhas, taxa de aprendizado e número de iterações.

Por fim, destaca-se que o *LightGBM* também é compatível com métodos de explicabilidade, como os valores de SHAP (*SHapley Additive exPlanations*), ampliando sua aplicabilidade em contextos que demandam transparência nas decisões, como auditorias públicas, diagnósticos médicos e sistemas de crédito.

Tabular Prior–data Fitted Network (TabPFN)

O *Modelo Tabular Prior–data Fitted Network* (TabPFN) (Hollmann, Müller, Eggensperger, & Hutter, 2023) representa um avanço significativo na aprendizagem de máquina para dados tabulares, pois transforma o processo de ajuste de modelos em um problema pré-treinado de inferência bayesiana. A ideia central é treinar, uma única vez, um *Transformer* com o objetivo de aproximar a distribuição preditiva posterior $p(y \mid x_{\text{test}}, D_{\text{train}})$ por meio de *in-context learning*.

Durante o treinamento "off-line", milhões de conjuntos sintéticos são gerados a partir de um *prior* que combina dois mecanismos: (i) *Modelos Causais Estruturais (Structural Causal Models* - SCM), que introduzem dependências causais explícitas entre atributos e o alvo, e (ii) *Redes Neurais Bayesianas (Bayesian Neural Networks* - BNN), que adicionam relações altamente não lineares e incerteza paramétrica. A cada amostra do prior, constrói-se um grafo acíclico direcionado, sorteiam-se nós a serem observados como características (features) e um nó como rótulo; ruídos são propagados através do grafo e, ao final, aplica-se uma quantização aleatória para mapear o valor contínuo do nó-rótulo em classes discretas. Essa estratégia gera tarefas com um número variável de atributos, classes, frações de valores ausentes e escalas de variância, forçando o modelo a aprender um algoritmo versátil.

A rede recebe todo o conjunto de dados como uma sequência bidimensional: primeiro, as atenções atuam sobre as colunas de cada linha; em seguida, atuam sobre as linhas de cada coluna. Esse desenho assegura a invariância por permutação tanto de amostras quanto de atributos, permitindo que um único *forward pass* produza, simultaneamente, distribuições preditivas para todas as observações não rotuladas. No treinamento descrito em Hollmann et al. (2023), utiliza-se um *Transformer* com 12 camadas e 512 unidades ocultas, totalizando aproximadamente 26 milhões de parâmetros, otimizado através de 18.000 lotes de 512 conjuntos de dados sintéticos cada. O critério de otimização consiste

na expectativa, sobre o *prior*, da negativa do logaritmo da densidade predita avaliada no rótulo verdadeiro.

$$\mathcal{L} = E_{(x_{\text{test}}, y_{\text{test}}, D_{\text{train}}) \sim p(D)} \Big[-\log q_\theta (y_{\text{test}} \mid x_{\text{test}}, D_{\text{train}}) \Big], \tag{2.13}$$

de forma que a minimização de $\mathcal{L}$ aproxima a predição bayesiana dada na integral

$$p(y \mid x_{\text{test}}, D_{\text{train}}) \propto \int_{\Phi} p(y \mid x_{\text{test}}, \varphi)\, p(D_{\text{train}} \mid \varphi)\, p(\varphi)\, d\varphi. \tag{2.14}$$

Finalizada essa etapa, o modelo torna-se um algoritmo de classificação "universal" para conjuntos tabulares de até dez mil linhas: basta apresentar D_{train} concatenado a X_{test} e obter, em menos de 0,05 s de consumo em GPU, densidades preditivas calibradas – desempenho que, em testes padronizados do AutoML-benchmark, supera *árvores de* decisão com *boosting* de gradiente ajustadas por quatro horas, apresentando um aumento de velocidade superior a 5.000× (Hollmann et al., 2023). A representação probabilística rica também possibilita tarefas adicionais: geração de novas linhas sintéticas por meio de amostragem autoregressiva das distribuições internas; estimação de densidade para detecção de anomalias; e extração de *embeddings* para transferência de aprendizado.

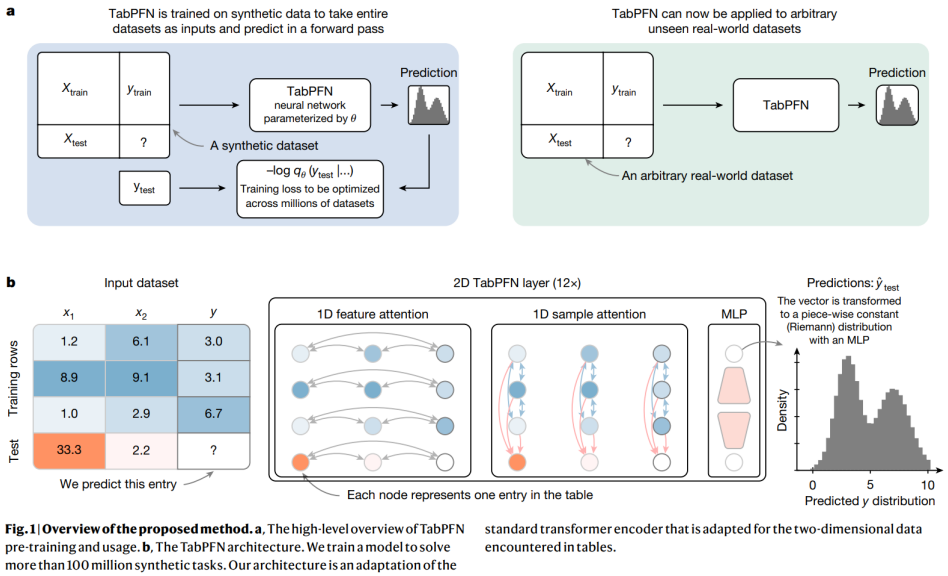

Fig. 1 | Overview of the proposed method. a, The high-level overview of TabPFN pre-training and usage. **b**, The TabPFN architecture. We train a model to solve more than 100 million synthetic tasks. Our architecture is an adaptation of the standard transformer encoder that is adapted for the two-dimensional data encountered in tables.

Figura 2.7: Estrutura do modelo TabPFN
Fonte: Adaptado de Hollmann et al. (2023)

Para ilustrar a arquitetura, a Figura 2.7 do artigo original mostra o diagrama de pré-treinamento e uso: à esquerda, o fluxo de datasets sintéticos alimentando o *Transformer*; à direita, o layout 2-D de atenção por linhas e colunas. A Figura 2.8 detalha o processo de geração de conjuntos de dados a partir do grafo causal, enquanto a Figura 3 compara a modelagem de funções-brinquedo — seno desalinhado, função degrau e dispersão hetero-

cedástica — revelando que o TabPFN aprende tanto padrões suaves quanto abruptos, algo que está além do alcance de regressões lineares ou *de Perceptrons* Multicamadas (MLPs) convencionais.

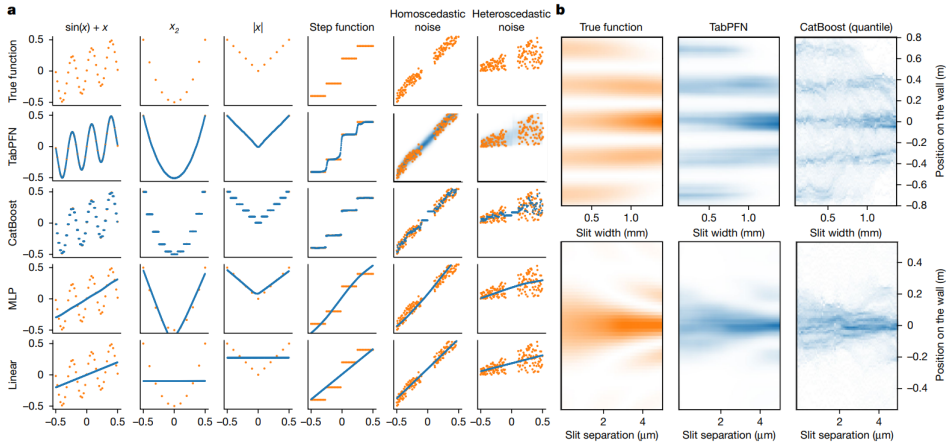

Fig. 3 | The behaviour of TabPFN and a set of baselines on simple functions. In all plots, we use orange for the ground truth and blue for model predictions. **a**, Each column represents a different toy function, each having a single feature (along the *x*-axis) and a target (along the *y*-axis). TabPFN can model a lot of different functions, including noisy functions. **b**, TabPFN can model distributions over outputs out of the box, which is exemplified by predicting the light intensity pattern in a double-slit experiment after observing the positions of 1,000 photons.

Figura 2.8: Exemplos de aprendizado de padrões pelo modelo TabPFN em um conjunto de funções diferentes.
Fonte: Adaptado de Hollmann et al. (2023)

No contexto de auditoria de contratos públicos, essas propriedades são especialmente relevantes: a modelagem explícita da incerteza reduz falsos positivos, e o treinamento prévio permite analisar bases históricas heterogêneas sem reajustes demorados.

2.3.1 Métodos de Ensemble e Arquiteturas Híbridas

O Aprendizado de Máquina por Comitês, ou *Ensemble Learning*, consiste em um paradigma metodológico no qual múltiplos modelos, denominados aprendizes base ou *weak learners*, são combinados estrategicamente para resolver um mesmo problema computacional. O pressuposto teórico fundamental dessa técnica reside na redução do erro de generalização por meio da agregação de hipóteses distintas, o que explora a diversidade estatística e estrutural dos modelos individuais (Dietterich, 2000).

Matematicamente, o erro esperado de um modelo pode ser decomposto em três componentes: viés (*bias*), variância (*variance*) e erro irredutível (ruído). Os métodos de *ensemble* atuam na mitigação dos dois primeiros componentes. A decomposição da ambiguidade, formalizada por Krogh and Vedelsby (1995), estabelece que o erro quadrático médio de um *ensemble* (E) é garantidamente menor ou igual à média dos erros quadráticos médios dos modelos individuais ($\bar{E}$), sendo essa diferença quantificada por um termo de diversidade (D):

$$E = \bar{E} - D \tag{2.15}$$

Onde D representa a ambiguidade ou a variância das predições dos modelos individuais em torno da média do *conjunto*. Essa relação matemática demonstra que o desempenho do sistema composto é maximizado não apenas pela acurácia individual dos modelos base, mas, significativamente, pela discordância (diversidade) entre eles. Se os modelos cometem erros idênticos, $D \to 0$, não há ganho no *ensemble*; entretanto, se os erros forem não-correlacionados, o ganho é maximizado.

Os métodos de *ensemble* são categorizados, primordialmente, em homogêneos e heterogêneos.

Os ensembles homogêneos utilizam o mesmo algoritmo base aplicado a diferentes distribuições dos dados de treinamento ou por meio de reponderação iterativa.

O Bagging (Bootstrap Aggregating),proposto por Breiman (1996), foca na redução da variância. Geram-se M subconjuntos de dados por meio de amostragem com reposição, e treina-se um modelo independente para cada um deles. A predição final é a média (em regressão) ou a moda (em classificação) das predições. O *Random Forest* é o exemplo canônico desta classe.

O Boosting foca na redução do viés e da variância. Os modelos são treinados sequencialmente, onde o modelo m busca corrigir os erros residuais do modelo $m-1$. Algoritmos como *Gradient Boosting Machine* (GBM), formalizado por Friedman (2001), e suas implementações otimizadas (*XGBoost*, *LightGBM*), pertencem a esta categoria, otimizando uma função de perda diferenciável no espaço de funções.

Diferentemente das abordagens anteriores, os *ensembles* heterogêneos ou híbridos combinam modelos gerados por diferentes vieses indutivos (e.g., Árvores de Decisão, Redes Neurais baseadas em Transformers, Modelos Lineares). A premissa, fundamentada na teoria da Generalização Empilhada (*Stacked Generalization*) de Wolpert (1992), é que algoritmos distintos tendem a convergir para ótimos locais diferentes e a falhar em regiões distintas do espaço de características.

A combinação das predições em arquiteturas híbridas ocorre frequentemente por meio de técnicas de *Voting* (Votação). No contexto da classificação probabilística e detecção de fraudes, o método de *Soft Voting* demonstra-ser teoricamente superior ao *Hard Voting* (votação majoritária simples), pois preserva a informação sobre a incerteza do modelo (Kittler, Hatef, Duin, & Matas, 1998). Seja um conjunto de M classificadores, onde cada modelo m estima uma probabilidade posterior $P_m(y_j|\mathbf{x})$ para a classe j dado o vetor de entrada $\mathbf{x}$. A probabilidade final do *ensemble* é dada pela média ponderada das probabilidades individuais:

$$\hat{P}_{ensemble}(y_j|\mathbf{x}) = \sum_{m=1}^{M} w_m P_m(y_j|\mathbf{x}) \tag{2.16}$$

sujeito a $\sum w_m = 1$ e $w_m \geq 0$. A classe predita $\hat{y}$ é aquela que maximiza a probabilidade agregada:

$$\hat{y} = \arg\max_j \left[\sum_{m=1}^{M} w_m P_m(y_j|\mathbf{x}) \right] \tag{2.17}$$

Esta abordagem é particularmente eficaz quando os modelos base são bem calibrados, permitindo que classificadores com maior "certeza"em determinadas instâncias influenciem mais fortemente a decisão final e suavizem erros de superconfiança de modelos individuais.

Uma extensão avançada de *ensembles* híbridos envolve a diversificação não apenas dos algoritmos, mas também das distribuições de dados de treinamento, uma estratégia crucial em domínios de aprendizado desbalanceado (He & Garcia, 2009a). Diferentes técnicas de reamostragem alteram a distribuição a priori das classes, modificando a fronteira de decisão aprendida.

Um modelo treinado com *Undersampling* tende a maximizar a sensibilidade (*Recall*) da classe minoritária, porém com um maior risco de falsos positivos. Em contraste, um modelo treinado com a distribuição original (*Imbalance*) tende a preservar a Precisão. A combinação híbrida desses modelos em um único *ensemble* permite construir uma fronteira de decisão que equilibra o *trade-off* entre precisão e revocação, estabilizando a variância associada às técnicas de amostragem e resultando em um estimador de risco mais robusto.

2.3.2 Métricas de Validação do Modelo

Ao realizarmos a avaliação de um modelo, este pode apresentar baixo viés e alta variância, sugerindo um sinal indicativo de Overfitting, ou apresentar baixa variância e alto viés, sugerindo um sinal indicativo de Underfitting. Também é importante ressaltar que, em machine learning, trabalhamos com os termos acurácia e precisão, onde uma alta acurácia corresponde a uma baixa variância e uma alta precisão corresponde a um baixo viés (Koehrsen, 2018).

Compreender como esses termos se relacionam ao realizarmos o ajuste do modelo é fundamental. para isso, primeiro precisamos entender o conceito de Matriz de Confusão.

Segundo Visa, Ramsay, Ralescu, and Van Der Knaap (2011), a Matriz de Confusão é composta por quatro valores:

- Verdadeiro positivo (VP): quantidade de classificações verdadeiramente positivas. O modelo obteve sucesso na classificação positiva.

- Falso positivo (FP): quantidade de classificações incorretamente positivas. O modelo classificou incorretamente um registro negativo como positivo.
- Verdadeiro negativo (VN): quantidade de classificações verdadeiramente negativas. O modelo obteve sucesso na classificação negativa.
- Falso negativo (FN): quantidade de classificações incorretamente negativas. O modelo classificou incorretamente um registro positivo como negativo.

Uma Tabela 2.4 representa como os dados se comportam na Matriz de Confusão.

Tabela 2.4: Matriz de Confusão

	Valor Previsto	
Valor Verdadeiro	Positivo	Negativo
Positivo	*Verdadeiro Positivo (VP)*	*Falso Positivo (FP)*
Negativo	*Falso Negativo (FP)*	*Verdadeiro Negativo (VN)*

Fonte: Elaborado pelo autor

Por meio da avaliação dos valores apresentados na Matriz de Confusão, são calculadas as medidas de avaliação: Acurácia, Precisão, Especificidade, Revocação, F1-Score e Curva ROC. Essas medidas serão detalhadas a seguir, conforme Powers (2020) e Sokolova, Japkowicz, and Szpakowicz (2006).

É importante destacar que a avaliação de modelos de classificação em cenários desbalanceados — como na detecção de fornecedores fraudulentos — exige métricas especializadas. Medidas tradicionais, como a acurácia, tornam-se enganosas nesses contextos, exigindo abordagens que capturem o desequilíbrio entre classes (He & Garcia, 2009b).

Acurácia

Medida básica para calcular o desempenho do modelo. Quanto maior for o valor da acurácia, maior será o número de acertos do modelo e menos erros serão cometidos. Representa quão próximas um conjunto de medições está do valor real. A Figura 2.9 ilustra o foco pretendido da Acurácia.

É considerada um indicador de cálculo simples, obtido pela divisão do total de acertos pelo total geral. Sua fórmula pode ser representada da seguinte forma:

$$\text{Acurácia} = \frac{vp + vn}{vp + vn + fp + fn} \tag{2.18}$$

Após a obtenção da Acurácia, calcula-se o erro do modelo com a fórmula a seguir:

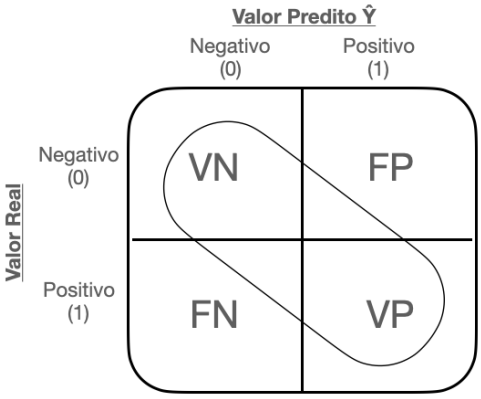

Figura 2.9: Acurácia
Fonte: Elaborado pelo autor

$$\text{Erro} = 1 - \textit{Acurácia} \tag{2.19}$$

É importante ressaltar que a Acurácia fornece valores globais e não é suficiente para elucidar o que ocorre em determinados modelos. Podemos identificar uma acurácia elevada; no entanto, isso pode ocorrer devido ao baixo balanceamento dos dados. Ou seja, se tivermos um conjunto de dados com classes predominantes, o modelo, ao classificar corretamente os dados predominantes, obterá uma alta acurácia, mesmo que tenha falhado na classificação de todos os dados não predominantes.

Precisão

É a proporção das predições corretas de um conjunto em relação a todas as previsões feitas desse conjunto. Representa a fração recuperada considerada relevante; seu foco é verificar a proporção de acertos dos dados classificados como positivos, conforme indicado na Figura 2.10.

A fórmula pode ser apresentada da seguinte forma:

$$\text{Precisão} = \frac{vp}{vp + fp} \tag{2.20}$$

Assim como na Acurácia, o desbalanceamento dos dados também causa anomalias na avaliação, por meio da Precisão; ao não considerar os falsos negativos (fn), um modelo pode apresentar uma precisão alta, mesmo que sua taxa de verdadeiros positivos (tp) seja baixa.

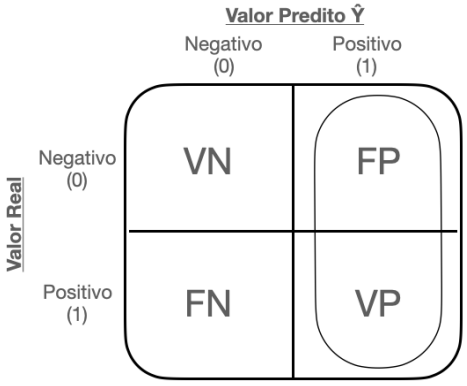

Figura 2.10: Precisão
Fonte: Elaborado pelo autor

Especificidade

Visa identificar a capacidade do modelo de prever a classificação negativa ao apresentar a proporção de dados classificados como verdadeiros negativos retornados pelo modelo em relação ao total de dados realmente negativos. Assim como é representado na Figura 2.11 a seguir.

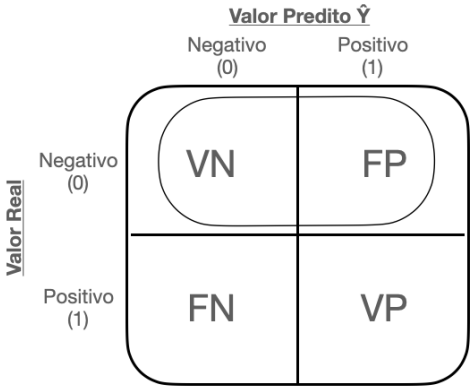

Figura 2.11: Especificidade
Fonte: Elaborado pelo autor

A fórmula pode ser representada da seguinte forma:

$$\text{taxa de verdadeiros negativos} = \frac{vn}{vn + fp} \tag{2.21}$$

Revocação (*Recall*) ou Sensibilidade

Visando mitigar os efeitos da Acurácia e da Precisão causados pelo desbalanceamento dos dados, esta métrica é utilizada para avaliar os valores recuperados entre os valores classifi-

cados como positivos reais. Seu foco visa quantificar quanto o modelo está capturando do que realmente deveria capturar. Ou seja, sua proporção de acertos. É indicado quando o objetivo é otimizar os verdadeiros positivos, mesmo que isso prejudique os falsos positivos. A Figura 2.12 ilustra o foco pretendido da Revocação.

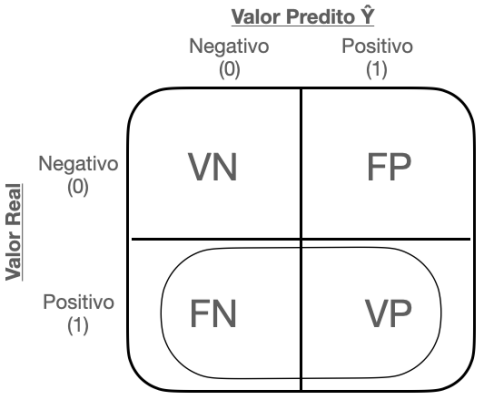

Figura 2.12: Revocação (Recall)
Fonte: Elaborado pelo autor

A fórmula é expressa da seguinte forma:

$$\text{Revocação} = \frac{vp}{vp + fn} \tag{2.22}$$

F1-Score

Utilizada como uma boa sugestão para mitigar problemas relacionados a conjuntos de dados desproporcionais. É considerada uma média harmônica entre a precisão e o recall. Ou seja, é uma forma de observar, em um único número, a relação entre a precisão e o recall, proporcionando uma visão da qualidade geral do modelo.

A fórmula pode ser representada da seguinte forma:

$$F1score = 2 \cdot \frac{\text{precis} * \text{revoc}}{\text{precis} + \text{revoc}} \tag{2.23}$$

Curva ROC

Utilizada para problemas de classificação binária, é gerada a partir da medição da Sensibilidade (*Recall*) e Especificidade. Permite acompanhar de forma gráfica a variação da Especificidade e Sensibilidade da performance do modelo, auxiliando, assim, na escolha do melhor ponto de corte para otimizar o desempenho. A Figura 2.13 a seguir descreve, de forma gráfica, como as curvas ROC se apresentam para um modelo de casos hipotéticos.

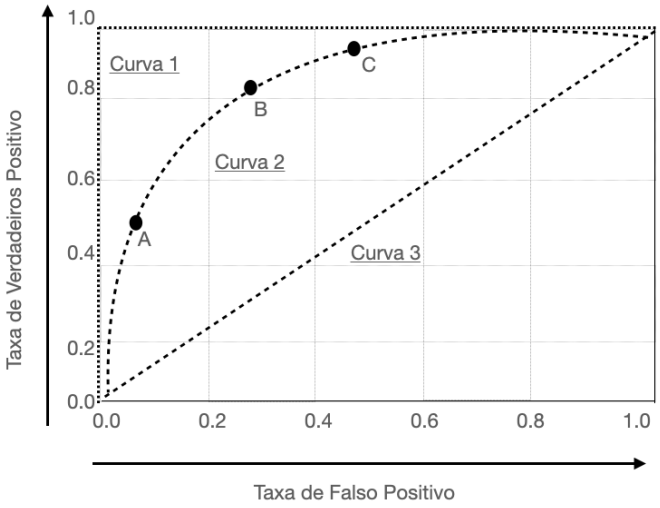

Figura 2.13: Curva ROC
Fonte: Elaborado pelo autor

A Figura 2.14 ilustra graficamente a distribuição dos dados presentes na área da Curva 1, onde não há falsos positivos e há cem por cento de verdadeiros positivos.

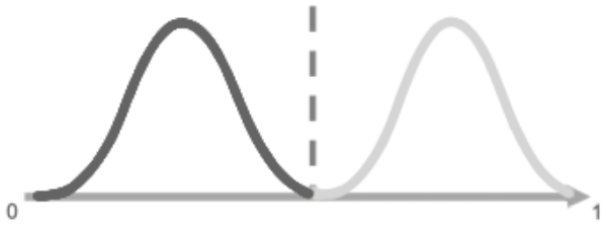

Figura 2.14: Curva 1
Fonte: Elaborado pelo autor

A Figura 2.15 ilustra graficamente a distribuição dos dados presentes na área da Curva 2, na qual existem possibilidades de falsos positivos e falsos negativos.

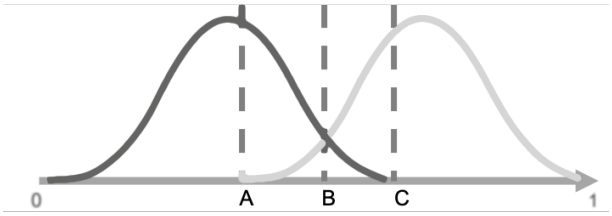

Figura 2.15: Curva 2
Fonte: Elaborado pelo autor

A Figura 2.16 ilustra graficamente a distribuição dos dados presentes na área da Curva 3, sendo que os valores são aleatórios e não agregam ao modelo.

Figura 2.16: Curva 3
Fonte: Elaborado pelo autor

A curva ROC permite discriminar o ponto de corte ideal, ou seja, quantificar a proporção dos resultados presentes na área sob a curva. Dessa forma, permite escolher a melhor distribuição para o problema proposto, como evidenciado pelos pontos A, B e C.

Em alguns casos, a curva ROC pode não ser facilmente interpretada, necessitando da aplicação do método AUC (Area Under the Curve), que apresenta valores entre 0 e 1; quanto mais próximo de 1, melhor. Ou seja, em um modelo cujas previsões são 100% corretas, terá AUC igual a 1. No entanto, para previsões 100% erradas, terá AUC igual a 0. Esse método é utilizado para simplificar a análise da curva ROC ao computar todos os pontos limiares da curva e apresentar a área sob a curva.

A fórmula pode ser representada da seguinte forma:

$$AUC(f) = \frac{\sum_{t_0 \in \mathcal{D}^0} \sum_{t_1 \in \mathcal{D}^1} \mathbf{1}[f(t_0) < f(t_1)]}{|\mathcal{D}^0| \cdot |\mathcal{D}^1|} \tag{2.24}$$

A curva ROC é considerada uma ferramenta comum de avaliação de modelos de classificação, devido à sua capacidade de avaliar a relação entre sensibilidade e especificidade. Dessa forma, é possível avaliar o ponto de corte que melhor atende a solução do problema proposto (Brown & Mues, 2012).

A AUC pode ser entendida como a eficiência do modelo, devido ao seu poder de discriminar os classificadores em todos os pontos de corte possíveis, sem considerar a distribuição das classes ou o custo de classificações erradas (Baesens et al., 2003).

LogLoss e Weighted LogLoss

A *Logarithmic Loss*, ou simplesmente LogLoss (também conhecida como Entropia Cruzada Binária), é uma métrica de avaliação utilizada em problemas de classificação em que a saída do classificador é uma probabilidade $p \in [0, 1]$, em vez de um rótulo de classe discreto. Diferentemente da acurácia, que contabiliza apenas se a predição final estava

correta ou incorreta com base em um limiar de corte, a LogLoss avalia a incerteza da predição, penalizando desvios entre a probabilidade estimada e o valor real da classe.

Matematicamente, para um conjunto de dados com N observações, a LogLoss é definida como o negativo da log-verossimilhança média:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{2.25}$$

Onde:

- N é o número total de observações;
- y_i representa o rótulo real da classe binária ($y_i \in \{0, 1\}$);
- p_i é a probabilidade predita de $y_i = 1$.

A Figura 2.17 ilustra o comportamento desta função. Observa-se que a penalidade cresce exponencialmente à medida que a probabilidade predita diverge da classe real. Por exemplo, se a classe verdadeira é 1 (Fraude) e o modelo prediz uma probabilidade próxima de 0 (confiança na regularidade), o valor da perda tende ao infinito. Essa característica força o modelo a ser calibrado, desencorajando predições confiantes que estejam incorretas.

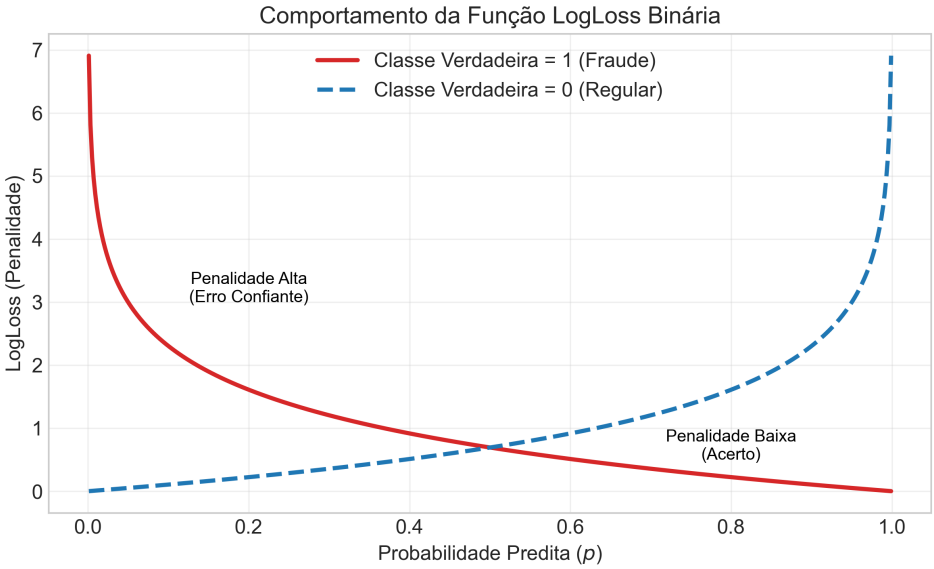

Figura 2.17: Comportamento da função LogLoss para classes binárias.
Fonte: Elaborado pelo autor.

Em cenários de detecção de fraudes em contratações públicas, os dados apresentam, invariavelmente, um severo desbalanceamento de classes, onde a quantidade de contratos regulares ($y = 0$) excede, em larga medida, a de contratos fraudulentos ($y = 1$). A utilização da LogLoss padrão nesses contextos pode induzir o modelo a priorizar a classe

majoritária, uma vez que a contribuição total do erro provém predominantemente dos exemplos negativos.

Para mitigar esse viés e alinhar a função de custo aos objetivos da auditoria — em que o custo de um Falso Negativo (não detectar uma fraude) é considerado superior ao de um Falso Positivo —, adota-se o *Weighted LogLoss*. Essa variação introduz um coeficiente de ponderação w, aplicado especificamente à classe positiva.

A formulação ajustada é apresentada da seguinte maneira:

$$\text{Weighted LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}[w \cdot y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{2.26}$$

O peso w é geralmente definido como o inverso da frequência da classe positiva ou é determinado pelo conhecimento de domínio sobre o custo do erro.

A Figura 2.18 compara as curvas de perda para uma instância positiva ($y = 1$). Nota-se que, ao aplicar o peso $w > 1$, a penalidade para erros na identificação de fraudes (baixa probabilidade predita para uma fraude real) torna-se significativamente maior do que na versão padrão.

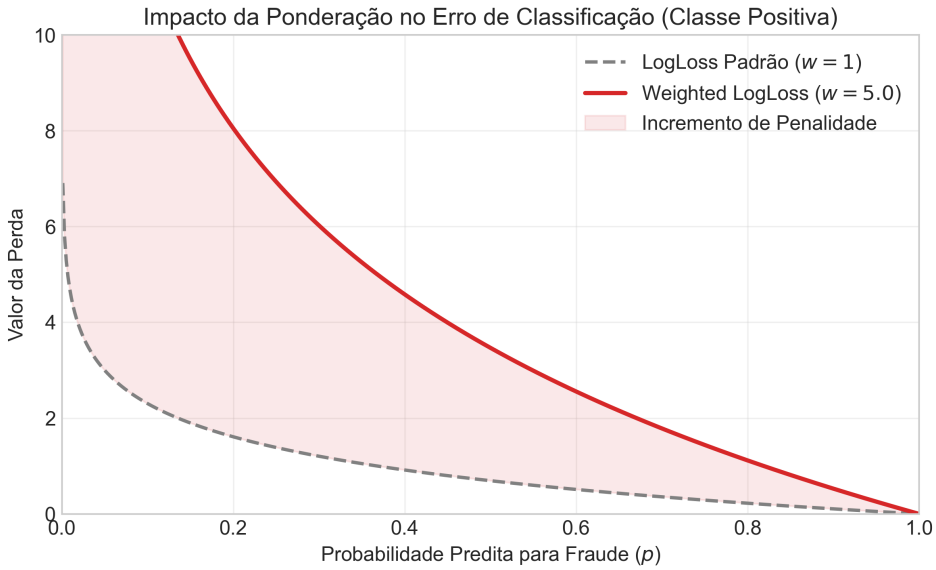

Figura 2.18: Comparação entre LogLoss Padrão e Ponderado para a classe positiva.
Fonte: Elaborado pelo autor.

Essa modificação na superfície de erro orienta o algoritmo de otimização (como o gradiente descendente em redes neurais ou *boosting*) a ajustar os parâmetros do modelo para recuperar com maior ênfase os exemplos da classe minoritária, resultando em um aumento da sensibilidade (*Recall*) sem a necessidade de alterar artificialmente o limiar de decisão *a posteriori* (He & Garcia, 2009b).

Métricas Compostas para Desbalanceamento e Trade-offs Operacionais

Cada métrica apresentada possui suas próprias peculiaridades que devem ser consideradas ao escolher o modelo de classificação a ser utilizado para resolver o problema concreto. Para isso, deve-se avaliar o problema e identificar o conjunto de métricas que melhor se adéque a solução, a fim de definir o ponto de corte que reflita a necessidade de maior otimização para os resultados tanto de sensibilidade quanto de especificidade. Em alguns casos, haverá necessidade de que o modelo acerte todos os verdadeiros positivos, mesmo que isso resulte em falsos positivos, como ocorre no diagnóstico de doenças. Assim, encontrar o equilíbrio é um processo iterativo que requer treinamento com diferentes combinações de parâmetros e conjuntos de dados.

Em cenários de classificação com dados fortemente desbalanceados – por exemplo, as métricas convencionais, como a acurácia, podem levar a interpretações equivocadas sobre o desempenho do modelo. Conforme discutido por Davis and Goadrich (2006), torna-se imperativo utilizar métricas compostas que integrem as dimensões de precisão e revocação, fornecendo uma avaliação mais realista, especialmente quando a classe de interesse é minoritária.

Uma métrica central para esses contextos é o Fβ-Score, que generaliza o F1-Score por meio da introdução de um parâmetro β, o qual permite ponderar de maneira diferenciada a precisão (P) e o recall (R). Essa métrica é definida pela seguinte equação:

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R}. \tag{2.27}$$

No caso específico da detecção de fraudes, em que os falsos negativos apresentam um custo operacional significativamente superior ao dos falsos positivos, adota-se o F2-Score ($\beta = 2$). Essa escolha reflete a necessidade de maximizar o recall, ou seja, identificar a maior quantidade possível de fraudes, mesmo que isso implique um aumento nos falsos positivos. Assim, o F2-Score é expresso da seguinte forma:

$$F_2 = 5 \cdot \frac{P \cdot R}{4P + R}. \tag{2.28}$$

A justificativa para essa ponderação reside no fato de que o custo de uma fraude não detectada é muito superior ao custo de uma investigação adicional decorrente de um falso positivo, situação comumente observada em sistemas de detecção de irregularidades. Em termos operacionais, o uso de $\beta = 2$ confere um peso quatro vezes maior ao recall em comparação à precisão, o que se alinha aos custos típicos relatados na literatura, na qual o custo de um falso negativo pode ser, por exemplo, dez vezes maior do que o de um falso positivo (Davis & Goadrich, 2006).

Outra abordagem importante em ambientes desbalanceados é o uso da curva Precision-Recall, cuja área sob a curva (AUC-PR) tem se mostrado mais sensível do que a *Area Under the Curve - Receiver Operating Characteristic* (AUC-ROC) para avaliar o desempenho nessas condições. Sejam $\text{Prec}(t)$ e $\text{Rec}(t)$ a precisão e o recall para um determinado threshold t; a AUC-PR é definida pela integral:

$$\text{AUC-PR} = \int_0^1 \text{Prec}(t)\, d\text{Rec}(t). \tag{2.29}$$

Estudos recentes, como os de Saito and Rehmsmeier (2015), indicam que, para dados com baixa prevalência da classe positiva, a AUC-PR tende a ter uma relação assintótica com o Fβ-Score, particularmente com o F2-Score, quando se prioriza a detecção de casos críticos.

Em termos operacionais, a otimização do ponto de corte (threshold) para a classificação é crucial para minimizar os custos associados aos erros. O custo total pode ser modelado pela seguinte equação:

$$\text{Custo Total} = C_{FP} \times FP + C_{FN} \times FN, \tag{2.30}$$

na qual C_{FP} e C_{FN} representam, respectivamente, os custos associados a falsos positivos e falsos negativos. Dado que, na detecção de fraudes, C_{FN} geralmente é muito superior a C_{FP}, o threshold ótimo t^* é determinado pela minimização do custo total:

$$t^* = \arg\min_t \left(C_{FP} \times FP(t) + C_{FN} \times FN(t) \right). \tag{2.31}$$

Ao derivar a função de custo em relação ao threshold, a condição de otimização pode ser expressa da seguinte maneira:

$$\frac{\partial \text{Custo}}{\partial t} = C_{FP} \cdot \frac{\partial FP}{\partial t} + C_{FN} \cdot \frac{\partial FN}{\partial t} = 0, \tag{2.32}$$

o que permite identificar o ponto de equilíbrio ideal entre precisão e recall para o problema em questão.

A integração das métricas AUC-PR e F2-Score com a análise dos trade-offs entre precisão e recall possibilita uma avaliação mais robusta do desempenho do modelo em ambientes de alto desbalanceamento. Dessa forma, a seleção do ponto de corte adequado torna-se um componente essencial para garantir que o modelo alcançue um equilíbrio que minimize os custos operacionais e maximize a eficácia na detecção de fraudes.

As análises realizadas por Provost and Fawcett (2013) reforçam a importância de combinar essas métricas para a tomada de decisão, evidenciando que um modelo com

AUC-PR elevada e F2-Score robusto é geralmente mais apto a lidar com os desafios impostos por conjuntos de dados desbalanceados.

2.3.3 Técnicas de Balanceamento e Pré-processamento de Dados

Em problemas de classificação, especialmente no contexto da detecção de fraudes em fornecedores, é comum enfrentar conjuntos de dados altamente desbalanceados. Nesses cenários, a classe que representa casos de fraude é significativamente menos representada em comparação com a classe dos registros legítimos, o que pode levar os modelos preditivos a favorecer a classe majoritária. Essa disparidade pode mascarar a capacidade do modelo de identificar corretamente os casos positivos, resultando em baixa sensibilidade e elevado risco de falsos negativos. Assim, torna-se necessária a aplicação de técnicas de balanceamento e pré-processamento dos dados, de modo a proporcionar um treinamento mais equilibrado e robusto dos algoritmos.

Nesse sentido, a estratégia de balanceamento visa ajustar a distribuição das classes, permitindo que o modelo aprenda de maneira mais eficaz os padrões associados à classe minoritária. Esse processo pode ser alcançado por meio de abordagens que aumentam a representatividade dos dados raros ou que reduzem a incidência de exemplos na classe dominante. A importância dessa etapa é ressaltada em estudos como os deBatista, Prati, and Monard (2004) e He and Garcia (2009b), que demonstraram que a aplicação de técnicas de geração de exemplos sintéticos pode melhorar significativamente o desempenho na detecção de fraudes, reduzindo o viés inerente aos dados desbalanceados.

As técnicas de balanceamento se dividem em duas abordagens principais: métodos de *oversampling* e métodos de *undersampling*. No *oversampling*, o objetivo é aumentar o número de instâncias da classe minoritária, podendo-se optar tanto pela replicação direta quanto pela criação de novos exemplos, o que permite uma representação mais diversificada e menos redundante dos dados. Essa abordagem, entretanto, deve ser aplicada com cautela, pois a simples replicação pode levar a problemas de *overfitting*, enquanto a geração de exemplos sintéticos deve preservar as características estatísticas originais para evitar a introdução de ruídos artificiais.

Por outro lado, o *undersampling* busca equilibrar o conjunto de dados, reduzindo o número de exemplos da classe majoritária. Essa técnica pode ser vantajosa em termos de custo computacional e na remoção de dados redundantes, mas pode acarretar a perda de informações importantes se não for realizada de forma criteriosa. Batista et al. (2004) discutem estratégias de *undersampling* que preservam a variabilidade dos dados, destacando

que a escolha dos exemplos a serem eliminados deve ser baseada em critérios estatísticos que assegurem a representatividade da classe majoritária.

Além das abordagens tradicionais de *oversampling* e *undersampling*, técnicas mais sofisticadas têm sido propostas para gerar novos exemplos de forma inteligente. O SMOTE (*Synthetic Minority Over-sampling Technique*), introduzido por Chawla, Bowyer, Hall, and Kegelmeyer (2002a), cria amostras sintéticas a partir da interpolação entre exemplos vizinhos da classe minoritária, ampliando a diversidade dos dados sem simplesmente replicar os casos existentes. Essa técnica tem se mostrado eficaz para reduzir o *overfitting* e melhorar a capacidade do modelo de detectar fraudes. Em complemento, o ADASYN (*Adaptive Synthetic Sampling*), proposto por He and Garcia (2009b) , adapta o processo de geração de exemplos sintéticos às regiões do espaço de características onde o modelo apresenta maior dificuldade, direcionando a criação de novos dados para áreas com maior densidade de erros. Essa abordagem adaptativa contribui para um aprimoramento da sensibilidade do modelo sem comprometer a integridade dos dados.

Dessa forma, a aplicação combinada dessas técnicas de balanceamento e pré-processamento torna-se indispensável para o desenvolvimento de modelos preditivos que sejam capazes de lidar com o desbalanceamento intrínseco a problemas de fraude. A literatura aponta que a escolha cuidadosa e a implementação criteriosa dessas metodologias, acompanhada de validação cruzada e de métricas de avaliação robustas como AUC-PR e F2-Score, podem resultar em modelos significativamente mais eficazes e operacionais (Chawla et al., 2002a; He & Garcia, 2009b; Batista et al., 2004).

Esta seção apresenta, de forma aprofundada, as principais técnicas de balanceamento – com ênfase em *oversampling, undersampling*, SMOTE e ADASYN – e discute suas aplicações, benefícios e limitações para a detecção de fraudes em fornecedores. Essa abordagem busca alinhar os métodos de pré-processamento aos desafios específicos do domínio, garantindo que os modelos desenvolvidos apresentem uma performance robusta mesmo em contextos de alta assimetria de classes.

Oversampling

O *oversampling* é uma técnica de balanceamento que consiste em aumentar o número de exemplos da classe minoritária, de modo a igualar ou aproximar o número de instâncias das classes presentes no conjunto de dados. Essa abordagem pode ser realizada por meio da replicação direta de exemplos existentes ou através da criação de novos exemplos sintéticos, mantendo a distribuição dos dados originais. No contexto de problemas de detecção de fraudes, onde os casos positivos representam uma fração muito pequena do total, o *oversampling* permite que o modelo tenha mais exemplos para aprender, reduzindo o viés a favor da classe majoritária.

Uma das vantagens do *oversampling* é que, ao aumentar a representatividade da classe minoritária, o modelo pode capturar melhor os padrões que caracterizam os eventos de fraude. Contudo, a replicação direta de exemplos pode levar a um problema de *overfitting*, pois o modelo tende a memorizar os dados replicados em vez de generalizar o conhecimento. Alternativamente, métodos que geram novos exemplos, mesmo que de maneira simples, podem ajudar a mitigar esse risco; no entanto, é necessário ter cautela para não distorcer a distribuição original dos dados. A literatura enfatiza que, embora o *oversampling* seja uma estratégia simples e intuitiva, sua aplicação deve ser combinada com outras técnicas de validação para evitar que o modelo se ajuste excessivamente aos dados sintetizados (Chawla et al., 2002a; He & Garcia, 2009b).

Além disso, o *oversampling* pode ser considerado parte de uma estratégia de pré-processamento que inclui a normalização e a transformação dos dados, de modo a preservar as características estatísticas relevantes enquanto se aumenta a quantidade de dados da classe minoritária. Esse processo é particularmente importante em conjuntos de dados onde a disparidade entre as classes pode comprometer a capacidade de aprendizado do modelo, resultando em alta variância para a classe minoritária. Estudos empíricos demonstram que a aplicação adequada de *oversampling* pode levar a melhorias significativas nas métricas de avaliação, especialmente em termos de *Recall* e F2-Score, que são cruciais em aplicações onde a detecção de fraudes é priorizada (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015).

Ao implementar *oversampling*, é recomendável utilizar métodos que minimizem a redundância dos dados e mantenham a diversidade dos exemplos, de forma a garantir que o modelo não seja induzido a aprender ruídos ou padrões artificiais. Assim, a técnica de *oversampling* deve ser empregada de maneira criteriosa, acompanhada de uma validação cruzada rigorosa, para assegurar que os ganhos em performance não sejam apenas reflexo de uma replicação desnecessária, mas sim de uma melhoria real na capacidade do modelo de generalizar a partir de um conjunto de dados balanceado.

Undersampling

O *undersampling* é uma técnica de balanceamento que consiste em reduzir o número de exemplos da classe majoritária, a fim de equilibrar o conjunto de dados. Em vez de aumentar artificialmente a classe minoritária, o *undersampling* retira casos da classe dominante, o que pode ser vantajoso quando há uma grande quantidade de dados redundantes ou quando a classe majoritária contém muitos exemplos semelhantes entre si.

A principal vantagem do *undersampling* reside na diminuição do custo computacional, pois, ao reduzir o tamanho do conjunto de dados, o tempo de treinamento do modelo é significativamente diminuído. Além disso, a remoção de exemplos redundantes pode

ajudar a reduzir o ruído presente no conjunto de dados, melhorando a capacidade do modelo de capturar os padrões essenciais para a classificação. Contudo, essa técnica também apresenta desvantagens: ao descartar dados, pode-se perder informações relevantes que contribuam para a construção de um modelo robusto e capaz de generalizar. Essa perda de informação é particularmente crítica em situações onde a diversidade da classe majoritária é importante para a identificação de fronteiras de decisão complexas.

Em contextos onde a quantidade de dados é extremamente elevada, o *undersampling* pode ser combinado com outras técnicas de pré-processamento, como a análise de componentes principais (PCA), para identificar e preservar as variáveis mais informativas. A escolha dos exemplos a serem removidos deve ser feita de forma criteriosa, frequentemente utilizando algoritmos que detectam redundâncias ou que priorizam a representatividade estatística dos dados. Referências como as de Batista et al. (2004) discutem métodos avançados de *undersampling*, ressaltando que a técnica pode ser ajustada para manter um equilíbrio entre a redução de dados e a preservação das informações essenciais, minimizando o risco de deterioração da performance do modelo.

No caso de problemas de detecção de fraude, em que a classe majoritária está muito bem representada, o *undersampling* pode ser uma abordagem eficaz para reduzir o viés do modelo, permitindo que os exemplos da classe minoritária tenham uma influência mais significativa na aprendizagem. No entanto, essa estratégia deve ser implementada com cautela para evitar a perda de variabilidade e garantir que o modelo continue a capturar a complexidade inerente à classe majoritária.

Links de Tomek (TomekLinks)

Além das abordagens genéricas de subamostragem (descritas na Seção 2.3.12), existem métodos de limpeza de dados que atuam de forma mais precisa na fronteira de decisão. A técnica de Links de Tomek (*TomekLinks*), introduzida por Tomek (1976), é um desses métodos, frequentemente aplicada para remover ruídos e ambiguidades na região onde as classes se sobrepõem (J.-w. Li, Chang, & Wang, 2017).

Um "Link de Tomek"é formalmente definido como um par de instâncias (x, y) que pertencem a classes distintas e satisfazem a condição de serem vizinhos mais próximos mútuos (Chawla, Bowyer, Hall, & Kegelmeyer, 2002b). Matematicamente, dado $d(x, y)$ como a função de distância entre as duas instâncias, o par (x, y) forma um Link de Tomek se, e somente se, para qualquer outra instância z:

$$d(x, y) < d(x, z) \quad \text{e} \quad d(x, y) < d(y, z) \tag{2.33}$$

Onde x e y são de classes opostas. A presença de tais links indica que as instâncias estão na fronteira de decisão ou que uma delas pode ser ruído, comprometendo a separação das classes.

Como técnica conjugada com *undersampling*, o método consiste em identificar todos os Links de Tomek no conjunto de dados e, em seguida, remover a instância pertencente à classe majoritária de cada par (J.-w. Li et al., 2017; Chawla et al., 2002b). O objetivo dessa remoção não é, primariamente, alterar a proporção entre as classes de forma drástica, mas sim "limpar"a fronteira de decisão, tornando a separação entre as classes mais clara e definida para o algoritmo classificador (J.-w. Li et al., 2017).

Na prática, a técnica *TomekLinks* é frequentemente utilizada não de forma isolada, mas como um método de limpeza de dados aplicado em conjunto com outras técnicas, como *oversampling* também. Em abordagens híbridas, como o *SMOTE+Tomek*, o SMOTE é aplicado primeiramente para gerar novas instâncias da classe minoritária, e os *TomekLinks* são usados subsequentemente para remover instâncias da classe majoritária que estejam causando sobreposição na fronteira (Chawla et al., 2002b; Kovács, 2019).

SMOTE (Synthetic Minority Over-sampling Technique) e Variantes

O SMOTE (*Synthetic Minority oversampling Technique*) é uma técnica avançada de *oversampling* que busca resolver os problemas associados à replicação direta de exemplos na classe minoritária. Proposto por Chawla et al. (2002a), o SMOTE cria exemplos sintéticos ao interpolar pontos entre instâncias vizinhas da classe minoritária. Essa abordagem tem a vantagem de introduzir diversidade nos dados e reduzir o risco de *overfitting*, pois os exemplos gerados não são cópias exatas dos existentes, mas sim novas amostras que preservam as características estatísticas da classe original.

A técnica consiste em, para cada exemplo da classe minoritária, identificar seus k vizinhos mais próximos e, a partir deles, gerar novos exemplos ao combinar aleatoriamente as diferenças entre o exemplo original e seus vizinhos. Esse processo resulta em uma expansão do conjunto de dados minoritário de maneira contínua, o que pode levar a uma melhor definição das fronteiras entre as classes. Estudos demonstram que o uso do SMOTE pode melhorar significativamente a capacidade do modelo de identificar casos da classe minoritária, sobretudo em termos de *Recall* e AUC-PR, métricas essenciais em ambientes com alto desbalanceamento (Chawla et al., 2002a; Saito & Rehmsmeier, 2015).

Apesar de suas vantagens, o SMOTE também apresenta limitações. A interpolação entre exemplos pode introduzir ruído se as amostras vizinhas não forem representativas ou se a classe minoritária for extremamente dispersa. Além disso, o método não considera a densidade local dos dados, o que pode resultar em exemplos sintéticos que não refletem adequadamente a complexidade do espaço de características. Por essas razões, a aplicação

do SMOTE deve ser acompanhada de uma análise cuidadosa dos dados e, frequentemente, deve ser combinada com técnicas de validação cruzada para assegurar que os benefícios em termos de balanceamento não comprometam a qualidade do modelo.

No contexto de fornecedores com histórico de fraudes, onde os dados positivos são escassos, o SMOTE oferece uma solução robusta para aumentar a representatividade da classe minoritária, permitindo ao modelo aprender melhor os padrões associados a irregularidades, sem recorrer à simples replicação de dados.

DSMOTE (Density-Based SMOTE)

Visando aprimorar o SMOTE padrão, que pode gerar amostras sintéticas em regiões de ruído ou sobreposição com a classe majoritária, também foram propostas variantes baseadas em densidade. O *DSMOTE* (*Density-Based Synthetic Minority Oversampling Technique*), por vezes referido como *DBSMOTE*, é uma dessa abordagens, proposta por Bunkhumpornpat, Sinapiromsaran, and Lursinsap (2012).

A lógica do DSMOTE consiste na integração do algoritmo de clusterização DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) ao processo de *oversampling* (Bunkhumpornpat et al., 2012). A técnica opera inicialmente aplicando o DBSCAN exclusivamente sobre as instâncias da classe minoritária. Essa etapa permite identificar clusters com formas arbitrárias (*arbitrarily shaped clusters*) e, fundamentalmente, filtrar as instâncias classificadas como ruído (*noise*) pelo DBSCAN (Bunkhumpornpat et al., 2012).

A título de explicação, o DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) é um algoritmo de clusterização não supervisionado fundamental, projetado para descobrir clusters em grandes bancos de dados espaciais e, simultaneamente, identificar dados discrepantes como ruído (Bunkhumpornpat et al., 2012). Diferentemente dos algoritmos baseados em centróides, como o K-Means, o DBSCAN utiliza uma noção de densidade, permitindo a identificação de clusters de formatos arbitrários (*arbitrarily shaped clusters*). O algoritmo opera com base em dois parâmetros principais: o raio da vizinhança (ϵ ou *Eps*) e o número mínimo de pontos necessários para formar uma região densa. Essa capacidade de isolar ruído torna-o particularmente útil nas etapas de pré-processamento e em técnicas híbridas de rebalanceamento, como o DSMOTE, que o utiliza para identificar regiões densas da classe minoritária antes de aplicar a geração de amostras (Bunkhumpornpat et al., 2012).

Após a clusterização, o DSMOTE aplica diferentes estratégias para as amostras identificadas como pontos centrais (*core points*) e pontos de fronteira de (*borderline points*) dentro dos clusters minoritários. A geração de novas instâncias sintéticas não ocorre aleatoriamente entre vizinhos, ao contrário do SMOTE. Em vez disso, o algoritmo calcula

um "pseudo-centróide"para cada cluster minoritário denso e gera as novas amostras ao longo do caminho mais curto (*shortest path*) entre uma instância minoritária selecionada e esse pseudo-centróide (Bunkhumpornpat et al., 2012).

Ao focar a geração de amostras em regiões minoritárias densas, validadas pelo DBSCAN, o DSMOTE evita a interpolação em direção a instâncias ruidosas ou isoladas, mitigando o risco de sobregeneralização e a introdução de ruído na fronteira de decisão.

BorderlineSMOTE

O algoritmo SMOTE (descrito na Seção anterior) demonstrou ser eficaz ao gerar novas instâncias sintéticas por meio da interpolação; no entanto, sua estratégia de seleção de instâncias é aleatória, tratando todos os exemplos da classe minoritária de forma idêntica. Isso pode gerar fragilidades, pois o SMOTE pode produzir amostras em regiões "seguras"(distantes da fronteira de decisão) ou reforçar instâncias que são ruído, o que pode prejudicar o modelo.(Kovács, 2019).

Para endereçar essa limitação, Han et al. propuseram o *BorderlineSMOTE*, uma técnica de *oversampling* adaptativa que foca na geração de amostras exclusivamente nas instâncias minoritárias que estão na fronteira da decisão — ou seja, aquelas mais propensas a serem mal classificadas. (H. Han, Wang, & Mao, 2005) propuseram o BorderlineSMOTE, uma técnica de *oversampling* adaptativa que foca a geração de amostras exclusivamente nas instâncias minoritárias que estão na fronteira da decisão — ou seja, aquelas mais prováveis de serem mal classificadas (J.-w. Li et al., 2017; Kovács, 2019).

O BorderlineSMOTE opera em duas fases. Primeiro, ele classifica cada instância da classe minoritária x_i em uma de três categorias: *segura* (safe), *perigo* (danger) ou ruído (noise). Essa classificação é feita analisando os k-vizinhos mais próximos (k-NN) de x_i. Seja k' o número de vizinhos que pertencem à classe majoritária:

- Se $k' < k/2$, a maioria dos vizinhos é da classe minoritária e x_i é considerada *segura*.
- Se $k/2 \leq k' < k$, a instância encontra-se na fronteira (*borderline*) e é classificada como *perigosa*.
- Se $k' = k$, todos os vizinhos são da classe majoritária e x_i é considerada *ruído*.

Na segunda fase, o algoritmo ignora as instâncias *classificadas como safe* e *noise* e aplica o procedimento de geração SMOTE apenas às instâncias classificadas como *danger* (H. Han et al., 2005). As amostras sintéticas são criadas ao interpolar uma instância *danger* com seus vizinhos mais próximos que pertencem à classe minoritária.

Ao concentrar os esforços de *oversampling* nas instâncias da fronteira, o *BorderlineSMOTE* visa fortalecer a definição da fronteira de decisão, melhorando a capacidade de

separação do classificador precisamente onde o risco de erro é maior, sem gerar amostras desnecessárias em regiões que já estão bem definidas.

ADASYN (Adaptive Synthetic Sampling)

O ADASYN (*Adaptive Synthetic Sampling*) é uma extensão do SMOTE, introduzida por He and Garcia (2009b), que visa aprimorar a geração de exemplos sintéticos por meio da adaptação à distribuição dos dados. Ao contrário do SMOTE, que gera novos exemplos de maneira uniforme, o ADASYN concentra a criação de amostras sintéticas em regiões onde a densidade da classe minoritária é menor ou onde o modelo enfrenta maior dificuldade de aprendizado. Essa abordagem adaptativa permite que o método se concentre em "áreas problemáticas" do espaço de características, facilitando a aprendizagem de padrões que podem ser críticos para a detecção de fraudes.

O funcionamento do ADASYN envolve a avaliação da dificuldade de cada exemplo da classe minoritária, com base na proporção de vizinhos pertencentes à classe majoritária. Exemplos que apresentam uma maior incidência de vizinhos de outra classe recebem um peso maior e, consequentemente, geram mais amostras sintéticas. Essa estratégia não apenas aumenta a quantidade de dados da classe minoritária, mas também melhora a qualidade desses dados ao direcionar a atenção para regiões onde a discriminação entre as classes é mais desafiadora. Pesquisas indicam que o ADASYN pode melhorar a performance do modelo em termos de recall e F2-Score, métricas críticas para problemas em que a detecção de fraudes é prioritária (He & Garcia, 2009b; Saito & Rehmsmeier, 2015).

Embora o ADASYN ofereça uma abordagem mais refinada do que o SMOTE, também exige uma análise cuidadosa dos dados, uma vez que a geração de amostras sintéticas em regiões muito ruidosas pode introduzir inconsistências no modelo. A seleção adequada dos parâmetros, como o número de vizinhos e o grau de adaptação, é crucial para garantir que as amostras geradas contribuam para a melhoria do desempenho do modelo, sem introduzir viés adicional ou ruído excessivo.

Em suma, as técnicas de *oversampling*, como SMOTE e ADASYN, representam estratégias avançadas para o balanceamento de conjuntos de dados desbalanceados. Enquanto o SMOTE gera novos exemplos por meio de interpolação simples, o ADASYN aprimora esse processo ao focar em regiões de difícil aprendizagem, proporcionando, assim, modelos mais sensíveis à detecção de fraudes em contextos nos quais a classe minoritária é escassa.

2.3.4 Regularização e Seleção de Variáveis *(Feature Selection)*

O processo de seleção de variáveis visa reduzir a dimensionalidade dos dados e destacar os atributos mais pertinentes à construção do modelo, ao mesmo tempo em que aprimora o

desempenho preditivo e diminui o custo computacional. Além disso, favorece a interpretabilidade dos resultados, especialmente em cenários que apresentam um grande número de atributos, muitos dos quais podem ser irrelevantes ou redundantes.

Desta forma, a seleção de variáveis pode ser formulada como a busca de um subconjunto $S \subseteq \{X_1, X_2, \ldots, X_p\}$ que maximize uma função objetivo f, frequentemente associada a métricas como acurácia ou poder de generalização.

$$f(S) = \text{Performance}(M, S), \tag{2.34}$$

em que M é o modelo preditivo treinado sobre o subconjunto S.

As três principais classes de métodos de seleção são Filtros, *Wrappers* e Métodos Embutidos (*Embedded*). Os Filtros analisam propriedades estatísticas dos dados, como correlação ou medidas de informação, sem levar em conta o modelo. Em relação aos *Wrappers*, esse método avalia conjuntos distintos de atributos com base no desempenho de um modelo preditivo, gerando resultados mais precisos a um custo computacional mais elevado, como o método *Recursive Feature Elimination* (RFE). Por fim, os Métodos embutidos (*Embedded*) integram a seleção de variáveis ao próprio treinamento do modelo, como ocorre em técnicas de regularização do tipo LASSO (*Least Absolute Shrinkage and Selection Operator*).

Após a seleção de variáveis, a eficácia do modelo é aferida por meio de métricas como Acurácia, Precisão, *Recall* e área sob a curva ROC (AUROC). A adoção de validação cruzada (*cross-validation*) em conjunto com técnicas de seleção adequadas é essencial para assegurar que o modelo resultante tenha boa capacidade de generalização em aplicações reais.

Regularização Rigde ou L2

Introduzida por (Hoerl & Kennard, 1970), sua concepção parte do princípio de que a regressão de mínimos quadrados falha quando o número de variáveis usadas para predição é superior ao número de observações. isso ocorre pela falta de diferenciação na importância das variáveis preditoras; assim, todas as variáveis são incluídas durante o processo. O modelo Ridge foi desenvolvido para incorporar fatores que podem priorizar parcelas da equação, visando mitigar a imprecisão dos estimadores em modelos que possuam variáveis independentes altamente correlacionadas, ou seja, com alto nível de multicolinearidade. A regularização L2 acrescenta penalidades à soma dos quadrados dos coeficientes diferentes de zero, com a finalidade de adicionar viés suficiente para aproximar as estimativas dos resultados reais, evitar o *overfitting* e mitigar grandes oscilações.

Segundo (McDonald, 2009), a regressão de Ridge adiciona uma penalidade igual ao quadrado da magnitude dos coeficientes, reduzindo todos os coeficientes ao mesmo fator sem eliminá-los. Dessa forma, estimula os valores a se aproximarem de zero ao introduzir uma nova informação λI à diagonal produzida por $(X^T X)$, com a finalidade de aumentar a variância de forma artificial e, consequentemente, aprimorar sua generalização. Onde λ representa a quantidade da penalidade.

A regularização de Ridge pode ser apresentada da seguinte forma:

$$L_{\text{ridge}}(\widehat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^t\widehat{\beta})^2 + \lambda\sum_{j=1}^{p} b_i^2 \tag{2.35}$$

Também pode ser reescrita em notação matricial:

$$(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta \tag{2.36}$$

A notação matricial apresentada pode ser minimizada ao ser derivada, na qual o termo $\hat{\beta}$ é isolado ao igualar a relação a zero, apresentando-se da seguinte forma:

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y \tag{2.37}$$

Tendo o exposto, a regressão de Ridge visa extrair melhores resultados dos modelos com a introdução do fator de minimização que, ao ser inserido no primeiro termo da equação, reduz o erro no conjunto de treino, enquanto sua inclusão no segundo termo penaliza a complexidade do modelo. No entanto, conforme relatado, essa técnica não zera os coeficientes e, portanto, não elimina variáveis, mas auxilia na compreensão dos dados do modelo.

Regularização Lasso ou L1

Também conhecida como norma de penalidade L1, ela adiciona uma penalidade à soma dos valores absolutos dos coeficientes diferentes de zero. Dessa forma, penaliza coeficientes altos, podendo reduzir seus valores a zero. Desse modo, este método possui a capacidade de simplificar o modelo ao remover variáveis. (Tibshirani, 1996a).

A fórmula em questão assemelha-se à regressão Ridge; no entanto, utiliza valores absolutos, conforme demonstrado na fórmula a seguir:

$$L_{\text{lasso}}(\widehat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^j\widehat{\beta})^2 + \lambda\sum_{j=1}^{n} |\widehat{\beta}|_j \tag{2.38}$$

Além de poder ser representada em sua forma matricial, essa relação também pode ser expressa de outras maneiras.

$$(Y - X\beta)^T(Y - X\beta) + \lambda|\beta|_1 \tag{2.39}$$

onde λ é o parâmetro de regularização que controla o trade-off entre o ajuste do modelo e sua simplicidade. Assim como na regressão Ridge, o LASSO empurra os coeficientes para valores menores; no entanto, a característica distintiva da norma l_1 é que ela pode forçar coeficientes irrelevantes a se tornarem exatamente zero quando λ atinge um valor apropriado. Dessa forma, o LASSO gera uma solução esparsa, na qual somente um subconjunto dos coeficientes apresenta valores diferentes de zero, facilitando a interpretação e a seleção das variáveis mais importantes.

A principal força do LASSO reside em dois efeitos. O primeiro é o "Efeito de Regularização"; ao limitar a soma dos valores absolutos dos coeficientes, o LASSO controla a complexidade do modelo e reduz o risco de *overfitting*. Esse mecanismo é especialmente útil em cenários com um grande número de preditores ou quando há multicolinearidade entre eles. O segundo é o "Efeito de Esparsidade", no qual a penalização l_1 impõe um custo não linear para a inclusão de variáveis. Com o aumento de λ, os coeficientes dos preditores menos relevantes são reduzidos exatamente a zero, efetivamente removendo essas variáveis do modelo. Essa propriedade contrasta com a regressão Ridge, que apenas diminui os coeficientes sem eliminá-los.

A formulação do problema de otimização do LASSO pode ser expressa de duas maneiras equivalentes. A forma penalizada é:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.40}$$

enquanto a formulação restrita (constrangida) é:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2, \quad \text{tal que} \quad \sum_{j=1}^{p} |\beta_j| \leq s, \tag{2.41}$$

onde s é uma constante que depende de λ. Conforme s diminui, a magnitude dos coeficientes é mais fortemente limitada, promovendo uma solução mais esparsa.

Assim, conforme destacado anteriormente, o parâmetro λ desempenha um papel central no desempenho do LASSO:

- $\lambda = 0$: O modelo LASSO se reduz à regressão OLS, sem qualquer penalização, mantendo todos os preditores.
- Conforme λ aumenta, Um número maior de coeficientes é forçado a zero, resultando em uma solução esparsa e em um modelo mais simples. Contudo, um valor

excessivamente alto de λ pode levar ao *underfitting*, comprometendo a capacidade preditiva.

- λ ótimo: Geralmente determinado por meio da validação cruzada, este valor representa o equilíbrio ideal entre o ajuste do modelo e sua simplicidade.

Comparação entre Ridge e Lasso

Embora tanto o LASSO quanto a regressão Ridge promovam a diminuição dos coeficientes, a penalização l_1 do LASSO possui a propriedade única de gerar soluções esparsas, ou seja, selecionar explicitamente um subconjunto de variáveis. Em contraste, a penalização l_2 da Ridge (conforme introduzida por Hoerl and Kennard (1970)) apenas encolhe os coeficientes, sem zerá-los, mantendo todas as variáveis no modelo.

Estudos comparativos, como os apresentados em H. L. Li (2024) e James, Witten, Hastie, Tibshirani, and Taylor (2023), demonstram que, para um conjunto de preditores com $p = 2$, os estimadores do LASSO alcançam o menor Resíduo da Soma dos Quadrados (RSS) dentro do conjunto de pontos que satisfaz $|\beta_1| + |\beta_2| \leq s$, enquanto os estimadores da Ridge são ótimos para a restrição $\beta_1^2 + \beta_2^2 \leq s$. Em diagramas de contorno, a restrição do LASSO tem formato de diamante, enquanto a da Ridge forma um círculo. Essa diferença geométrica ilustra por que o LASSO tende a produzir soluções esparsas, eliminando variáveis com coeficientes pequenos. Assim como é possível ver na 2.19,

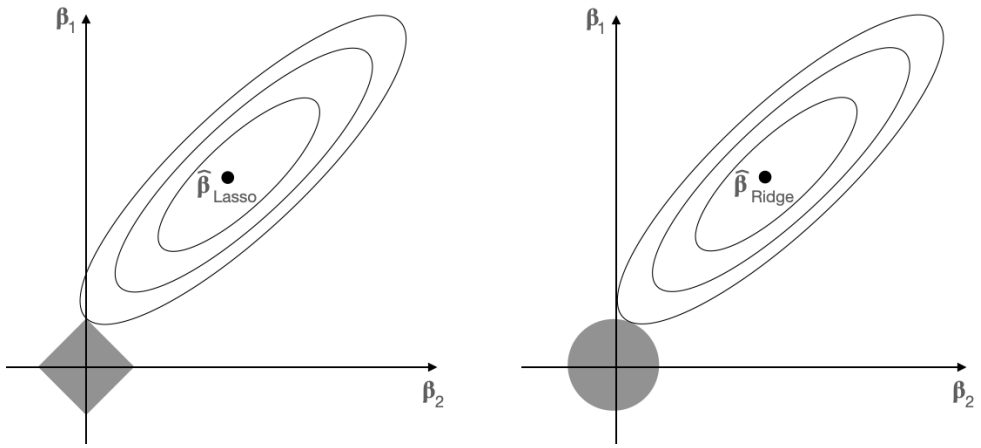

Figura 2.19: Ridge e Lasso
Fonte: Elaborado pelo autor

A consistência do LASSO na identificação do modelo correto depende de certas condições, dentre as quais se destaca a *strong irrepresentable condition* (Chandrashekar & Sahin, 2014). Essa condição impõe limites às correlações entre os preditores, de modo que o LASSO possa distinguir adequadamente as variáveis relevantes das irrelevantes. Além

disso, é necessário que o parâmetro de regularização λ diminua a uma taxa apropriada conforme o número de observações n aumenta, garantindo a consistência da seleção de variáveis.

Em resumo, o LASSO é uma ferramenta poderosa para a regularização e seleção de variáveis, capaz de produzir modelos esparsos que melhoram a interpretabilidade e reduzem o risco de *overfitting*. No entanto, sua aplicação deve ser cuidadosamente avaliada em contextos de alta correlação entre preditores, nos quais suas limitações podem demandar o uso de métodos complementares.

Elastic Net

O método *Elastic Net* foi proposto como uma solução híbrida que combina as vantagens das penalizações L1 (LASSO) e L2 (Ridge) para superar as limitações inerentes a cada técnica quando aplicadas isoladamente (Wang, Zhu, & Zou, 2006). Essa abordagem é especialmente útil em cenários onde há alta correlação entre os preditores ou quando o número de variáveis é grande em relação ao número de observações.

No *Elastic Net*, a função de custo é modificada pela adição simultânea de duas penalizações, de forma que a estimativa dos coeficientes β é obtida resolvendo o problema de otimização:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right), \tag{2.42}$$

onde:

- $\|y - X\beta\|_2^2$ representa o erro quadrático entre os valores observados e os preditos,
- $\lambda_2 \|\beta\|_2^2$ é a penalização do tipo L2, que encolhe os coeficientes de forma uniforme sem zerá-los,
- $\lambda_1 \|\beta\|_1$ é a penalização do tipo L1, que pode levar alguns coeficientes a exatamente zero, promovendo a esparsidade.

A presença simultânea de ambos os termos de regularização permite que o *Elastic Net* mantenha a capacidade de seleção de variáveis característica do LASSO, enquanto preserva a estabilidade e a robustez em situações de multicolinearidade proporcionada pela Ridge. Em termos práticos, os hiperparâmetros λ_1 e λ_2 são ajustados — frequentemente via validação cruzada — para alcançar um balanço adequado entre complexidade do modelo e capacidade preditiva. A Figura 2.20 ilustra, de forma esquemática, a interação entre as restrições impostas pelas penalizações L1 e L2.

A solução do *Elastic Net* é geralmente obtida por métodos iterativos, como o *Coordinate Descent*, que se mostra eficiente mesmo em espaços de alta dimensionalidade. Assim,

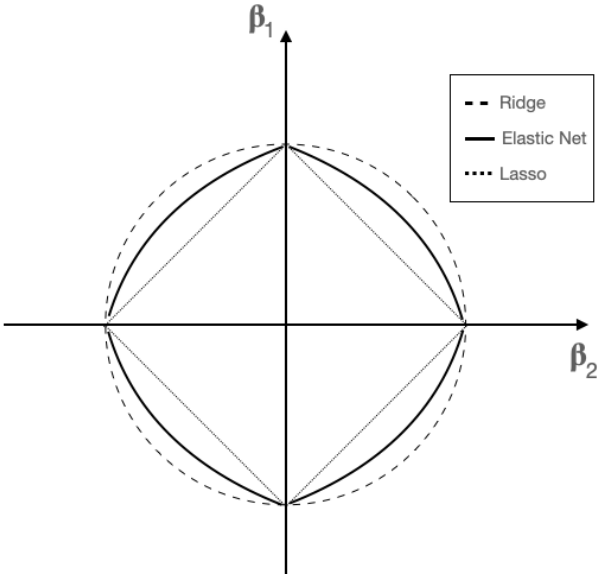

Figura 2.20: Comparação esquemática entre os métodos LASSO, Ridge e Elastic Net
Fonte: Elaborado pelo autor

o *Elastic Net* é amplamente utilizado em problemas de regressão e classificação, onde a interpretação dos modelos e a redução do *overfitting* são requisitos essenciais.

Recursive Feature Elimination (RFE)

O *Recursive Feature Elimination* (RFE) é uma abordagem do tipo wrapper destinada à seleção de variáveis, que busca identificar o subconjunto de atributos que contribuem de maneira mais significativa para o desempenho preditivo de um modelo. Ao contrário dos métodos baseados apenas em penalização, o RFE utiliza o desempenho do modelo como critério direto para a remoção iterativa de variáveis.

O procedimento do RFE consiste nas seguintes etapas:

1. Treinar o modelo com o conjunto completo de atributos.
2. Avaliar a importância de cada variável, que pode ser mensurada pelos coeficientes estimados, pela redução da impureza em modelos de árvores ou por outras métricas internas.
3. Eliminar o(s) atributo(s) considerado(s) de menor relevância.
4. Repetir o processo com o conjunto reduzido de variáveis até que se alcance o número desejado de atributos.

Formalmente, a importância de uma variável X_j pode ser quantificada pela função de ranking:

$$\text{Rank}(X_j) = \text{Importância}(X_j) \quad \forall\, X_j \in S, \tag{2.43}$$

na qual Importância(X_j) reflete a contribuição do atributo X_j na predição realizada pelo modelo. Esse valor pode ser derivado, por exemplo, dos coeficientes de uma regressão linear ou dos ganhos de informação em algoritmos de árvores.

O RFE é particularmente útil em contextos de alta dimensionalidade, pois permite a redução progressiva do conjunto de variáveis, resultando em modelos mais simples, com menor risco de *overfitting* e maior interpretabilidade. Além disso, quando combinado com validação cruzada, o RFE pode identificar de forma robusta o subconjunto de atributos que otimiza o desempenho do modelo em dados não vistos, garantindo que a seleção não seja apenas fruto de uma particularidade do conjunto de treinamento.

Em síntese, enquanto os métodos de regularização como LASSO e Ridge incorporam a seleção de variáveis de maneira implícita através da penalização dos coeficientes, o *Recursive Feature Elimination* adota uma abordagem explícita, eliminando recursivamente as variáveis menos relevantes com base no desempenho do modelo. Essa característica torna o RFE uma ferramenta valiosa para problemas em que a interpretabilidade e a redução de dimensionalidade são prioridades.

2.3.5 Otimização de Hiperparâmetros

Visando otimizar os resultados do modelo, uma opção possível é seleção e alteração de hiperparâmetros consiste em definir parâmetros internos ao modelo (e que não são ajustados diretamente no processo de treinamento) de modo a melhorar seu desempenho e evitar problemas como *overfitting*. Em geral, métodos de aprendizado estatístico apresentam ao menos um hiperparâmetro relacionado à penalização da complexidade do modelo (Bergstra & Bengio, 2012).

Posto isso, a adequada definição de hiperparâmetros visa alcançar um equilíbrio entre *overfitting* (ajuste excessivo aos dados de treinamento) e *underfitting* (modelo simples demais). O valor de λ em métodos de regularização e a taxa de aprendizado em algoritmos de *boosting* são exemplos de hiperparâmetros que podem exercer influência decisiva no desempenho do modelo.

Dentre as principais implementações de técnicas de otimização de hiperparâmetros, destacam-se:

- *Grid Search*: realiza uma busca exaustiva sobre um espaço pré-definido de combinações de hiperparâmetros.

- *Randomized Search*: escolhe aleatoriamente combinações de hiperparâmetros em um espaço determinado, reduzindo o custo computacional comparado ao *Grid Search* e geralmente obtendo resultados satisfatórios (Bergstra & Bengio, 2012).
- *Bayesian Optimization*: modela a função de *loss* como uma função desconhecida e utiliza *aquisição functions* para direcionar as avaliações a regiões promissoras do espaço de busca (Snoek, Larochelle, & Adams, 2012).
- *K-Fold Cross-Validation*: amplamente empregado para avaliar o efeito de diferentes hiperparâmetros no desempenho do modelo. Nesse processo, os dados de treinamento são divididos em K partes, e o treinamento é repetido K vezes, cada vez com um conjunto distinto como *fold* de validação.
- *Genetic Algorithms*: baseia-se em conceitos de seleção natural, realizando mutações e cruzamentos de configurações de hiperparâmetros para buscar configurações de melhor desempenho.

Para este estudo, optou-se pelo uso de *Grid Search* na otimização dos hiperparâmetros, em conjunto com o método de *K-Fold Cross-Validation* (com $K = 10$) na base de treinamento. Esse procedimento possibilita avaliar de forma sistemática as combinações de hiperparâmetros e identificar aquelas que minimizam consistentemente as estatísticas de erro.

A Figura 2.21 ilustra o processo de *grid search* para dois hiperparâmetros, no qual o ponto de interseção das setas verdes indica os valores que produzem o menor erro. A área em verde representa a distribuição dos valores de erro obtidos de acordo com cada combinação.

A título demonstrativo, a Tabela 2.5 apresenta os principais hiperparâmetros otimizados neste trabalho para cada modelo de aprendizado de máquina. Observa-se que esses parâmetros atuam diretamente no controle da complexidade e no equilíbrio entre *overfitting* e *underfitting*.

O resultado da otimização para cada combinação de hiperparâmetros é avaliado de acordo com métricas específicas do modelo, como o RMSE em regressão ou a AUC em classificação.

Por fim, ressalta-se que o método de *Grid Search* pode se tornar computacionalmente oneroso à medida que o número de combinações cresce de forma exponencial.

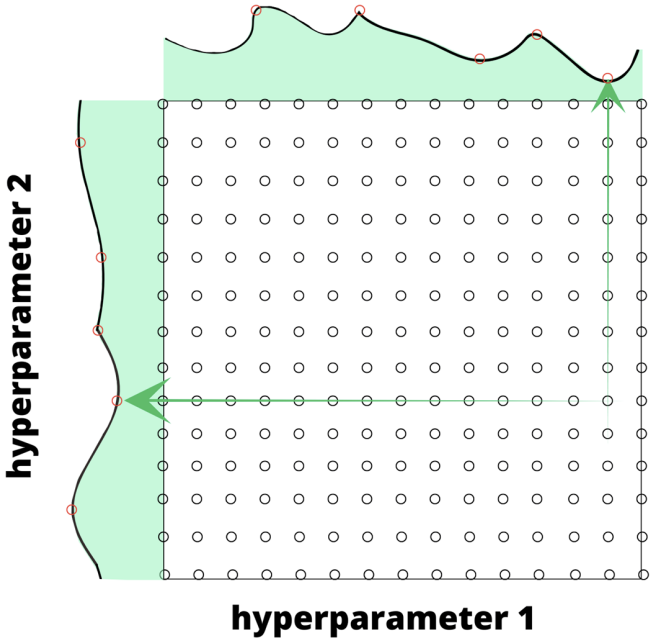

Figura 2.21: Distribuição das estatísticas de erro
Fonte: Adaptado de (Bergstra & Bengio, 2012)

Tabela 2.5: Hiperparâmetros Otimizados nos Modelos de *Machine Learning*

Modelo de ML	Hiperparâmetro Otimizado
Elastic Net	Regularização λ e α
Gradient Boosting	Número de estágios de *boosting*,
	Taxa de aprendizado (learning rate),
	Profundidade máxima das árvores de regressão
Random Forest	Número de árvores (estimators),
	Profundidade máxima das árvores
XGBoost	Número de árvores,
	Profundidade máxima das árvores,
	Taxa de aprendizado

Fonte: Adaptado de Pereira (2024).

2.3.6 Testes Estatísticos para Comparação de Modelos - *Model Confidence Set* (MCS)

A avaliação do desempenho preditivo de modelos a partir de uma amostra de validação é usualmente realizada por meio de métricas quantitativas; no entanto, para a comparação robusta de múltiplos modelos, a aplicação de testes estatísticos formais é imprescindível. Tradicionalmente, o Teste de Diebold and Mariano (1995) tem sido empregado para comparar a acurácia preditiva de um modelo com um benchmark, sob a hipótese nula de que os desempenhos entre dois modelos são iguais. Contudo, essa abordagem se revela

limitada quando se deseja comparar simultaneamente diversos modelos sem a necessidade de definir um modelo de referência.

Para suprir essa limitação, este trabalho adota o método denominado Modelo de Conjunto(MCS), proposto por Hansen, Lunde, and Nason (2011). O MCS permite a comparação conjunta de um conjunto inicial de modelos, M_0, com base na análise de suas funções de perda e não requer a definição prévia de um benchmark. Em essência, o método classifica os modelos de acordo com a perda esperada e, por meio de um procedimento sequencial de testes, elimina aqueles que apresentam desempenho inferior, até que se obtenha um subconjunto M^* que, com um determinado nível de confiança, contenha os melhores modelos disponíveis.

Sejam $L_{i,t}$ e $L_{j,t}$ as funções de perda, como os erros quadráticos, dos modelos i e j no tempo t. A comparação entre dois modelos é realizada por meio da diferença $d_{ij,t} \equiv L_{i,t} - L_{j,t}$. A hipótese nula para o conjunto de modelos $M \subset M_0$ é formalizada da seguinte forma:

$$H_{0,M} : E\left[d_{ij,t}\right] = 0, \quad \forall\, i, j \in M. \tag{2.44}$$

Essa hipótese implica que, sob $H_{0,M}$, os modelos apresentam desempenho preditivo equivalente. Caso essa hipótese seja rejeitada, conclui-se que pelo menos um dos modelos difere em termos de acurácia preditiva, o que demanda a aplicação de uma regra de eliminação.

O procedimento MCS é estruturado da seguinte forma: inicialmente, define-se $M = M_0$ e procede-se com um teste de igualdade de desempenho, chamado de teste de Equivalência de Capacidade Preditiva (EPA). Se a hipótese de EPA não for rejeitada, o conjunto M é considerado o conjunto de confiança M^* com um nível de confiança $(1 - \alpha)$. Por outro lado, se a hipótese for rejeitada, calcula-se a estatística para cada modelo $i \in M$.

$$d_i = \frac{1}{|M|} \sum_{j \in M} d_{ij,t}, \tag{2.45}$$

a qual representa a performance média do modelo i em relação aos demais. Em seguida, estima-se a variância de d_i, denotada por $\mathrm{var}(d_i)$, e define-se o estatístico

$$t_i = \frac{d_i}{\sqrt{\mathrm{var}(d_i)}}, \tag{2.46}$$

identificando-se o modelo com pior desempenho como aquele que maximiza t_i. Esse modelo é, então, eliminado do conjunto M e o teste de EPA é repetido com o conjunto reduzido. O processo iterativo é finalizado quando a hipótese de igualdade de desem-

penho não é rejeitada, resultando no conjunto final M^* de modelos cuja performance é estatisticamente equivalente.

Matematicamente, o conjunto dos melhores modelos pode ser definido da seguinte forma:

$$M^* \equiv \{i \in M_0 \, : \, E(d_{ij,t}) \leq 0, \, \forall \, j \in M_0\}. \tag{2.47}$$

Dessa forma, o MCS identifica, com um dado nível de confiança, um subconjunto M^* que certamente contém os modelos de melhor desempenho.

Em resumo, o método MCS oferece uma abordagem estatisticamente fundamentada para a seleção de modelos em contextos de previsão, permitindo a comparação simultânea de múltiplos modelos sem a necessidade de um benchmark pré-estabelecido. Essa técnica é especialmente valiosa em cenários com alta heterogeneidade de modelos, pois possibilita identificar, com um determinado nível de confiança, o subconjunto de modelos que apresenta o melhor desempenho preditivo.

2.4 Explicabilidade de IA (XAI)

A explicabilidade em inteligência artificial (IA), ou Explainable AI (XAI), busca abordar a necessidade de interpretabilidade em modelos de aprendizado de máquina, particularmente em aplicações nas quais o processo decisório deve ser compreendido e auditado. Conforme de Moraes Souza et al. (2024), a interpretabilidade permite que os *stakeholders* compreendam os modelos e avaliem suas implicações, especialmente quando os resultados são obtidos de maneira automatizada.

Modelos de aprendizado de máquina, como Redes Neurais Profundas e Máquinas de Vetores de Suporte (SVM), podem aproximar padrões complexos. Contudo, sua interpretabilidade frequentemente diminui à medida que a complexidade aumenta, resultando no que é conhecido como "modelos de caixa-preta"(*black-box models*). Essa opacidade apresenta desafios em contextos onde as decisões precisam ser auditáveis e justificáveis.

Nesse contexto, a XAI provê mecanismos para interpretar os impactos de cada variável na variável-alvo (interpretação global) e para compreender a predição de uma observação específica (interpretação local). De acordo com Fisher, Rudin, and Dominici (2019) apud de Moraes Souza et al. (2024), uma métrica de importância global das características pode ser obtida por meio de abordagens baseadas em permutação. Para uma característica relevante ao modelo, a permutação de seus valores deve resultar em uma redução no desempenho preditivo. A diferença nos valores da função de perda antes e depois da permutação pode ser utilizada como uma medida para sua importância geral.

A interpretabilidade apresenta desafios em modelos como as Redes Neurais Profundas (*Deep Learning*), onde os parâmetros das camadas ocultas representam hierarquias

abstratas que influenciam a variável-alvo, mas que são de difícil expressão em termos dos preditores originais. Como apontado por Leo and et al. (2019), conciliar algoritmos computacionalmente eficientes com interpretações compreensíveis é um dos desafios para a gestão do risco bancário e para outras aplicações de aprendizado de máquina em finanças.

No contexto da auditoria financeira, Schreyer, Sattarov, and Borth (2022) propuseram a utilização do *Shapley value* para explicar redes neurais autoencoders na identificação de anomalias contábeis, fornecendo informações direcionadas aos auditores. A abordagem foi utilizada para esclarecer anomalias em auditorias de demonstrações financeiras, buscando mitigar a dificuldade na interpretação de modelos altamente parametrizados (Schreyer et al., 2022).

A explicabilidade dos modelos de IA é, portanto, um campo de estudo ativo. Em contextos como o de compras públicas, a justificativa das decisões é um requisito para a transparência e auditoria. Ferramentas como LIME (*Local Interpretable Model-agnostic Explanations*) e SHAP (baseado no Shapley value) são utilizadas para gerar explicações das predições, detalhando como cada variável contribui para o resultado. Isso permite aos tomadores de decisão analisar os modelos de forma mais detalhada.

Shapley Value

Os valores de Shapley (Shapley, 1953), originados da teoria dos jogos, são utilizados para quantificar a contribuição de cada variável em modelos de previsão. Essa metodologia atribui valores que capturam o impacto marginal de cada variável no resultado final do modelo. Matematicamente, o valor de Shapley para uma variável i é definido por:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{i\}) - v(S) \right], \tag{2.48}$$

Onde S é um subconjunto de variáveis que não contém i, N representa o conjunto total de variáveis, e $v(S)$ é a função de valor (ou predição) do modelo, utilizando o subconjunto S. Este método calcula a média ponderada de todas as contribuições marginais possíveis à variável i.

Os valores de Shapley satisfazem propriedades axiomáticas, notavelmente Eficiência (ou precisão local), Nulidade (*dummy*) e Consistência (Molnar, 2022). A propriedade de Eficiência estabelece que a soma das contribuições das variáveis deve ser igual à diferença entre a predição do modelo e a predição média (valor base). A Nulidade determina que variáveis que não alteram o resultado do modelo recebem contribuição zero. A Consistência assegura que, se a contribuição marginal de uma variável aumenta (ou permanece a mesma) independentemente de outras variáveis, seu valor de Shapley atribuído não deve diminuir.

Adicionalmente, os valores de Shapley podem ser aplicados de maneira agnóstica ao modelo (*model-agnostic*), pois requerem apenas acesso às entradas (*inputs*) e saídas (*outputs*) do modelo, e não à sua estrutura interna. Implementações computacionais, como a biblioteca SHAP em Python, disponibilizam algoritmos para estimar esses valores, incluindo otimizações específicas para diferentes classes de modelos, como árvores de decisão e redes neurais (Lundberg & Lee, 2017).

Por outro lado, O cálculo exato dos valores de Shapley é computacionalmente intensivo, apresentando uma complexidade que cresce exponencialmente com o número de variáveis, M, (ex: $O(M2^M)$). Para contornar essa limitação, foram desenvolvidos diferentes algoritmos que atuam como componentes de um ecossistema de explicabilidade mais amplo, como o disponibilizado pela biblioteca 'shap' (Lundberg & Lee, 2017). Este *framework* unificado implementa múltiplos algoritmos otimizados para diferentes classes de modelos. Para modelos baseados em árvores (ex: XGBoost), o 'TreeExplainer' (e sua versão acelerada 'GPUTreeExplainer') oferece uma estimativa de alta velocidade. Para modelos de *deep learning*, são disponibilizados o 'DeepExplainer' e o 'GradientExplainer'. Modelos LinearE são tratados pelo 'LinearExplainer', que pode opcionalmente contabilizar correlações. O 'KernelExplainer' serve como a implementação universal agnóstica. Outras abordagens, como o 'AdditiveExplainer' para Modelos Aditivos Generalizados (GAMs) ou encapsulamentos de LIME (via 'LimeTabular'), também são fornecidas.

A título demonstrativo, vamos aprofundar no KernelShap, o explicador mais agnóstico em relação a modelos.O Kernel SHAP unifica a abordagem de regressão local do LIME (*Local Interpretable Model-agnostic Explanations*) com os fundamentos teóricos dos valores de Shapley. A metodologia opera por meio da amostragem de coalizões (subconjuntos de variáveis) e da construção de um modelo de regressão linear local ponderado.

O processo de estimação para uma instância específica x, conforme descrito por Molnar (2022), envolve cinco etapas: (1) Amostrar coalizões $z'_k \in \{0,1\}^M$, onde M é o número total de variáveis e z'_k é um vetor binário indicando a presença (1) ou ausência (0) de uma variável; (2) Converter essas coalizões binárias z'_k em instâncias de dados x_k, onde $z'_{k,i} = 1$ significa usar o valor original da instância (x_i) e $z'_{k,i} = 0$ significa usar um valor amostrado do *background dataset*; (3) Obter a predição do modelo $f(x_k)$ para cada instância amostrada x_k; (4) Calcular o peso $\pi_x(z'_k)$ de cada instância usando o "Shapley kernel"; e (5) Ajustar um modelo de regressão linear ponderado $g(z'_k)$ com esses dados.

Esse modelo de regressão linear g é definido da seguinte forma:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{2.49}$$

Os coeficientes otimizados ϕ_i dessa regressão correspondem aos valores de Shapley

estimados (Molnar, 2022). O "Shapley kernel"π_x é o componente central que assegura a aderência às propriedades de Shapley, como a Eficiência. Ele atribui pesos maiores às coalizões com poucas variáveis, próximas do conjunto vazio $|z'| \approx 0$, ou com quase todas as variáveis, próximas do conjunto completo $|z'| \approx M$. A fórmula do kernel é:

$$\pi_x(z') = \frac{(M-1)}{\binom{M-1}{|z'|}|z'|(M-|z'|)} \tag{2.50}$$

onde $|z'|$ é o número de variáveis presentes na coalizão (Lundberg & Lee, 2017). Covert and Lee (2021) analisou essa formalização como uma solução de regressão, propondo técnicas para reduzir a variância da estimativa e calcular intervalos de confiança para os valores.

A principal limitação teórica do Kernel SHAP reside na forma como simula a "ausência"de uma variável (Molnar, 2022). Ao substituir valores por amostras aleatórias do *background dataset*, o método estima a expectativa marginal, $E_{X_C}[f(X_S, X_C)]$, tratando efetivamente as variáveis como independentes (Lundberg & Lee, 2017). Como demonstrado por Aas, Jullum, and Løland (2021) e discutido por Molnar (2022), quando as variáveis estão correlacionadas (multicolinearidade), essa amostragem de perturbações gera instâncias de dados não realistas ou "improváveis"(por exemplo: um paciente com "idade=5"e "status=aposentado").

Isso pode levar a explicações que atribuem peso a instâncias que não existem na distribuição real, resultando em interpretações imprecisas. A alternativa teoricamente mais precisa seria usar a expectativa condicional, $E[f(X)|X_S = x_S]$, modelando explicitamente a dependência das variáveis; no entanto, essa abordagem apresenta um custo computacional significativamente maior (Aas et al., 2021).

Capítulo 3

Metodologia

3.1 Metodologia

Seguindo a classificação sugerida por Prodanov and de Freitas (2013), a presente pesquisa, quanto à sua natureza, classifica-se como aplicada, pois visa gerar conhecimento para aplicação prática e para a solução de problemas específicos. O objetivo é propor uma metodologia para a seleção e priorização de contratos para auditoria no contexto governamental, utilizando técnicas de aprendizado de máquina. Quanto aos objetivos, a pesquisa se classifica como exploratória, buscando obter informações e insights por meio de uma revisão bibliográfica abrangente. A abordagem da pesquisa é predominantemente quantitativa, pois emprega técnicas estatísticas e matemáticas para identificar padrões, relações e pesos entre variáveis e critérios, utilizando técnicas de mineração de dados e aprendizado de máquina.

A metodologia deste trabalho consiste na integração de técnicas de mineração de dados, visando a criação de um score de risco para a priorização de auditorias de contratos públicos. As etapas da metodologia foram organizadas com base no CRISP-DM (Cross-Industry Standard Process for Data Mining).

A metodologia adotada neste estudo foi estruturada para abordar o problema de priorização de contratos públicos para auditoria, utilizando técnicas de aprendizado de máquina e mineração de dados. O processo metodológico foi segmentado em etapas principais, conforme ilustrado na Figura 3.1. Cada etapa será detalhada a seguir.

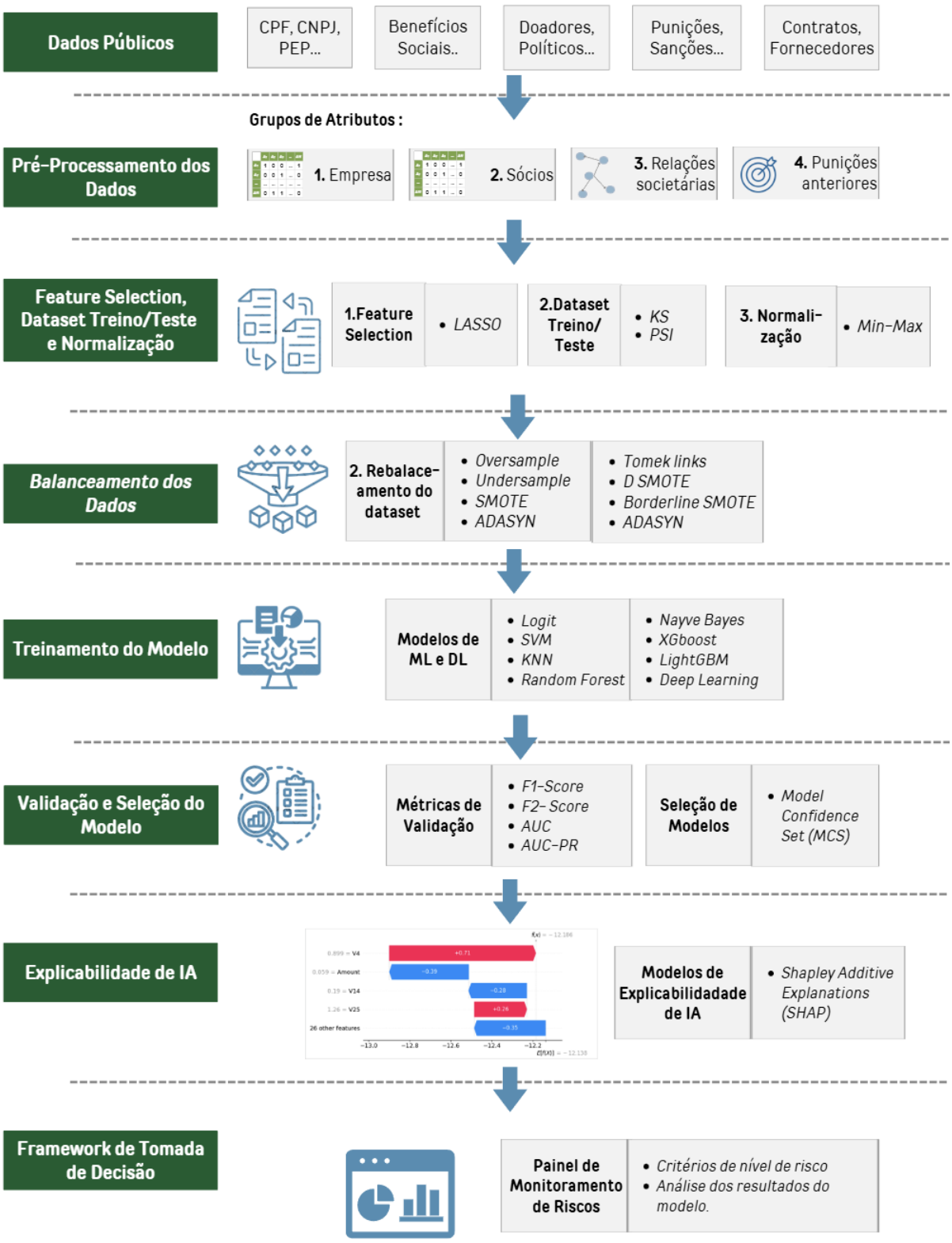

Figura 3.1: Fluxo Metodológico Adotado na Pesquisa

Fonte: Elaborado pelo autor.

A primeira fase da metodologia está centrada na compreensão do contexto das compras

públicas e nos desafios associados ao processo de auditoria de contratos. Essa etapa é realizada por meio da revisão bibliográfica, com a análise de estudos existentes sobre a detecção de fraudes nas compras públicas e *o benchmarking* das melhores práticas em auditorias de contratos públicos. O entendimento das normas legais e do processo de licitação também é fundamental, dada a complexidade da legislação brasileira, incluindo a Lei 14.133/2021.

Durante essa etapa, foi analisado o escopo da pesquisa e definidos os objetivos principais: melhorar a eficiência dos controles internos para auditoria, priorizando contratos que apresentam maior risco de irregularidades com base na aplicação de técnicas de mineração de dados. Além disso, realiza-se a definição dos critérios de risco para o mapeamento das bases de dados necessárias.

A segunda fase consistiu na coleta e preparação de dados, que se baseou na extração de dados públicos do Portal de Compras do Governo Federal e de outras bases de dados governamentais. Os dados incluem informações sobre fornecedores, contratos, processos de licitação, sanções administrativas e dados financeiros e eleitorais. Após a extração, as bases de dados foram submetidas a processos de limpeza e normalização, com o objetivo de garantir a integridade dos dados e reduzir redundâncias.

A terceira fase consistiu no desenvolvimento do modelo de detecção de contratos com maior potencial de riscos, baseado em algoritmos de aprendizado de máquina. Durante a modelagem, foram utilizados atributos de risco para construir o escore de risco. Os atributos foram categorizados em quatro principais categorias: atributos da empresa, atributos dos sócios, atributos do relacionamento (utilizando técnicas de grafos) e atributos relacionados a licitações e contratos.

Foram utilizados cinco algoritmos de aprendizado de máquina: Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree e Support Vector Machine (SVM), entre outros. Os algoritmos foram avaliados em combinação com técnicas de rebalanceamento de dados, incluindo Undersampling, Oversampling, SMOTE (e suas variações) e ADASYN, considerando que os dados de fraude são geralmente desbalanceados.

A quarta fase de avaliação consistiu em testar a ferramenta com um conjunto específico de dados de fornecedores com histórico de sanções. Além disso, a validação foi realizada por meio da comparação dos resultados com os modelos selecionados na revisão da literatura. As métricas utilizadas para avaliar o desempenho dos modelos foram: Acurácia, Área Sob a Curva (AUC), Precisão, Recall, F1-Score , F2-Score e AUC-PR(Precision-Recall Curve). Houve ainda a comparação do desempenho dos modelos utilizando o Model Confidence Set (MSC).

Após a validação, a quinta fase consistiu na prova de conceito, utilizando um painel de governança que apresentou as predições de riscos nas contratações de uma empresa pública

federal de tecnologia, empregando critérios de explicabilidade. Houve o mapeamento de informações de sistemas internos e a realização de testes adicionais com os auditores para que estes informassem suas percepções qualitativas sobre a ferramenta.

3.2 Compreensão do negócio

Conforme informado no item anterior, o campo de aplicação da pesquisa foi em uma empresa de grande porte do setor de tecnologia da informação, com receita anual de R$ 3,6 bilhões, mais de 6000 empregados e um plano de contratação anual de R$ 973 milhões, refletindo a complexidade e o volume de suas operações de aquisição.

Conforme ilustrado na figura Figura 3.2, as principais fontes de demanda incluem tanto o contexto interno (áreas de governança e gestão interna) quanto o contexto externo (órgãos fiscalizadores e demandas de entidades externas). A auditoria de contratações conta com 4 auditores responsáveis pela análise de 2495 documentos de aquisição e 487 contratos de despesas vigentes, além de atender a denúncias internas e externas e alertas de órgãos reguladores.

Assim, o processo de planejamento de auditoria é dividido em várias etapas. Primeiro, é realizada a coleta de informações preliminares, que envolve uma análise do escopo, da legislação aplicável, do fluxo do processo e das bases de dados relevantes. Em seguida, o levantamento do universo amostral ocorre, onde são extraídos dados sobre licitações e contratos existentes nos sistemas internos, além de levantar alertas de órgãos reguladores e demandas da alta administração.

Conforme apresentado na figura Figura 3.2, os componentes em preto representam o estado atual do processo, havendo fragilidade em relação: a um critério objetivo de seleção de contratos a serem auditados, cruzamento com bases externas para aprimorar a seleção.

Os efeitos potenciais dessas falhas incluem o aumento da vulnerabilidade da empresa a fraudes, questionamentos por órgãos reguladores e danos à imagem institucional, sendo caracterizado um risco elevado, refletindo seu impacto significativo.

Para mitigar esses riscos, a inclusão de controles como a implementação de uma metodologia estruturada para identificação de contratos com perfil de risco mais elevado e a integração de ferramentas analíticas para o cruzamento de informações de risco, componente TO-BE da Figura 3.2, pode auxiliar os auditores a identificar e a priorizar dos contratos a serem auditados, aumentando a consistência e a transparência do processo de auditoria,

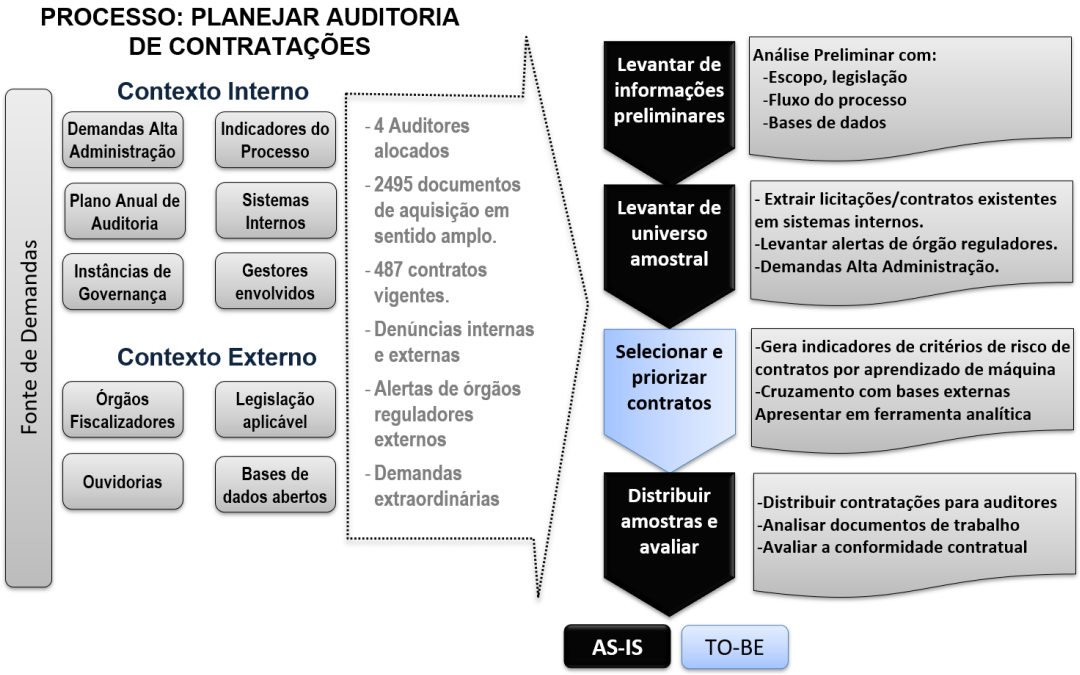

Figura 3.2: Processo de Planejamento de Auditorias de Contratações
Fonte: Elaborado pelo autor

3.3 Preparação e Modelagem dos Dados

Uma das principais dificuldades enfrentadas no contexto das compras públicas é a grande quantidade de dados não estruturados e a dificuldade de integrá-los para análise. Portanto, um mapeamento das bases de dados foi realizado, incluindo informações provenientes de bases públicas como o Portal de Compras e Transparência do Governo Federal do Brasil, tais como dados financeiros, eleitorais, sanções administrativas e outros. Esse mapeamento auxiliou na identificação quais fontes são possíveis e relevantes para a construção de um modelo preditivo de risco, permitindo maior precisão na priorização de contratos para auditoria.

A partir dos critérios de riscos identificados na revisao bibliográfica, foram mapeadas algumas das bases utilizadas no trabalhos, tais como:

- Sanções aplicadas pela Administração Pública: Base de dados de empresas penalizadas, inidôneas ou suspensas, disponibilizada pela Controladoria Geral da União (CGU). Foi utilizada uma versão da base disponibilizada em abril de 2023.

- Cadastro Nacional de Pessoas Físicas e Jurídicas: Inclui informações públicas sobre pessoas físicas (nome, data de nascimento, CPF anonimizado e status do CPF, além de pessoas politicamente expostas) e empresas (filiais, CNAE, endereço, data de constituição, telefone, capital social, entre outros), assim como sócios e proprietários

(nomes e CPFs anonimizados). Essas informações foram obtidas a partir da Receita Federal do Brasil.

- Benefícios Sociais: Informações sobre o recebimento de benefícios sociais, como Bolsa Família, Auxílio Defeso, Benefício de Prestação Continuada, Garantia Safra, entre outros. As bases estão disponíveis no Portal da Transparência da CGU.

- Dados Eleitorais: Informações sobre políticos, candidatos, filiações partidárias e doações a partidos políticos, disponibilizados pelo Tribunal Superior Eleitoral (TSE).

- Fornecedores e Compras Públicas: Informações sobre compras públicas do Governo Federal, incluindo contratos, licitações, lances de licitantes, fornecedores, atas, valores contratados, entre outros.

- Cadastro de Dívida Ativa da União: Dados sobre empresas e pessoas com dívidas ativas na União, sejam elas tributárias, trabalhistas ou previdenciárias. Essa base é disponibilizada pela Procuradoria Geral da Fazenda Nacional (PGFN).

A vantagem desse trabalho está na possibilidade de consolidar todas essas informações em um data lake, permitindo uma análise global ao invés de múltiplos esquemas locais, reduzindo o custo marginal de implementação dos padrões de detecção de risco para cada uma das bases de dados de despesas públicas.

De posse desses dados, foi realizada uma operação ETL (extrair, transformar e carregar), a partir dos arquivos brutos extraídos dos portais governamentais originais, para um data lake unificado. Isso possibilita uma caracterização detalhada dos parceiros e das empresas com base nos dados coletados, carregando o que é relevante a partir dos registros de CPF e CNPJ das entidades a serem analisadas (Figura 3.3).

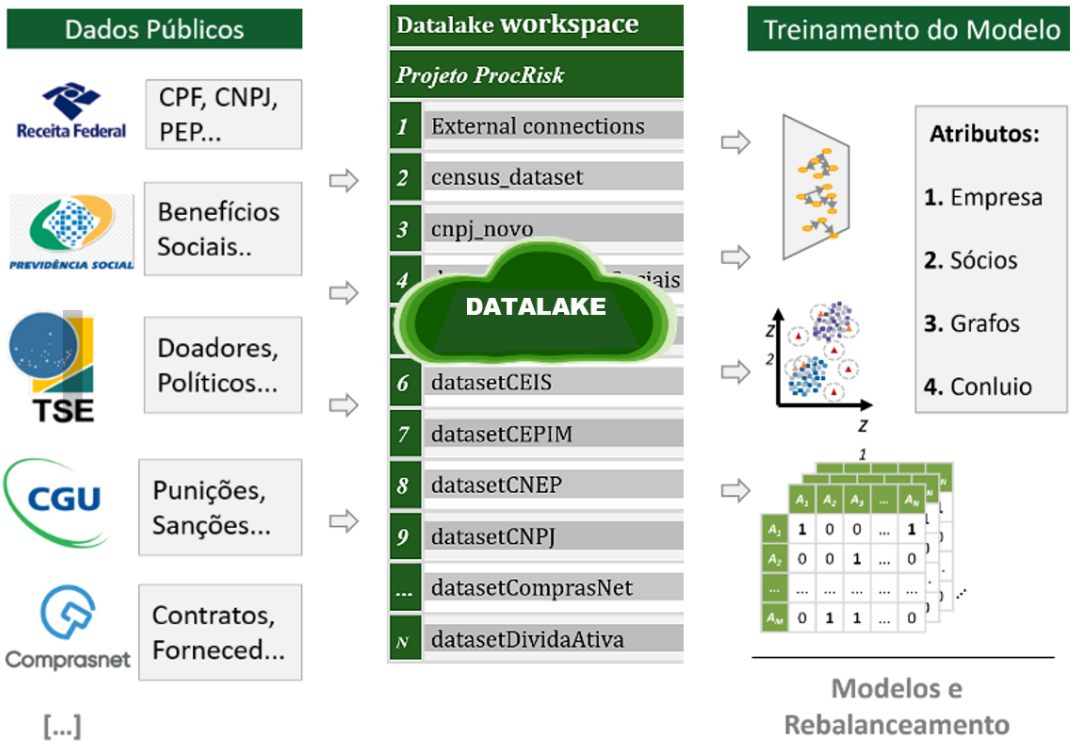

Figura 3.3: Estrutura de ETL para preparação dos dados
Fonte: Elaborado pelo autor

A modelagem dos dados considerou os critérios e indicadores selecionados com base na disponibilidade de dados, revisão teórica, relevância e na capacidade de refletir os riscos associados aos contratos, sendo validados pelos auditores. Os principais critérios e indicadores utilizados para a modelagem foram divididos em três categorias principais: Risco da Empresa, Risco dos Sócios, e Risco do Contrato/Licitação, relacionados nas tabelas 3.1, 3.2 e 3.3.

Tabela 3.1: Critério: Risco Empresa

Critério	**Indicador**
Empresa < 1 ano de criada. (Jesus, 2020; Arista, Fazekas, & Volkotrub, 2024; Sun & Sales, 2018)	O indicador será igual a 1 se a empresa tiver menos de 1 ano de idade, caso contrário, o indicador será igual a 0. Base: CNPJ.

Continua na próxima página...

Continuação da Tabela 3.1

Critério	Indicador
Empresa com múltiplas atividades. (Jesus, 2020; J. L. R. Nascimento, 2022)	O indicador será igual a 1 se a empresa possuir mais de 10 atividades cadastradas no CNAE, distribuídas no mínimo em 3 seções diferentes. Caso contrário, o indicador será igual a 0. Base: CNPJ.
Empresa com contratações > R$ 1 M. (Jesus, 2020; Salazar, Partnership, Brown, & Neumann, 2024)	O indicador será igual a 1 se o valor total das compras vencidas da empresa, desde 2020, for superior a R$ 1 milhão. Caso contrário, o indicador será igual a 0. Base: SIASG.
Empresa com doação eleitoral. (Jesus, 2020; Khemakhem & Dicko, 2013); (J. L. R. Nascimento, 2022)	Empresa já doou recursos para campanha desde 2018. Se o valor da doação > 0, então o resultado é 1; caso contrário, é 0. Base: TSE.
Empresa com menos de 4 empregados (Jesus, 2020; U.S. General Services Administration Office of Inspector General, Office of Audits, n.d.) (Sales & Carvalho, 2016)	Se possui < 4 funcionários na tabela da RAIS = 1, senão = 0. Base: RAIS.
Empresa punição anterior. (Jesus, 2020); (Sales & Carvalho, 2016)	Se o CNPJ possui registro de punição anterior, então = 1, senão = 0. Bases: CEIS e CEPIM.
Empresa recebedora de pagamentos > R$ 1 milhão de outras fontes. (Jesus, 2020) (J. L. R. Nascimento, 2022)	Se empresa recebeu pagamentos de órgãos públicos acima de R$ 1 M desde 2020 de outras origens, o indicador é 1; caso contrário, é 0. Base: SIAFI, SICONV, FUNDEB.
Tipo da Empresa (Jesus, 2020) (J. L. R. Nascimento, 2022)	Gerada a partir do hot encoding da informação de tipo de empresa. Base: CNPJ.
Capital Social (Jesus, 2020) (J. L. R. Nascimento, 2022)	Gerada a partir do hot encoding da informação de capital social organizado em faixas de valor informação de tipo de empresa. Base: CNPJ.
Indicativo Matriz ou Regional (Jesus, 2020) (J. L. R. Nascimento, 2022)	Gerada a partir do hot encoding da informação de se o cnpj é matriz ou filial. Base: CNPJ.

Continua na próxima página...

Continuação da Tabela 3.1

Critério	Indicador
Rede de relacionamento com punição anterior (Jesus, 2020; Sales, 2016; J. L. R. Nascimento, 2022; Fazekas et al., 2021; Carneiro, Veloso, Ventura, Palumbo, & Costa, 2020)	Identifica se a rede relacionamento da empresa (sócios e empresas vinculadas à empresa em análise) já foram punidas anteriormente. Base: CEIS, CNPJ.
Empresa sem empregado RAIS (Jesus, 2020; U.S. General Services Administration Office of Inspector General, Office of Audits, n.d.; Sales & Carvalho, 2016)	Se não possui funcionários na tabela da RAIS = 1, senão = 0. Base: RAIS.

Tabela 3.2: Critério: Risco Sócio

Critério	**Indicador**
Sócio Doador Eleitoral (Jesus, 2020; Khemakhem & Dicko, 2013; Sales & Carvalho, 2016)	Se sócio realizou doação (CPF anonimizado + nome) > 0 então = 1, senão = 0. Base: TSE.
Sócio Inscrito em Programa Social (Jesus, 2020) (C. L. Nascimento, 2022)	Se sócio recebeu benefício social, então indicador = 1, senão = 0. Bases: BPC, Auxílio Emergencial, Defeso, Bolsa Família.
Sócio com Instrução Baixa (C. L. Nascimento, 2022)	Se sócio consta na RAIS com indicador de até nível médio, então indicador = 1, senão = 0. Base: RAIS.
Sócio com Salário Baixo na RAIS (C. L. Nascimento, 2022)	Se sócio consta na RAIS com salário base inferior a R$ 1.567,50 então indicador = 1, senão = 0. Base: RAIS.
Sócio Falecido (Jesus, 2020) (Sales & Carvalho, 2016); (C. L. Nascimento, 2022)	Se sócio consta como falecido, então indicador = 1, senão = 0. Base: CPF.
Sócio Filiado a Partido Político (Jesus, 2020); (Sales & Carvalho, 2016); (C. L. Nascimento, 2022)	Se a empresa tem sócio filiado ao partido político, então indicador = 1, senão = 0. Base: TSE(até 2022).
Sócio Funcionário Público Federal (Sales & Carvalho, 2016); (C. L. Nascimento, 2022))	Se empresa possui sócio servidor público federal = 1, senão = 0. Base: Servidores Federais.
Sócio recebedor de pagamentos de órgãos públicos > R$ 1 M (Sales & Carvalho, 2016) (C. L. Nascimento, 2022)	Se sócio recebeu de pagamentos de órgãos públicos acima de 1 milhão desde 2017 (SIAFI, SICONV, FUNDEB), então indicador = 1, senão = 0.

Tabela 3.3: Critério: Risco Contrato/Licitação

Critério	**Indicador**
Múltiplos aditivos de valor/vigência (Jesus, 2020) (Sales & Carvalho, 2016)	Se contrato possui quantidade de aditivos contratuais de valor ou vigência > 2, então = 1, senão = 0. Base: Sistemas corporativos internos.
Licitação de alta materialidade (Jesus, 2020) (Sales & Carvalho, 2016)	Se o valor da licitação excede R$ 1 M, então = 1, senão = 0. Base: Sistemas corporativos internos.
É dispensa ou inexigibilidade (Jesus, 2020) (Sales & Carvalho, 2016)	Se licitação foi modalidades dispensa ou inexigibilidade = 1, senão = 0. Base: Sistemas corporativos internos.
Alerta de órgãos reguladores (Jesus, 2020) (Sales & Carvalho, 2016)	Se contrato ou fornecedor recebeu de órgãos reguladores, então = 1, senão = 0. Base: Sistemas corporativos internos.

Posto isso, a literatura destaca que a integração de diferentes sistemas de alerta e bases de dados pode fornecer uma visão abrangente e em tempo real dos riscos associados às contratações públicas. Esses alertas permitem uma atuação preventiva e a correção de irregularidades antes que causem danos maiores ao erário público .

Merece destaque que o escopo de risco sócio, com base nos trabalhos relacionados a redes de relacionamentos, também haverá a informação se o sócio possui alguma empresa que tenha sido punida anteriormente.

Análise Exploratória do Dados

Para maior compreensão dos dados foi realizada análise exploratória de dados (AED), visando identificar padrões subjacentes e detectar possíveis inconsistências nos dados. Para facilitar a interpretação e garantir uma melhor comparabilidade entre variáveis de diferentes escalas, foi aplicada a normalização Min-Max Scaling, conforme recomendado em estudos como J. Han et al. (2011). Essa transformação ajusta os valores das variáveis para o intervalo $[0, 1]$, preservando a distribuição original e reduzindo o impacto de diferenças de magnitude.

A Tabela 3.4 apresenta um resumo estatístico das principais variáveis do conjunto de dados após a normalização. Os valores incluem a média (μ), o desvio padrão (σ), os valores mínimo e máximo ($[0, 1]$) e o 1º quartil (25%) e 3º Quartil (75%). Essa análise permite identificar características importantes da distribuição dos dados, como a presença de assimetrias ou variáveis com concentração em determinados intervalos.

Ressalta-se que, tendo em vista que as variáveis são derivadas de outras fontes, a base não possui informações de ausentes(missing).

Tabela 3.4: Estatísticas descritivas dos atributos utilizados no modelo

Atributo	Quant.	Média	Desvio Padrão	Mínimo	25%	75%	Máximo
Empresa Sem Empregado RAIS	12090.0	0.440612	0.496481	0.0	0.0	1.0	1.00
Empresa Menos 4 Empregados Rais	12090.0	0.592721	0.491348	0.0	0.0	1.0	1.00
Empresa Punida	12090.0	0.156824	0.363650	0.0	0.0	0.0	1.00
Empresa Doadora	12090.0	0.115219	0.319299	0.0	0.0	0.0	1.00
Socio ou Responsável Doador	12090.0	0.288172	0.452930	0.0	0.0	1.0	1.00
Empresa com Múltiplas Atividades	12090.0	0.124897	0.330615	0.0	0.0	0.0	1.00
Empresa com Menos de 1 ano de Criada	12090.0	0.000744	0.027275	0.0	0.0	0.0	1.00
Empresa com Valor em Compras (SIASG) Maior que R$ 1 Milhão	12090.0	0.173366	0.378579	0.0	0.0	0.0	1.00
Empresa recebedora de pagamentos de órgaos públicos com valor maior que 1 milhão desde 2017	12090.0	0.157734	0.364506	0.0	0.0	0.0	1.00
Sócio ou responsável recebedor de pagamentos de órgaos públicos com valor maior que 1 milhão desde 2017	12090.0	0.000248	0.015751	0.0	0.0	0.0	1.00
Primeira Compra da Empresa (SIASG) em Menos de 90 Dias da Abertura	12090.0	0.065261	0.246995	0.0	0.0	0.0	1.00
Socio ou Responsável Funcionário Público Federal	12090.0	0.065095	0.246704	0.0	0.0	0.0	1.00
Socio Filiado Partido Político	12090.0	0.132366	0.362093	0.0	0.0	0.0	2.40
Responsavel Filiado Partido Político	12090.0	0.101175	0.317906	0.0	0.0	0.0	1.10
Socio ou Responsável com Mandado de Prisão em Aberto	12090.0	0.001572	0.039613	0.0	0.0	0.0	1.00
Municipio Ex-Sócio Muito Distante Municipio Empresa	12090.0	0.102316	0.678751	0.0	0.0	0.0	18.00
Municipio Sócio Muito Distante Município Empresa	12090.0	0.042804	0.332075	0.0	0.0	0.0	10.50
CEP Ex-Sócio Igual CEP Empresa	12090.0	0.052316	0.259354	0.0	0.0	0.0	6.00
CEP Sócio Igual CEP Empresa	12090.0	0.053846	0.241738	0.0	0.0	0.0	5.00
Ex-Sócio Inscrito em Programa Social	12090.0	0.059326	0.247602	0.0	0.0	0.0	8.00
Sócio Inscrito em Programa Social	12090.0	0.028122	0.193911	0.0	0.0	0.0	12.00
Ex-Sócio com Instrução Baixa na RAIS	12090.0	0.015281	0.109562	0.0	0.0	0.0	2.50
Sócio com Instrução Baixa na RAIS	12090.0	0.006266	0.069019	0.0	0.0	0.0	2.00
Ex-Sócio com Salario Baixo na RAIS	12090.0	0.065302	0.866602	0.0	0.0	0.0	62.50
Sócio com Salario Baixo na RAIS	12090.0	0.041253	0.680496	0.0	0.0	0.0	51.25
Sócio Falecido	12090.0	0.016543	0.141103	0.0	0.0	0.0	6.00
Ex-Socio sem Veiculo	12090.0	0.800434	3.383829	0.0	0.0	1.0	208.75
Socio sem Veiculo	12090.0	0.450806	2.483421	0.0	0.0	0.5	150.00
Município Responsável Muito Distante Município Empresa	12090.0	0.030190	0.195169	0.0	0.0	0.0	1.50
CEP Responsável Igual CEP Empresa	12090.0	0.058768	0.168932	0.0	0.0	0.0	1.00
Responsável Inscrito em Programa Social	12090.0	0.030811	0.125621	0.0	0.0	0.0	1.00
Responsável com Instrução Baixa na RAIS	12090.0	0.004053	0.050681	0.0	0.0	0.0	1.25
Responsável com Salario Baixo na RAIS	12090.0	0.013730	0.120005	0.0	0.0	0.0	1.25
Responsável Falecido	12090.0	0.007899	0.088820	0.0	0.0	0.0	1.25
Responsável sem Veiculo	12090.0	0.198222	0.308054	0.0	0.0	0.5	1.00
Rede em 2 nível punida	12090.0	0.004549	0.067297	0.0	0.0	0.0	1.00
Rede em 3 nível punida	12090.0	0.043011	0.240225	0.0	0.0	0.0	3.00

Fonte: Elaborado pelo autor.

A matriz de correlação de Pearson fornece uma visão estatística das relações lineares entre os atributos do conjunto de dados. A correlação varia de -1 a 1, onde valores próximos de 1 indicam uma forte correlação positiva, valores próximos de -1 representam uma correlação negativa e valores próximos de zero sugerem ausência de relação linear entre as variáveis. A partir da análise dos coeficientes de correlação, foi possível identificar padrões relevantes que auxiliam na compreensão da estrutura dos dados e na seleção das variáveis para modelagem preditiva.

Os resultados revelam que algumas variáveis apresentam forte correlação positiva, sugerindo que quando um atributo aumenta, o outro também tende a crescer. Destaca-se a relação entre a variável "Empresa recebedora de pagamentos de órgãos públicos com valor maior que 1 milhão desde 2017"e "Empresa com Valor em Compras (SIASG) Maior que R$ 1 Milhão", com coeficiente de correlação de aproximadamente 0,3. Essa relação indica que empresas que recebem valores elevados de órgãos públicos tendem a ter também altos valores em compras via SIASG, sugerindo um padrão de fornecedores recorrentes que mantêm contratos públicos ao longo do tempo.

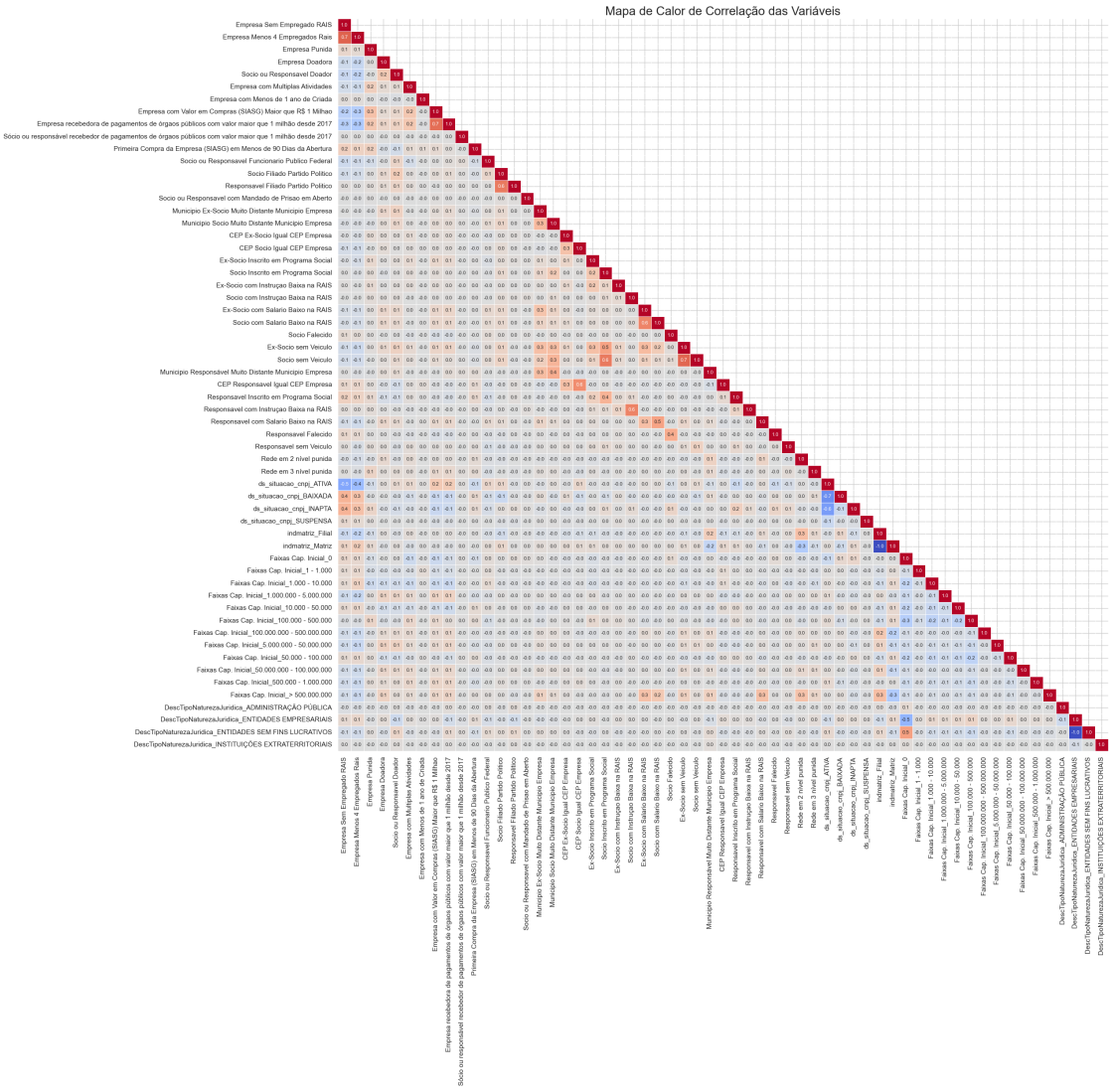

Figura 3.4: Matriz de Correlação Pearson entre os atributos
Fonte: Elaborado pelo autor

A análise da matriz de correlação reforça a importância de selecionar variáveis que apresentam forte relação com os riscos associados às empresas punidas. A identificação de padrões de comportamento permite avaliar com maior precisão quais fatores devem ser considerados na construção do modelo preditivo. Os resultados indicam que a experiência da empresa no mercado, o volume financeiro de compras públicas e a relação entre doações eleitorais e recebimento de contratos públicos são aspectos altamente correlacionados e merecem atenção na análise de riscos. Além disso, a relação entre punições administrativas e continuidade na prestação de serviços ao governo sugere fragilidades nos mecanismos de controle e aplicação de sanções.

Os achados desta análise contribuem para a fundamentação das escolhas metodológicas empregadas na modelagem preditiva e fornecem subsídios para a definição de critérios

de risco, permitindo que o modelo desenvolvido reflita com maior precisão os padrões observados no conjunto de dados. A interpretação detalhada dessas relações permite que os auditores compreendam o impacto de cada variável na priorização de contratos de alto risco, garantindo maior transparência e eficácia no processo de auditoria.

3.4 Desenvolvimento do Modelo

3.4.1 Seleção de Atributos (*Feature Selection*)

A etapa de seleção de atributos foi conduzida com o objetivo de reduzir a dimensionalidade do conjunto inicial de 59 variáveis, identificando um subconjunto otimizado de *features* que preservasse a capacidade preditiva e reduzisse a complexidade dos modelos subsequentes. Dada a natureza de classificação binária do problema (risco vs. não risco), optou-se pela aplicação de uma Regressão Logística com regularização L1 (LASSO). Esta técnica é metodologicamente eficaz para a seleção, pois a penalidade L1 tem a propriedade de zerar os coeficientes das variáveis consideradas menos relevantes (Tibshirani, 1996b).

Para a implementação, foi utilizada a função LogisticRegressionCV do scikit-learn, que integra a seleção de hiperparâmetros com a validação cruzada. O processo foi configurado com a penalidade L1 (penalty='l1') e o solucionador *'liblinear'*. A validação cruzada (CV) empregou a estratégia validação cruzada com 5 *folds*, garantindo que a proporção das classes fosse mantida em cada *fold*, em consistência com a separação treino-teste. Adicionalmente, para mitigar o viés gerado pelo desbalanceamento dos dados durante a própria calibração da seleção, foi ativado o parâmetro *class_weight='balanced'*.

O procedimento de CV otimizou o hiperparâmetro `C`, que representa o inverso da força de regularização, onde valores menores de "C"implicam em maior penalização. A busca foi realizada em uma grade logarítmica de 30 pontos, variando de 10^{-2} a 10^{4}. A Figura 3.5 ilustra a curva de validação resultante, demonstrando o desempenho em função do parâmetro `C`. Diferente da abordagem anterior que minimizava o MSE, esta otimização maximizou a Acurácia Médiaobtida na validação cruzada. O valor de `C` ótimo encontrado (indicado pela linha vertical no gráfico) representa o ponto de melhor equilíbrio entre *underfitting* (valores de `C` muito baixos) e *overfitting* (valores de `C` muito altos).

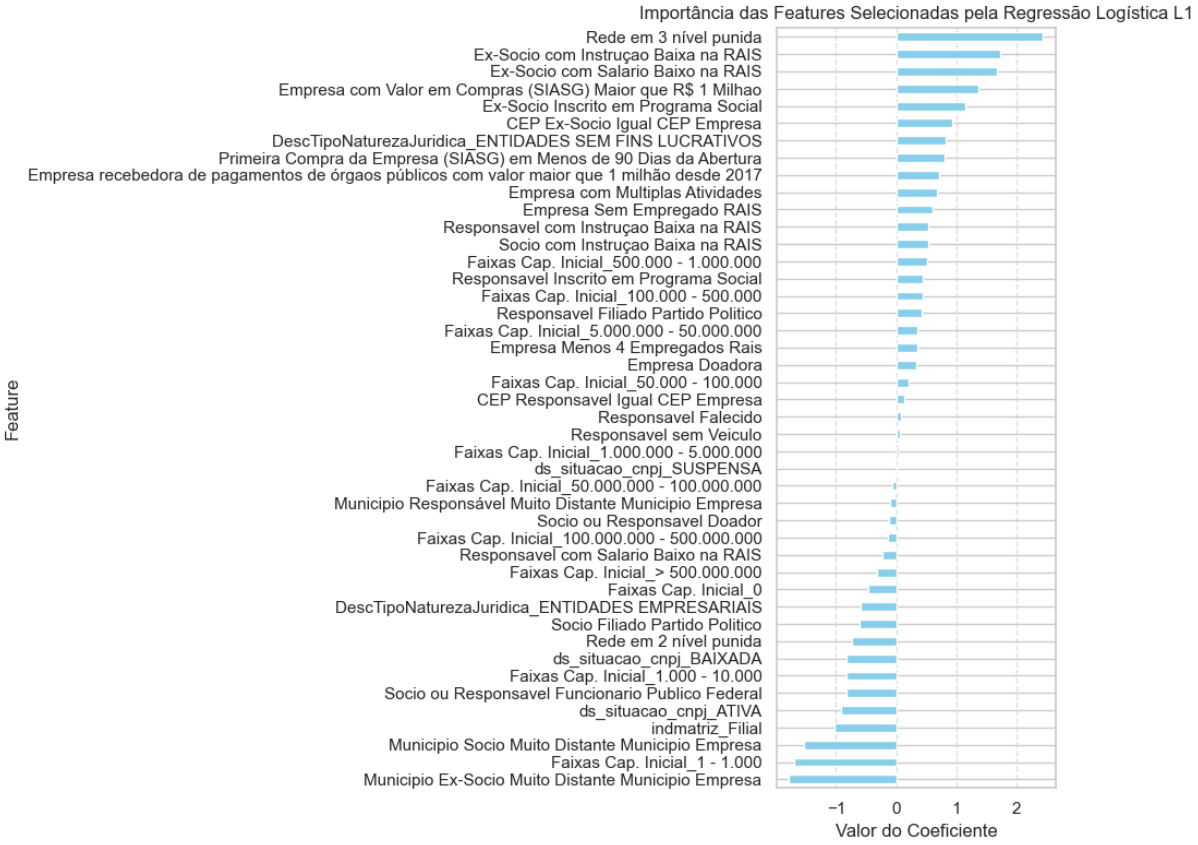

Figura 3.5: Importância das Features Selecionadas pela Regressão Logística L1

Como resultado deste processo, foram selecionadas, correspondendo a todos os coeficientes que permaneceram não-nulos. A Figura 3.6 detalha estas variáveis e a magnitude de seus respectivos coeficientes, que indicam a força e a direção (positiva ou negativa) da contribuição de cada *feature* para a predição do risco. O uso deste conjunto enxuto de *features* é utilizado nas etapas subsequentes de modelagem.

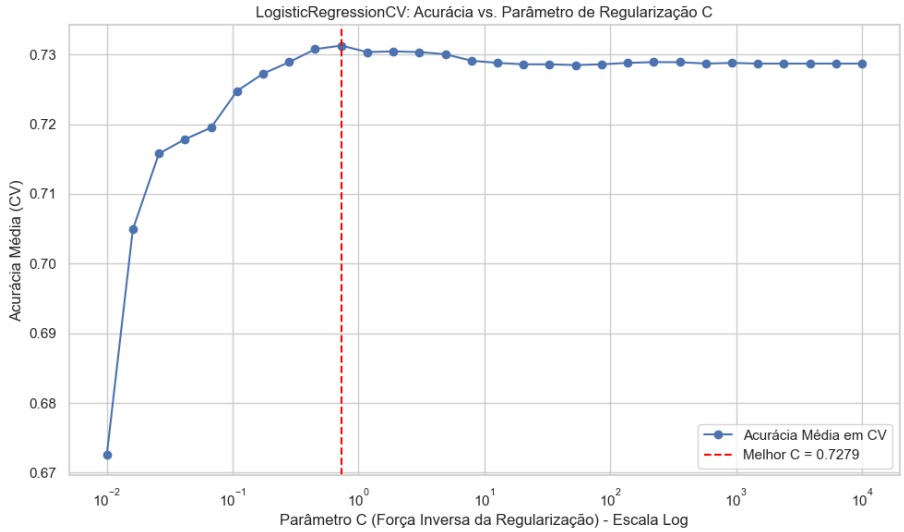

Figura 3.6: Regressão Logística L1(CV): Acurácia Média vs. Parâmetro C

3.4.2 Preparação da Base de Treino e Teste

Para o desenvolvimento dos modelos supervisionados, foi realizada a segmentação do conjunto de dados em subconjuntos de treinamento (treino) e validação (teste). Neste estudo, foi definida uma proporção de 80% dos dados para treino e 20% para teste.

Conforme discutido, o conjunto de dados é caracterizado pelo desbalanceamento da variável-alvo "Empresa com Punição". Este é um cenário comum em problemas de detecção de risco, onde a classe positiva (minoritária) é geralmente em menor número. A prevalência geral de casos positivos (empresas punidas) identificada no conjunto de dados completo foi de 15,68%.

Visando assegurar que esta proporção crítica fosse mantida em ambas as partições, mitigando o risco de viés na avaliação do desempenho, foi empregada a técnica de **amostragem estratificada** (*stratified sampling*). Tal método garante que a distribuição da variável-alvo nos conjuntos de treino e teste constitua uma representação fidedigna da distribuição original.

A eficácia da estratificação foi validada por meio da comparação da prevalência da classe minoritária entre as partições, conforme detalhado na Tabela 3.5.

Tabela 3.5: Prevalência da Variável-Alvo nas bases Treino-Teste

Conjunto	**Proporção da Classe Minoritária**
Geral (Completo)	15,68%
Treinamento (Treino)	15,68%
Validação (Teste)	15,67%

Fonte: Elaborado pelo autor.

Como observado, a prevalência manteve-se estatisticamente idêntica, confirmando o sucesso da estratificação.

Adicionalmente à estratificação da variável-alvo, verificou-se se as 44 variáveis independentes (*features*) mantêm suas distribuições estatísticas entre os conjuntos de treino e teste. A divergência nessas distribuições, fenômeno conhecido como *data drift* ou *covariate shift*, comprometeria a avaliação de generalização, uma vez que o modelo seria testado em dados com características distintas daquelas utilizadas em seu treinamento.

Para a validação da estabilidade das *features*, foram aplicados dois testes estatísticos robustos. O Teste de Kolmogorov-Smirnov (KS) foi utilizado para avaliar a hipótese nula de que as duas amostras (treino e teste) provêm da mesma distribuição (Siddiqi, 2017), onde p-valores elevados indicam similaridade. Concomitantemente, aplicou-se o *Population Stability Index (PSI)*, métrica que quantifica a magnitude da mudança, na qual valores inferiores a 0.10 são considerados indicadores de alta estabilidade. Ambas as métricas são ferramentas metodológicas estabelecidas para o monitoramento e validação de modelos, essenciais para detectar o *data drift* entre a população de desenvolvimento e a de aplicação (Anderson, 2007).

Os resultados obtidos para a totalidade das 44 *features* confirmaram a alta estabilidade das partições. Conforme demonstrado na Tabela 3.6, que resume uma amostra dos resultados, os p-valores do teste KS permaneceram significativamente elevados, enquanto os valores de PSI mostraram-se próximos de zero, muito abaixo do limiar de alerta.

Tabela 3.6: Amostra da Análise de Estabilidade Distribucional das Features (Treino vs. Teste)

Feature (Amostra de 25 Variáveis)	**Teste KS (p-value)**	**PSI**	**Resultado**
DescTipoNaturezaJuridica_ENTIDADES EMPRESARIAIS	1.000000	0.000000	Estável
DescTipoNaturezaJuridica_ENTIDADES SEM FINS LUCRATIVOS	1.000000	0.000000	Estável
Ex-Socio com Instruçao Baixa na RAIS	1.000000	0.000000	Estável
Faixas Cap. Inicial_1 - 1.000	1.000000	0.000000	Estável
Faixas Cap. Inicial_100.000 - 500.000	1.000000	0.000000	Estável
Faixas Cap. Inicial_100.000.000 - 500.000.000	1.000000	0.000000	Estável
Faixas Cap. Inicial_> 500.000.000	1.000000	0.000000	Estável
Primeira Compra da Empresa (SIASG) em Menos de 1 ano	1.000000	0.000000	Estável
Rede em 2 nível punida	1.000000	0.000000	Estável
Responsavel Falecido	1.000000	0.000000	Estável
Responsavel com Salario Baixo na RAIS	1.000000	0.000000	Estável
Socio ou Responsavel Funcionario Publico Federal	1.000000	0.000000	Estável
ds_situacao_cnpj_BAIXADA	1.000000	0.000000	Estável
indmatriz_Filial	1.000000	0.000000	Estável
Empresa com Valor em Compras (SIASG) Maior que R$1M	1.000000	0.000000	Estável
Empresa recebedora de pagamentos de órgaos públicos.	1.000000	0.000000	Estável
Empresa com Multiplas Atividades	0.999952	0.000000	Estável
Faixas Cap. Inicial_5.000.000 - 50.000.000	0.999656	0.000000	Estável
Empresa Menos 4 Empregados Rais	0.941717	0.000000	Estável
ds_situacao_cnpj_ATIVA	0.937742	0.000000	Estável
Empresa Doadora	0.933615	0.000000	Estável
Responsavel sem Veiculo	0.267149	0.000385	Estável
Socio Filiado Partido Politico	0.998165	0.001730	Estável

Fonte: Elaborado pelo autor.

Este rigor metodológico na segmentação e validação dos dados auxilia o treinamento, garantindo que o subconjunto de teste constitui uma amostra representativa, conferindo um dado melhor qualidade para o modelo.

3.4.3 Balanceamento da Base de Dados

Para lidar com o desbalanceamento dos dados, foram aplicadas quatro técnicas distintas: *oversampling, undersampling, SMOTE (Synthetic Minority Oversampling Technique) e ADASYN (Adaptive Synthetic Sampling).* Cada uma dessas abordagens possui características específicas, sendo escolhidas com base nas necessidades do modelo e na complexidade do problema.

A técnica de *oversampling* consistiu em replicar os exemplos da classe minoritária, aumentando sua representação no conjunto de dados. Embora simples, essa abordagem pode levar a problemas como *overfitting*, uma vez que repete informações sem adicionar novas variações. Para mitigar esses efeitos, foram aplicadas técnicas mais avançadas, como o SMOTE, que cria exemplos sintéticos da classe minoritária ao interpolar os dados existentes. O SMOTE foi escolhido por sua capacidade de gerar amostras mais diversas, reduzindo o risco de *overfitting* e melhorando a generalização do modelo.

Por outro lado, o *undersampling* foi utilizado para reduzir a quantidade de exemplos da classe majoritária, equilibrando as proporções entre as classes. Essa abordagem é útil para simplificar a base de dados e acelerar o treinamento dos modelos, mas pode resultar na perda de informações importantes. Por fim, o ADASYN foi implementado como uma técnica adaptativa, que gera amostras sintéticas de maneira mais concentrada em regiões onde a classe minoritária é mais sub-representada, aumentando a eficácia do balanceamento em cenários complexos.

A Tabela 3.7 apresenta a distribuição de instâncias por classe no conjunto de treino antes e após a aplicação de diferentes técnicas de rebalanceamento. No conjunto original observa-se forte desequilíbrio (82% da classe 0). A subamostragem igualou as classes reduzindo a classe majoritária, enquanto a sobreamostragem e SMOTE replicaram ou geraram amostras sintéticas para a classe minoritária, mantendo o total de 1 636 instâncias em cada classe. O ADASYN, por sua vez, gerou 1 555 exemplos sintéticos, resultando em leve desequilíbrio residual.

Tabela 3.7: Conjunto de treino antes e após técnicas de rebalanceamento.

Técnica	**Classe 0**	**Classe 1**
Imbalanced (Original)	1.636	364
Undersampling (Subamostragem)	364	364
Oversampling (Sobreamostragem)	1.636	1 636
SMOTE	1.36	1.636
ADASYN	1.636	1.555

O processo de balanceamento foi cuidadosamente validado para assegurar que os dados resultantes mantivessem sua representatividade original e refletissem adequadamente a

realidade dos casos analisados. Essa etapa foi fundamental para a construção de um conjunto de dados robusto, apto a suportar as fases subsequentes de seleção de atributos e modelagem preditiva.

3.4.4 Treinamento e Validação do Modelo

Após a etapa de seleção de atributos, o próximo passo da metodologia consistiu no treinamento de modelos de aprendizado de máquina utilizando os dados pre-processados e equilibrados. Esta fase teve como objetivo construir um modelo preditivo capaz de identificar contratos com maior risco de irregularidades, considerando os critérios definidos previamente.

Para esta tarefa, foram utilizados algoritmos clássicos de aprendizado supervisionado, conhecidos por sua eficácia em problemas de classificação. Os algoritmos escolhidos incluem *Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine* (SVM) e Redes Neurais. A seleção desses algoritmos foi fundamentada em suas características complementares e capacidades de capturar padrões complexos em dados desbalanceados.

O algoritmo *Random Forest* foi escolhido devido à sua robustez na modelagem de variáveis categóricas e contínuas, bem como à sua capacidade de lidar com interações não lineares entre os atributos. Ele cria um conjunto de árvores de decisão independentes e utiliza a média ou a votação majoritária para fornecer uma previsão final, reduzindo o risco de overfitting observado em árvores de decisão isoladas.

O *K-Nearest Neighbors* (KNN) foi incluído por sua simplicidade e eficácia em contextos onde a separação entre classes é bem definida. Este algoritmo classifica uma instância com base na maioria das classes de seus vizinhos mais próximos, o que pode ser útil para identificar contratos com características similares aos de maior risco.

Naive Bayes foi utilizado devido à sua eficiência em conjuntos de dados com muitas variáveis categóricas. Este modelo probabilístico baseia-se no Teorema de Bayes e na suposição de independência condicional entre os atributos, o que permite uma rápida construção e aplicação em sistemas de detecção de risco.

O *Support Vector Machine* (SVM) foi aplicado para separar classes com margens máximas, tornando-se uma escolha apropriada para problemas de classificação em conjuntos de dados de alta dimensionalidade. Este modelo utiliza funções de kernel para projetar os dados em um espaço de maior dimensionalidade, permitindo identificar fronteiras não lineares.

Por fim, as Redes Neurais foram incorporadas para capturar padrões mais complexos e não lineares nos dados. Este modelo utiliza camadas de neurônios artificiais conectados

para aprender representações hierárquicas dos dados, sendo particularmente eficaz em problemas com alta variabilidade nas entradas.

Para consolidar os resultados obtidos pelos modelos individuais, implementou-se uma arquitetura de *Hybrid Ensemble*. A estratégia de combinação adotada foi o *Soft Voting*, onde a probabilidade final é calculada pela média aritmética das probabilidades estimadas pelos classificadores base (*TabPFN*, *XGBoost* e *LightGBM*), treinados sob diferentes estratégias de amostragem (*Undersampling* e dados originais). Esta abordagem visa reduzir a variância das predições e mitigar o viés de confiança excessiva (*overconfidence*) observado em modelos isolados.

Durante o treinamento, técnicas de validação cruzada foram utilizadas para avaliar o desempenho dos modelos em diferentes partições do conjunto de dados, garantindo a generalização das previsões. Além disso, foram aplicadas estratégias de otimização de hiperparâmetros, como a busca em grade (*GridSearch*), para ajustar os parâmetros dos algoritmos e maximizar sua capacidade preditiva. A tabela 3.8 sintetiza os parâmetros considerados para os modelos.

Tabela 3.8: Modelos, hiperparâmetros, métodos de amostragem e validação para detecção de fraudes

Modelo	**Hiperparâmetros**
Logistic Regression	Solver=liblinear, penalty=[l1,l2], C=[0.01..100], class_weight=[None,'balanced']
Random Forest	n_estimators=[50..500], max_depth=[3..None], min_samples_split=[2..20], min_samples_leaf=[1..10], class_weight=[None,'balanced','balanced_subsample']
K-Nearest Neighbors	n_neighbors=[1,30], weights=[uniform,distance], p=[1,2]
SVM	kernel=[linear,rbf], C=[0.1..10], gamma=[0.001..1], class_weight=[None,'balanced']
Gaussian Naive Bayes	var_smoothing=[1e-12..1e-6]
XGBoost	n_estimators=[50..400], max_depth=[3..9], learning_rate=[0.001..0.3], subsample=[0.6..1.0], colsample_bytree=[0.6..1.0], gamma=[0,5] ,scale_pos_weight para classe minoritária
LightGBM	n_estimators=[50..500], max_depth=[3..15], learning_rate=[0.01..0.2], num_leaves=[20..150], class_weight='balanced'

Para o treinamento, utilizou-se uma proporção de 20% dos dados para teste e 80% para treinamento, com validação cruzada em cinco folds. Todos os modelos foram avaliados com base em métricas específicas para dados desbalanceados, como F2-Score, AUC-PR e Recall, que priorizam a detecção de falsos negativos. Esse enfoque foi essencial para garantir que o modelo fosse capaz de identificar contratos de alto risco sem comprometer excessivamente a taxa de falsos positivos, atendendo às necessidades específicas do contexto de auditoria.

3.4.5 Teste de Seleção do Modelo (MCS)

A fase de validação e seleção do modelo foi uma etapa crucial para garantir a robustez e a confiabilidade do modelo preditivo. O objetivo principal desta etapa foi avaliar o desempenho dos modelos de aprendizado de máquina treinados anteriormente e selecionar aquele que apresentasse o melhor equilíbrio entre as métricas de avaliação, considerando as peculiaridades dos dados desbalanceados e as exigências do contexto de auditoria.

Para avaliar a eficácia dos modelos, foram utilizadas múltiplas métricas de desempenho. As principais métricas selecionadas foram F2-Score, Área Sob a Curva *Precision-Recall* (AUC-PR), Área Sob a Curva ROC (AUC-ROC), Precisão e Recall. A métrica *Weighted LogLoss* foi utilizada como função de perda devido à sua sensibilidade ao desbalanceamento dos dados, avaliando o desempenho e a calibração do modelo em prever a classe minoritária.

Além disso, a validação foi realizada por meio de uma abordagem de validação cruzada em *k-fold*, na qual o conjunto de dados foi dividido em k partições (neste caso, k=5). Em cada iteração, uma partição foi utilizada como conjunto de teste, enquanto as demais foram empregadas para treinamento, permitindo uma avaliação robusta e minimizando o risco de sobreajuste (*overfitting*). Essa abordagem assegurou que o desempenho do modelo fosse testado em diferentes amostras do conjunto de dados, refletindo sua capacidade de generalização.

Uma ferramenta adicional para a seleção do modelo foi o uso do método do Conjunto de Confiança do Modelo (MCS), baseado em testes estatísticos para identificar o conjunto de modelos que, com um nível de confiança especificado (neste caso, 95%), incluísse o modelo ou modelos mais eficazes. O MCS compara modelos simultaneamente com base em suas funções de perda, como erro quadrático médio ou erro absoluto médio, e exclui iterativamente aqueles com desempenho inferior. Este método permitiu uma avaliação formal e estatisticamente fundamentada, garantindo que o modelo selecionado fosse robusto e confiável.

Com base nos resultados da validação, o modelo final foi escolhido considerando não apenas as métricas quantitativas, mas também a interpretabilidade e a facilidade de in-

tegração com o fluxo de trabalho existente na auditoria governamental. Essa abordagem equilibrada garantiu que o modelo fosse tanto estatisticamente sólido quanto operacionalmente viável, maximizando seu impacto na detecção de contratos com maior risco de irregularidades.

3.5 Explicabilidade do Modelo de IA

A explicabilidade do modelo é uma etapa relevante no desenvolvimento de sistemas baseados em aprendizado de máquina, especialmente em contextos sensíveis, como a auditoria governamental. Nesta pesquisa, foram implementadas técnicas de inteligência artificial explicável (XAI) para garantir que o modelo preditivo identificasse não apenas os contratos com maior risco de irregularidades, mas também fornecesse justificativas claras e compreensíveis para suas predições. A transparência é essencial para aumentar a confiança dos auditores e gestores na ferramenta, bem como para atender aos requisitos de prestação de contas e de governança.

Entre as técnicas de explicabilidade disponíveis, os valores de Shapley (SHAP) (Shapley, 1953) foram selecionados como a abordagem principal devido à sua capacidade de atribuir impactos quantitativos a cada variável de entrada em uma predição específica. O SHAP é baseado em conceitos da teoria dos jogos e calcula a contribuição marginal de cada variável para a predição final, considerando todas as possíveis combinações de variáveis. Essa abordagem permite tanto explicações globais, fornecendo insights sobre os fatores que mais influenciam o modelo como um todo, quanto explicações locais, detalhando a influência das variáveis em uma instância específica.

No contexto desta pesquisa, os valores de SHAP foram aplicados para interpretar os resultados do modelo de Gradient Boosting, que foi identificado como o mais eficaz na fase de validação. O processo envolveu a análise dos impactos dos atributos em diferentes níveis: atributos relacionados à empresa (como a idade da empresa, o número de empregados e o histórico de sanções), atributos dos sócios (como o histórico de doações eleitorais e os vínculos com benefícios sociais), e atributos relacionados ao contrato ou à licitação (como os valores contratados e as modalidades de licitação).

Os resultados da análise de explicabilidade revelaram que alguns atributos tiveram um impacto significativamente maior nas predições do modelo. Por exemplo, as empresas que possuem múltiplas atividades cadastradas, sócios com histórico de doações eleitorais e contratos com valores elevados emergem como os principais fatores de risco. A visualização dos valores de SHAP, por meio de gráficos como histogramas e gráficos de dispersão, permitiu uma interpretação clara de como cada atributo contribuiu para a classificação de risco. Além disso, a análise local forneceu explicações detalhadas para casos específicos,

permitindo que os auditores investigassem mais a fundo os contratos que apresentaram altas pontuações de risco.

Uma das principais vantagens do uso de SHAP é a capacidade de detectar potenciais vieses no modelo. Por exemplo, foi possível identificar situações em que o modelo atribuía um peso excessivo a atributos menos relevantes ou negligenciava variáveis importantes em decorrência de interações complexas entre os dados. Esses insights contribuíram para refinar o modelo e a aumentar sua confiabilidade.

Por fim, a explicabilidade do modelo foi integrada ao framework de tomada de decisão, facilitando a comunicação dos resultados aos auditores e gestores. A interface final inclui visualizações dos valores de SHAP para cada contrato, permitindo uma análise transparente e fundamentada. Esse componente reforça a utilidade prática da ferramenta, assegurando que as decisões sejam compreensíveis, justificáveis e alinhadas às necessidades da auditoria governamental.

3.6 Framework de Tomada de Decisão

O framework de tomada de decisão desenvolvido nesta pesquisa foi projetado para traduzir os resultados do modelo preditivo em insights acionáveis, proporcionando aos auditores ferramentas que auxiliem na priorização e análise de contratos públicos. Este framework integra os resultados da análise preditiva com painéis de visualização e critérios qualitativos, promovendo uma abordagem holística para o planejamento de auditorias.

A base do framework é composta por um painel de monitoramento de riscos que consolida informações provenientes do modelo preditivo e do módulo de explicabilidade. Este painel foi desenvolvido para apresentar os contratos priorizados de maneira clara e estruturada, classificando-os de acordo com os níveis de risco atribuídos pelo modelo. As visualizações incluem gráficos de barras e mapas de calor que destacam os contratos com maior risco, permitindo a rápida identificação de padrões e tendências.

Além disso, o framework incorpora critérios qualitativos definidos em conjunto com os auditores da organização, garantindo que aspectos que não são capturados pelos dados também sejam considerados. Por exemplo, contratos com maior visibilidade pública ou um histórico de problemas em auditorias anteriores podem receber um peso adicional na classificação final. Este componente qualitativo é integrado ao framework por meio de uma interface interativa, onde os auditores podem ajustar os critérios e simular diferentes cenários de priorização.

Outro aspecto central do framework é o apoio na interpretação dos valores de SHAP, que são utilizados para justificar as classificações atribuídas pelo modelo. Esse recurso foi implementado com o objetivo de fornecer explicações claras e acessíveis sobre os fatores

que influenciam a pontuação de risco de cada contrato. Para contratos classificados como de alto risco, o painel apresenta os principais atributos que contribuíram para a classificação, facilitando a compreensão das causas subjacentes e a formulação de estratégias de auditoria direcionadas.

O framework também inclui funcionalidades para a exportação de relatórios e a integração com sistemas internos da organização, como ferramentas de planejamento e monitoramento de auditorias. Os relatórios gerados incluem um resumo dos contratos priorizados, as justificativas baseadas nos valores de SHAP e uma análise de tendências de risco, permitindo que os resultados sejam compartilhados de forma eficiente com gestores e partes interessadas.

Finalmente, a implementação deste framework foi validada por meio de uma prova de conceito com dados reais de uma empresa pública federal de tecnologia, conforme descrito nas etapas anteriores. A validação demonstrou que o framework não apenas melhora a eficiência na seleção de contratos para auditoria, mas também aumenta a transparência e a confiança no processo ao fornecer justificativas claras e fundamentadas para as decisões tomadas.

Com a integração de análise preditiva, explicabilidade e recursos interativos, o framework de tomada de decisão desenvolvido nesta pesquisa representa uma ferramenta robusta para a gestão de riscos em auditorias governamentais, alinhando-se às melhores práticas de governança e transparência.

Capítulo 4

Resultados e Discussão

Este capítulo apresenta os principais resultados obtidos a partir da aplicação da metodologia abordada no tópico anterior, detalhando o desempenho dos modelos, os critérios estatísticos de comparação, as técnicas de explicabilidade e, por fim, as aplicações práticas em auditoria pública.

Com a aplicação da metodologia desenvolvida, espera-se alcançar um aumento na eficácia dos auditores em relação à identificação e priorização de contratos com maior perfil de risco para avaliação nos trabalhos de auditoria governamental. Por conseguinte, espera-se uma alocação mais eficaz dos recursos humanos e financeiros das equipes de auditoria.

4.1 Desempenho dos Modelos

A Tabela 4.1 apresenta os resultados quantitativos obtidos nos experimentos. A avaliação do desempenho dos classificadores fundamentou-se em um conjunto abrangente de métricas, incluindo acurácia, AUC (*Area Under the Curve*), precisão, revocação (recall), F1-score, F2-score, AUCPR (*Area Under the Precision-Recall Curve*) e *Weighted LogLoss* (LogLoss Ponderado). Esses indicadores foram utilizados como métricas de custo para avaliar a calibração das probabilidades preditas, penalizando erros na classe minoritária (fraude) com maior severidade, proporcionalmente ao desbalanceamento das classes.

No contexto da detecção de irregularidades em contratações públicas, a métrica F2-score foi priorizada para a seleção e ranqueamento dos modelos, dado que essa medida atribui um peso superior à revocação (*Recall*) em detrimento da precisão. Essa abordagem metodológica alinha-se à necessidade de minimizar a ocorrência de falsos negativos, visto que a não detecção de um contrato irregular acarreta custos institucionais superiores aos custos operacionais de auditoria de um falso positivo.

Para assegurar a reprodutibilidade e a organização dos experimentos, a nomenclatura dos modelos adota as seguintes siglas: LR (*Logistic Regression*), SVM (*Support Vector Machine*), RF (*Random Forest*), XGB (XGBoost), LGBM (LightGBM), KNN (K-Nearest Neighbors) e TabPFN (Tabular Prior-data Fitted Network). As estratégias de reamostragem são denotadas pelos sufixos: SMOTE, Borderline SMOTE, D-SMOTE, Tomek Links, ADASYN, Oversampling, Undersampling e *Imbalance* (dados originais). Os métodos de otimização de hiperparâmetros são identificados pelos prefixos GS (Grid Search) e RS (Random Search). O modelo proposto, denominado Hybrid Ensemble, consiste na combinação por soft voting dos classificadores base que apresentaram melhor desempenho individual em diferentes distribuições de dados.

De acordo com os resultados apresentados, os modelos que atingiram os maiores valores de F2-score foram o TabPFN com Undersampling (F2 = 0.635), o *Hybrid Ensemble* (F2 = 0.632), *XGBoost com Grid Search e Undersampling* (F2 = 0.629) e LighGBM com Grid Search e Imbalance (F2 = 0.627). Esses modelos também apresentaram desempenho competitivo em métricas complementares, como AUC e Recall, evidenciando sua robustez na identificação de casos positivos, ao mesmo tempo em que penalizam os falsos negativos.

Tabela 4.1: Detalhamento da Performance dos Modelos

Model	AUC ↑	AUCPR ↑	Precision ↑	Recall ↑	F1 ↑	F2 ↑	Weight LogLoss ↓
TabPFN Undersampling	0.836	0.503	0.365	0.778	0.497	0.634	0.843
Hybrid Ensemble	0.836	0.495	0.369	0.767	0.498	0.631	0.846
XGB GS Undersampling	0.833	0.501	0.375	0.757	0.501	0.629	0.857
LGBM GS Imbalance	0.831	0.494	0.377	0.752	0.502	0.627	0.864
LGBM GS Undersamp.	0.832	0.498	0.365	0.739	0.488	0.613	0.861
LGBM RS Undersamp.	0.829	0.495	0.366	0.752	0.493	0.621	0.862
LR RS Tomek	0.829	0.479	0.361	0.762	0.490	0.624	0.863
LR GS Tomek	0.829	0.478	0.361	0.762	0.490	0.624	0.863
LGBM GS Tomek	0.831	0.492	0.368	0.744	0.493	0.618	0.863
LR RS Oversampling	0.829	0.482	0.366	0.752	0.492	0.621	0.864
LR GS Oversampling	0.828	0.483	0.366	0.752	0.493	0.621	0.864
LR RS Undersampling	0.829	0.481	0.353	0.767	0.484	0.621	0.864
LGBM RS Tomek	0.827	0.497	0.366	0.741	0.490	0.615	0.864
LR RS Imbalance	0.828	0.479	0.364	0.760	0.492	0.624	0.864
LR CV Oversampling	0.829	0.482	0.366	0.754	0.492	0.622	0.864
LR GS imbalance	0.829	0.479	0.362	0.760	0.490	0.623	0.865
LGBM RS Imbalance	0.827	0.501	0.378	0.744	0.501	0.623	0.865
LR CV Undersampling	0.828	0.481	0.353	0.767	0.484	0.621	0.865
LR GS Undersampling	0.828	0.481	0.353	0.767	0.484	0.621	0.865
LR SMOTE	0.826	0.477	0.350	0.781	0.483	0.626	0.868
LR CV Border. SMOTE	0.825	0.470	0.342	0.770	0.473	0.615	0.869

Continua na próxima página

Continuação da Tabela 4.1

Model	AUC ↑	AUCPR ↑	Precision ↑	Recall ↑	F1 ↑	F2 ↑	Weight LogLoss ↓
LR RS SMOTE	0.826	0.476	0.351	0.778	0.483	0.625	0.869
LR GS SMOTE	0.825	0.477	0.350	0.778	0.483	0.625	0.870
LR RS Border. SMOTE	0.825	0.471	0.343	0.767	0.474	0.615	0.870
LR GS Borderline SMOTE	0.825	0.471	0.342	0.765	0.472	0.613	0.870
TabPFN RF Undersamp.	0.823	0.473	0.349	0.768	0.479	0.619	0.870
SVM Oversampling	0.826	0.467	0.354	0.757	0.482	0.616	0.875
SVM GS Undersampling	0.828	0.478	0.359	0.728	0.481	0.604	0.877
LR CV ADASYN	0.827	0.479	0.334	0.789	0.469	0.620	0.878
SVM SMOTE	0.821	0.468	0.340	0.757	0.469	0.608	0.878
SVM ADASYN	0.827	0.467	0.349	0.768	0.479	0.618	0.878
SVM RS Undersampling	0.825	0.480	0.362	0.749	0.488	0.617	0.879
LR RS ADASYN	0.826	0.479	0.331	0.783	0.465	0.615	0.879
LR GS ADASYN	0.826	0.479	0.332	0.783	0.466	0.615	0.879
RF RS Oversampling	0.823	0.490	0.379	0.694	0.490	0.595	0.879
SVM Border. SMOTE	0.819	0.467	0.317	0.783	0.451	0.605	0.882
SVM Undersampling	0.820	0.466	0.351	0.741	0.476	0.606	0.887
LGBM Undersampling	0.814	0.475	0.348	0.746	0.475	0.607	0.892
RF GS Undersampling	0.826	0.491	0.367	0.760	0.495	0.625	0.899
RF RS Undersampling	0.827	0.497	0.360	0.754	0.487	0.618	0.900
LGBM imbalance	0.810	0.479	0.359	0.731	0.481	0.605	0.902
XGB RS Tomek	0.830	0.489	0.446	0.588	0.507	0.552	0.907
LGBM Oversampling	0.808	0.468	0.361	0.699	0.476	0.589	0.909
RF GS Oversampling	0.824	0.483	0.376	0.720	0.494	0.608	0.909
XGB RS Imbalance	0.832	0.499	0.460	0.596	0.518	0.562	0.914
RF GS Borderline SMOTE	0.821	0.468	0.406	0.657	0.502	0.584	0.920
RF GS SMOTE	0.821	0.476	0.412	0.601	0.489	0.550	0.921
RF GS ADASYN	0.821	0.474	0.403	0.588	0.478	0.538	0.933
RF RS SMOTE	0.824	0.477	0.439	0.525	0.478	0.505	0.937
RF RS ADASYN	0.821	0.475	0.445	0.501	0.471	0.489	0.943
XGB Undersampling	0.800	0.459	0.345	0.738	0.470	0.601	0.947
XGB RS SMOTE	0.784	0.447	0.327	0.654	0.436	0.545	0.982
XGB RS Borderline SMOTE	0.783	0.446	0.339	0.646	0.444	0.547	0.988
XGB RS ADASYN	0.778	0.441	0.318	0.659	0.429	0.543	0.989
XGB RS D-SMOTE	0.826	0.498	0.460	0.596	0.519	0.562	0.994
XGB Oversampling	0.780	0.444	0.362	0.644	0.463	0.557	1.009
RF CV Undersampling	0.793	0.405	0.321	0.717	0.443	0.575	1.060
TabPFN ADASYN	0.817	0.478	0.523	0.395	0.450	0.416	1.080

Continua na próxima página

Continuação da Tabela 4.1

Model	AUC ↑	AUCPR ↑	Precision ↑	Recall ↑	F1 ↑	F2 ↑	Weight LogLoss ↓
LGBM Borderline SMOTE	0.801	0.458	0.473	0.432	0.452	0.440	1.081
TabPFN SMOTE	0.821	0.476	0.456	0.390	0.420	0.402	1.085
SVM GS Oversampling	0.758	0.378	0.334	0.559	0.418	0.492	1.088
TabPFN Borderline SMOTE	0.822	0.474	0.460	0.406	0.431	0.415	1.091
LGBM ADASYN	0.797	0.458	0.497	0.393	0.438	0.410	1.091
LGBM SMOTE	0.797	0.462	0.489	0.390	0.433	0.406	1.093
TabPFN Oversampling	0.773	0.395	0.366	454	0.405	0.433	1.105
LGBM RS ADASYN	0.792	0.453	0.478	0.358	0.409	0.377	1.121
LGBM GS ADASYN	0.782	0.448	0.494	0.398	0.441	0.414	1.156
LGBM GS SMOTE	0.783	0.440	0.497	0.395	0.440	0.412	1.157
LGBM GS Borderline SMOTE	0.784	0.444	0.469	0.422	0.444	0.430	1.161
XGB GS ADASYN	0.786	0.452	0.512	0.369	0.429	0.391	1.164
XGB GS Borderline SMOTE	0.793	0.455	0.491	0.398	0.439	0.413	1.165
XGB SMOTE	0.787	0.448	0.485	0.385	0.428	0.401	1.168
XGB ADASYN	0.786	0.455	0.526	0.356	0.424	0.380	1.169
XGB Borderline SMOTE	0.791	0.455	0.509	0.406	0.451	0.423	1.177
LGBM RS Borderline SMOTE	0.786	0.439	0.477	0.398	0.433	0.411	1.184
LGBM RS SMOTE	0.783	0.439	0.494	0.382	0.430	0.400	1.187
XGB GS SMOTE	0.787	0.446	0.468	0.366	0.411	0.383	1.193
SVM RS Oversampling	0.728	0.332	0.323	0.509	0.395	0.456	1.235
XGB GS Tomek	0.834	0.494	0.577	0.219	0.316	0.249	1.236
LR CV Tomek	0.829	0.484	0.577	0.208	0.304	0.238	1.242
SVM RS Tomek	0.827	0.473	0.359	0.754	0.486	0.618	1.251
TabPFN Tomek	0.838	0.510	0.622	0.237	0.342	0.270	1.254
LGBM Tomek	0.822	0.489	0.611	0.263	0.367	0.297	1.258
SVM Tomek	0.824	0.465	0.362	0.760	0.490	0.622	1.259
XGB GS imbalance	0.834	0.507	0.643	0.169	0.266	0.197	1.264
SVM GS Tomek	0.821	0.474	0.350	0.760	0.479	0.615	1.265
LR CV Imbalance	0.829	0.482	0.599	0.179	0.274	0.208	1.271
SVM RS Imbalance	0.826	0.473	0.361	0.757	0.488	0.620	1.278
RF RS Borderline SMOTE	0.802	0.419	0.430	0.398	0.413	0.404	1.281
TabPFN imbalance	0.838	0.510	0.667	0.19	0.293	0.221	1.284
SVM GS Imbalance	0.821	0.476	0.349	0.762	0.479	0.616	1.288
RF RS Tomek	0.827	0.498	0.777	0.126	0.216	0.151	1.289

Continua na próxima página

Continuação da Tabela 4.1

Model	AUC ↑	AUCPR ↑	Precision ↑	Recall ↑	F1 ↑	F2 ↑	Weight LogLoss ↓
SVM imbalance	0.823	0.463	0.362	0.757	0.490	0.621	1.290
RF GS Tomek	0.824	0.495	0.748	0.113	0.195	0.136	1.291
XGB Tomek	0.809	0.471	0.568	0.263	0.359	0.295	1.305
LGBM GS D-SMOTE	0.829	0.492	0.615	0.182	0.280	0.211	1.317
LGBM RS D-SMOTE	0.830	0.494	0.590	0.211	0.310	0.242	1.318
RF GS Imbalance	0.826	0.498	0.81	0.097	0.172	0.118	1.319
RF RS Imbalance	0.825	0.498	0.774	0.092	0.163	0.111	1.319
LGBM RS Oversampling	0.745	0.384	0.343	0.512	0.410	0.465	1.333
LR CV D-SMOTE	0.767	0.412	0.445	0.414	0.428	0.419	1.336
LR GS D-SMOTE	0.768	0.410	0.444	0.422	0.432	0.426	1.339
XGB imbalance	0.809	0.469	0.592	0.242	0.343	0.274	1.340
SVM RS SMOTE	0.718	0.354	0.313	0.477	0.378	0.432	1.344
SVM RS ADASYN	0.713	0.354	0.313	0.517	0.390	0.457	1.348
XGB GS D-SMOTE	0.829	0.499	0.623	0.224	0.329	0.257	1.350
XGB RS Undersampling	0.823	0.482	0.187	1	0.315	0.534	1.351
LR RS D-SMOTE	0.767	0.409	0.447	0.422	0.434	0.426	1.351
SVM D-SMOTE	0.763	0.410	0.465	0.337	0.390	0.356	1.357
RF GS D-SMOTE	0.820	0.496	0.803	0.081	0.146	0.099	1.365
TabPFN RF Tomek	0.818	0.473	0.566	0.253	0.349	0.284	1.370
SVM RS Borderline SMOTE	0.715	0.351	0.313	0.496	0.383	0.443	1.374
TabPFN RF imbalance	0.811	0.433	0.481	0.184	0.266	0.210	1.380
RF RS D-SMOTE	0.820	0.489	0.750	0.071	0.129	0.086	1.383
TabPFN DSMOTE	0.832	0.509	0.657	0.187	0.290	0.218	1.390
LGBM GS Oversampling	0.736	0.381	0.355	528	0.424	0.480	1.407
LGBM D-SMOTE	0.821	0.480	0.608	0.226	0.329	0.259	1.409
XGB GS Oversampling	0.732	0.398	0.238	0.744	0.361	0.522	1.410
XGB D-SMOTE	0.819	0.466	0.593	0.248	0.349	0.280	1.425
SVM GS SMOTE	0.702	0.338	0.309	0.422	0.356	0.393	1.432
XGB RS Oversampling	0.721	0.381	0.241	0.715	0.361	0.513	1.440
SVM GS ADASYN	0.695	0.331	0.299	438	0.355	0.400	1.441
TabPFN RF Borderline SMOTE	0.762	0.360	0.362	0.356	0.359	0.357	1.442
TabPFN RF SMOTE	0.756	0.407	0.389	0.366	0.377	0.371	1.442
SVM GS D-SMOTE	0.748	0.391	0.423	0.237	0.303	0.260	1.452
SVM GS Border. SMOTE	0.695	0.317	0.296	0.430	0.350	0.394	1.462
SVM RS D-SMOTE	0.749	0.366	0.402	0.163	0.232	0.185	1.484
TabPFN RF D-SMOTE	0.813	0.459	0.528	0.211	0.300	0.239	1.538
TabPFN RF ADASYN	0.741	0.366	0.372	0.324	0.346	0.332	1.581

Continua na próxima página

Continuação da Tabela 4.1

Model	AUC ↑	AUCPR ↑	Precision ↑	Recall ↑	F1 ↑	F2 ↑	Weight LogLoss ↓
TabPFN RF Oversamp.	0.711	0.306	0.301	0.303	0.302	0.302	1.636
RF CV Border. SMOTE	0.785	0.382	0.405	0.377	0.391	0.382	1.997
RF CV Imbalance	0.758	0.384	0.364	0.361	0.362	0.361	2.000
RF CV SMOTE	0.780	0.393	0.383	0.366	0.374	0.369	2.004
RF CV ADASYN	0.776	0.386	0.399	0.353	0.374	0.361	2.006
RF CV Oversampling	0.762	0.343	0.343	0.485	0.401	0.448	2.107
RF CV Tomek	0.756	0.389	0.372	0.401	0.385	0.394	2.133
KNN GS Undersampling	0.783	0.392	0.342	0.601	0.436	0.522	2.317
KNN RS Undersampling	0.783	0.392	0.342	0.601	0.436	0.522	2.317
Total Resultado	0.785	0.423	0.407	0.523	0.402	0.453	2.526
NB GS Oversampling	0.764	0.350	0.319	0.686	0.435	0.557	3.098
NB RS Oversampling	0.764	0.350	0.319	0.686	0.435	0.557	3.098
NB Oversampling	0.764	0.350	0.319	0.686	0.435	0.557	3.098
RF CV D-SMOTE	0.803	0.432	0.448	0.235	0.307	0.259	3.330
NB GS Imbalance	0.759	0.349	0.366	0.612	0.457	0.538	3.451
NB RS Imbalance	0.759	0.349	0.366	0.612	0.457	0.538	3.451
NB imbalance	0.759	0.349	0.366	0.612	0.457	0.538	3.451
NB RS Undersampling	0.755	0.346	0.330	0.686	0.445	0.564	3.457
NB GS Undersampling	0.755	0.346	0.330	0.686	0.445	0.564	3.458
NB Undersampling	0.755	0.346	0.330	0.686	0.445	0.564	3.458
NB GS Tomek	0.759	0.346	0.369	0.620	0.462	0.545	3.560
NB RS Tomek	0.759	0.346	0.369	0.620	0.462	0.545	3.560
NB Tomek	0.759	0.346	0.369	0.620	0.462	0.545	3.561
KNN GS Tomek	0.772	0.396	0.575	0.190	0.284	0.219	4.495
KNN RS Tomek	0.772	0.396	0.575	0.190	0.284	0.219	4.495
KNN GS Imbalance	0.769	0.392	0.554	0.163	0.250	0.189	4.580
KNN RS imbalance	0.769	0.392	0.554	0.163	0.250	0.189	4.580
NB GS ADASYN	0.743	0.368	0.196	0.878	0.320	0.518	4.858
NB RS ADASYN	0.743	0.368	0.196	0.878	0.320	0.518	4.858
NB ADASYN	0.743	0.368	0.196	0.878	0.320	0.518	4.858
NB GS Border. SMOTE	0.743	0.368	0.196	0.870	0.320	0.515	5.016
NB RS Border. SMOTE	0.743	0.368	0.196	0.870	0.320	0.515	5.016
NB Borderline SMOTE	0.743	0.368	0.196	0.870	0.320	0.515	5.016
KNN Undersampling	0.751	0.347	0.348	0.543	0.424	0.488	5.113
NB GS SMOTE	0.747	0.372	0.197	0.878	0.322	0.520	5.131
NB RS SMOTE	0.747	0.372	0.197	0.878	0.322	0.520	5.131
NB SMOTE	0.747	0.372	0.197	0.878	0.322	0.520	5.131
KNN GS Oversampling	0.734	0.322	0.346	0.443	0.388	0.419	7.300
KNN RS Oversampling	0.734	0.322	0.346	0.443	0.388	0.419	7.300
KNN imbalance	0.733	0.331	0.483	0.226	0.308	0.253	7.327
KNN Tomek	0.735	0.337	0.477	0.250	0.326	0.276	7.422

Continua na próxima página

Continuação da Tabela 4.1

Model	AUC ↑	AUCPR ↑	Precision ↑	Recall ↑	F1 ↑	F2 ↑	Weight LogLoss ↓
KNN ADASYN	0.734	0.312	0.344	0.480	0.400	0.444	7.707
KNN Border. SMOTE	0.733	0.314	0.370	0.461	0.410	0.439	7.781
KNN SMOTE	0.732	0.319	0.365	0.477	0.413	0.449	7.888
KNN GS ADASYN	0.737	0.331	0.400	0.387	0.393	0.390	8.032
KNN RS ADASYN	0.737	0.331	0.400	0.387	0.393	0.390	8.032
KNN GS Border. SMOTE	0.737	0.331	0.414	0.377	0.393	0.383	8.121
KNN RS Border. SMOTE	0.737	0.331	0.414	0.377	0.393	0.383	8.121
KNN GS SMOTE	0.735	0.330	0.394	0.369	0.380	0.373	8.207
KNN RS SMOTE	0.735	0.330	0.394	0.369	0.380	0.373	8.207
KNN Oversampling	0.724	0.310	0.356	446	0.395	0.424	8.756
KNN GS D-SMOTE	0.730	0.338	0.450	0.216	0.291	0.241	8.851
KNN RS D-SMOTE	0.730	0.338	0.450	0.216	0.291	0.241	8.851
KNN D-SMOTE	0.684	0.304	0.486	0.205	0.286	0.231	13.42
NB RS D-SMOTE	0.567	0.210	0.181	0.414	0.252	0.329	16.53
NB GS D-SMOTE	0.567	0.210	0.181	0.414	0.252	0.329	16.53
NB D-SMOTE	0.567	0.210	0.181	0.414	0.252	0.329	16.53

Verificou-se a influência das técnicas de rebalanceamento e da otimização de hiperparâmetros no desempenho preditivo. Embora a literatura frequentemente sugira o aumento da representatividade da classe minoritária (via *Oversampling* ou SMOTE)(Chawla et al., 2002a; He & Garcia, 2009b; Phua, Alahakoon, & Lee, 2004), os experimentos realizados indicaram que, para as arquiteturas não lineares de alto desempenho (TabPFN e *Gradient Boosting*), a técnica de *Undersampling* apresentou resultados superiores na métrica de *Weighted LogLoss*. Essa observação sugere que, neste domínio específico de fraude em contratações, a redução do ruído presente na classe majoritária contribui de maneira mais eficaz para a definição das fronteiras de decisão do que a geração de dados sintéticos.

Destaca-se o desempenho do modelo baseado em *Deep Learning* para dados tabulares (TabPFN) e da abordagem de *Hybrid* Ensemble, que superaram estatisticamente os modelos tradicionais baseados em árvores (*XGBoost e Random Forest*) isolados. A métrica AUC apresentou correlação positiva com o F2-score, corroborando a consistência dos classificadores selecionados. Esses resultados indicam que a seleção da arquitetura do modelo, combinada com a estratégia de amostragem adequada, impacta a capacidade do sistema de auditoria. A priorização de modelos com alto *Recall* e baixo *Weighted LogLoss*, mesmo com variações na precisão, alinha-se às estratégias de mitigação de risco, reduzindo

a probabilidade de omissão de contratos irregulares. Tal configuração é coerente com as práticas de sistemas de alerta no setor público, onde a sensibilidade deve ser privilegiada para assegurar a cobertura dos casos relevantes.

Dessa forma, o estudo evidencia a relevância da integração entre técnicas de pré-processamento, otimização de hiperparâmetros e construção de ensembles híbridos. A abordagem proposta, validada pelo teste estatístico MCS, proporciona maior reforço metodológico para o desenvolvimento de sistemas de apoio à decisão orientados ao risco, com probabilidades calibradas para auxiliar a priorização de auditorias.

4.1.1 Comparação e Seleção de Modelos

A determinação do modelo ótimo para a implantação em ambiente produtivo demanda uma análise que transcenda as estimativas pontuais de desempenho apresentadas na Tabela 4.1. É necessário avaliar a estabilidade das predições e a significância estatística das diferenças observadas entre os classificadores, especialmente considerando o severo desbalanceamento dos dados.

Para formalizar a comparação e mitigar o risco de seleção baseada no acaso, aplicou-se o procedimento *Model Confidence Set* (MCS), proposto por Hansen et al. (2011). O MCS consiste em um teste de hipóteses sequencial que elimina modelos estatisticamente inferiores até que reste um subconjunto, denotado por M^*, que contém o(s) melhor(es) modelo(s) com um nível de confiança $(1 - \alpha)$.

Neste estudo, o teste foi configurado com um nível de significância $\alpha = 0.05$ e 10.000 replicações de *bootstrap*. Diferentemente de abordagens tradicionais que utilizam erro quadrático ou acurácia em trabalhos de regressão, a função de perda adotada para o MCS foi a *Weighted LogLoss* (LogLoss Ponderado). Essa métrica penaliza os erros de classificação na classe minoritária (fraude) com um peso proporcional ao desbalanceamento da base de dados (≈ 5.37), forçando o teste estatístico a priorizar modelos que apresentem probabilidades bem calibradas para os casos de interesse.

O procedimento MCS resultou na identificação do Conjunto Superior de Modelos ($M^*_{95\%}$), cujos detalhes são apresentados na Tabela 4.2. A tabela está ordenada pelo p-valor do MCS e, secundariamente, pelo menor erro ponderado médio.

A análise da Tabela 4.2 evidencia a robustez do modelo Hybrid Ensemble, que obteve um p-valor máximo ($p = 1.000$), consolidando-se como o modelo de referência na minimização do risco ponderado. Observa-se que, embora outros modelos, como o `TabPFN Undersampling` e o *Gradient Boosting* (*XGBoost* e *LightGBM*), permaneçam no conjunto de confiança, estes apresentam um p-valor limítrofe ($p = 0.0606$), próximo ao nível de corte de 0,05. Isso indica que o Ensemble apresenta, estatisticamente, uma consistência superior na redução da função de perda ponderada.

Tabela 4.2: Resultados do Model Confidence Set (MCS) — Modelos no Conjunto de Confiança 95%

Modelo	**Incluído MCS**	**Pvalor MCS**	**AUC**	**F2**	**Weighted LogLoss**
Hybrid Ensemble	Sim	1.000	0.836	0.631	0.846
TabPFN Undersampling	Sim	0.061	0.836	0.634	0.843
XGB GS Undersampling	Sim	0.061	0.833	0.629	0.857
LGBM GS Imbalance	Sim	0.061	0.831	0.627	0.864
LR GS Tomek	Sim	0,061	0.829	0.624	0.863
LR RS Imbalance	Sim	0.061	0.829	0.624	0.864
LR RS Tomek	Sim	0.061	0.829	0.624	0.863
LGBM RS Imbalance	Sim	0.061	0.828	0.623	0.865
LR CV Oversampling	Sim	0.061	0.829	0.622	0.864
LGBM RS Undersampling	Sim	0.061	0.829	0.621	0.862
LR GS Oversampling	Sim	0.061	0.829	0.621	0.864
LR RS Oversampling	Sim	0.061	0.829	0.621	0.864
LGBM GS Tomek	Sim	0.061	0.831	0.618	0.863
TabPFN RF Undersampling	Sim	0.061	0.823	0.616	0.870
LGBM RS Tomek	Sim	0.061	0.827	0.615	0.864
LGBM GS Undersampling	Sim	0.061	0.832	0.613	0.861
LR GS SMOTE	Sim	0.056	0.825	0.625	0.870
LR RS Borderline SMOTE	Sim	0.056	0.825	0.615	0.870
LR GS Borderline SMOTE	Sim	0.056	0.825	0.613	0.870
LR RS SMOTE	Sim	0.055	0.826	0.625	0.869

Fonte: Elaborado pelo autor.

Dois pontos merecem destaque na interpretação desses resultados. O desempenho superior do Hybrid Ensemble valida a hipótese de que a combinação de arquiteturas distintas (Transformers e Árvores), treinadas sob diferentes distribuições de dados (balanceadas e desbalanceadas), resulta em uma melhor calibração das probabilidades de risco. O Ensemble consegue mitigar a confiança excessiva (*overconfidence*) observada em modelos individuais, resultando em um *Weighted LogLoss* competitivo.

É interessante observar a presença de modelos de Regressão Logística (`LR`) no conjunto de confiança, que superam ou igualam modelos mais complexos em termos estatísticos de perda, devido à natureza da métrica *LogLoss* (Tibshirani, 1996a). Os modelos lineares tendem a gerar probabilidades bem calibradas, sendo penalizados de forma menos severa pela função de perda ponderada em comparação com modelos de árvores, que podem apresentar alta variância nas probabilidades extremas.

Diante do exposto, a seleção do modelo final considera o equilíbrio entre desempenho estatístico e complexidade operacional. O *Hybrid Ensemble* é recomendado para aplicações em que a precisão da estimativa de risco é crítica, servindo como base para sistemas de *scoring*. No entanto, para cenários com severas restrições computacionais, o *TabPFN Undersampling* permanece como uma alternativa viável de modelo único, visto que não foi estatisticamente rejeitado e apresenta o maior F2-Score (0.6391) entre os modelos

analisados, garantindo alta sensibilidade na detecção de fraudes.

4.1.2 Explicabilidade do Modelo (XAI)

Após a seleção de um modelo com desempenho estatístico robusto na Seção anterior, a etapa subsequente consiste em mover a análise de uma avaliação de desempenho (caixa-preta) para a interpretação dos mecanismos de decisão do modelo. A explicabilidade, portanto, é um componente metodológico determinante para a governança, como nos trabalhos de auditoria, em que a justificativa para uma predição é tão relevante quanto a própria predição (Ribeiro, Singh, & Guestrin, 2016).

Conforme delineado na metodologia (Seção 3.5), esta análise emprega o framework SHAP (*SHapley Additive exPlanations*) (Lundberg & Lee, 2017), que se baseia em valores da teoria dos jogos para atribuir a contribuição de cada atributo (*feature*) à saída do modelo. A análise foi conduzida sobre o modelo Ensemble, utilizando o método "shap.KernelExplainer".

Primeiramente, a análise de explicabilidade global identifica os atributos que possuem o maior impacto médio na magnitude da predição de risco, considerando todo o conjunto de dados. A Figura 4.1 apresenta o gráfico de resumo (summary plot) dos valores SHAP para a classe de alto risco (Risco = 1).

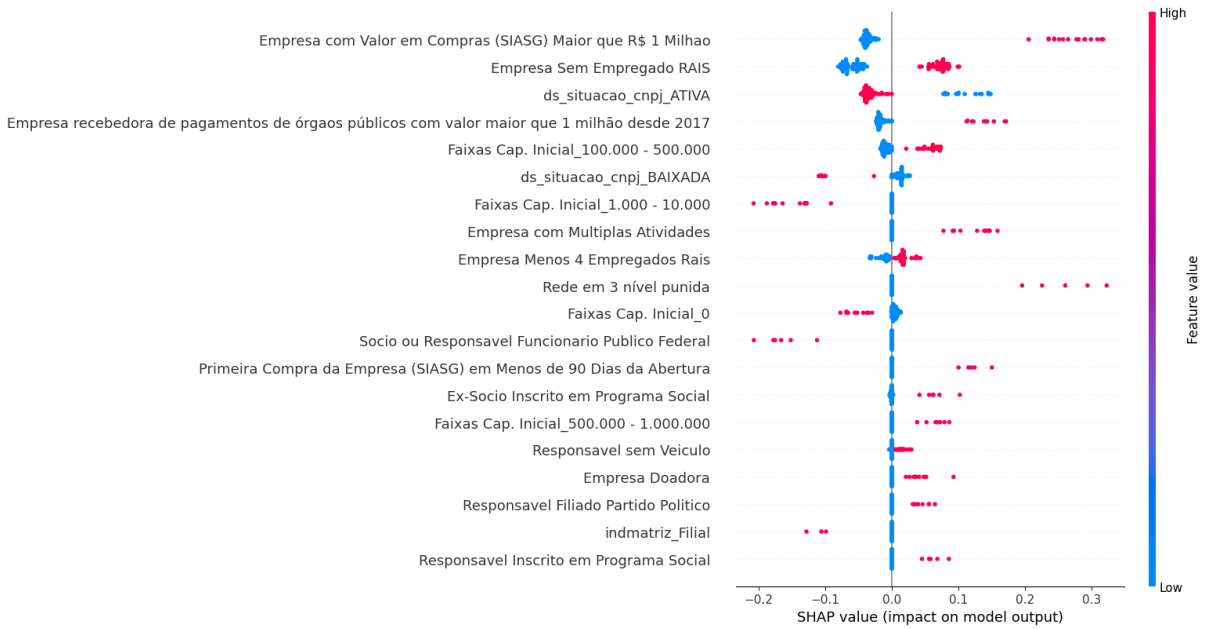

Figura 4.1: Importância Global dos Atributos (SHAP Summary Plot)
Fonte: Elaborado pelo autor.

A análise da Figura 4.1 revela os preditores de risco mais relevantes, destacando 'Empresa com Valor em Compras (SIASG) Maior que R$ 1 Milhão', 'Empresa sem Empregado

RAIS' e "Rede em 3 níveis punida'. O gráfico demonstra que valores elevados de 'Empresa com Valor em Compras (SIASG) Maior que R$ 1 Milhão' (pontos vermelhos à direita do eixo zero) estão fortemente associados a um aumento no escore de risco. Similarmente, o atributo "Empresa Sem Empregado RAIS", quando presente, desloca a predição para a direita (maior risco). Inversamente, empresas que possuem empregados (pontos azuis) tendem a ter o risco reduzido, o que demonstra a capacidade do modelo de detectar empresas sem estrutura operacional formal (possíveis empresas de fachada). Esta visualização valida a intuição de auditoria e confirma quais indicadores de risco, derivados das bases de dados públicas, são mais determinantes para o modelo.

O modelo também demonstrou sensibilidade em relação à regularidade cadastral junto aos órgãos de controle. O atributo "ds_situacao_cnpj_ATIVA"evidencia que empresas sem situação cadastral ativa (representadas pelos pontos azuis deslocados à direita) estão fortemente associadas a um maior risco de irregularidade.

No que se refere à capacidade econômica, representada pelas faixas de capital social, observa-se uma relação não linear. Faixas de capital intermediárias (de R$ 100.000 a R$ 500.000) tendem a contribuir positivamente para o risco, enquanto faixas de capital muito reduzido (de R$ 1.000 a R$ 10.000) apresentam uma contribuição negativa, resultando em uma redução do risco. Esse comportamento pode indicar que fraudes mais complexas e de maior valor financeiro, capturadas pelo modelo, exigem um capital social mínimo para habilitação em processos licitatórios, excluindo microempresas com capacidade de capital extremamente baixa.

Embora com menor frequência no conjunto de dados, os atributos relacionais funcionam como determinantes de alto impacto quando estão presentes. As variáveis Rede em 3º nível punida' apresentam um comportamento de "gatilho": a ocorrência desses fatores (pontos vermelhos) resulta em um incremento substancial na probabilidade de fraude.

Em segunda análise, enquanto a análise global aborda de forma ampla as características mais relevantes, a análise local é a ferramenta para a execução da auditoria, permitindo a interpretabilidade da predição de um contrato ou fornecedor específico. A Figura 4.2 ilustra um *waterfall plot* para uma instância de alto risco selecionada do conjunto de teste.

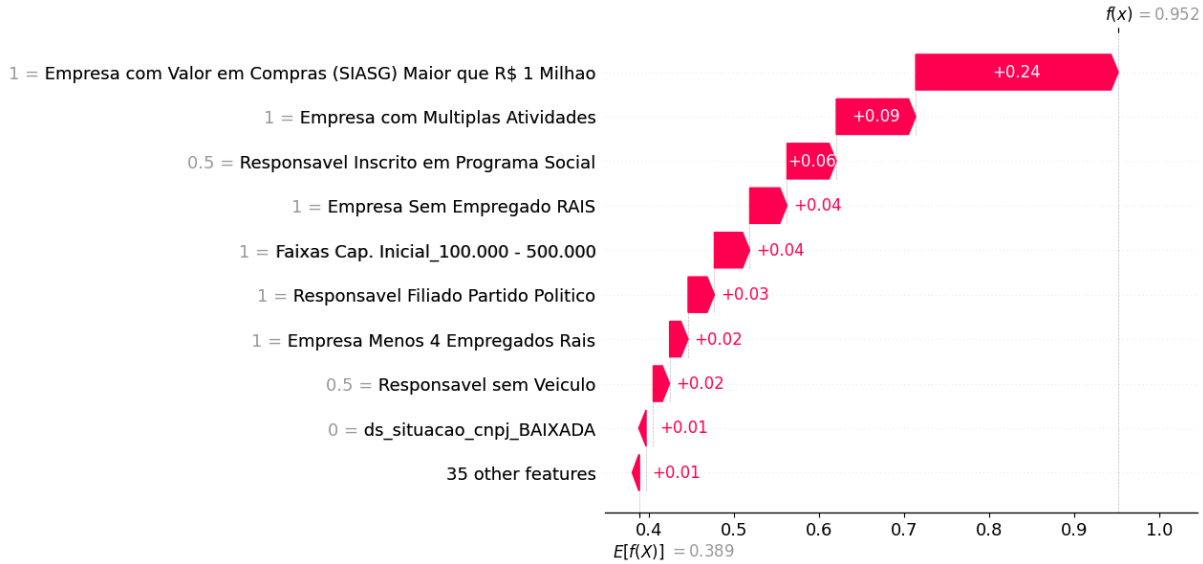

Figura 4.2: Explicação Local (SHAP Waterfall) para uma Instância de Alto Risco (Probabilidade predita $> 90\%$).

Fonte: Elaborado pelo autor.

O *waterfall plot* detalha o balanço de forças para esta predição específica. O valor base ($E[f(X)]$) representa a média do risco predito para o conjunto de dados. Fatores destacados em vermelho, como "Compras no SIASG Maior que 1 Milhão", "Responsável inscrito em Programa Social"e "Empresa com Menos de 4 Empregados na RAIS", entre outros, ampliam a predição para um escore mais elevado (alto risco). Fatores destacados em azul, como "Sócio ou Responsável Funcionário Público Federal = 1"e um baixo capital social, entre outros, atuam como mitigadores,gerando a predição para um escore menor (baixo risco). Esta decomposição fornece ao auditor uma justificativa clara e acionável, indicando exatamente quais atributos daquele fornecedor devem ser investigados.

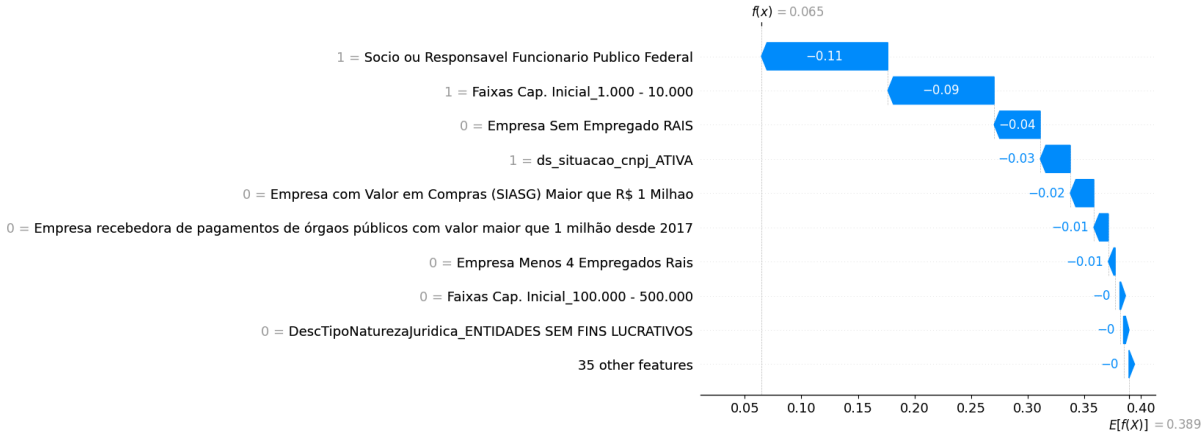

Figura 4.3: Explicação Local (SHAP Force Plot) para uma Instância de Alto Risco (Probabilidade predita $< 20\%$).

Fonte: Elaborado pelo autor.

A integração dessas duas visões de explicabilidade (global e local) fornece a transparência necessária para a implementação do modelo no *framework* de tomada de decisão (Seção 3.6), permitindo que os auditores confiem nas recomendações do modelo e ajam com base nelas.

4.1.3 Definição das Regras de Decisão

A operacionalização de modelos de aprendizado de máquina em contextos de auditoria requer a tradução de predições probabilísticas em ações de controle objetivas. A saída bruta dos classificadores supervisionados — obtida, por exemplo, por meio do método *predict_proba()* — fornece um *score* de risco contínuo $S \in [0, 1]$, que reflete a probabilidade estimada de que uma contratação apresente irregularidades. Embora informativo, esse valor numérico precisa ser segmentado em categorias discretas para se integrar aos fluxos de trabalho das equipes de fiscalização.

Conforme preconizado na literatura sobre *Continuous Auditing* e *Risk Assessment* (Chan & Vasarhelyi, 2011), a definição dos limiares (*thresholds*) de decisão deve equilibrar a precisão estatística do modelo à capacidade operacional do órgão de controle. Uma abordagem puramente estatística pode gerar um volume de alertas superior à força de trabalho disponível, enquanto uma abordagem estritamente operacional pode negligenciar riscos materiais.

Neste trabalho, adotou-se uma abordagem híbrida para a definição das regras de decisão. A distribuição empírica dos *scores* de risco gerados pelo modelo foi analisada (ver Figura 4.4), permitindo identificar zonas de concentração de probabilidade. Com base nesta análise e em diretrizes de auditoria fundamentadas no risco, estabeleceu-se um *framework* de priorização estratificado em três níveis, detalhado a seguir:

- Risco Muito Alto ($S \geq 0,80$): Esta categoria representa o topo da pirâmide de risco, contendo contratos com indícios contundentes de irregularidade. A densidade de fraudes neste estrato é máxima, indicando uma alta precisão do modelo. Nessa situação, recomenda-se a prioridade crítica para auditoria imediata, preferencialmente por auditores seniores, devido ao elevado potencial de retorno sobre o esforço fiscalizatório.
- Risco Alto ($0,60 \leq S < 0,80$): Contratos que exibem padrões severos de atipicidade e forte correlação com o histórico de fraudes, embora com uma probabilidade marginalmente inferior ao nível crítico. Nessa situação, recomenda-se a inclusão em filas de alta prioridade, devendo ser auditados assim que os casos de risco "Muito Alto"forem tratados.
- Risco Médio ($0,40 \leq S < 0,60$): Caracteriza-se como uma zona de transição e incerteza, onde a probabilidade de irregularidade é intermediária. Os sinais de anomalia presentes podem indicar tanto fraudes complexas quanto contratos regulares atípicos. Nessa situação, a decisão de auditar deve ser estritamente contextual, ponderada por fatores exógenos ao modelo (materialidade, relevância estratégica, denúncias em ouvidoria) e pela disponibilidade de força de trabalho excedente.
- Risco Baixo ($0,20 \leq S < 0,40$): Contratos com uma probabilidade reduzida de irregularidade, apresentando maior conformidade com os padrões lícitos da base de treinamento, embora ainda possuam características limítrofes. Nessa situação, recomenda-se a verificação apenas em casos de auditorias temáticas específicas ou quando houver indícios externos robustos que contradigam o modelo.
- Risco Muito Baixo ($S < 0,20$): Categoria que abrange a ampla maioria dos contratos regulares, nos quais o modelo não identificou padrões de risco significativos. Nessa situação, recomenda-se a aplicação de automação na aprovação ou auditoria por amostragem aleatória simples, visando apenas o controle de qualidade rotineiro e a monitoração.

A Figura 4.4 apresenta a distribuição dos contratos do conjunto de teste classificados pelo modelo *Hybrid Ensemble*, segmentados em uma escala de cinco níveis de risco (Muito Baixo, Baixo, Médio, Alto e Muito Alto). O eixo vertical representa a proporção de cada classe (Regular e Fraude) contida em cada faixa de score, permitindo a avaliação simultânea da cobertura (*Recall*) e da Precisão do modelo em diferentes limiares de decisão.

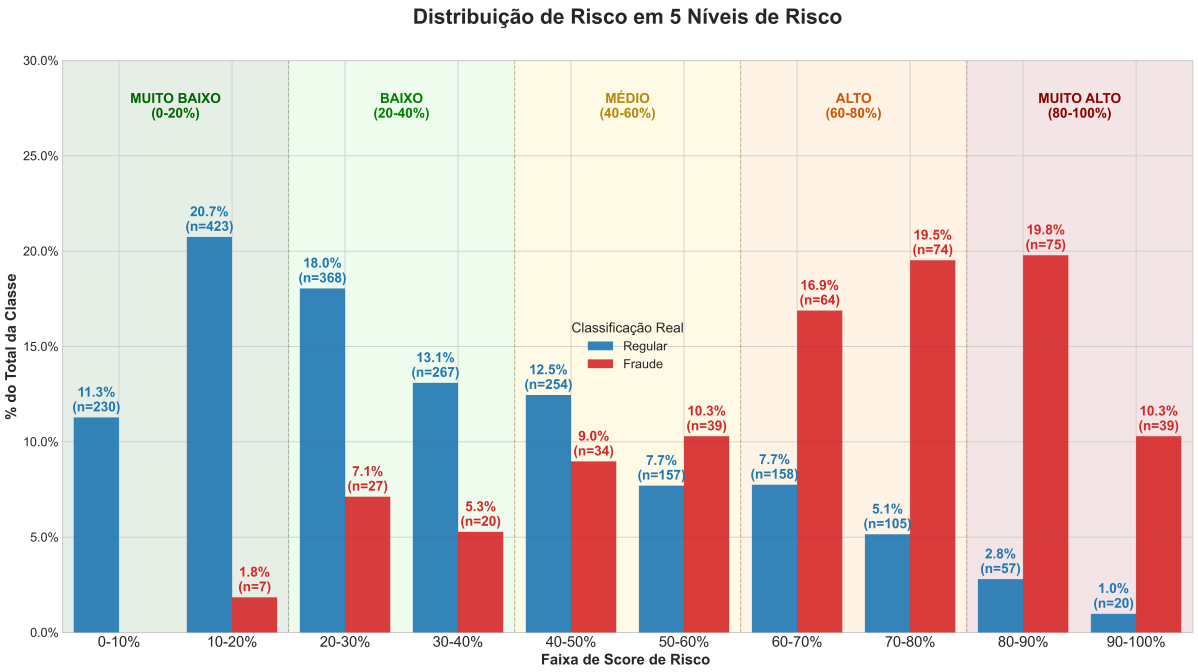

Figura 4.4: Distribuição de Contratos Regulares e Fraudulentos por % da Classe e Faixa de Score de Risco.

A análise dos dados evidencia uma correlação positiva entre a faixa de score atribuída e a assertividade em casos de fraude, validando a calibração e a capacidade do modelo de classificar os contratos conforme a probabilidade de irregularidade.

Nos estratos inferiores, especificamente na faixa de risco Muito Baixo (0-20%), observa-se uma concentração expressiva de contratos regulares (32,0% do total da classe regular), em contraste com a presença residual de fraudes (apenas 1,8% do total da classe fraude, correspondendo a $n = 7$ casos). Esse comportamento sugere que a automação da aprovação ou a aplicação de auditoria amostral mínima neste segmento implica um risco reduzido de falsos negativos, otimizando o fluxo de trabalho ao liberar recursos humanos da análise de contratos com baixa probabilidade de erro.

Na extremidade oposta, as faixas de risco Alto (60-80%) e Muito Alto (80-100%) demonstram o enriquecimento da amostra em relação à classe de interesse. O estrato Muito Alto, isoladamente, captura 30,1% de todas as fraudes confirmadas no conjunto de teste ($n = 75 + n = 39$), enquanto contém uma parcela reduzida de contratos regulares (3,8%). A concentração de irregularidades nessas faixas superiores indica que a priorização fiscalizatória baseada no modelo tende a maximizar a taxa de acerto (precisão) das auditorias, assegurando que o esforço analítico seja direcionado aos casos com maior materialidade de risco.

A faixa de risco Médio (40-60%) caracteriza-se como uma zona de transição e incerteza, onde as distribuições das classes Regular e Fraude apresentam maior sobreposição. Neste

intervalo, a probabilidade de um contrato ser irregular se aproxima da probabilidade de ser regular, o que justifica a necessidade de intervenção humana (análise contextual) para desempatar, visto que os padrões detectados pelo algoritmo não são, por si sós, conclusivos para uma classificação binária assertiva.

Em suma, a estratificação proposta demonstra que o *Hybrid Ensemble* não apenas classifica, mas também ordena o risco de forma coerente. A implementação desta escala de cinco níveis permite à gestão pública modular a intensidade do controle, maximizando a automação nos 40% dos contratos com menor risco e concentrando a expertise dos auditores nos 40% dos contratos com maiores indícios de anomalia.

A discretização do risco facilita a comunicação dos resultados aos gestores e permite a implementação de políticas de governança mais transparentes. Por fim, o modelo proposto pode evoluir, incorporando o feedback das auditorias realizadas (confirmação de fraude ou falso positivo) e recalibrando periodicamente os limiares de decisão, mantendo o sistema adaptado às mudanças nas táticas de fraude e na legislação.

Para viabilizar a utilização prática das regras de decisão definidas, foi desenvolvido um artefato tecnológico de suporte: o "Painel de Governança e Priorização de Auditorias". Essa ferramenta traduz a complexidade dos modelos preditivos e da análise de explicabilidade (SHAP) em uma interface visual intuitiva para auditores.

O painel, ilustrado na Figura 4.5, oferece as seguintes funcionalidades críticas para o processo de auditoria: a segmentação dos contratos por nível de risco, com indicadores visuais de alerta e explicabilidade integrada, apresentando os principais fatores de risco (features) que contribuíram para o score de cada contrato, permitindo ao auditor compreender o "porquê"da classificação (conforme a Seção 4.1.2).

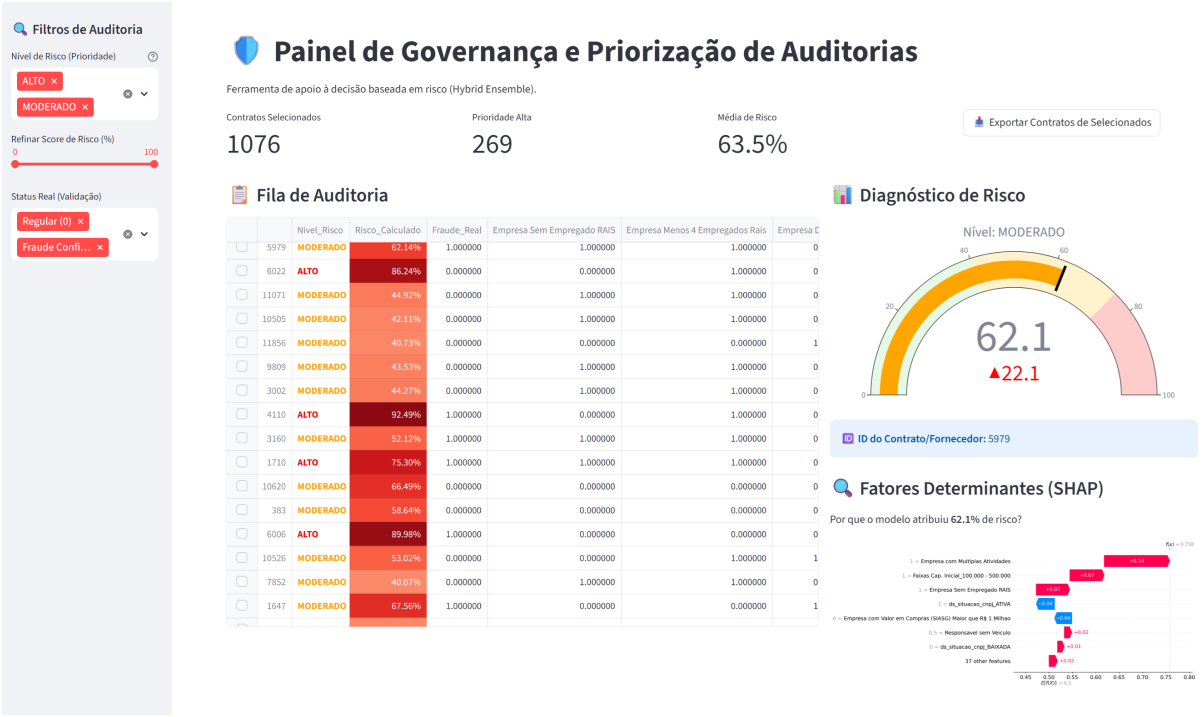

Figura 4.5: Painel de Governança para Priorização de Auditorias Baseada em Risco.

Fonte: Elaborado pelo autor.

Esta camada de visualização conclui o ciclo da pesquisa atual, convertendo a inteligência computacional em valor público tangível por meio da melhoria na alocação dos recursos de fiscalização.

Capítulo 5

Considerações Finais

5.1 Contribuições do Trabalho

Esta pesquisa situa-se na fronteira entre a Ciência de Dados e a Auditoria Governamental (*Audit Analytics*), oferecendo avanços substanciais para a literatura sobre detecção de fraudes e para a prática de controle externo. As contribuições deste estudo podem ser categorizadas em três dimensões principais: inovações metodológicas na modelagem preditiva, rigor estatístico na avaliação de modelos e aplicabilidade institucional de inteligência artificial explicável.

A principal contribuição metodológica reside na avaliação pioneira do modelo *Tabular Prior-Data Fitted Network* (TabPFN) (Hollmann et al., 2023, 2025), utilizado em *Ensemble* com *Xgboost e LightGBM* no domínio de compras públicas. Enquanto a literatura predominante em detecção de fraudes se concentra em métodos de *ensemble* baseados em árvores (como *Random Forest e Gradient Boosting*), este trabalho inclui uma arquitetura baseada em *Transformers* aplicada a dados tabulares. A pesquisa evidencia que a execução do modelo TabPFN em Ensemble oferece desempenho superior ou competitivo em cenários de dados tabulares de alta dimensionalidade e proporciona um tratamento mais eficaz do desbalanceamento de classes, características intrínsecas às bases de dados de fraude.

Outro ponto é que o estudo prospecta um *pipeline* completo e reprodutível para a construção de *scores* de risco em auditoria, sistematizando etapas críticas que frequentemente são tratadas de forma isolada na literatura. O *framework* proposto integra:

- Estratégias avançadas de pré-processamento e engenharia de atributos (*feature engineering*) específicas para o ecossistema de dados governamentais brasileiros;
- Técnicas de reamostragem supervisionada (ex: SMOTE, ADASYN e suas variantes) para mitigar o viés da classe majoritária, utilizando critérios rigorosos para Valida-

ção cruzada estratificada, a fim de garantir a generalização do modelo em dados não vistos;

- Rigor científico na seleção dos melhores modelos por meio do model confidence set.
- Definição de regras de decisão operacionais baseadas em limiares de probabilidade, facilitando a tradução de *outputs* matemáticos em ações de controle.

Este framework serve como um guia de referência para Órgãos de Controle que buscam modernizar suas matrizes de risco com base em evidências empíricas.

Diferentemente das abordagens convencionais que selecionam o "melhor modelo"baseando-se apenas em estimativas pontuais de métricas de desempenho (como a média do F2-Score), este trabalho incorpora o *Model Confidence Set (MCS)*. Ao aplicar testes de hipóteses sequenciais para identificar o subconjunto de modelos estatisticamente indistinguíveis do ótimo, a pesquisa amplia a confiabilidade da seleção. Essa abordagem é crucial em contextos de alta responsabilidade (*high-stakes decision making*), mitigando o risco de adotar um modelo cujo desempenho superior seja fruto de aleatoriedade amostral e não de uma capacidade preditiva real.

Por fim, a pesquisa contribui para a ética e a transparência na adoção de IA no setor público por meio da implementação de técnicas de *Explainable AI* (XAI), especificamente utilizando os valores de Shapley (SHAP). Ao decompor a predição de risco em contribuições marginais de cada atributo, o modelo proposto supera a "caixa-preta", permitindo que auditores compreendam o raciocínio por trás de cada alerta. Isso não apenas aumenta a confiança e a adoção do sistema pelos usuários finais, mas também garante a conformidade com os princípios de *due dilligence*, transparência e a motivação dos atos administrativos, que são essenciais na esfera pública.

Em síntese, este trabalho propõe não apenas uma ferramenta técnica de detecção, mas também estabelece um paradigma metodológico para auditorias mais preditivas, eficientes e transparentes.

5.2 Limitações

Apesar dos avanços proporcionados pela metodologia proposta, este estudo apresenta algumas limitações que devem ser consideradas tanto na interpretação dos resultados quanto nas futuras extensões do trabalho.

Primeiramente, a base de dados utilizada foi composta por informações públicas; no entanto, algumas delas estão disponíveis apenas por meio de acordos de cooperação entre órgãos públicos. Adicionalmente, informações sobre filiações e doações partidárias não estão mais disponíveis publicamente devido à restrição imposta pela Lei 13.709/2018 - Lei

Geral de Proteção de Dados (LGPD). Destaca-se que os atributos relacionados ao contrato não puderam ser realizados no treinamento por não estarem disponíveis. Contudo, para inclusão em produção, essas informações estarão disponíveis.

Outro ponto refere-se à dinâmica dos contratos públicos, em que a instituição que promove a contratação não pode criar diferenciações ou restrições à competição, mesmo que tenha um score de risco mensurado como elevado. Além disso, a evolução das políticas públicas e mudanças na legislação — como a entrada em vigor da nova Lei de Licitações (Lei nº 14.133/2021) — podem impactar a aplicabilidade dos critérios de risco definidos, exigindo revisão periódica do modelo proposto.

Por fim, embora a interpretabilidade tenha sido considerada por meio de técnicas como SHAP, é necessário ter cuidado na extrapolação das explicações, pois isso pode gerar possíveis vieses.

Essas limitações ressaltam a importância de ampliar a base de dados com fontes adicionais e de refinar periodicamente os modelos com base em novos dados e realidades normativas.

5.3 Conclusões

Este trabalho teve como objetivo propor e avaliar uma abordagem baseada em aprendizado de máquina explicável para a priorização de auditorias em contratações públicas, com foco na construção de um framework de classificação de risco fundamentado em dados históricos, métricas estatísticas robustas e interpretabilidade.

A partir de uma base de dados estruturada contendo informações contratuais e indicadores públicos, foram testadas diferentes estratégias de balanceamento de dados, técnicas de seleção de atributos e algoritmos de classificação supervisionada. O modelo TabPFN, utilizado em Ensemble com XGboost e LightGBM, representa um avanço na área de deep learning para dados tabulares. Destaca-se por seu desempenho competitivo, mesmo em contextos com classes desbalanceadas, ampliando o leque de ferramentas disponíveis para auditoria baseada em dados.

A aplicação do Model Confidence Set (MCS) permitiu identificar, com base estatística, os modelos que apresentam desempenho consistentemente alto, contribuindo para a seleção fundamentada de modelos preditivos. Adicionalmente, a utilização de técnicas de explicabilidade, como o SHAP, garantiu a transparência dos critérios de decisão dos modelos, um elemento essencial para a sua adoção institucional em ambientes regulados e públicos.

A proposta culminou na definição de um framework de decisão para a classificação de risco em contratações, incluindo regras operacionais para a categorização por nível de

risco e sua implementação em um Painel de Governança, voltado ao suporte à decisão de auditores e gestores públicos. Tal painel representa um passo concreto na transformação digital do controle público, viabilizando o uso responsável da inteligência artificial em políticas de integridade.

Como implicações práticas, os resultados podem auxiliar instituições de controle a otimizar a alocação de recursos, direcionar auditorias de forma mais eficiente e fortalecer a prevenção de fraudes. Ao mesmo tempo, contribuem para o avanço do campo de auditoria orientada por dados (data-driven audit), integrando inteligência artificial explicável aos princípios de governança e responsabilidade.

Como limitações, destacam-se a dependência da disponibilidade de dados públicos, a ausência de rótulos completos para todas as irregularidades e a necessidade de revalidação contínua dos modelos com novas informações contratuais e normativas.

Para trabalhos futuros, sugere-se: (i) a expansão do framework para novas bases de dados; (ii) o uso de feedback contínuo por parte de auditores para o refinamento dos modelos; (iii) o estudo de abordagens semi-supervisionadas ou com detecção de anomalias para casos não rotulados; (iv) a análise das predições falso-positivas e falso-negativas que apresentaram elevado erro para o refinamento do modelo; e (v) o aprofundamento da integração entre os scores de risco e os fluxos de trabalho nos sistemas de controle interno e auditoria.

Conclui-se que a metodologia proposta é viável e promissora, oferecendo um caminho realista para fortalecer a governança pública, com base em ciência de dados e inteligência artificial explicável.

Referências

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Applied Computing and Intelligence*, *1*(1), 1–20. 69

Anderson, R. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*. Oxford: Oxford University Press. 88

Arista, R., Fazekas, M., & Volkotrub, G. (2024). *Using beneficial ownership data for large-scale risk assessment in public procurement: The example of 6 european countries* (Tech. Rep.). Government Transparency Institute. Retrieved from https://www.govtransparency.eu/wp-content/uploads/2024/07/Arista-Fazekas-Volkotrub_BO-CRI_GTI_WP_2024.pdf (GTI Working Paper) 76

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, *54*, 627–635. 43

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20–29. 48, 49, 51

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(1), 281–305. 62, 63, 64

Boueri, R., Rocha, F., & Rodopoulos, F. (2015). *Avaliação da qualidade do gasto público e mensuração da eficiência*. Ministério da Fazenda. 1, 2

Brasil. (2021). *Lei nº 14.133, de 1º de abril de 2021. Lei de Licitações e Contratos Administrativos.* (https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14133.htm) 15

Brasil. (2023). *Portal de compras do governo federal.* Retrieved 2023, from https://www.gov.br/compras/pt-br (Página inicial) 1

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140. 36

Breiman, L. (1997). *Arcing the edge* (Tech. Rep.). Technical Report 486, Statistics Department, University of California at 29

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32. 27

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. 43

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2012). Dbsmote: Density-based synthetic minority over-sampling technique. In *Applied intelligence* (Vol. 36, p. 664-677). doi: 10.1007/s10489-011-0287-y 53, 54

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, *57*(1), 203–216. 13

Carneiro, D., Veloso, P., Ventura, A., Palumbo, G., & Costa, J. (2020). Network analysis for fraud detection in portuguese public procurement. In *Intelligent data engineering and automated learning–ideal 2020: 21st international conference, guimaraes, portugal, november 4–6, 2020, proceedings, part ii* (p. 390-401). Springer International Publishing. 78

Carpanese, I., Velasco, R. B., Interian, R., Paulo Neto, O. C., & Ribeiro, C. C. (2021). A decision support system for fraud detection in public procurement. *International Transactions in Operational Research*, *28*(1), 27-47. 3, 19

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning* (pp. 161–168). 26

Carvalho, M. (2019). *Manual de direito administrativo* (7th ed.). Salvador: Editora JusPODIVM. 15, 16

CFC. (2009). *Resolução do conselho federal de contabilidade nº 1.207/2009.* Conselho Federal de Contabilidade (CFC). Retrieved 30 mar. 2021, from http://www.cfc.org.br 1

CGU. (2022). *Resultado do monitoramento dos recursos federais repassados a estados e municípios.* Ministério da Transparência, Fiscalização e Controladoria-Geral da União (CGU). Retrieved 15 de jan. de 2023, from https://www.gov.br/cgu/pt-br/coronavirus/cgu-monitora-aplicacao-dos-recursos-federais-repassados-a-estados-e-municipios 1

Chan, D. Y., & Vasarhelyi, M. A. (2011). Innovation and practice of continuous auditing. *International Journal of Accounting Information Systems*, *12*(2), 152–160. doi: 10.1016/j.accinf.2011.01.001 109

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. 59

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning.* MIT Press. 22

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Crisp-dm 1.0. step-by-step data mining guide* (Tech. Rep.). CRISP-DM Consortium. 2

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. 25, 26

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002a). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357. 49, 50, 52, 103

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002b). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357. doi: 10.1613/jair.953 51, 52

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). 30, 31

Covert, I., & Lee, S. I. (2021). Improving kernelshap: Practical shapley value estimation using linear regression. In *Proceedings of the 24th international conference on artificial intelligence and statistics (aistats)* (pp. 3457–3465). 69

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc

curves. In *Proceedings of the 23rd international conference on machine learning (icml 2006)* (pp. 233–240). 46, 50

Decarolis, F., & Giorgiantonio, C. (2022). Corruption red flags in public procurement: new evidence from italian calls for tenders. *EPJ Data Science*, *11*(16). doi: 10.1140/epjds/s13688-022-00325-x 11

de Moraes Souza, J. G., de Castro, D. T., Peng, Y., et al. (2024). A machine learning-based analysis on the causality of financial stress in banking institutions. *Computational Economics*, *64*, 1857–1890. Retrieved from https://doi.org/10.1007/s10614-023-10514-z doi: 10.1007/s10614-023-10514-z 13, 14, 66

Deng, H., Zhou, Y., Wang, L., & Zhang, C. (2021, 12). Ensemble learning for the early prediction of neonatal jaundice with genetic features. *BMC Medical Informatics and Decision Making*, *21*. doi: 10.1186/s12911-021-01701-9 29

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. 64

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple classifier systems*, *1857*, 1–15. 35

Fazekas, M., Dávid-Barrett, E., Hellmann, O., Márk, L., & McCorley, C. (2020). Controlling corruption in development aid: New evidence from contract-level data. *Studies in Comparative International Development*, *55*, 481-515. doi: 10.1007/s12116-020-09315-4 10, 14

Fazekas, M., & Tóth, I. J. (2016). From corruption to state capture: A new analytical framework with empirical applications from hungary. *Political Research Quarterly*, *69*(2), 320-334. doi: 10.1177/1065912916639137 19

Fazekas, M., Tóth, I. J., & King, L. P. (2016). An objective corruption risk index using public procurement data. *European Journal on Criminal Policy and Research*, *22*, 369-397. 9, 13, 14

Fazekas, M., Wachs, J., & Kertész, J. (2021). Corruption risk in contracting markets: a network science perspective. *International Journal of Data Science and Analytics*, *12*(45-60). doi: 10.1007/s41060-019-00204-1 9, 78

Ferwerda, J., Deleanu, I., & Unger, B. (2017). Corruption in public procurement: finding the right indicators. *European Journal on Criminal Policy and Research*, *23*, 245-267. 18

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable importance metric through model agnostic measures. *Journal of Machine Learning Research*, *20*, 1-81. 66

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232. 36

Guégan, D., & Hassani, B. (2018). Regulatory learning: How to supervise machine learning models? an application to credit scoring. *The Journal of Finance and Data Science*, *4*(3), 157-171. Retrieved from https://www.sciencedirect.com/science/article/pii/S2405918817300648 doi: https://doi.org/10.1016/j.jfds.2018.04.001 26

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In D.-S. Huang, X.-P. Zhang, & G.-B. Huang (Eds.), *Advances in intelligent computing, proceedings of the international conference on intelligent computing (icic 2005)* (Vol. 3644, pp. 878–887). Springer-

Verlag. doi: 10.1007/11538059_91 54
Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Morgan Kaufmann. 20, 81
Hansen, L. P., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497. 65, 104
He, H., & Garcia, E. A. (2009a). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263–1284. 37
He, H., & Garcia, E. A. (2009b). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. doi: 10.1109/TKDE.2008.239 38, 45, 48, 49, 50, 55, 103
Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. 56, 59
Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2023). TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. *Proceedings of the 11th International Conference on Learning Representations (ICLR)*. Retrieved from https://github.com/automl/TabPFN (Conference version arXiv:2207.01848) 33, 34, 35, 114
Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., ... Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, *637*(7974), 319–326. doi: 10.1038/s41586-024-08328-6 114
Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons. 22, 23, 24
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer. 21
James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in python* (2nd ed.). Springer Nature. 59
Jesus, M. B. (2020). *Modelo preditivo de risco de irregularidades em compras públicas no estado de goiás* (Mestrado Profissional em Computação Aplicada). Universidade de Brasília, Instituto de Ciências Exatas, Brasília, Brasil. (Orientador: Prof. Dr. Gladston Luiz da Silva) 12, 18, 19, 76, 77, 78, 79, 80
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems (nips)* (pp. 3146–3154). 32
Khemakhem, H., & Dicko, S. (2013). Directors' political connections and compliance with board of directors regulations: The case of s&p/tsx 300 companies. *International Journal of Business and Management*, *8*. Retrieved from https://doi.org/10.5539/ijbm.v8n24p117 doi: 10.5539/ijbm.v8n24p117 77, 79
Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. *"Keele University and Durham University Joint Report"*. 6
Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, *20*(3), 226–239. 36
Koehrsen, W. (2018). Overfitting vs. underfitting: A complete example. *Towards Data Science*. 37
Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*, 249–268. 20

Kovács, G. (2019). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*, *83*, 105662. doi: 10.1016/j.asoc.2019.105662 52, 54

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In *Advances in neural information processing systems* (Vol. 7, pp. 231–238). 35

LeCun, Y., Misra, I., & Ba, J. (2021). Self-supervised learning: The dark matter of intelligence. *arXiv preprint arXiv:2103.14746*. 22

Leo, M., & et al. (2019). Conciliating efficient algorithms with interpretations: Applications in finance. *Journal of Financial Data Science*, *5*(2), 45-67. 67

Li, C. (2016). A gentle introduction to gradient boosting. *URL: http://www. ccs. neu. edu/home/vip/teach/MLcourse/4_ boosting/slides/gradient_boosting. pdf*. 29

Li, H. L. (2024). *10.3 variable selection property of the lasso | introduction to data science.* Data Science Workshop. (Available online) 59

Li, J.-w., Chang, K.-c., & Wang, C.-c. (2017). A novel imbalanced data classification approach based on smote and tomek links. In *2017 ieee international conference on innovations in intelligent systems and applications (inista)* (p. 322-327). doi: 10.1109/INISTA.2017.8001178 51, 52, 54

Lopes, C. R., & de Jesus, P. A. G. (2024). Licitações e contratos na administração pública: aspectos, desafios e melhores práticas. *Studies in Multidisciplinary Review*, *5*(1), 57-78. doi: 10.55034/smrv5n1-004 17

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*. 68, 69, 106

Lyra, M. S., Curado, A., Damásio, B., Bação, F., & Pinheiro, F. L. (2021). Characterization of the firm–firm public procurement co-bidding network from the state of ceará (brazil). *Applied Network Science*, *6*, 77. doi: 10.1007/s41109-021-00418-y 10, 14

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 93–100. 57

Mishina, Y., Murata, R., Yamauchi, Y., Yamashita, T., & Fujiyoshi, H. (2015). Boosted random forest. *IEICE TRANSACTIONS on Information and Systems*, *98*(9), 1630–1636. 27

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill. 20

Mlondo, N. J. (2013). Effectiveness of knowing your customer policy in combating money laundering in commercial banks in tanzania: a case of bank of africa (t) limited. 3

Modrušan, N., & Rabuzin, K. (2019). Prediction of public procurement corruption indices using machine learning methods. In *Proceedings of the 11th international joint conference on knowledge discovery, knowledge engineering and knowledge management (ic3k 2019)* (p. 333-340). SCITEPRESS. doi: 10.5220/0008353603330340 9

Modrušan, N., Rabuzin, K., & Mršic, L. (2021). Review of public procurement fraud detection techniques powered by emerging technologies. , *12*(2). Retrieved 2024-10-08, from http://thesai.org/Publications/ViewPaper?Volume=12&Issue=2&Code=IJACSA&SerialNo=72 doi: 10.14569/IJACSA.2021.0120272 8, 14

Molnar, C. (2022). *Interpretable machine learning*. https://christophm.github.io/interpretable-ml-book/. (Acessado em 15 de novembro de 2025) 67, 68, 69

Munkhdalai, T., Lkhagvadorj, Namsrai, O.-E., Lee, J. Y., & Ryu, K. H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability*, *11*(3), 699. 26, 27

Nascimento, C. L. (2022). *O controle interno preventivo à luz do sistema alice: Propostas de trilhas para detecção de anomalias na execução de programas sociais* (Mestrado Profissional em Administração Pública). Fundação Getulio Vargas, Escola Brasileira de Administração Pública e de Empresas, Piauí, Brasil. (Orientador: Kaizô Iwakami Beltrão) 11, 14, 17, 19, 79

Nascimento, J. L. R. (2022). *Índice de priorização de objetos de auditoria: Um estudo de caso para municípios sergipanos* (Unpublished master's thesis). Escola Nacional de Administração Pública. 19, 77, 78

OECD. (2015). *Preventing corruption in public procurement*. Organização para a Cooperação e Desenvolvimento Econômico. 1

Padhi, S., & Mohapatra, P. (2011). Detection of collusion in government procurement auctions. *J Purch Supply Management*, *17*, 207–221. 3

Pereira, F. G. (2024). *Forecasting inflation in brazil with machine learning methods: Integrating shrinkage method for variable selection with shapley value interpretation* (Master's dissertation). University of Brasília, Brasília, Brazil. (Dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in Applied Computing) 64

Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *Acm sigkdd explorations newsletter*, *6*(1), 50–59. 103

Powers, D. M. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*. 38

Prodanov, C. C., & de Freitas, E. C. (2013). *Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico* (2nd ed.). Novo Hamburgo: Feevale. 70

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. "O'Reilly Media, Inc.
. 47

PwC. (2013). *Identifying and reducing corruption in public procurement in the eu*. European Commission. 1

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144). New York, NY, USA: ACM. doi: 10.1145/2939672.2939778 106

Rodrigues, N. C. S., & Lima Filho, R. N. (2017). Modalidades licitatórias e o risco de ocorrência de fraudes nos municípios baianos fiscalizados pela controladoria geral da união. In *Xvi congresso de controladoria e contabilidade da usp* (pp. 10, 12, 13, 30, 82). 15, 16, 17

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. doi: 10.1371/journal.pone.0118432 47, 50, 52, 55

Salazar, C., Partnership, O. C., Brown, S., & Neumann, G. (2024). *Red flags in public procurement: A guide to using data to detect and mitigate risks*. Guia Metodológico. Open Contracting Partnership. Retrieved from https://www.open-contracting

.org/wp-content/uploads/2024/12/OCP2024-RedFlagProcurement-1.pdf 77
Sales, L. J. (2016). *Proposta de modelo de classificação do risco de contratos públicos* (Mestrado em Economia do Setor Público). Universidade Nacional de Brasília, Faculdade de Economia, Administração e Contabilidade – FACE/UnB, Brasília, Brasil. (Orientador: Prof. Dr. Rafael Terra de Menezes) 12, 78
Sales, L. J., & Carvalho, R. N. (2016, June). Measuring the risk of public contracts using bayesian classifiers. In *Bma@ uai* (p. 7-13). 19, 77, 78, 79, 80
Santos, F. B., & Souza, K. R. (2023). *Como combater a corrupção em licitações: Detecção e prevenção de fraudes* (1st ed.). Belo Horizonte: FORUM. Retrieved from https://books.google.com.br/books?id=4i5qswEACAAJ 17
Schreyer, M., Sattarov, T., & Borth, D. (2022). Reshape: Explaining accounting anomalies in financial statement audits by enhancing shapley additive explanations. In (p. 1-9). 67
Shapley, L. S. (1953). A value for n-person games. In *Contribution to the theory of games 2*. Princeton, NJ: Princeton University Press. 67, 94
Siddiqi, N. (2017). *Credit risk scorecards: Developing and implementing intelligent credit scoring* (2nd ed.). Hoboken, NJ: John Wiley & Sons. 88
Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems* (pp. 2951–2959). 63
Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence* (pp. 1015–1021). 38
Souza, R. V. F. (2023). *Identificação automática de conluio em pregões do comprasnet com aprendizado de máquina* (Dissertação de Mestrado Profissional em Engenharia Elétrica). Universidade de Brasília, Faculdade de Tecnologia, Departamento de Engenharia Elétrica, Brasília, Brasil. (Orientador: Alexandre Solon Nery, Coorientador: Fábio Lúcio Lopes de Mendonça) 11, 14, 17
Soylu, A., Óscar Corcho, Elvesæter, B., Badenes-Olmedo, C., Yedro-Martínez, F., Kovacic, M., ... Roman, D. (2022). Data quality barriers for transparency in public procurement. *Information*, *13*, 99. doi: 10.3390/info13020099 10
Sun, T., & Sales, L. J. (2018, March). Predicting public procurement irregularity: An application of neural networks. *Journal of Emerging Technologies in Accounting*, *15*(1), 141–154. 76
Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press. 22
Tibshirani, R. (1996a). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. 57, 105
Tibshirani, R. (1996b). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x 85
Tomek, I. (1976). Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-6*(11), 769-772. doi: 10.1109/TSMC.1976.4309452 51
Trueck, S., & Rachev, S. T. (2009). *Rating based modeling of credit risk: theory and application of migration matrices*. Academic press. 23
U.S. General Services Administration Office of Inspector General, Office of Audits. (n.d.).

PROCUREMENT FRAUD HANDBOOK. Handbook / Relatório Institucional. Retrieved from https://www.gsaig.gov/sites/default/files/misc-reports/ProcurementFraudHandbook_0.pdf (n.d. (Referencia eventos de 2009 e 2010)) 77, 78

Velasco, R. B., Carpanese, I., Interian, R., Paulo Neto, O. C. G., & Ribeiro, C. C. (2021). A decision support system for fraud detection in public procurement. , *28*(1), 27–47. Retrieved 2024-10-08, from https://onlinelibrary.wiley.com/doi/10.1111/itor.12811 doi: 10.1111/itor.12811 8

Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *MAICS*, *710*, 120–127. 37

Wang, L., Zhu, J., & Zou, H. (2006). The doubly regularized support vector machine. *Statistica Sinica*, 589–615. 60

Wensink, W., & Vet, J. M. (2006). Identifying and reducing corruption in public procurement in the eu.
3

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241–259. 36