SÉRGIO HENRIQUE DE CARVALHO PEDROSO

DETECÇÃO DE FRAUDES EM AUDITORIA INTERNA POR MEIO DE ALGORITMOS DE APRENDIZADO SEMI-SUPERVISIONADO

LAVRAS – MG

2025

SÉRGIO HENRIQUE DE CARVALHO PEDROSO

DETECÇÃO DE FRAUDES EM AUDITORIA INTERNA POR MEIO DE ALGORITMOS DE APRENDIZADO SEMI-SUPERVISIONADO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, para a obtenção do título de mestre.

Prof. DSc. Wilian Soares Lacerda
Orientador

LAVRAS – MG
2025

Ficha Catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

de Carvalho Pedroso, Sérgio Henrique.
Detecção de fraudes em auditoria interna por meio de algoritmos de aprendizado semi-supervisionado / Sérgio Henrique de Carvalho Pedroso. - 2025.
59 p. : il.

Orientador: Wilian Soares Lacerda

Dissertação (Mestrado Acadêmico) - Universidade Federal de Lavras, 2025.
Bibliografia.

1. Aprendizado de máquina. 2. Auditoria. 3. Fraudes. 4. Semi-supervisionado. 5. CAATs. I. Soares Lacerda, Wilian. II. Universidade Federal de Lavras. III. Título.

SÉRGIO HENRIQUE DE CARVALHO PEDROSO

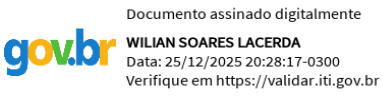
DETECÇÃO DE FRAUDES EM AUDITORIA INTERNA POR MEIO DE ALGORITMOS DE APRENDIZADO SEMI-SUPERVISIONADO

FRAUD DETECTION IN INTERNAL AUDITING THROUGH SEMI-SUPERVISED LEARNING ALGORITHMS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, para a obtenção do título de mestre.

APROVADA em 29 de setembro de 2025.

Prof. DSc. Wilian Soares Lacerda	UFLA
Prof. DSc. Bruno Henrique Groenner Barbosa	UFLA
Prof. DSc. Thiago de Souza Rodrigues	CEFET-MG



Prof. DSc. Wilian Soares Lacerda
Orientador

LAVRAS – MG
2025

A Deus. Aos meus pais, Sérgio e Silene.

AGRADECIMENTOS

Agradeço à Universidade Federal de Lavras (UFLA) pela excelência no ensino e pela estrutura disponibilizada ao longo de toda a minha formação. Ao Programa de Pós-Graduação em Engenharia de Sistemas e Automação (PPGESISA) e ao grupo de pesquisa AIA – Artificial Intelligence and Automation, expresso minha gratidão pelo espaço concedido e pelo suporte acadêmico oferecido.

Aos meus pais, Sérgio e Silene, registro meu mais profundo reconhecimento por todo amor, apoio e encorajamento incondicional, que foram fundamentais em cada etapa desta jornada.

Aos meus amigos, que sempre demonstraram apoio e companheirismo, fazendo com que os momentos de dificuldade se tornassem mais leves. Pela presença constante, pelas risadas que aliviaram o peso dos dias e pelo acolhimento genuíno em cada etapa desta jornada. Sem vocês, o caminho teria sido muito mais difícil.

Ao meu orientador, professor Dr. Wilian Soares Lacerda, agradeço pela orientação sempre atenta e pela confiança depositada em mim para a realização deste trabalho. Estendo também minha gratidão aos demais professores do programa, pela constante disponibilidade em contribuir com minha formação.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

“Existe algo de bom neste mundo, e vale a pena lutar por isso.”

J.R.R. Tolkien, O Senhor dos Anéis (Samwise Gamgee)

RESUMO

Este estudo investigou a aplicação de algoritmos de aprendizado de máquina para apoiar a detecção de fraudes e anomalias contábeis em uma empresa do setor logístico. A base de dados foi extraída do ERP (*Enterprise Resource Planning*), abrangendo registros de serviços vinculados à frota rodoviária adquiridos entre 2023 e 2025. Foram comparadas três estratégias de aprendizado semi-supervisionado, *Active Learning*, *Pseudo-Labeling* e *Label Propagation*, aplicadas a diferentes modelos, incluindo regressão logística, MLP, XGBoost e CatBoost. O protocolo experimental consistiu em executar cada combinação de modelo e estratégia por 10 repetições independentes, com amostragem estratificada e uso de um conjunto de teste fixo para aferição final. Os resultados indicaram que o *Active Learning*, em especial combinado ao CatBoost, alcançou o melhor desempenho, com F1 de 0,94 e PR-AUC de 0,98. Como segunda melhor estratégia, o CatBoost com *Label Propagation* obteve F1 de 0,80 e PR-AUC de 0,87. O *Pseudo-Labeling*, por sua vez, apresentou limitações em iterações posteriores devido à propagação de ruído. Em termos práticos, os achados demonstram o potencial de integração de técnicas semi-supervisionadas aos processos de auditoria interna, promovendo maior eficiência na priorização de casos suspeitos, redução de custos operacionais e fortalecimento da governança corporativa.

Palavras-chave: aprendizado de máquina; auditoria; fraudes; semi-supervisionado; CAATs.

ABSTRACT

This study investigated the application of machine learning algorithms to support fraud and anomaly detection in accounting data from a logistics company. The dataset was extracted from the ERP (Enterprise Resource Planning) system, covering road fleet service records acquired between 2023 and 2025. Three semi-supervised learning strategies were compared, *Active Learning*, *Pseudo-Labeling*, and *Label Propagation*, applied to different models, including Logistic Regression, MLP, XGBoost, and CatBoost. The experimental protocol consisted of running each model-strategy combination over 10 independent repetitions, with stratified sampling and a fixed test set used exclusively for final evaluation. Results showed that *Active Learning*, particularly when combined with CatBoost, achieved the best performance, with an F1 score of 0.94 and PR-AUC of 0.98. As the second-best approach, CatBoost with *Label Propagation* reached an F1 score of 0.80 and PR-AUC of 0.87. In contrast, *Pseudo-Labeling* exhibited limitations in later iterations due to noise propagation from pseudo-labels. In practical terms, the findings demonstrate the potential of integrating semi-supervised learning techniques into internal auditing processes, enabling greater efficiency in prioritizing suspicious cases, reducing operational costs, and strengthening corporate governance.

Keywords: machine learning; auditing; fraud; semi-supervised; CAATs

INDICADORES DE IMPACTO

O trabalho apresenta impactos potenciais e concretos em diferentes dimensões. No campo tecnológico, contribui para o avanço da auditoria assistida por computador, oferecendo soluções escaláveis de detecção de fraudes por meio de algoritmos de aprendizado de máquina semi-supervisionados. No aspecto econômico, possibilita a redução de perdas financeiras decorrentes de irregularidades e maior eficiência no uso de recursos humanos, ao direcionar auditores para casos mais relevantes.

Em termos sociais e institucionais, fortalece a governança corporativa e a transparência, aumentando a confiança em processos contábeis e promovendo práticas mais seguras nas organizações. O caráter extensionista se evidencia pela parceria com empresa do setor logístico, ampliando o impacto para além do ambiente acadêmico e favorecendo a transferência de conhecimento para o setor produtivo.

Os principais territórios impactados incluem o setor logístico e cadeias produtivas associadas, beneficiando gestores, auditores, profissionais de tecnologia e, indiretamente, clientes e parceiros. Quanto às áreas da Política Nacional de Extensão, a pesquisa se insere em Tecnologia e produção e Trabalho. Em alinhamento com os Objetivos de Desenvolvimento Sustentável (ODS), destaca-se a contribuição para o ODS 8 (Trabalho decente e crescimento econômico), 9 (Indústria, inovação e infraestrutura) e 16 (Paz, justiça e instituições eficazes).

Assim, este trabalho se estende além da contribuição acadêmica, gerando impactos sociais, econômicos e tecnológicos que fortalecem a auditoria interna, aumentam a integridade dos processos corporativos e promovem inovação no ambiente organizacional.

IMPACT INDICATORS

This work presents both potential and concrete impacts across different dimensions. In the technological field, it contributes to the advancement of computer-assisted auditing by providing scalable fraud detection solutions through semi-supervised machine learning algorithms. From an economic perspective, it enables the reduction of financial losses caused by irregularities and increases efficiency in the use of human resources by directing auditors to the most relevant cases.

In social and institutional terms, it strengthens corporate governance and transparency, enhancing trust in accounting processes and promoting safer practices within organizations. The extensionist character is evidenced by the partnership with a company in the logistics sector, expanding the impact beyond the academic environment and fostering knowledge transfer to the productive sector.

The main territories impacted include the logistics sector and associated production chains, benefiting managers, auditors, technology professionals and, indirectly, clients and partners. Regarding the thematic areas of the National Extension Policy, the research falls under Technology and Production and Work. In alignment with the United Nations Sustainable Development Goals (SDGs), this work contributes to SDG 8 (Decent Work and Economic Growth), SDG 9 (Industry, Innovation, and Infrastructure) and SDG 16 (Peace, Justice, and Strong Institutions).

Therefore, this research goes beyond academic contribution, generating social, economic, and technological impacts that strengthen internal auditing, enhance the integrity of corporate processes, and foster innovation in organizational environments.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Na era digital, as operações contábeis se tornaram cada vez mais complexas e volumosas, exigindo ferramentas avançadas para garantir a integridade e a confiabilidade das informações financeiras (Khorunzhak; Lukanovska, 2019). Fraudes e anomalias contábeis podem resultar em prejuízos significativos para as empresas e afetar a confiança de investidores e partes interessadas. Nesse contexto, métodos tradicionais de auditoria interna frequentemente se mostram insuficientes para lidar com a crescente quantidade de dados e a sofisticação das fraudes. A aplicação de técnicas de aprendizado de máquina surge, portanto, como uma solução promissora para identificar padrões anômalos e fraudes em grandes volumes de dados contábeis, oferecendo uma abordagem mais precisa e eficiente.

Estudos recentes reforçam essa perspectiva ao destacar que a inteligência artificial pode transformar a função de auditoria interna, reduzindo procedimentos manuais, ampliando a capacidade de análise para populações inteiras de dados, não apenas amostras, e permitindo a entrega de serviços de auditoria mais estratégicos e de maior valor agregado (Wassie; Lakatos, 2024).

A detecção de anomalias é um campo abrangente que se dedica a identificar instâncias de dados ou eventos que não seguem o comportamento esperado (Borges *et al.*, 2023; Chandola; Banerjee; Kumar, 2009). O termo detecção de *outliers* é frequentemente utilizado como sinônimo de detecção de anomalias. Embora existam nuances técnicas entre os dois termos em certos contextos, como em séries temporais, na prática eles são usados de forma intercambiável em diversas áreas, incluindo ciência de dados, aprendizado de máquina, segurança cibernética e monitoramento de sistemas. De acordo com (Olteanu; Rossi; Yger, 2023), há um consenso emergente na literatura de que os termos "*outlier*" e "anomalia" são tratados como equivalentes em muitas abordagens e algoritmos.

Devido ao grande volume de dados processados diariamente em sistemas de contabilidade, torna-se necessária a auditoria dessas operações de forma escalável e confiável. A literatura se refere aos CAATs (*Computer Assisted Audit Tools*) como sistemas que oferecem diversos benefícios na detecção de fraudes. Esses benefícios incluem a capacidade de analisar grandes volumes de dados de forma rápida e precisa, aumentando a confiança na identificação de anomalias e irregularidades. Além disso, os CAATs melhoram a eficiência do processo de auditoria, reduzindo o tempo necessário para realizar testes e verificações rotineiras, permitindo

que os auditores se concentrem em análises mais complexas e estratégicas. A integração desses sistemas também contribui para a sustentabilidade corporativa, promovendo transparência e responsabilidade nas operações empresariais, além de ajudar a garantir a conformidade com regulamentos e normas legais (Samagaio; Diogo, 2022).

Diante desse cenário, a aplicação de técnicas de aprendizado de máquina para auditoria interna se apresenta como uma solução promissora, oferecendo maior eficiência na identificação de padrões suspeitos e redução do esforço humano em análises repetitivas.

1.1 Objetivo Geral

Investigar e avaliar estratégias de aprendizado semi-supervisionado aplicadas à detecção de fraudes e anomalias em dados contábeis de uma empresa do setor logístico, validando seu potencial como suporte à auditoria interna.

1.2 Objetivos Específicos

- Preparar a base de dados contábeis extraída do ERP da empresa, realizando seleção de variáveis e transformações necessárias para a análise.
- Implementar modelos de aprendizado de máquina supervisionados e semi-supervisionados (*Active Learning*, *Pseudo-Labeling* e *Label Propagation*) para detecção de fraudes em pedidos de compra e serviços.
- Avaliar o desempenho dos modelos por meio de métricas adequadas, considerando o desbalanceamento das classes.
- Discutir os impactos práticos e organizacionais da adoção dessas técnicas na auditoria interna e na governança corporativa.

1.3 Motivação

Na empresa do setor logístico analisada neste trabalho, o processo de compras envolve valores expressivos e grande volume de registros, mas até recentemente não havia um procedimento estruturado de auditoria para monitorar essas operações. A ausência de rótulos confiáveis

para a maioria das transações dificulta a identificação de irregularidades e expõe a organização a riscos financeiros relevantes.

Esse contexto torna inviável a adoção de controles manuais, evidenciando a necessidade de ferramentas que permitam análises escaláveis e automatizadas. Assim, este projeto busca desenvolver e validar modelos capazes de auxiliar os auditores na detecção de irregularidades, fortalecendo os processos contábeis, a segurança financeira e a governança da empresa.

Além disso, observa-se uma lacuna de pesquisas aplicadas no contexto brasileiro, especialmente com dados reais do setor logístico. Este estudo pretende preencher parte dessa lacuna, aproximando avanços acadêmicos das demandas práticas de empresas nacionais.

1.4 Organização do Documento

Este trabalho está estruturado em seis capítulos. O capítulo 1 apresenta a introdução, os objetivos, a motivação e a contextualização do tema. O capítulo 2 revisa a literatura relacionada a fraudes contábeis, auditoria interna, anomalias e técnicas de aprendizado de máquina. O capítulo 3 descreve a metodologia adotada, incluindo a extração e preparação dos dados, os modelos e as estratégias utilizadas, além das métricas de avaliação. O capítulo 4 apresenta os resultados experimentais, bem como a discussão e implicações no contexto de negócio. Por fim, o capítulo 5 reúne as conclusões, limitações do estudo e sugestões para trabalhos futuros.

2 TRABALHOS RELACIONADOS

A literatura recente aponta que os modelos de aprendizado supervisionado ainda dominam as pesquisas em detecção de fraudes financeiras, representando mais da metade dos estudos publicados. Entre eles, destacam-se regressão logística, máquinas de vetor de suporte, árvores de decisão, florestas aleatórias, k-vizinhos mais próximos, Naive Bayes e redes neurais artificiais, amplamente aplicados pela robustez e adaptabilidade a diferentes cenários de fraude. Em paralelo, métodos de *ensemble* como XGBoost e técnicas baseadas em autoencoders também vêm sendo explorados. Já as abordagens não supervisionadas, como Isolation Forest e DBSCAN, respondem por uma parcela menor, mas têm se mostrado úteis em contextos com escassez de rótulos. Embora menos frequentes, estratégias híbridas que combinam aprendizado supervisionado, não supervisionado e profundo surgem como alternativas promissoras para ampliar a detecção de padrões complexos em bases altamente desbalanceadas (Aros *et al.*, 2024).

O trabalho apresentado por Schreyer *et al.* (2017) explora cinco abordagens de métodos de aprendizado de máquina não supervisionados para a detecção de anomalias em dados contábeis de grande escala. Entre os métodos analisados, as redes neurais *autoencoders* e o OC-SVM (*One Class - Support Vector Machines*) obtiveram os melhores resultados, com valores de F1 de 0,72 e 0,43, respectivamente. Além disso, foi aplicado o conceito de anomalias globais e locais. As anomalias globais correspondem a compras raramente realizadas e com valores significativamente diferentes, enquanto as anomalias locais referem-se a compras feitas frequentemente, mas com pequenas variações de valor ou frequência, tornando-as mais difíceis de identificar.

O estudo proposto por Alampay & Abu (2022) apresenta um modelo de *autoencoder* estocástico para detecção de anomalias, que ajusta dinamicamente seus parâmetros para evitar ajustes manuais durante a inferência. O modelo foi testado contra outros modelos de detecção de anomalias em um cenário de aprendizado semi-supervisionado, mostrando desempenho competitivo em diferentes cenários de detecção de anomalias.

No trabalho proposto por Bay *et al.* (2006), foram utilizados métodos de Naive Bayes com algoritmo de expectativa-maximização (EM) e Naive Bayes Positivo para identificar contas e transações suspeitas, no contexto de dados não rotulados, avaliando atributos derivados de lançamentos contábeis de diversas companhias. Neste trabalho, os resultados indicam que o tratamento explícito da incerteza dos rótulos pode melhorar o desempenho, dadas as suposições apropriadas.

Em Rahman & Zhu (2024), é relatada uma investigação de algoritmos *ensemble* para a detecção de fraudes em companhias chinesas, também validando o impacto de bancos de dados desbalanceados no treinamento e eficiência destes modelos. Foram aplicadas as técnicas *Bagging*, Florestas Aleatórias, RUSBoost e CusBoost, obtendo a melhor performance no último com o resultado AUC médio de 0,7827. Ao comparar com os demais resultados, conclui-se que métodos *ensemble* para dados desbalanceados conseguem capturar as características intrínsecas de problemas de detecção de anomalias em dados contábeis.

Um estudo utilizou um sistema híbrido de detecção de fraudes em cartões de crédito (HCCFD) que aplicou um algoritmo genético e distribuição normal multivariada para identificar transações fraudulentas. Ao testar o sistema, ele obteve uma precisão de 93,5%, superando redes neurais artificiais (84,2%), árvores de decisão (80,0%) e máquinas de vetor de suporte (68,5%) quando treinados no mesmo conjunto de dados (Makolo; Adeboye, 2021).

O trabalho proposto por Meenu *et al.* (2020) focou na detecção de anomalias em transações de cartão de crédito utilizando o algoritmo Isolation Forest. O desempenho foi avaliado com base em métricas como precisão e *recall*, mostrando que o Isolation Forest foi eficiente na identificação de padrões fraudulentos em um conjunto de dados altamente desbalanceado.

Em Leevy *et al.* (2023) comparam diretamente classificadores binários (CatBoost, XGBoost, Random Forest, entre outros) com métodos de classe única (OC-SVM, GMM, redes one-class) em tarefas de detecção de fraude e mostram vantagem consistente dos binários em termos de AUPRC quando rótulos das duas classes estão disponíveis. Por outro lado, métodos one-class continuam úteis quando apenas a classe normal é rotulável ou quando o custo de anotação é proibitivo.

Mais recentemente, abordagens baseadas em *Active Learning* (AL) têm ganhado destaque. Em Carcillo *et al.* (2018), aplicou-se AL em fluxos de dados contábeis desbalanceados e não estacionários, demonstrando que a estratégia de *High Risk Querying* melhora a adaptação a mudanças nos padrões de fraude. De forma semelhante, em contextos distintos como seguridade social Vlasselaer *et al.* (2015) e demonstrações financeiras Karlos *et al.* (2017), verificou-se que o aprendizado ativo supera abordagens supervisionadas tradicionais, mesmo com poucos exemplos rotulados, reduzindo o custo de anotação e aumentando a precisão.

Diversos estudos reforçam a importância de incorporar dados não rotulados de forma incremental e iterativa em cenários de aprendizado semi-supervisionado (SSL). Guo *et al.* (2024) propuseram o método Incremental Self-Training (IST), no qual o modelo inicia com um con-

junto reduzido de exemplos rotulados e, a cada iteração, incorpora lotes de exemplos não rotulados classificados com maior confiança. Essa estratégia resultou em ganhos expressivos de desempenho em comparação ao treinamento em lote único, alcançando melhorias médias de até +6,7% em F1-score em *benchmarks* de texto e imagem, além de maior eficiência computacional.

Viegas *et al.* (2018) exploraram um modelo de *committee-based co-training* para identificar fraudes em consumo de energia elétrica. Nesse trabalho, múltiplos classificadores eram treinados e, a cada iteração, amostras não rotuladas com maior consenso eram adicionadas ao conjunto de treino. O método apresentou desempenho superior ao supervisionado tradicional, atingindo AUC de 0,96 contra 0,89 dos modelos de referência, evidenciando a robustez do processo incremental.

Além disso, pesquisas recentes ressaltam o potencial da combinação entre *Active Learning* e aprendizado semi-supervisionado (SSL). Essa integração permite explorar melhor os dados não rotulados e selecionar amostras mais informativas, ampliando a cobertura da detecção de fraudes sem comprometer a precisão. Aplicações em domínios como monitoramento aduaneiro e seguros médicos confirmam a eficácia dessa abordagem (Kim *et al.*, 2020; Zhang, 2024).

Em síntese, a literatura sobre detecção de fraudes apresenta um amplo leque de técnicas, desde métodos não supervisionados até abordagens híbridas baseadas em *Active Learning* e SSL. Contudo, observa-se que a maior parte dos estudos concentra-se em domínios como cartões de crédito e seguros, enquanto aplicações em auditoria interna de compras corporativas permanecem pouco exploradas. Essa lacuna, aliada ao desafio da escassez de rótulos, motiva a presente pesquisa, que investiga a aplicação de estratégias semi-supervisionadas com foco em *Active Learning*, *Pseudo-Labeling* e *Label Propagation* no contexto de uma empresa do setor logístico.

3 REFERENCIAL TEÓRICO

O processo de detecção de fraudes em dados contábeis consiste na identificação de anomalias de valores, frequências e tipos de compras ou serviços contratados. Dessa forma, a abordagem proposta visa adotar técnicas de aprendizado de máquina, bem como suas estratégias de aprendizado, para tratar desta questão.

Alguns conceitos importantes da literatura serão apresentados a fim de proporcionar melhor entendimento do trabalho realizado.

3.1 Fraudes em Contabilidade

A contabilidade tem como objetivo principal fornecer informações sobre a posição patrimonial, financeira e econômica das organizações. No entanto, tais informações podem ser manipuladas intencionalmente, resultando em fraudes contábeis, ou seja, ações deliberadas com o propósito de enganar usuários, geralmente visando ganho financeiro ou ocultação de má gestão.

De acordo com Albrecht, Albrecht & Albrecht (2011), fraude contábil pode ser definida como "uma manipulação intencional de registros, demonstrações financeiras ou transações com o intuito de criar uma falsa aparência da situação financeira da empresa". Essa prática compromete não apenas os *stakeholders* diretamente envolvidos, mas também o funcionamento dos mercados e a confiança no sistema financeiro.

O chamado problema contábil refere-se à dificuldade em garantir a veracidade das demonstrações contábeis, dada a possibilidade de manipulação por parte da própria gestão, que frequentemente possui acesso privilegiado às informações e aos sistemas internos. Esse problema é discutido por Healy & Wahlen (1999), que destacam o uso de práticas contábeis oportunistas, como o gerenciamento de resultados (*earnings management*), para atingir metas específicas ou atender expectativas de mercado.

As fraudes contábeis mais comuns incluem:

- Reconhecimento prematuro de receitas: quando empresas registram receitas antes que elas sejam efetivamente realizadas.
- Subavaliação de passivos: ocultação de obrigações para melhorar indicadores de solvência.

- Superavaliação de ativos: manipulação de valores patrimoniais para inflar resultados ou patrimônio líquido.
- Transações fora de balanço: estruturação de operações complexas que não aparecem nas demonstrações principais.

A detecção dessas práticas de forma manual é desafiadora, uma vez que fraudes são, por natureza, eventos raros, intencionais e sofisticados. Assim, métodos automatizados baseados em dados, como algoritmos de detecção de anomalias, têm ganhado destaque na literatura de auditoria e ciência de dados aplicada à contabilidade. Estudos como os de Perols (2011) e Dechow *et al.* (2011) mostram que técnicas de aprendizado de máquina podem superar abordagens tradicionais ao detectar padrões sutis em dados contábeis que indicam comportamento fraudulento.

Essa abordagem computacional, aliada ao conhecimento contábil, é fundamental para mitigar o problema contábil e fortalecer os mecanismos de controle, auditoria e conformidade regulatória.

3.2 Auditoria Interna e Detecção de Fraudes

A auditoria interna é uma atividade independente e objetiva de avaliação e consultoria, projetada para agregar valor e melhorar as operações de uma organização. Seu papel principal é ajudar a entidade a atingir seus objetivos, fornecendo uma abordagem sistemática e disciplinada para avaliar e melhorar a eficácia dos processos de governança, gerenciamento de riscos e controles internos (The Institute of Internal Auditors, 2020).

No contexto da contabilidade, a auditoria interna exerce uma função crítica na identificação de falhas operacionais, inconsistências contábeis e, sobretudo, na prevenção e detecção de fraudes. Ao realizar revisões periódicas e análises de documentos, lançamentos contábeis e processos operacionais, o auditor interno pode identificar padrões atípicos ou indícios de manipulação, contribuindo para a integridade das informações financeiras.

A literatura destaca que, embora a auditoria interna tradicionalmente utilize abordagens baseadas em amostragem e verificação manual, essas técnicas podem não ser suficientes para detectar fraudes sofisticadas ou intencionais, que muitas vezes se camuflam em grandes volumes de dados. Dessa forma, a integração de técnicas automatizadas, como modelos de detecção de

anomalias e algoritmos de aprendizado de máquina, tem-se mostrado uma estratégia promissora para ampliar a capacidade de análise da auditoria (Choi; Lin; Walker, 2019).

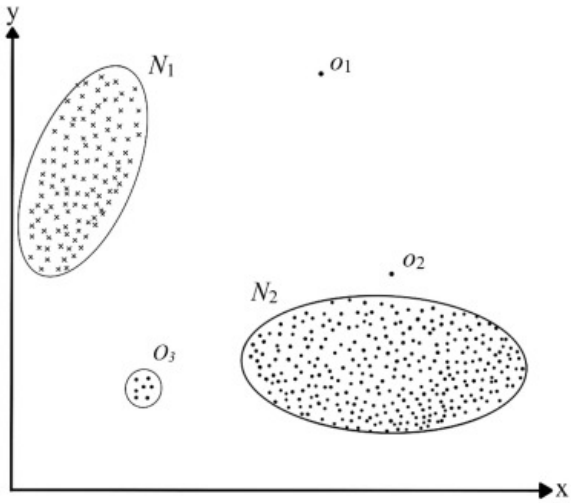
Estudos como o de Braun (2015) evidenciam que o uso de sistemas de apoio à decisão baseados em dados pode melhorar significativamente o julgamento dos auditores em relação à presença de fraudes. Além disso, segundo Liou (2008), a auditoria interna eficaz está diretamente associada à robustez dos controles internos e ao uso de ferramentas analíticas que permitam uma visão abrangente das operações contábeis e financeiras.

3.3 Anomalias

O conceito de anomalias pode ser definido conforme proposto por Hawkins (1980), em que um *outlier* é considerado uma observação que se desvia significativamente das demais, a ponto de levantar suspeitas de que tenha sido gerado por mecanismos distintos.

De forma ilustrativa, a Figura 3.1 apresenta um conjunto de dados representados em um espaço bidimensional, exemplificando diferentes tipos de anomalias. Os conjuntos agrupados em N1 e N2 são regiões de maior frequência de eventos, o que podemos considerar como não anomalias. Observações individuais mais distantes como em O1 e O2, ou agrupamentos pequenos como O3, são classificados como anomalias por diferentes critérios.

Figura 3.1 – Representação de agrupamentos e anomalias em um espaço bidimensional



Fonte: (Chandola; Banerjee; Kumar, 2009)

Anomalias podem ser classificadas em três categorias: pontuais, contextuais e coletivas, conforme Hilal, Gadsden & Yawney (2022). Tais técnicas são essenciais para a detecção de

fraudes no contexto do trabalho proposto, além de auxiliar na construção de possíveis cenários de fraude.

Anomalias pontuais são instâncias individuais de dados que são significativamente diferentes do restante. Por exemplo, se um funcionário da empresa realiza uma compra interna de suprimentos com um valor muito maior que o usual, essa compra seria considerada uma anomalia pontual, como nos pontos O1 e O2.

Já as anomalias contextuais são instâncias de dados que são anômalas em um contexto específico. Por exemplo, se um funcionário normalmente compra peças de reposição apenas durante a semana, mas faz uma compra no fim de semana, isso pode ser considerado uma anomalia contextual, dado o contexto do período em que foi feita, independente do montante gasto.

Por fim, as anomalias coletivas são uma coleção de instâncias relacionadas que são anômalas em conjunto, dado todo o espaço de observações realizadas. Por exemplo, várias pequenas compras de combustível feitas por um grupo de funcionários em um curto período, que individualmente podem parecer normais, mas coletivamente indicam um padrão anômalo, como no conjunto O3. Neste caso seriam uma anomalia coletiva.

3.4 Aprendizado de Máquina

O Aprendizado de Máquina (ou *machine learning*) no contexto de inteligência artificial (IA) abrange uma ampla gama de técnicas, aplicações e desafios. Aprendizado de máquinas é uma subárea da IA que se concentra em desenvolver algoritmos que permitem aos sistemas aprender a partir de dados, aprimorando suas capacidades de forma automática sem precisar de uma programação explícita para cada tarefa. Isso contrasta com abordagens tradicionais de programação, onde as regras são definidas manualmente (Schneider; Guo, 2018).

O conceito de aprendizado de máquinas consolidou-se como uma ferramenta essencial para a IA, com aplicações que vão desde tarefas simples, como filtragem de e-mails, até problemas complexos em áreas como medicina, finanças e ciência de dados. Conforme discutido em Dietterich (1996), a evolução dos algoritmos e a maior disponibilidade de dados foram fatores determinantes para o surgimento de avanços e desafios que moldaram a área, ressaltando a necessidade de uma análise crítica das metodologias e de suas limitações.

3.4.1 Técnicas de aprendizado

As técnicas de aprendizado de máquina podem ser agrupadas em três grandes categorias: aprendizado supervisionado, não supervisionado e semi-supervisionado (Dutton; Conroy, 1997).

No aprendizado supervisionado, utilizam-se dados históricos rotulados para treinar modelos capazes de prever valores de saída a partir de novas entradas. O treinamento consiste em ajustar uma função preditiva a partir de pares entrada–saída, de modo que o algoritmo minimize os erros em relação aos rótulos conhecidos. Esse paradigma é amplamente aplicado em tarefas como classificação e regressão (Bishop, 2006).

O aprendizado não supervisionado, por sua vez, é empregado quando os dados não possuem rótulos disponíveis. O objetivo principal é identificar estruturas ou padrões ocultos, como agrupamentos ou representações latentes, permitindo explorar regularidades nos dados sem depender de saídas previamente conhecidas. Métodos como *clustering* e análise de componentes principais (PCA) são exemplos clássicos (Hastie; Tibshirani; Friedman, 2009).

3.4.1.1 Aprendizado Semi-Supervisionado (SSL)

Entre os dois extremos supervisionado e não supervisionado encontra-se o aprendizado semi-supervisionado (SSL), que combina um pequeno conjunto rotulado $X_L = \{(x_i, y_i)\}$ e um grande conjunto não rotulado $X_U = \{x_j\}$. A ideia é explorar a estrutura dos dados não rotulados para refinar a fronteira de decisão aprendida. Um levantamento recente organiza esse campo em cinco famílias principais: modelos generativos, regularização por consistência, métodos baseados em grafos, pseudo-rotulagem (*self-training*) e abordagens híbridas (Yang *et al.*, 2023).

Algumas hipóteses teóricas ajudam a explicar quando o semi-supervisionado tende a trazer ganhos (Chapelle; Schölkopf; Zien, 2006): a hipótese de *clusters*, de baixa densidade, de *manifold* e a generativa, cada uma oferecendo fundamentos geométricos ou probabilísticos para o aproveitamento dos dados não rotulados.

3.4.1.1.1 Métodos baseados em grafos

Os métodos baseados em grafos representam a geometria de $X_L \cup X_U$ por meio de um grafo de vizinhança (k-NN) e propagam rótulos ao longo das conexões. Essa propagação se dá pela hipótese de que nós conectados tendem a compartilhar o mesmo rótulo, o que permite inferir classes para os exemplos não rotulados. O *Label Propagation* é um exemplo clássico dessa categoria, sendo um método transdutivo que assume que os dados não rotulados vistos no treino são exatamente aqueles que se deseja prever (Yang *et al.*, 2023). Essa abordagem é vantajosa em cenários com estrutura de similaridade bem definida, mas depende fortemente da qualidade da representação e da conectividade do grafo.

3.4.1.1.2 Métodos de pseudo-rotulagem

A pseudo-rotulagem, também chamada de *self-training*, utiliza um classificador inicial treinado em X_L para rotular exemplos de X_U acima de um limiar de confiança, reincorporando-os ao treinamento. O processo é iterativo e visa expandir o conjunto rotulado de forma automática, reduzindo a dependência de anotações humanas (Zhao *et al.*, 2022; Pedregosa *et al.*, 2011). Apesar de simples e eficaz em bases moderadamente limpas, essa técnica é sensível a ruídos, ou seja, erros de pseudo-rotulagem podem se propagar ao longo das iterações, deteriorando o desempenho do modelo.

3.4.1.1.3 Aprendizado Ativo (Active Learning)

O *Active Learning* (AL) é uma abordagem voltada para cenários em que obter rótulos é caro ou demorado, como em tarefas de detecção de fraudes. Em vez de rotular grandes volumes de dados aleatoriamente, o modelo seleciona iterativamente os exemplos mais informativos para anotação por um especialista (Settles, 2009). Entre as estratégias clássicas estão:

- **Uncertainty Sampling**: o modelo solicita rótulos para exemplos em que sua incerteza é maior;
- **Query by Committee (QBC)**: múltiplos modelos são treinados e exemplos com maior desacordo entre eles são priorizados;

- **Stream-Based Selective Sampling**: amostras são avaliadas sequencialmente e selecionadas conforme um critério de informatividade;
- **Pool-Based Sampling**: o modelo escolhe amostras mais relevantes em um conjunto fixo de dados não rotulados.

Essa estratégia reduz o custo de anotação e acelera o aprendizado, sendo especialmente útil em contextos de desbalanceamento, como a identificação de fraudes ou anomalias financeiras (Leevy *et al.*, 2023).

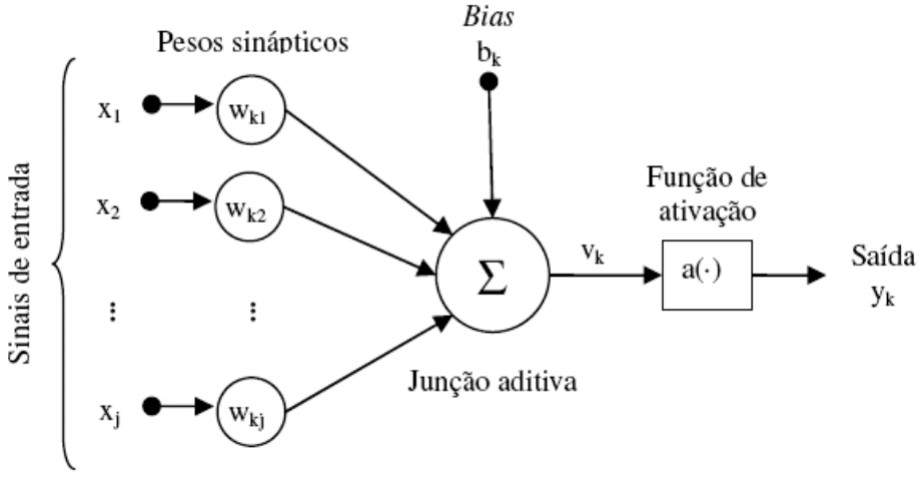
3.4.2 Modelos de aprendizado de máquina

Serão apresentados nesta seção os modelos de aprendizado de máquina, que serão aplicados neste projeto, e seus principais conceitos e características.

3.4.2.1 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) foram inicialmente desenvolvidas por McCulloch & Pitts (1943), quando descreveram formalmente o conceito de neurônio artificial. Nesse modelo, o neurônio é representado como uma unidade computacional que recebe valores de entrada e produz uma saída. A transformação das entradas em saída ocorre por meio de uma função de ativação, conforme ilustrado na Figura 3.2.

Figura 3.2 – Representação de um neurônio artificial



Fonte: (Frascaroli, 2006)

De forma geral, seu funcionamento pode ser expresso de acordo com a Equação 3.1.

$$z = \sum_{i=1}^{n} w_i x_i + b, \tag{3.1}$$

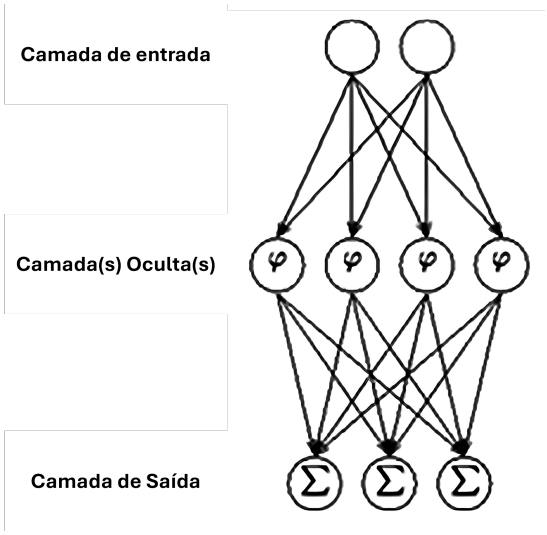
Em que x_i são as entradas, w_i os pesos associados e b o viés. A saída do neurônio é obtida aplicando-se uma função de ativação $a(\cdot)$ sobre z, conforme Equação 3.2.

$$y = a(z). \tag{3.2}$$

Funções de ativação não lineares, como a sigmoide ou a ReLU (*Rectified Linear Unit*), permitem que as redes neurais representem relações complexas entre as variáveis de entrada e saída (Goodfellow; Bengio; Courville, 2016).

A RNA de perceptron multicamada (MLP, do inglês) consiste em muitos neurônios organizados em camadas, com pelo menos uma camada oculta, conforme Figura 3.3, cujos neurônios utilizam técnicas como *backpropagation* para aprendizagem supervisionada (Desai; Shah, 2021).

Figura 3.3 – Rede neural com multicamadas



Fonte: (Desai; Shah, 2021)

O conceito de RNA é inspirado na rede neural biológica do cérebro humano, sendo constituída por um sistema de nós conectados organizados em camadas: uma camada de entrada, uma ou mais camadas intermediárias ocultas e uma camada de saída. Uma vez interligados, são formados links ponderados associados que enviam sinais entre si em forma numérica. A saída de cada neurônio é formada como uma função da soma ponderada de sua entrada. Os pesos nas

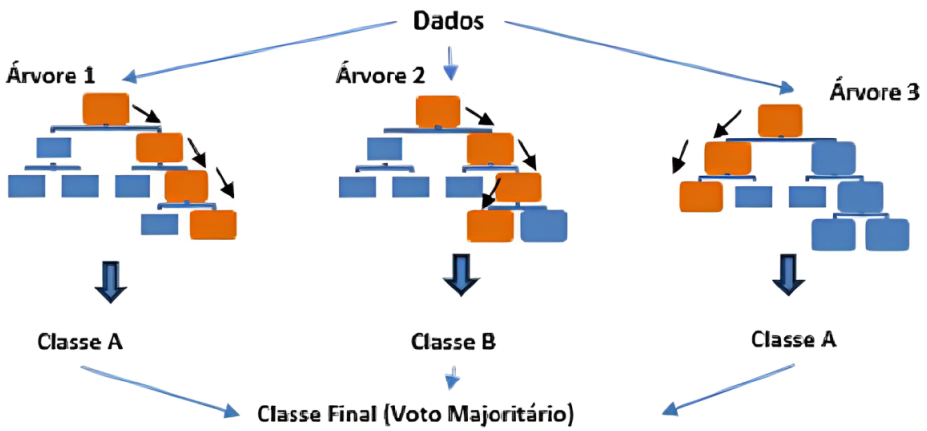
conexões são modificados na fase de aprendizagem para representar a força das conexões em relação aos nós (Bishop, 1995).

3.4.2.2 Métodos baseados em árvores de decisão

Modelos de *ensemble learning* baseados em árvores de decisão têm se mostrado extremamente eficazes em tarefas de classificação e regressão, explorando a combinação de múltiplos classificadores para aumentar a robustez e o poder preditivo. Entre as abordagens mais consolidadas destacam-se as Florestas Aleatórias, que utilizam a técnica de *bagging*, e os métodos de *boosting*, dos quais o Gradient Boosting e sua implementação otimizada, o XGBoost, são exemplos notáveis.

O classificador Florestas Aleatórias combina múltiplas árvores de decisão construídas a partir de amostras do conjunto de dados históricos, aplicando a técnica de *bagging* (amostragem com reposição). Após o treinamento, a classificação final de um novo caso é obtida por meio de um voto majoritário entre as árvores individuais, garantindo resultados mais robustos e generalizáveis (Breiman, 2001). Esse método reduz a chance de *overfitting* ao aproveitar a diversidade das árvores e apresenta boa acurácia em diversos domínios, como saúde, finanças e sensoriamento remoto (An *et al.*, 2023; Liaw; Wiener, 2001). A Figura 3.4 ilustra a lógica de construção de uma floresta de decisão.

Figura 3.4 – Representação de um modelo de Florestas Aleatórias

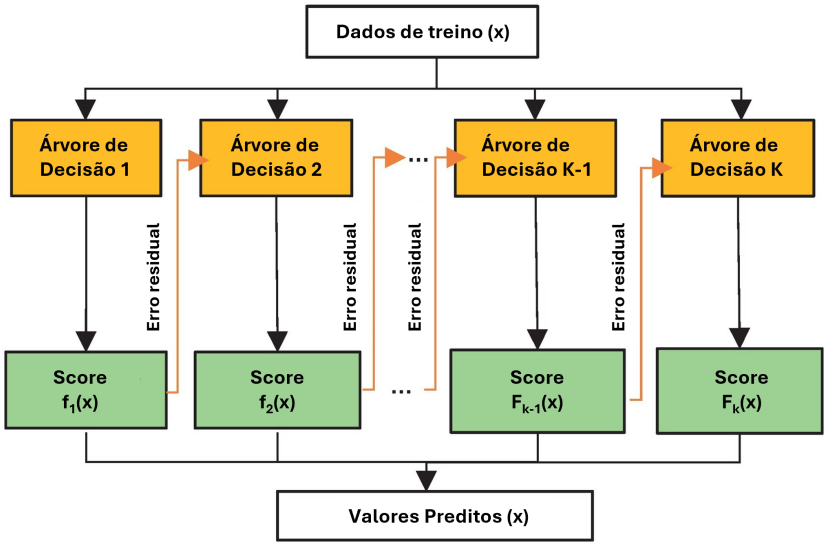


Fonte: (Farhat *et al.*, 2023)

Por outro lado, o Gradient Boosting segue uma abordagem sequencial, na qual cada novo modelo busca corrigir os erros dos anteriores, minimizando uma função de perda por

meio da descida de gradiente em espaço funcional (Friedman, 2001). Essa técnica mostrou-se fundamental para o avanço dos métodos de *ensemble*, proporcionando ganhos significativos em acurácia, embora com maior sensibilidade a parâmetros e custos computacionais mais elevados. A Figura 3.5 ilustra o fluxo de treinamento do método.

Figura 3.5 – Fluxo de treinamento do Gradient Boosting



Fonte: (Manoharan *et al.*, 2022)

Nesse contexto surge o XGBoost (*eXtreme Gradient Boosting*), uma implementação otimizada do Gradient Boosting amplamente reconhecida por sua eficiência computacional e alto desempenho preditivo em dados estruturados. O XGBoost introduz regularização L1 e L2 para mitigar *overfitting*, paralelização na construção das árvores, além de técnicas como *pruning*, que remove divisões pouco relevantes, e *early stopping*, que interrompe o treinamento quando não há melhora consistente em validação (Chen; Guestrin, 2016). Tais características tornam o XGBoost uma das ferramentas mais eficazes para aprendizado supervisionado, sendo recorrente em aplicações práticas e competições de ciência de dados.

Mais recentemente, surgiu o CatBoost, também baseado em Gradient Boosting, desenvolvido com foco no tratamento eficiente de variáveis categóricas e na redução do viés decorrente do processo de ordenação durante o treinamento. O CatBoost utiliza uma técnica denominada *ordered boosting*, que evita o uso de informações futuras ao construir as árvores, mitigando problemas de *overfitting* comuns em implementações tradicionais (Prokhorenkova *et al.*, 2019). Além disso, o algoritmo incorpora estratégias de codificação categórica nativas, dispensando pré-processamentos adicionais e tornando-se especialmente eficaz em conjuntos de dados tabulares com variáveis desse tipo.

3.4.2.3 Regressão Logística

A Regressão Logística é um modelo estatístico amplamente utilizado para problemas de classificação binária. Ela estima a probabilidade de um evento ocorrer com base em uma combinação linear das variáveis independentes. Para isso, utiliza a função logística (ou sigmoide), Equação 3.3, que garante que os valores previstos estejam sempre no intervalo entre 0 e 1.

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}} \tag{3.3}$$

Apesar de ser um modelo linear, sua performance pode ser bastante satisfatória em *datasets* com boa separabilidade, sendo a interpretabilidade um de seus pontos fortes. A sua interpretação probabilística permite que cada coeficiente seja associado ao impacto relativo de uma variável na chance de ocorrência do evento, tornando-se uma escolha comum em diversas áreas, como epidemiologia, ciências sociais e aprendizado de máquina (Hosmer; Lemeshow; Sturdivant, 2013).

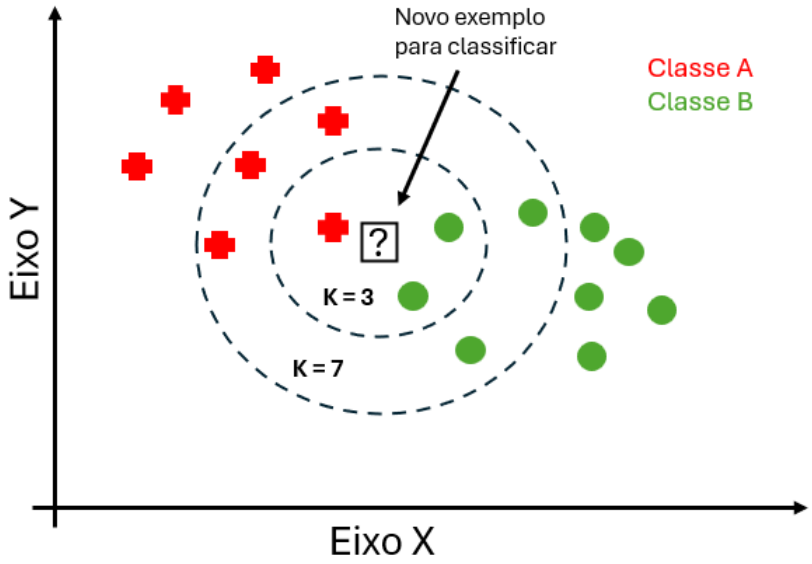
No contexto de detecção de fraudes, a regressão logística é frequentemente utilizada como modelo de referência (*baseline*), dada sua robustez, facilidade de implementação e capacidade de fornecer interpretações claras sobre fatores de risco associados a comportamentos fraudulentos. Apesar de modelos mais complexos como árvores de decisão ou redes neurais frequentemente superarem a regressão logística em termos de acurácia, sua transparência e interpretabilidade fazem com que ela continue sendo amplamente aplicada em auditoria e análise contábil (Ngai *et al.*, 2011).

3.4.2.4 *K-Nearest Neighbors*

O método *K-Nearest Neighbors* (KNN) é um algoritmo de aprendizado supervisionado amplamente utilizado devido à sua simplicidade conceitual e à boa performance em diversos domínios. A ideia central consiste em classificar uma nova instância de acordo com a maioria das classes presentes entre seus "k" vizinhos mais próximos em um espaço de características.

A Figura 3.6 representa a lógica de classificação do KNN baseado na definição de "k" vizinhos mais próximos. Neste exemplo, o novo exemplo seria classificado como classe B caso k = 3, e como classe A quando k = 7.

Figura 3.6 – Exemplo de classificação de uma nova instância no KNN considerando diferentes valores de k



Fonte: (Kumar *et al.*, 2020)

Apesar de sua eficácia, o KNN apresenta desafios como alta complexidade computacional em bases de grande porte e forte sensibilidade à escolha do parâmetro k, além de ser influenciado por características ruidosas ou irrelevantes nos dados (Ferreira; Cardoso; Mendes-Moreira, 2020).

Para mitigar essas limitações, diversas extensões foram propostas, como o uso de métricas de distância ponderadas, seleção de variáveis relevantes e técnicas de redução de dimensionalidade (Guo *et al.*, 2003). Ainda assim, o KNN permanece como um dos algoritmos fundamentais em aprendizado de máquina, tanto em sua forma original quanto como componente em diferentes métodos.

Um exemplo dessa aplicação ampliada ocorre em algoritmos de aprendizado semi-supervisionado baseados em grafos, como o *Label Propagation*. Nesse contexto, o KNN é utilizado para construir o grafo de vizinhança que conecta instâncias similares, servindo como *kernel* de proximidade. Assim, a decisão sobre quais pontos são vizinhos influencia diretamente o processo de propagação de rótulos, tornando a escolha de k e da métrica de distância elementos críticos para a qualidade do aprendizado. Dessa forma, o KNN não apenas atua como um classificador independente, mas também como um bloco de construção essencial em técnicas modernas de propagação de rótulos.

3.5 Métricas de desempenho

As métricas de desempenho para classificadores são essenciais para avaliar a eficácia de modelos de classificação e guiar sua seleção e configuração apropriadas. A seguir, são apresentadas algumas das métricas que serão utilizadas para a avaliação dos modelos deste projeto.

A avaliação de classificadores em problemas de aprendizado supervisionado geralmente é baseada na matriz de confusão, que resume os acertos e erros do modelo em relação às classes reais. A partir dessa matriz, derivam-se as principais métricas de avaliação utilizadas neste trabalho.

A matriz organiza as predições em quatro categorias principais:

- Verdadeiros Positivos (TP): casos positivos corretamente classificados como positivos;
- Falsos Positivos (FP): casos negativos incorretamente classificados como positivos;
- Verdadeiros Negativos (TN): casos negativos corretamente classificados como negativos;
- Falsos Negativos (FN): casos positivos incorretamente classificados como negativos.

A acurácia mede a proporção de previsões corretas em relação ao total de previsões feitas. É uma métrica simples e amplamente utilizada, especialmente quando as classes estão balanceadas. No entanto, sua eficácia diminui em cenários de desequilíbrio de classes, onde classes majoritárias podem distorcer os resultados (Liu *et al.*, 2014). Como mostrado na Equação 3.4, a acurácia leva em conta todas as categorias da matriz de confusão.

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

Precisão refere-se à proporção de verdadeiros positivos entre todas as instâncias classificadas como positivas, conforme definido na Equação 3.5, enquanto o *recall*, ou taxa de verdadeiro positivo (TPR, na sigla em inglês), mede a proporção de verdadeiros positivos entre todas as instâncias que são realmente positivas, conforme mostrado na Equação 3.6. Essas métricas são particularmente úteis em situações de desequilíbrio de classes, onde a acurácia pode ser enganosa (Seliya; Khoshgoftaar; Hulse, 2009).

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{3.5}$$

$$\text{Recall ou TPR} = \frac{TP}{TP+FN} \tag{3.6}$$

O F1-Score é a média harmônica da precisão e do *recall*, fornecendo uma medida balanceada entre ambas, como ilustrado na Equação 3.7, sendo especialmente útil quando há um desequilíbrio significativo entre as classes. Essa métrica é ideal para cenários onde tanto falsos positivos quanto falsos negativos precisam ser minimizados (Luque *et al.*, 2019).

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{3.7}$$

A área sob a curva ROC (*Receiver Operating Characteristic*), ou AUC-ROC, traça a taxa de verdadeiros positivos (sensibilidade) contra a taxa de falsos positivos, conforme Equação 3.8. A AUC (Área Sob a Curva), definida na Equação 3.9, fornece uma medida da capacidade do classificador em distinguir entre classes. Uma AUC próxima de 1 indica um bom desempenho, enquanto uma AUC próxima de 0,5 indica um desempenho aleatório (Deist *et al.*, 2018).

$$\text{FPR} = \frac{FP}{FP+TN} \tag{3.8}$$

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(FPR)\, d(\text{FPR}) \tag{3.9}$$

Por fim, uma métrica especialmente relevante em cenários com classes desbalanceadas é a área sob a curva de Precisão-*Recall*, conhecida como PR-AUC (*Precision-Recall Area Under the Curve*). Diferente da AUC-ROC, que considera a taxa de verdadeiros e falsos positivos, a PR-AUC foca diretamente nas métricas de precisão e *recall*, representando a relação entre elas em diferentes limiares de decisão, conforme definido na Equação 3.10. Essa métrica é particularmente sensível à classe positiva, sendo mais adequada quando o interesse está concentrado em detectar corretamente a minoria. Valores de PR-AUC próximos de 1 indicam que o modelo mantém alta precisão mesmo com alto *recall*, enquanto valores baixos sugerem que o modelo sofre para equilibrar essas duas dimensões (Saito; Rehmsmeier, 2015).

$$\text{PR-AUC} = \int_0^1 \text{Precisão}(\text{Recall})\, d(\text{Recall}) \tag{3.10}$$

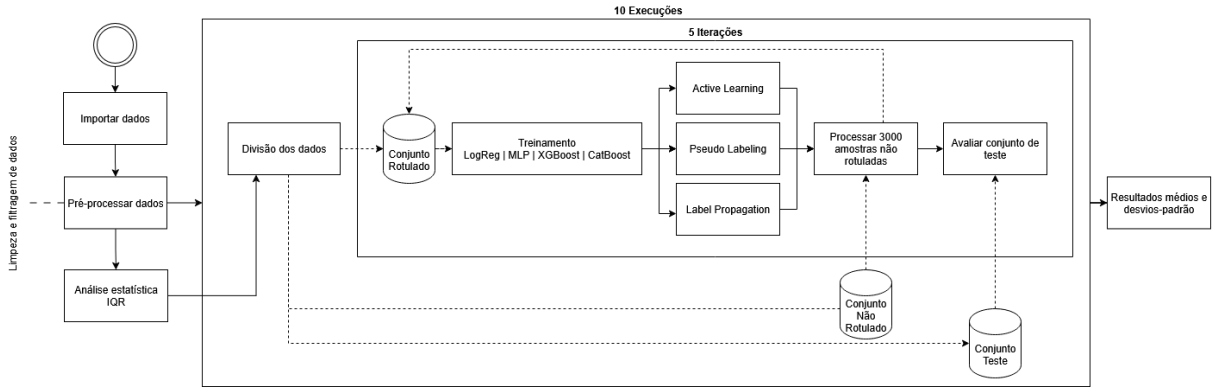
Em cenários altamente desbalanceados, a ROC-AUC pode ser pouco informativa, pois não reflete diretamente a relação entre precisão e *recall* da classe minoritária. Por isso, neste trabalho, F1 e PR-AUC recebem maior ênfase na comparação entre modelos.

4 METODOLOGIA

Este capítulo apresenta a metodologia adotada para o desenvolvimento e avaliação do sistema de detecção de fraudes em cenário de rótulos escassos. O processo combina diferentes etapas, a preparação e exploração inicial dos dados, definição de rótulos a partir de dois racionais (denúncias e anomalias estatísticas), amostragem estratificada para criação dos conjuntos, aplicação de estratégias semi-supervisionadas iterativas com orçamento controlado, ajuste de limiar orientado à métrica F1 e repetição de todo o ciclo experimental para estimar desempenho médio e variabilidade.

A Figura 4.1 sintetiza esse processo em um fluxograma, destacando o fluxo de dados desde a extração no ERP até a avaliação final dos modelos. As etapas apresentadas serão detalhadas ao longo deste capítulo.

Figura 4.1 – Fluxograma da metodologia aplicada



Fonte: Do autor (2025)

4.1 Ambiente de implementação

Neste trabalho, foram utilizados dados reais extraídos do ERP da empresa do setor logístico. As consultas SQL foram executadas em servidor local, com os resultados sendo exportados para análise em ambiente externo.

As etapas de preparação, modelagem e avaliação dos dados foram implementadas através da IDE (*Integrated Development Environment*) VSCode na linguagem Python, escolhida por sua ampla adoção em ciência de dados e pela disponibilidade de bibliotecas para aprendizado de máquina, manipulação de dados e visualização. Os principais pacotes empregados foram:

- **Pandas e NumPy**: manipulação, estruturação e transformação dos dados tabulares;
- **Polars**: manipulação de grandes volumes de dados com foco em desempenho;
- **Scikit-Learn**: implementação de modelos de regressão logística e MLP, além de ferramentas de pré-processamento, imputação e cálculo de métricas de desempenho;
- **XGBoost e CatBoost**: treinamento de modelos baseados em árvores de decisão;
- **Matplotlib**: geração de gráficos e visualização dos resultados experimentais.

Os experimentos foram executados em um computador equipado com processador AMD Ryzen 7 5700X3D 8-Core (3.0 GHz), 32 GB de memória RAM, SSD de 1 TB e GPU NVIDIA RTX 4070, com sistema operacional Windows 11 (64 bits). Esse ambiente proporcionou capacidade suficiente para o processamento dos conjuntos de dados (aproximadamente 170 mil registros) e para a execução repetida dos experimentos sem comprometer o tempo de simulação. Para assegurar reprodutibilidade, as execuções foram controladas por sementes aleatórias fixadas em cada repetição.

4.2 Extração de dados

A empresa do setor logístico rodoviário utiliza o ERP Protheus (TOTVS)[1] com armazenamento em servidor local. Para evitar sobrecarga no ambiente transacional, foi criada uma consulta SQL que seleciona apenas os campos relevantes e materializa o resultado em um *dataflow* no Power BI[2].

A base histórica abrange lançamentos contábeis desde 2017, incluindo compras, serviços, movimentações de estoque, ativos, aluguéis e eventos, totalizando mais de 3,5 milhões de linhas e 53 colunas. Neste estudo, o recorte selecionado contempla serviços adquiridos entre abril de 2023 e maio de 2025, considerando os dez mais frequentemente comprados. Essa filtragem buscou garantir quantidade suficiente de observações por item e evitar distorções causadas por produtos raros. O conjunto final resultou em aproximadamente 170 mil registros, empregados nas etapas de preparação, extração de *features* e modelagem. A classe positiva corresponde

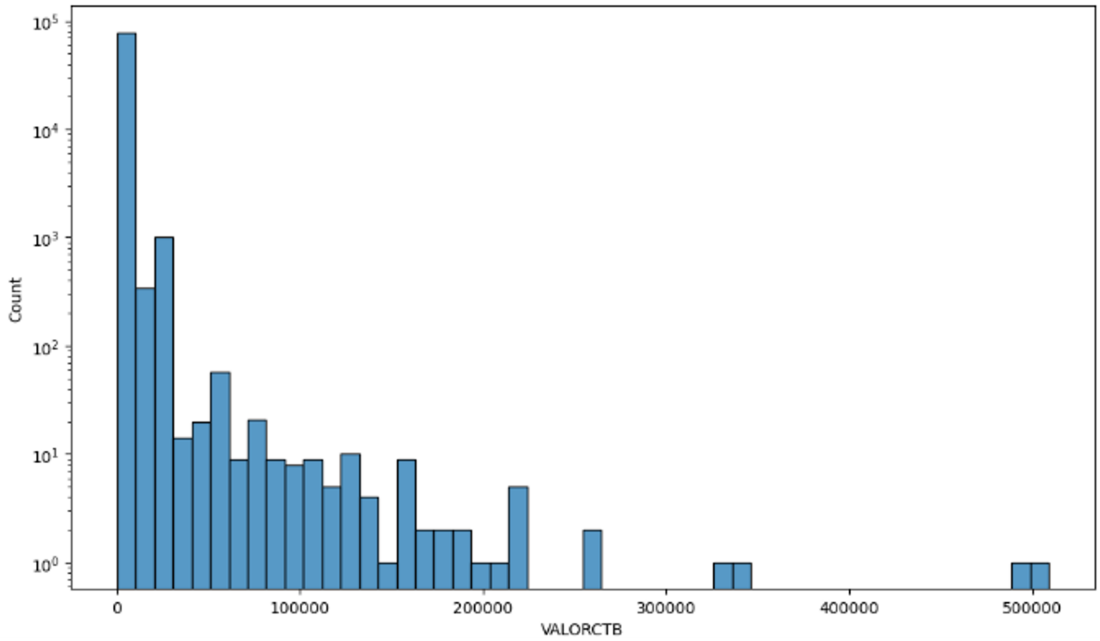
[1] TOTVS é uma empresa brasileira especializada em softwares de gestão empresarial. Mais informações em: https://www.totvs.com/

[2] Power BI é uma ferramenta de análise e visualização de dados da Microsoft. Mais informações em: https://powerbi.microsoft.com

às compras marcadas como fraudulentas pelo canal interno de denúncias, totalizando 3.935 ocorrências.

Na etapa de exploração inicial dos dados, foram avaliados fatores como a distribuição das variáveis numéricas, quantidade de categorias distintas existentes no conjunto, presença de valores nulos e demais padrões descritivos dos dados. As Figuras 4.2 e 4.3 apresentam as distribuições de valor e quantidade em escala logarítmica, que revelam forte assimetria à direita, a maior parte das observações concentra-se em valores baixos, enquanto apenas poucas ocorrências assumem valores extremamente elevados. Esse comportamento motivou a adoção de normalização e o uso de métricas robustas, de forma a reduzir os efeitos de distorções causadas por *outliers*.

Figura 4.2 – Distribuição logarítmica de valor contábil



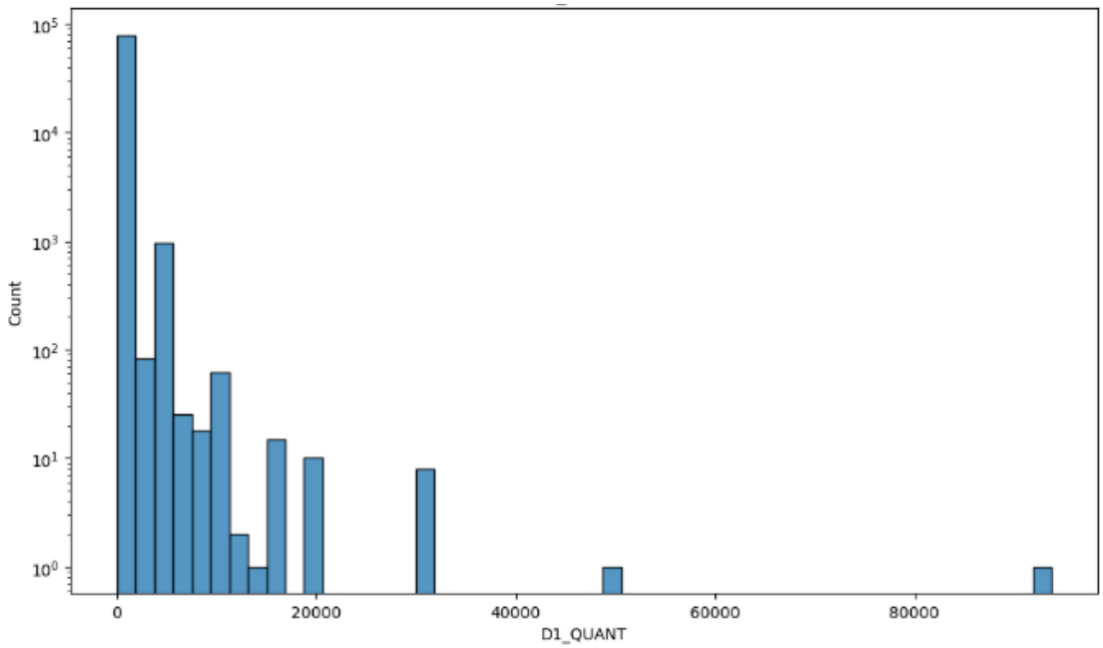
Fonte: Do autor (2025)

A base de dados utilizada contém registros de pedidos associados a indicadores de anomalia e denúncias de fraude. A partir da data de digitação foram derivadas variáveis de calendário (ano, mês, dia, dia da semana) e suas transformações cíclicas, utilizadas como preditores.

4.3 Pré-processamento de dados

O pré-processamento envolveu diversas etapas complementares. Inicialmente, foram padronizadas as *strings*, com remoção de acentos e conversão para letras minúsculas, evitando

Figura 4.3 – Distribuição logarítmica de quantidade



Fonte: Do autor (2025)

distinções artificiais entre valores equivalentes. Na sequência, foi realizada a conversão de tipos e a extração de variáveis temporais, como ano, mês e dia da semana. Também foram definidos grupos de operação e aplicado um controle de qualidade, que incluiu a deduplicação (remoção de registros repetidos) e a validação de domínios, ou seja, a verificação de que os valores de cada variável se encontram dentro dos intervalos ou categorias permitidos.

As variáveis categóricas foram codificadas e os atributos numéricos normalizados, garantindo consistência para diferentes algoritmos de modelagem. Modelos como XGBoost e CatBoost lidam bem com escalas heterogêneas, mas em abordagens lineares a padronização é essencial para a estabilidade do treinamento.

Além disso, foram criadas variáveis temporais em formato cíclico, como `mes_sin`, `mes_cos`, `week_sin` e `week_cos`, de modo a representar sazonalidades sem rupturas artificiais. A seleção das *features* considerou tanto o conhecimento dos especialistas da empresa quanto uma análise de autocorrelação, com o objetivo de reduzir a multicolinearidade e manter apenas variáveis mais independentes, aumentando a robustez e a interpretabilidade dos modelos.

O conjunto final foi então estruturado em variáveis numéricas e categóricas, apresentadas nas Tabelas 4.1 e 4.2, respectivamente. Para ilustrar a aplicação dessas variáveis na prática, a Figura 4.4 exibe exemplos reais de registros de compra de serviços que compuseram a base de dados do estudo.

Figura 4.4 – Exemplo de registros de compra de serviços

D1_FILIAL	D1_DTDIGIT	D1_QUANT	PRODUTO	VALORCTB	MODELO_VEICULO	freq_desvio_modelo	freq_mediana_modelo	val_mediana_modelo	val_desvio_modelo	date_dias_intervalo
02	11/02/2025	1	Serv. A	R$ 600,00	CAMINHAO VW 17.190 WORKER 2014	0,98	1	R$ 600,00	R$ 1.906,37	54
09	10/03/2025	1	Serv. B	R$ 400,00	CM VOLVO FH13 520 TRACADO	2,63	4	R$ 300,00	R$ 240,05	0
61	26/08/2024	1	Serv. B	R$ 100,00	CM VOLVO FH 540 6X4T 2022 2022	2,93	6	R$ 280,00	R$ 188,83	18
09	01/08/2023	1	Serv. C	R$ 550,00	CM VOLVO FMX480 6X4T TRACADO	2,13	3	R$ 313,01	R$ 227,19	67
09	09/01/2024	1	Serv. C	R$ 327,59	CM VOLVO FMX480 6X4T TRACADO	2,13	3	R$ 313,01	R$ 227,19	161

Fonte: Do autor (2025)

Tabela 4.1 – Variáveis numéricas selecionadas

Variável	**Descrição**
`D1_QUANT`	Quantidade de produtos/serviços
`VALORCTB`	Valor contábil do pedido
`freq_desvio_modelo`	Desvio da frequência quanto à mediana por modelo de veículo
`freq_mediana_modelo`	Frequência mediana por modelo de veículo
`val_mediana_modelo`	Valor mediano por modelo de veículo
`val_desvio_modelo`	Desvio do valor quanto à mediana por modelo de veículo
`date_dias_intervalo`	Intervalo em dias entre serviços
`dt_ano`	Ano do registro
`dt_dia`	Dia do mês
`mes_sin`, `mes_cos`	Representação cíclica do mês
`week_sin`, `week_cos`	Representação cíclica da semana

Fonte: Do autor (2025)

Tabela 4.2 – Variáveis categóricas selecionadas

Variável	**Descrição**
`D1_FILIAL`	Filial da empresa
`PRODUTO`	Nome do produto ou serviço adquirido
`MODELO_VEICULO`	Modelo do veículo vinculado ao pedido

Fonte: Do autor (2025)

Adicionalmente, foi implementado um módulo de detecção preliminar de anomalias baseado em estatística descritiva, com o objetivo de fornecer um primeiro filtro analítico antes das etapas mais sofisticadas do pipeline de auditoria. Para isso, utilizou-se o método IQR (*Interquartile Range*), adotando-se um fator de multiplicação igual a 6. Esse fator ampliado foi definido de forma conservadora, buscando minimizar a geração de falsos positivos em um con-

texto no qual a variabilidade natural entre modelos de veículos e tipos de serviços poderia levar a detecções excessivamente sensíveis. Assim, apenas valores substancialmente discrepantes foram classificados como potenciais anomalias.

A análise considerou dois critérios principais:

- Frequência de serviços por veículo: para cada modelo, calculou-se a mediana do número de serviços realizados por veículo dentro do período analisado. Veículos cuja frequência observada ultrapassou o limite estatístico definido pelo IQR foram sinalizados por apresentarem um comportamento atípico em relação ao padrão de sua própria categoria.

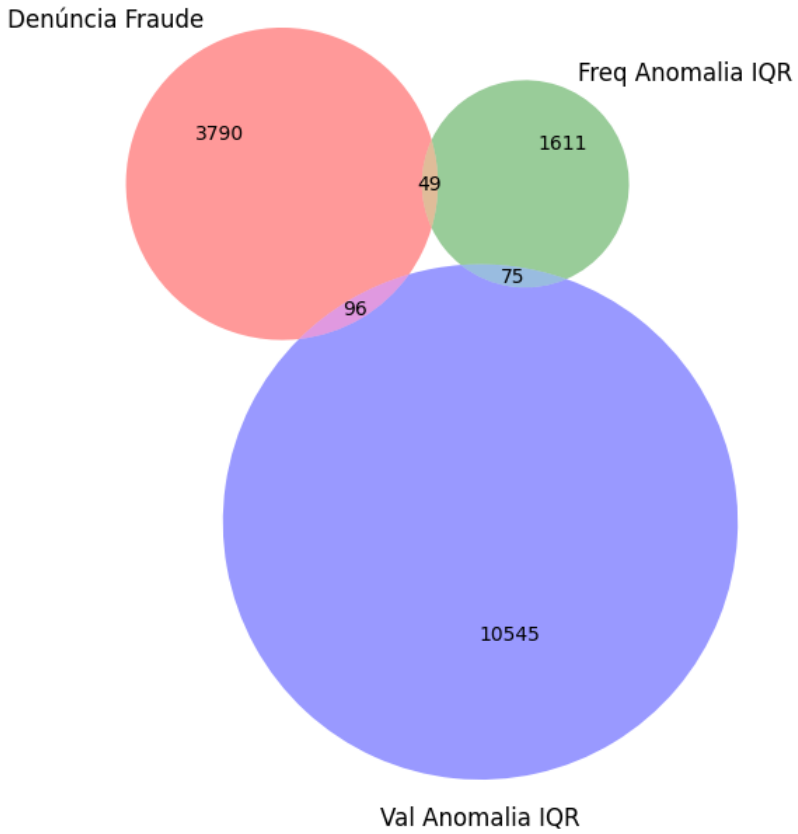
- Valores unitários de serviços: avaliou-se também o valor monetário unitário de cada serviço, novamente utilizando a mediana por modelo de veículo como referência. Valores acima do limite superior do IQR foram classificados como suspeitos, por indicarem possíveis distorções de custos, inconsistências de registro ou serviços inadequadamente precificados.

A Figura 4.5 apresenta a intersecção entre os rótulos gerados por esses critérios e os rótulos previamente existentes de "denúncia", evidenciando tanto a sobreposição quanto discrepâncias entre suspeitas identificadas empiricamente e aquelas reportadas manualmente.

Essa análise estatística não se limitou à identificação de meros *outliers* numéricos. Ao contrário, funcionou como uma camada inicial de detecção de anomalias nos registros, oferecendo uma visão quantitativa que subsidia a auditoria interna e contribui para priorização de casos. Dessa forma, o módulo estatístico atua como um mecanismo de triagem, aumentando a eficiência dos auditores ao destacar observações que, pela sua magnitude ou frequência, merecem investigação mais aprofundada.

Por fim, o conjunto de dados foi dividido em três partes: treino inicial com rótulos positivos e negativos confiáveis, conjunto não rotulado e conjunto de teste com fraudes rotuladas. A separação utilizou amostragem estratificada por classe (fraude/negativo). Em cada iteração, uma quantidade fixa de novas instâncias de dados é transferida do conjunto não rotulado para a base rotulada. Cada modelo evoluiu de forma independente, mantendo seu próprio conjunto rotulado e não rotulado, sem interferência entre abordagens.

Figura 4.5 – Intersecção entre rótulos de denúncia e anomalias estatísticas (quantidade de amostras)



Fonte: Do autor (2025)

4.4 Modelos

Foram avaliados quatro modelos: Regressão Logística, XGBoost, MLP e CatBoost. Para os modelos de Regressão Logística e MLP, da biblioteca Scikit-learn foi implementada a padronização de atributos numéricos e binarização para atributos categóricos por *one-hot encoding*. No XGBoost, as variáveis categóricas foram previamente codificadas no pré-processamento. Já para o CatBoost, as variáveis categóricas foram tratadas nativamente, com imputação simples e indicação explícita das colunas categóricas, sem a codificação binária. Para evitar vazamento de dados, o pré-processador foi ajustado somente no conjunto rotulado da iteração vigente e, depois, aplicado ao conjunto não rotulado e ao conjunto de teste.

Os modelos utilizados foram configurados com os seguintes hiperparâmetros principais, os demais foram configurados com seus valores padrões:

- **Regressão Logística**
  - *class_weight* = “balanced” (compensação do desbalanceamento de classe)
  - *solver* = “liblinear” (otimizador adequado para datasets menores)
  - *max_iter* = 1000 (iterações máximas para convergência)
- **XGBoost**
  - *n_estimators* = 800 (número de árvores)
  - *max_depth* = 6 (profundidade máxima das árvores)
  - *learning_rate* = 0.05 (taxa de aprendizado)
  - *subsample* = 0.9 (razão da proporção das instancias de treinamento)
  - *colsample_bytree* = 0.8 (controle de overfitting via amostragem)
  - *reg_lambda* = 2.0 (regularização l2)
  - *reg_alpha* = 0.0 (regularização l1)
  - *tree_method* = “hist” (otimização de construção das árvores)
  - *eval_metric* = “logloss” (métrica de avaliação interna)
- **Perceptron Multicamadas (MLP)**
  - *hidden_layer_sizes = (512, 256)*
  - *activation* = “relu” (função de ativação)
  - *alpha* = 1e-3 (regularização l2)
  - *solver* = “adam”
  - *batch_size* = 256
  - *learning_rate_init* = 0.001
  - *max_iter* = 300 (iterações máximas)
  - *early_stopping* = True *validation_fraction* = 0.1 (parada antecipada com 10% dos dados para validação)

- **CatBoost**
  - *iterations* = 1000 (número de árvores)
  - *depth* = 8 (profundidade máxima)
  - *learning_rate* = 0.1
  - *l2_leaf_reg* = 5.0 (regularização l2 nas folhas)
  - *loss_function* = "Logloss"
  - *eval_metric* = "AUC"
  - *od_type* = "Iter"
  - *od_wait* = 50 (early stopping após 50 iterações sem melhora)

4.5 Estratégias utilizadas

Em relação ao *Pseudo-Labeling* (PL), a cada iteração o modelo calcula a probabilidade de fraude para todas as amostras do conjunto e selecionam-se apenas as de alta confiança, muito próximas de 1 (provável fraude) ou de 0 (provável não fraude). Dentro do orçamento de cada iteração, priorizam-se esses extremos e busca-se manter equilíbrio entre as classes. As amostras escolhidas recebem pseudo-rótulos por regra: positivo se a probabilidade $\geq 0{,}9$ e falso se $\leq 0{,}1$. Finalmente, são movidas do conjunto não rotulado para o conjunto rotulado da próxima iteração, passando a integrar o treino subsequente.

No *Active Learning* (AL), adotou-se *uncertainty sampling* em lote. A cada iteração, são selecionadas as amostras em que o modelo está mais incerto, aquelas com probabilidade prevista próxima de 50%. Em seguida, utiliza-se uma simulação de auditor, que revela o rótulo real presente na base de dados como se um especialista tivesse analisado o caso. A adoção desse auditor simulado permite reproduzir o fluxo real do AL sem envolver um auditor humano a cada iteração, evitando custos operacionais e tornando o experimento viável. Os rótulos revelados são então incorporados ao conjunto rotulado da próxima iteração, permitindo o aprimoramento progressivo do modelo.

No *Label Propagation* (LP), os dados foram primeiro transformados em vetores numéricos usando o mesmo pré-processamento aplicado aos demais modelos. Em seguida, aplicou-se o algoritmo Label Spreading do Scikit-learn, baseado em k-vizinhos (k = 10), kernel "knn" e

com parâmetro de suavização $\alpha = 0{,}1$, sobre o conjunto que reúne exemplos rotulados e não rotulados. O método estima, para cada exemplo do conjunto, a probabilidade de ser fraude. Com essas probabilidades, são selecionados até o limite do orçamento, os casos de maior confiança, ou seja, com probabilidades mais distantes de 50%. Esses casos recebem pseudo-rótulos conforme a regra, probabilidade $\geq 50\%$ = positivo e $< 50\%$ = negativo, e são inclusos no conjunto de dados de treinamento para a iteração seguinte.

Cada estratégia foi aplicada durante 5 iterações sucessivas, adicionando até 3.000 novas amostras por iteração, de forma independente para cada modelo. Ao final de cada iteração, os modelos foram reavaliados em um conjunto de teste fixo.

4.6 Avaliação e Métricas

O desempenho dos modelos foi avaliado utilizando os seguintes indicadores, computados no conjunto de teste ao final de cada iteração:

- **F1-Score**: equilíbrio entre precisão e *recall*, sendo a principal métrica usada para comparação.
- **ROC AUC**: medida da capacidade de separação entre classes.
- **PR-AUC**: área sob a curva de precisão-recall, relevante em contextos desbalanceados.
- **Recall**: taxa de detecção de fraudes.

O limiar de decisão para cada modelo foi ajustado com base na maximização do F1-score em um conjunto de validação (20% dos dados de treino). Essa separação de validação foi realizada por amostragem estratificada a cada repetição.

4.7 Repetições e Análise Estatística

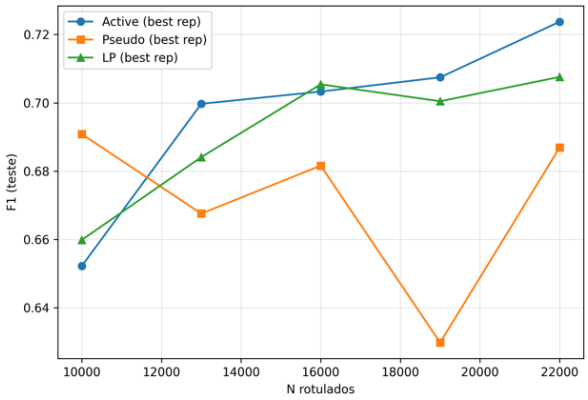
O experimento completo foi repetido 10 vezes com diferentes sementes de aleatoriedade, totalizando 120 execuções (4 modelos × 3 estratégias × 10 repetições). As métricas finais reportadas correspondem à média e desvio-padrão da última iteração de cada repetição, permitindo análise de estabilidade e robustez dos métodos.

5 RESULTADOS E DISCUSSÃO

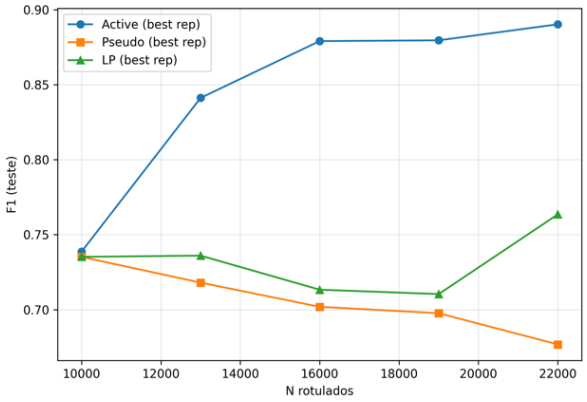
Esta seção apresenta os resultados e suas principais interpretações, comparando as estratégias Pseudo-Labeling (PL), Active Learning (AL) e Label Propagation (LP). São mostradas as curvas de F1-score ao longo das iterações, as métricas da melhor execução, os valores médios com desvios-padrão e as implicações práticas para auditoria interna.

A Figura 5.1 apresenta a evolução do F1-score no conjunto de teste, considerando a melhor execução (maior F1 na última iteração dentre as 10 repetições), à medida que novos exemplos são incorporados pelas diferentes estratégias.

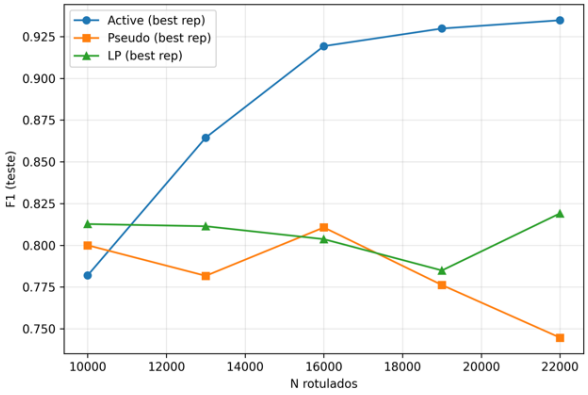
Figura 5.1 – F1-score (teste) ao longo das iterações por modelo base, na melhor execução, comparando PL, AL e LP.



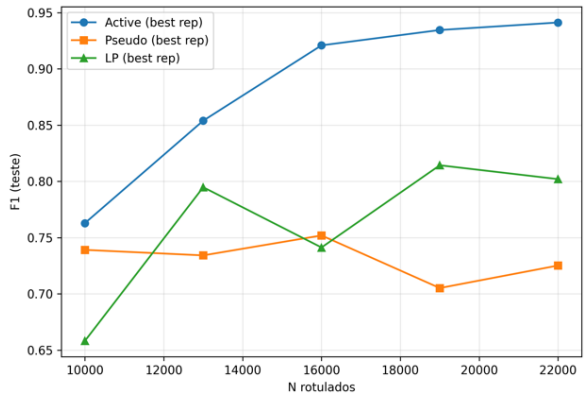
Logistic Regression



MLP



XGBoost



CatBoost

Fonte: Do autor (2025)

É possível observar que o *Active Learning* apresenta crescimento consistente em todas as arquiteturas, com destaque para XGBoost e CatBoost, que alcançam os maiores valores de F1. A MLP evolui de forma estável até se aproximar ao F1 de 0,9 e a Regressão Logística, em patamar inferior, também segue essa tendência positiva.

Por outro lado, o *Pseudo-Labeling* apresenta queda progressiva de desempenho, possível reflexo da propagação de ruído nos pseudo-rótulos. Entretanto, na Regressão Logística, observa-se maior variação dos resultados ao longo das iterações, embora sem ganhos consistentes. Já o *Label Propagation* mantém-se em nível intermediário entre as duas estratégias, aproximando-se do *Active Learning* na Regressão Logística e, de forma geral, superando o *Pseudo-Labeling*, ainda que sem apresentar evolução clara ao longo das iterações.

Na Tabela 5.1, que reúne os resultados da melhor execução, nota-se que CatBoost–AL (F1 = 0,941, *Recall* = 0,960) e XGBoost–AL (F1 = 0,935, *Recall* = 0,960) apresentam os maiores valores entre os modelos, seguidos pelo MLP–AL (F1 = 0,890, *Recall* = 0,908). O *Pseudo-Labeling* atinge valores elevados de *recall* (até 0,980), mas com F1 consideravelmente inferior, indicando desequilíbrio entre precisão e *recall*.

Tabela 5.1 – Métricas no teste para a melhor execução (maior F1 na última iteração) por modelo e estratégia

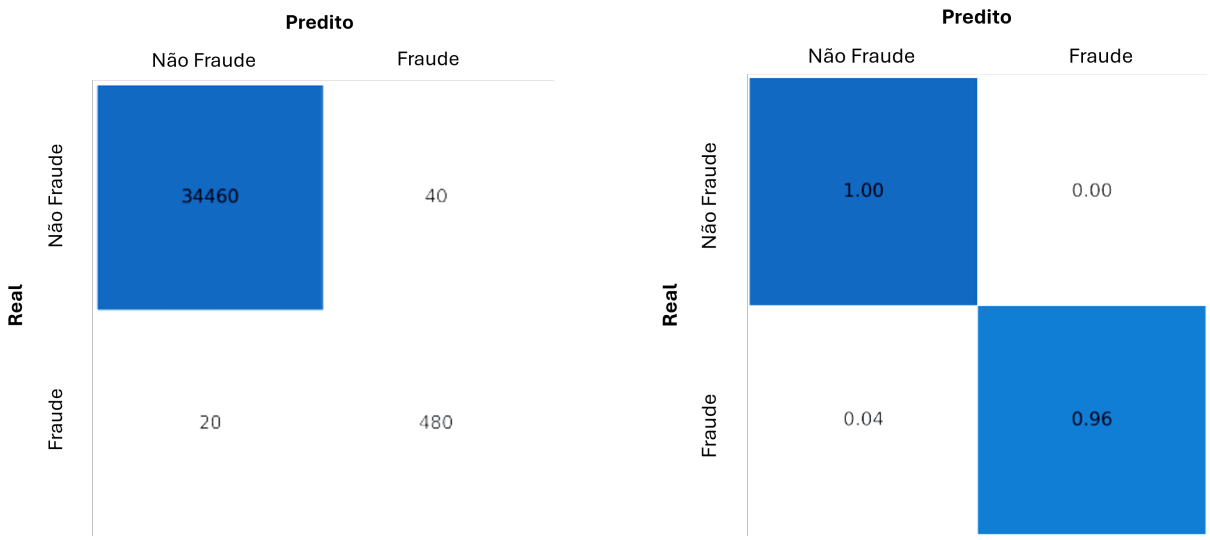
Modelo	Estratégia	F1	Recall	PR-AUC	ROC AUC
CatBoost	Active	**0,941**	0,960	**0,981**	**1,000**
	LP	0,802	0,964	0,869	0,998
	Pseudo	0,725	**0,966**	0,769	0,997
LogReg	Active	**0,724**	0,930	0,552	0,994
	LP	0,708	0,934	**0,622**	**0,995**
	Pseudo	0,687	**0,974**	0,599	0,995
MLP	Active	**0,890**	0,908	**0,922**	**0,999**
	LP	0,763	0,920	0,790	0,997
	Pseudo	0,677	**0,974**	0,736	0,996
XGBoost	Active	**0,935**	0,960	**0,980**	**1,000**
	LP	0,819	0,960	0,910	0,999
	Pseudo	0,745	**0,980**	0,858	0,998

Fonte: Do autor (2025)

As matrizes de confusão apresentadas na Figura 5.2 detalham os resultados do CatBoost. No AL, o número de falsos negativos foi reduzido (20 ocorrências), resultando em *recall* de 96%, enquanto os falsos positivos foram baixos (40 casos), refletindo em F1 de 0,941. No LP,

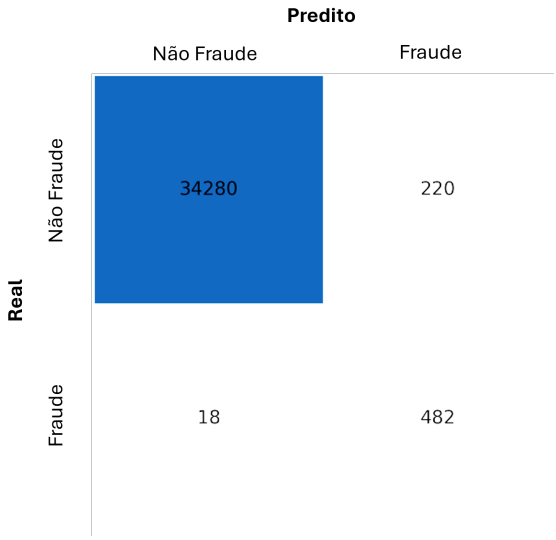
o *recall* manteve-se em patamar similar em 96,4%, mas com maior número de falsos positivos (220 casos), reduzindo o F1 para 0,802.

Figura 5.2 – Matrizes de confusão do CatBoost nas estratégias AL e LP, apresentadas em contagem absoluta e percentual, referentes à melhor execução

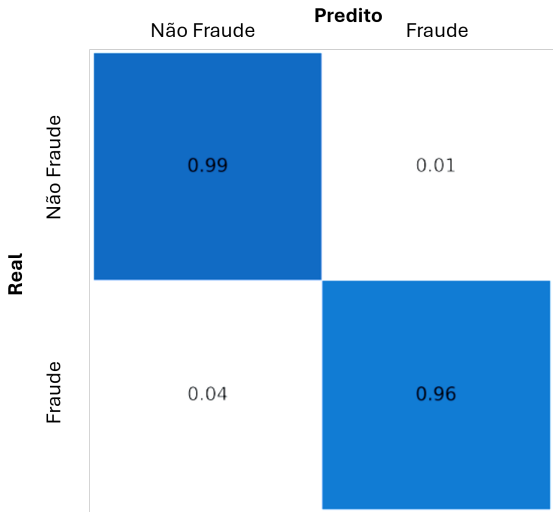


CatBoost-AL Contagem

CatBoost-AL Percentual



CatBoost-LP Contagem



CatBoost-LP Percentual

Fonte: Do autor (2025)

A Tabela 5.2 e Figura 5.3 apresentam os resultados médios de desempenho, com respectivos desvios-padrão, obtidos na fase de teste para cada modelo e estratégia. Embora o ROC AUC se mantenha praticamente saturado em todos os cenários, o detalhamento de F1 e recall evidencia diferenças entre os modelos e estratégias. A estratégia de *Active Learning* obteve os melhores valores médios de F1 em todos os modelos, com destaque para o CatBoost (F1 = 0,930, *Recall* = 0,958) e o XGBoost (F1 = 0,922, *Recall* = 0,959).

Tabela 5.2 – Métricas de desempenho no teste (média ± desvio padrão) por modelo e estratégia

Modelo	Estratégia	F1	Recall	PR-AUC	ROC AUC
CatBoost	Active	**0,930 ± 0,009**	0,958 ± 0,008	**0,975 ± 0,006**	**1,000 ± 0,000**
	LP	0,762 ± 0,034	0,970 ± 0,018	0,857 ± 0,026	0,998 ± 0,000
	Pseudo	0,690 ± 0,026	**0,982 ± 0,009**	0,783 ± 0,049	0,997 ± 0,001
XGBoost	Active	**0,922 ± 0,011**	0,959 ± 0,010	**0,978 ± 0,003**	**0,999 ± 0,000**
	LP	0,783 ± 0,028	0,974 ± 0,008	0,889 ± 0,017	0,998 ± 0,001
	Pseudo	0,718 ± 0,019	**0,981 ± 0,008**	0,814 ± 0,028	0,997 ± 0,001
MLP	Active	**0,863 ± 0,021**	0,915 ± 0,014	**0,922 ± 0,013**	**0,999 ± 0,000**
	LP	0,717 ± 0,028	0,948 ± 0,019	0,760 ± 0,044	0,997 ± 0,000
	Pseudo	0,633 ± 0,031	**0,977 ± 0,005**	0,723 ± 0,031	0,996 ± 0,000
LogReg	Active	**0,708 ± 0,009**	0,935 ± 0,022	0,565 ± 0,015	0,994 ± 0,000
	LP	0,691 ± 0,012	0,921 ± 0,023	**0,594 ± 0,028**	**0,994 ± 0,001**
	Pseudo	0,646 ± 0,034	**0,968 ± 0,007**	0,565 ± 0,035	0,994 ± 0,001

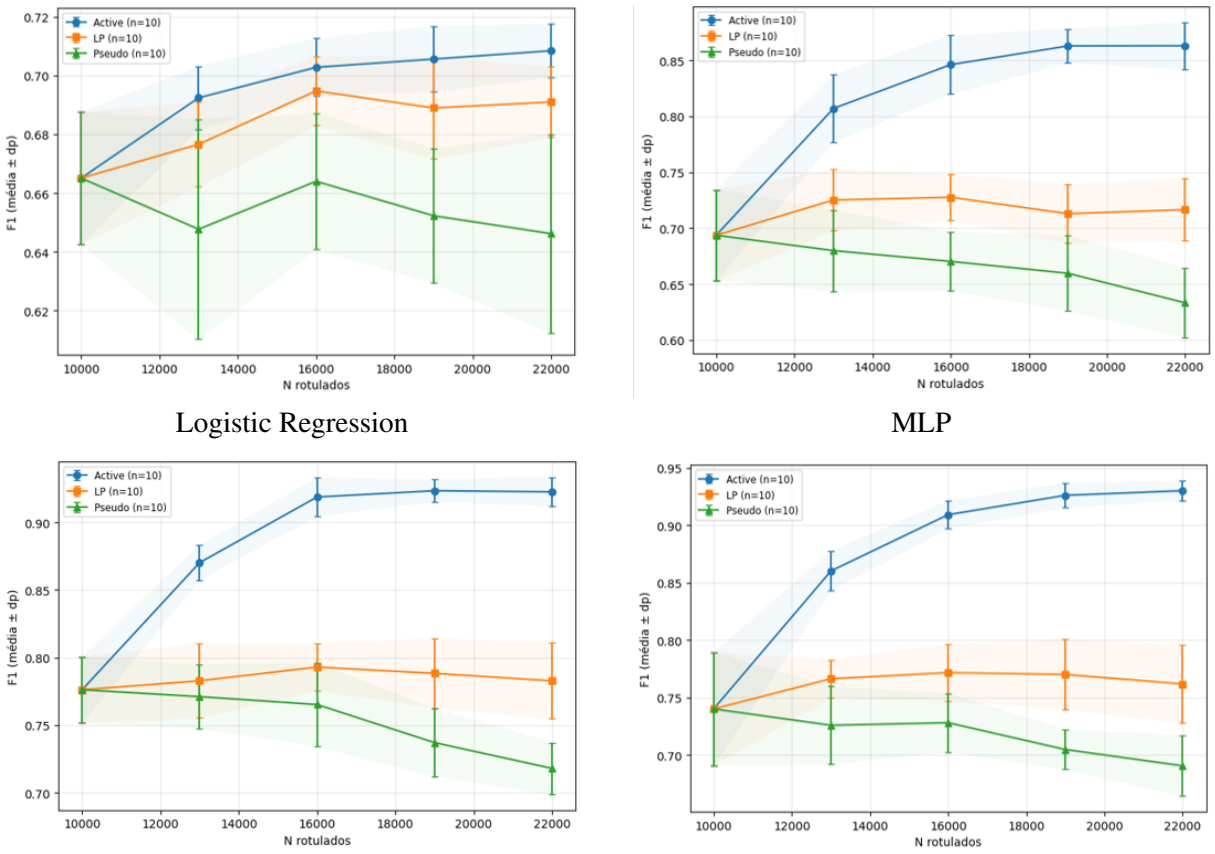
Fonte: Do autor (2025)

O *Pseudo-Labeling* apresentou os maiores valores médios de *recall*, alcançando até 0,982 no CatBoost e 0,981 no XGBoost, mas às custas de quedas expressivas no F1. Isso sugere maior propensão a falsos positivos, o que compromete a utilidade prática em auditoria interna, aumentando o volume de casos a serem auditados. O *Label Propagation*, por sua vez, manteve desempenho intermediário, garantindo *recalls* elevados e F1 moderados, posicionando-se de forma mais consistente que o *Pseudo-Labeling*, mas sem atingir os patamares do *Active Learning*.

Apesar dos altos desempenhos, observa-se que os desvios-padrão do F1 na estratégia *Active Learning* são ligeiramente superiores aos das demais abordagens, sinalizando maior sensibilidade a variações nos dados ou nas inicializações dos modelos.

De forma geral, a estratégia de *Active Learning* (AL) apresentou os maiores valores de F1 em todas as arquiteturas, com destaque para CatBoost e XGBoost, que superaram 0,93, mantendo *recalls* elevados em torno de 96%. Esse comportamento indica que o AL conseguiu equilibrar detecção abrangente de fraudes com baixo número de falsos positivos, o que se refletiu em métricas mais robustas. Contudo, o principal fator a ser considerado é que esse ganho

Figura 5.3 – F1-score (média ± desvio padrão) ao longo das iterações por modelo base comparando PL, AL e LP.



Logistic Regression | MLP

XGBoost | CatBoost

Fonte: Do autor (2025)

depende da participação de um auditor para validar os exemplos selecionados, o que implica maior custo operacional e necessidade de recursos humanos qualificados.

Já o *Pseudo-Labeling*, apesar de alcançar os maiores valores de *recall* em alguns cenários (até 98% no CatBoost e no XGBoost), o F1 foi consideravelmente inferior, revelando grande volume de falsos positivos. Esse resultado reforça a natureza dessa estratégia, que tende a propagar rótulos incorretos e aumentar a sensibilidade do modelo às instâncias suspeitas, ainda que em detrimento da precisão. Em um contexto prático, esse comportamento se traduz em maior número de casos a serem revisados, o que pode sobrecarregar a auditoria e reduzir a eficiência do processo. *Label Propagation* manteve desempenho intermediário entre as duas abordagens, com *recalls* elevados (acima de 92% em média) e valores de F1 que superaram o Pseudo-Labeling, mas não atingiram os patamares do *Active Learning*.

Outro aspecto importante é a análise das métricas utilizadas que, embora o ROC AUC tenha se mantido próximo de 1,0 em todos os casos, essa métrica mostrou-se pouco informativa

no contexto, já que não diferenciou adequadamente os métodos. As comparações mais significativas emergiram a partir da análise conjunta de F1 e *recall*, métricas que refletem melhor a realidade de auditoria interna. O F1 indicou a capacidade de manter equilíbrio entre precisão e abrangência, enquanto o *recall* mostrou-se essencial para assegurar que fraudes não fossem ignoradas. Nesse sentido, o AL mostrou-se mais consistente, ainda que com desvios-padrão ligeiramente superiores.

Durante a análise estatística dos dados, especialmente pela identificação de *outliers* via IQR, foram identificadas incongruências nos processos de compras, como a vinculação de vários serviços a um único veículo no sistema ou a seleção de itens incorretos apenas para agilizar o fluxo de aquisição. Embora tais inconsistências não configurem necessariamente fraudes, comprometem a qualidade da base e impactam a performance dos modelos, gerando maior número de falsos positivos.

A interpretação dos resultados deve considerar não apenas o desempenho estatístico, mas também a viabilidade operacional. O *Active Learning* apresentou métricas superiores, mas exige participação humana para rotulação iterativa, o que pode ser custoso em ambientes com grande volume de transações. Já o Pseudo-Labeling e o *Label Propagation*, embora apresentem desempenho inferior em F1, dispensam esse esforço manual, podendo ser aplicados em cenários com restrições de recursos humanos. Dessa forma, a escolha da estratégia mais adequada dependerá do equilíbrio entre acurácia técnica e viabilidade de implementação.

6 CONCLUSÃO

O presente trabalho investigou estratégias de aprendizado semi-supervisionado aplicadas à detecção de fraudes em registros contábeis e transações corporativas, com foco em cenários de escassez de rótulos, comuns na prática de auditoria interna. Foram comparados os métodos de *Active Learning*, *Pseudo-Labeling* e *Label Propagation*, avaliados em diferentes modelos de classificação, incluindo algoritmos baseados em árvores de decisão (CatBoost e XGBoost), redes neurais (MLP) e regressão logística.

Como contribuição, o estudo demonstrou o potencial do uso de métodos semi-supervisionados para ampliar a capacidade de identificação de anomalias em bases com poucos rótulos disponíveis, fornecendo subsídios para aplicações práticas em ambientes corporativos. A abordagem explorada contribui ainda para reduzir o esforço humano em auditoria, ao direcionar a análise para casos mais relevantes.

Além disso, a investigação estatística trouxe ganhos adicionais ao revelar incongruências nos processos de compras, evidenciando o papel da ciência de dados não apenas na detecção de fraudes, mas também como instrumento de diagnóstico organizacional. A empresa já planeja retificações nesses procedimentos, o que deverá reduzir falsos positivos e aumentar a confiabilidade da solução.

Entre as limitações do presente estudo, destaca-se o fato de ter sido conduzido a partir de dados de uma única organização, o que restringe a abrangência dos achados, já que os padrões de fraude e anomalia podem variar entre empresas distintas do mesmo setor. Soma-se a isso a escolha metodológica de avaliar apenas três estratégias de aprendizado (*Active Learning*, *Pseudo-Labeling* e *Label Propagation*) e quatro modelos (MLP, Regressão Logística, XGBoost e CatBoost), o que restringe a análise a essas configurações adotadas.

Outro aspecto relevante é a qualidade dos dados fornecidos, que pode ter sido afetada por inconsistências nos processos de compras, erros de registro ou práticas operacionais pouco padronizadas, introduzindo ruído que impacta diretamente o desempenho dos modelos. Além disso, a confirmação das fraudes selecionadas foi limitada, uma vez que não havia processos de validação totalmente consolidados para todos os casos. Essa condição restringe a robustez da validação, mas reflete um desafio comum em estudos aplicados em ambientes corporativos reais, nos quais a construção de rotinas de auditoria e governança tende a evoluir de forma contínua.

Diante dessas limitações, estudos futuros podem avançar em diferentes direções. Uma possibilidade é ampliar a diversidade de técnicas avaliadas, incorporando métodos como regularização por consistência, autoencoders ou estratégias semi-supervisionadas híbridas, que conciliem a escalabilidade do *Pseudo-Labeling* com a precisão do *Active Learning*. Outra frente relevante é a aplicação em múltiplos conjuntos de dados de diferentes organizações, permitindo avaliar a generalização dos resultados para além do padrão específico observado neste estudo. Além disso, investigações futuras podem explorar mecanismos de melhoria da qualidade dos dados, como integração com processos de governança e validação contínua de registros, de modo a mitigar ruídos decorrentes de inconsistências operacionais. Por fim, recomenda-se o desenvolvimento de estudos em parceria com áreas de auditoria e compliance, voltados a estruturar procedimentos sistemáticos de confirmação das fraudes identificadas, possibilitando mensurar com maior clareza o impacto prático das técnicas na detecção de irregularidades.

Em síntese, os resultados alcançados reforçam que, embora o *Active Learning* tenha apresentado superioridade técnica, a escolha da estratégia mais adequada deve considerar o equilíbrio entre desempenho e custo de implementação. Evidencia-se, assim, a relevância do uso de métodos semi-supervisionados como ferramenta de apoio à detecção de fraudes, tanto pela capacidade de identificar anomalias quanto pelo potencial de induzir melhorias nos processos corporativos.

REFERÊNCIAS

ALAMPAY, R.; ABU, P. Non-parametric stochastic autoencoder model for anomaly detection. **International Journal of Advanced Computer Science and Applications**, v. 13, n. 5, p. 407–414, 2022. ISSN 2158-107X.

ALBRECHT, W. S.; ALBRECHT, C. C.; ALBRECHT, C. O. **Fraud Examination**. 4. ed. Mason, OH: Cengage Learning, 2011. ISBN 978-0538453361.

AN, Q. *et al.* A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges. **Sensors**, MDPI, v. 23, n. 9, p. 4178, 2023.

AROS, L. H. *et al.* Financial fraud detection through the application of machine learning techniques: a literature review. **Humanities and Social Sciences Communications**, v. 11, 2024.

BAY, S. *et al.* Large scale detection of irregularities in accounting data. **Proceedings - IEEE International Conference on Data Mining, ICDM**, p. 75–86, 12 2006.

BISHOP, C. M. **Neural networks for pattern recognition**. USA: Oxford university press, 1995.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006. ISBN 9780387310732.

BORGES, F. E. de M. *et al.* One-class classifier based on principal curves. **Neural Computing and Applications**, v. 35, n. 26, p. 19015–19024, Sep 2023. ISSN 1433-3058. Acesso em: 19 set. 2025. Disponível em: https://doi.org/10.1007/s00521-023-08721-8.

BRAUN, R. L. Auditors' professional skepticism and fraud detection: The role of a forensic mindset. **Current Issues in Auditing**, American Accounting Association, v. 9, n. 1, p. A1–A8, 2015.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, Oct 2001. ISSN 1573-0565. Acesso em: 19 set. 2025. Disponível em: https://doi.org/10.1023/A:1010933404324.

CARCILLO, F. *et al.* Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization. **International Journal of Data Science and Analytics**, v. 5, n. 4, p. 285–300, 2018.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM Comput. Surv.**, v. 41, 07 2009.

CHAPELLE, O.; SCHöLKOPF, B.; ZIEN, A. **Semi-Supervised Learning**. The MIT Press, 2006. Acesso em: 19 set. 2025. ISBN 9780262255899. Disponível em: https://doi.org/10.7551/mitpress/9780262033589.001.0001.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Acesso em: 19 set. 2025. Disponível em: https://doi.org/10.1145/2939672.2939785.

CHOI, J.; LIN, S.; WALKER, M. Machine learning techniques in auditing research and practice: Current trends and future opportunities. **Journal of Accounting Literature**, Elsevier, v. 42, p. 1–34, 2019.

DECHOW, P. M. *et al.* Predicting material accounting misstatements. **Contemporary Accounting Research**, Wiley Online Library, v. 28, n. 1, p. 17–82, 2011.

DEIST, T. *et al.* Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. **Medical Physics**, v. 45, p. 3449 – 3459, 2018.

DESAI, M.; SHAH, M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (mlp) and convolutional neural network (cnn). **Clinical eHealth**, v. 4, p. 1–11, 2021. ISSN 2588-9141. Acesso em: 19 set. 2025. Disponível em: https://www.sciencedirect.com/science/article/pii/S2588914120300125.

DIETTERICH, T. G. Machine learning. **ACM Comput. Surv.**, v. 28, p. 3, 1996.

DUTTON, M.; CONROY, G. V. **A Review of Machine Learning**. 1997. 341–367 p.

FARHAT, R. S. *et al.* **Floresta Aleatória Isotrópica: Um algoritmo de classificação**. Trabalho de Conclusão de Curso (TCC), Florianópolis, SC., 2023.

FERREIRA, P. J. S.; CARDOSO, J. M.; MENDES-MOREIRA, J. knn prototyping schemes for embedded human activity recognition with online learning. **Comput.**, v. 9, p. 96, 2020.

FRASCAROLI, B. **Utilização de redes neurais artificiais para classificação de ratings de risco soberano**. Trabalho de Conclusão de Curso (Graduação), São Paulo, jan 2006.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189 – 1232, 2001. Acesso em: 19 set. 2025. Disponível em: https://doi.org/10.1214/aos/1013203451.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA: MIT Press, 2016. ISBN 9780262035613.

GUO, G. *et al.* Knn model-based approach in classification. In: PALADE, V.; HOWLETT, R.; JAIN, L. (Ed.). **Knowledge-Based Intelligent Information and Engineering Systems**. Berlin, Heidelberg: Springer, 2003, (Lecture Notes in Computer Science, v. 2774). p. 986–996.

GUO, Y. *et al.* Incremental self-training for efficient semi-supervised learning. **Pattern Recognition**, v. 146, p. 110048, 2024.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer, 2009. ISBN 9780387848570.

HAWKINS, D. **Identification of Outliers**. Chapman and Hall, 1980. (Monographs on applied probability and statistics). Acesso em: 19 set. 2025. ISBN 9780412219009. Disponível em: https://books.google.com.br/books?id=fb0OAAAAQAAJ.

HEALY, P. M.; WAHLEN, J. M. Managing earnings. **The accounting review**, American Accounting Association, v. 74, n. 4, p. 421–448, 1999.

HILAL, W.; GADSDEN, S. A.; YAWNEY, J. Financial fraud: A review of anomaly detection techniques and recent advances. **Expert Systems with Applications**, v. 193, p. 116429, 2022. ISSN 0957-4174. Acesso em: 19 set. 2025. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417421017164.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. 3rd. ed. Hoboken, NJ: Wiley, 2013. ISBN 9780470582473.

KARLOS, S. *et al.* Using active learning methods for predicting fraudulent financial statements. In: ILIADIS, L.; MAGLOGIANNIS, I.; PAPADOPOULOS, H. (Ed.). **Artificial Intelligence Applications and Innovations**. Cham: Springer, 2017, (IFIP Advances in Information and Communication Technology, v. 584). p. 351–362.

KHORUNZHAK, N.; LUKANOVSKA, I. Accounting in the digital economy: Problems and prospects. **Black Sea Economic Studies**, 2019.

KIM, S. *et al.* Take a chance: Managing the exploitation-exploration dilemma in customs fraud detection via online active learning. **arXiv preprint arXiv:2010.14282**, 2020. Acesso em: 19 set. 2025. Disponível em: https://arxiv.org/abs/2010.14282.

KUMAR, M. *et al.* Performance evaluation of classifiers for the recognition of offline handwritten gurumukhi characters and numerals: A study. **Artificial Intelligence Review**, v. 53, 03 2020.

LEEVY, J. L. *et al.* Investigating the effectiveness of one-class and binary classification for fraud detection. **Journal of Big Data**, v. 10, n. 157, 2023.

LIAW, A.; WIENER, M. **Classification and Regression by RandomForest**. 2001.

LIOU, F.-S. Fraud detection and prevention: A data mining approach. **Expert Systems with Applications**, Elsevier, v. 35, n. 3, p. 1319–1327, 2008.

LIU, Y. *et al.* A strategy on selecting performance metrics for classifier evaluation. **Int. J. Mob. Comput. Multim. Commun.**, v. 6, p. 20–35, 2014.

LUQUE, A. *et al.* Exploring symmetry of binary classification performance metrics. **Symmetry**, v. 11, p. 47, 2019.

MAKOLO, A.; ADEBOYE, T. Credit card fraud detection system using machine learning. **International Journal of Information Technology and Computer Science**, 2021.

MANOHARAN, A. *et al.* Artificial neural networks, gradient boosting and support vector machines for electric vehicle battery state estimation: A review. **Journal of Energy Storage**, v. 55, p. 105384, 2022. ISSN 2352-152X. Acesso em: 19 set. 2025. Disponível em: https://www.sciencedirect.com/science/article/pii/S2352152X22013780.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, p. 115–133, 1943.

MEENU *et al.* Anomaly detection in credit card transactions using machine learning. **Banking Insurance eJournal**, 2020.

NGAI, E. W. T. *et al.* The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. **Decision Support Systems**, Elsevier, v. 50, n. 3, p. 559–569, 2011.

OLTEANU, M.; ROSSI, F.; YGER, F. Meta-survey on outlier and anomaly detection. **Neurocomputing**, v. 555, p. 126634, 2023. ISSN 0925-2312. Acesso em: 19 set. 2025. Disponível em: https://www.sciencedirect.com/science/article/pii/S0925231223007579.

PEDREGOSA, F. *et al.* **scikit-learn: Machine Learning in Python - Semi-supervised learning**. [*S.l.*], 2011. Accessed: 2025-03-27. Disponível em: https://scikit-learn.org/stable/modules/semi_supervised.html.

PEROLS, J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. **Auditing: A Journal of Practice & Theory**, American Accounting Association, v. 30, n. 2, p. 19–50, 2011.

PROKHORENKOVA, L. *et al.* **CatBoost: unbiased boosting with categorical features**. 2019. Acesso em: 19 set. 2025. Disponível em: https://arxiv.org/abs/1706.09516.

RAHMAN, M. J.; ZHU, H. Detecting accounting fraud in family firms: Evidence from machine learning approaches. **Advances in Accounting**, v. 64, p. 100722, 2024. ISSN 0882-6110. Acesso em: 19 set. 2025. Disponível em: https://www.sciencedirect.com/science/article/pii/S0882611023000810.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. **PLOS ONE**, Public Library of Science, v. 10, n. 3, p. 1–21, 03 2015. Acesso em: 19 set. 2025. Disponível em: https://doi.org/10.1371/journal.pone.0118432.

SAMAGAIO, A.; DIOGO, T. A. Effect of computer assisted audit tools on corporate sustainability. **Sustainability**, v. 14, n. 2, 2022. ISSN 2071-1050. Acesso em: 19 set. 2025. Disponível em: https://www.mdpi.com/2071-1050/14/2/705.

SCHNEIDER, W.; GUO, H. Machine learning. **The journal of physical chemistry. B**, v. 122 4, p. 1347, 2018.

SCHREYER, M. *et al.* Detection of anomalies in large scale accounting data using deep autoencoder networks. **CoRR**, abs/1709.05254, 2017. Acesso em: 19 set. 2025. Disponível em: http://arxiv.org/abs/1709.05254.

SELIYA, N.; KHOSHGOFTAAR, T.; HULSE, J. V. A study on the relationships of classifier performance metrics. **2009 21st IEEE International Conference on Tools with Artificial Intelligence**, p. 59–66, 2009.

SETTLES, B. **Active Learning Literature Survey**. [*S.l.*], 2009. Acesso em: 19 set. 2025. Disponível em: http://burrsettles.com/pub/settles.activelearning.pdf.

The Institute of Internal Auditors. **International Professional Practices Framework (IPPF)**. 2020. https://www.theiia.org/en/standards/what-is-standards/. Acesso em: 19 set. 2025.

VIEGAS, F. *et al.* A novel approach for fraud detection in electricity consumption using a committee-based co-training. **Machine Learning**, v. 107, n. 8-10, p. 1611–1637, 2018.

VLASSELAER, V. V. *et al.* Afraid: Fraud detection via active inference in time-evolving social networks. In: **Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**. Paris, France: IEEE/ACM, 2015. p. 659–666.

WASSIE, F. A.; LAKATOS, L. P. Artificial intelligence and the future of the internal audit function. **Humanities and Social Sciences Communications**, Palgrave, v. 11, n. 1, p. 1–13, 2024.

YANG, X. *et al.* A survey on deep semi-supervised learning. **IEEE Transactions on Knowledge and Data Engineering**, v. 35, n. 9, p. 8934–8954, 2023.

ZHANG, J. Research on medical insurance fraud identification method based on multi-source datasets. **Journal of Electronics and Information Science**, v. 9, n. 3, p. 53–61, 2024.

ZHAO, Z. *et al.* Lassl: Label-guided self-training for semi-supervised learning. **The Thirty-Sixth AAAI Conference on Artificial Intelligence**, p. 9208–9216, 2022.