Universidade de Brasília

Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas Públicas

Programa de Pós-Graduação em Administração

Alex Cerqueira Pinto

Tese Doutorado

Aplicações de Inteligência Artificial na Mitigação de Fraudes e Identificação de Anomalias em Instituições Financeiras

Brasília

2025

Universidade de Brasília

Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas Públicas

Programa de Pós-Graduação em Administração

Alex Cerqueira Pinto

Aplicações de Inteligência Artificial na Mitigação de Fraudes e Identificação de Anomalias em Instituições Financeiras

Tese apresentada ao Programa de Pós-Graduação em Administração (PPGA) da Universidade de Brasília (UnB) como requisito à obtenção do título de Doutor em Administração.

Área de Concentração: Finanças e Métodos Quantitativos.

Orientador: Prof. Dr. Carlos Rosano Peña

Brasília

2025

Aplicações de Inteligência Artificial na Mitigação de Fraudes e Identificação de Anomalias em Instituições Financeiras

Alex Cerqueira Pinto

Tese apresentada ao Programa de Pós-Graduação em Administração (PPGA) da Universidade de Brasília (UnB) como requisito à obtenção do título de Doutor em Administração.

Área de Concentração: Finanças e Métodos Quantitativos.

Aprovada em ___ de outubro de 2025

Banca Examinadora

_____

Prof. Dr. Carlos Rosano Peña

Universidade de Brasília

_____

Prof. Dr. Victor Rafael Rezende Celestino

Universidade de Brasília

_____

Prof. Dr. Alexandre Xavier Ywata de Carvalho

Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa – IDP

_____

Prof. Dr. Pedro Henrique Melo Albuquerque

Bellevue College – Department Of Mathematics (Bellevue, WA, USA)

Brasília

2025

Alex Cerqueira Pinto
Aplicações de Inteligência Artificial na Mitigação de Fraudes e Identificação de Anomalias em Instituições Financeiras– Brasília, 2025, Xp.

Tese (Doutorado) - Programa de Pós-Graduação em Administração da Universidade de Brasília – UnB. Área de Concentração: Finanças e Métodos Quantitativos.

Orientador: Prof. Dr. Carlos Rosano Peña

1. Detecção de Fraudes. 2. Anomalias. 3. Dados Sintéticos. 4. Bancos. I. Orientador: Dr. Carlos Rosano Pena II. Universidade de Brasília. III. Programa de Pós-Graduação em Administração (PPGA). IV. Aplicações de Inteligência Artificial na Mitigação de Fraudes e Identificação de Anomalias em Instituições Financeiras.

Dedicatória

Dedico este trabalho à minha família, meu porto seguro e minha fonte de inspiração e de amor. À minha mãe, Marinete, cujo carinho, força e sabedoria sempre iluminou meu caminho e meu caráter; à minha esposa, Thais, companheira incansável que compartilha sonhos, lágrimas, alegrias e vitórias ao meu lado, ficaremos juntos para sempre; à minha filha, Aurora, minha princesa que me ensinou o verdadeiro sentido do amor, do cuidado, da felicidade, você é o maior presente na minha vida sou muito feliz por ter e conviver com você; e, especialmente, ao meu filho Thiago, que brilha no céu e cuja memória me inspira a buscar sempre o melhor de mim. Que cada página aqui escrita seja um reflexo do amor que recebi e da força que vocês representam para mim na minha vida.

Agradecimentos

Chegar até aqui foi um esforço muito grande, sem dúvida o maior desafio acadêmico de minha vida e com certeza não conseguiria sozinho, e é com gratidão profunda que reconheço aqueles que caminharam ao meu lado. À minha família, por me acolher nos dias de dúvidas e celebrar minhas conquistas mesmo à distância: obrigado pela paciência e pelo apoio incondicional. Peço desculpas pelos momentos de ausência e estresse, em especial, à minha doce Aurora, por aquelas vezes em que você me quis junto e eu não pude estar – foi doloroso demais, mas cada sacrifício fez valer a pena, pois terminou em realização e sucesso.

Ao meu orientador, Prof. Dr. Carlo Rosano, cujo olhar crítico e estímulo constante elevaram a qualidade deste trabalho, e a todos os professores que compartilharam saberes e práticas que enriqueceram meu percurso acadêmico: meu sincero reconhecimento. Aos membros da banca de qualificação, pelas sugestões precisas e pelo tempo dedicado em aprimorar este estudo.

Aos meus amigos, pelo companheirismo e pelas palavras de ânimo nas fases mais desafiadoras; e aos colegas de trabalho, que de alguma forma contribuíram com insights, trocas de ideias e gestos de solidariedade durante a elaboração deste texto. Cada um de vocês foi parte essencial desta jornada. Muito obrigado!

Epígrafe

"Todas as vitórias ocultam uma abdicação"
(Simone de Beauvoir)

Resumo

A detecção de fraudes em transações financeiras é um desafio crítico para assegurar a segurança e a integridade das instituições do setor. Com a crescente sofisticação das práticas fraudulentas, torna-se indispensável o aprimoramento contínuo dos modelos de detecção. Nesse contexto, os avanços em inteligência artificial assumem papel fundamental, oferecendo soluções capazes de identificar e mitigar ameaças de forma mais eficaz. Neste contexto, este trabalho tem como objetivo desenvolver aplicações e métodos de inteligência artificial voltados à melhoria de modelos de prevenção a fraudes e detecção de anomalias na indústria financeira, abordando lacunas e desafios identificados na área. Em destaque, apresenta, de forma inédita, a proposição de um modelo SLM para geração de dados sintéticos utilizando inteligência artificial generativa, visando o balanceamento de classes. Para alcançar o objetivo proposto, este trabalho está estruturado em um capítulo introdutório seguido de quatro artigos, apresentados nos capítulos subsequentes. O segundo capítulo estabelece o referencial teórico, abordando conceitos fundamentais sobre fraudes, ciência de dados e modelos analíticos. Na sequência, o terceiro capítulo apresenta uma revisão sistemática da literatura, utilizando técnicas de bibliometria e análise de redes complexas para mapear relações de citação entre estudos e evidenciar os principais desafios do tema: desbalanceamento de classes, necessidade de detecção em tempo real, interpretabilidade e escassez de dados rotulados. Com base nesses achados, o quarto capítulo propõe modelos de IA generativa voltados à criação de dados sintéticos, visando corrigir o desequilíbrio de classes. Nesse contexto, é introduzido o modelo Aurora de SLM, projetado especificamente para geração de dados sintéticos. Por fim, o quinto capítulo apresenta um estudo de caso que contempla o desenvolvimento e a aplicação empírica de um modelo de detecção de anomalias para pessoas físicas. A relevância desta pesquisa transcende o âmbito acadêmico, estendendo-se à indústria financeira e aos profissionais dedicados à prevenção de fraudes, oferecendo soluções para lacunas críticas do setor, especialmente no que se refere ao desbalanceamento de classes. A principal contribuição reside no aprimoramento de técnicas consolidadas, na aplicação empírica rigorosa, na análise crítica dos resultados e, sobretudo, na promoção da inovação por meio da proposição de uma solução original para os desafios de geração de dados sintéticos e *data augmentation*.

Palavras- chave: Inteligência Artificial (IA); Risco Operacional Bancário; Detecção de Fraudes; Anomalias; Desbalanceamento de Classes; Dados Sintéticos.

Abstract

Fraud detection in financial transactions is a critical challenge for ensuring the security and integrity of institutions in the sector. As fraudulent practices become increasingly sophisticated, continuous improvement of detection models is essential. In this context, advances in artificial intelligence play a fundamental role, providing solutions capable of identifying and mitigating threats more effectively. This work aims to develop AI-based applications and methods to enhance fraud prevention and anomaly detection models in the financial industry, addressing key gaps and challenges identified in the field. Notably, it introduces an innovative SLM model for synthetic data generation using generative AI, designed to tackle class imbalance. The study is structured into an introductory chapter followed by four articles. Chapter two establishes the theoretical framework, covering core concepts of fraud, data science, and analytical models. Chapter three presents a systematic literature review, employing bibliometric techniques and complex network analysis to map citation relationships and highlight major challenges: class imbalance, real-time detection requirements, interpretability, and scarcity of labeled data. Based on these findings, chapter four proposes generative AI models for synthetic data creation, introducing the Aurora SLM model specifically designed for this purpose. Finally, chapter five presents a case study involving the development and empirical application of an anomaly detection model for individual accounts. The relevance of this research extends beyond academia, offering practical solutions to critical gaps in the financial industry, particularly regarding class imbalance. Its main contributions include the refinement of established techniques, rigorous empirical application, in-depth critical analysis of results, and, most importantly, innovation through the proposal of an original solution for synthetic data generation and data augmentation.

Key-words: Artificial Intelligence (AI); Banking Operational Risk; Fraud Detection; Anomalies; Class Imbalance; Synthetic Data.

Lista de ilustrações

Lista de tabelas

Lista de quadros

Lista de abreviaturas e siglas

ANNs	Redes Neurais Artificiais
AUC	Area Under the Curve (Área sob a Curva)
BERT	Bidirectional Encoder Representations from Transformers
BIS	Bank of International Settlements (Banco de Compensações Internacionais)
CBOW	Continuous Bag-of-Words (Modelo de Saco de Palavras Contínuo)
CGAN	Redes Adversárias Generativas Condicionais
CMN	Conselho Monetário Nacional
CNN	Redes Neurais Convolucionais
COPOD	Copula-Based Outlier Detection
Crisp-DM	Cross Industry Standard Process for Data Mining
CTGAN	Redes Adversárias Generativas Tabulares Condicionais
DAGMM	Deep Autoencoding Gaussian Mixture Model
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DT	Decision Tree (Árvore de Decisão)
FNN	Redes Neurais Feed-forward
GAN	Generative Adversarial Networks (Redes Adversárias Generativas)
GBM	Gradient Boosting Machine
GenAI	Inteligência Artificial Generativa
GPT	Generative Pre-trained Transformer
IA	Inteligência Artificial
ICL	In-Context Learning
k-NN	K-Nearest Neighbors (Vizinhos Mais Próximos)
LLM	Large Language Models (Grandes Modelos de Linguagem)
LOF	Local Outlier Factor
LR	Regressão Logística
MAD	Mean Absolute Difference (Diferença Média Absoluta)
MAE	Mean Absolute Error (Erro Absoluto Médio)
MAPE	Mean Absolute Percentual Error (Erro Percentual Absoluto Médio)
ML	Machine Learning (Aprendizado de Máquina)
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error (Erro Quadrático Médio)
OC-SVM	One-Class Support Vector Machine
PCA	Análise de Componentes Principais
PNL	Processamento de Linguagem Natural
RAG	Retrieval-Augmented Generation (Geração Aumentada por Recuperação)
RMSE	Root Mean Square Error (Raíz do Erro Médio Quadrático)
RNNs	Redes Neurais Recorrentes
RO	Risco Operacional
ROC	Receiver Operating Characteristic
SDV	Synthetic Data Vault
SHAP	Shapley Additive Explanations
SLM	Small Language Models (Modelos de Linguagem Pequenos)
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine (Máquina de Vetores de Suporte)

T5	Text-To-Text Transfer Transformer
VAE	Autoencoders Variacionais

Sumário

Capítulo 1 - Introdução

Um ambiente de negócios cada vez mais digitalizado oferece oportunidades e desafios às instituições financeiras. A digitalização permitiu ofertar uma gama mais ampla de serviços e produtos financeiros durante as 24 horas dos 7 dias da semana, atendendo às necessidades dos clientes de forma mais eficiente e conveniente. No entanto, a digitalização também aumenta a exposição a ameaças cibernéticas, principalmente no setor financeiro de varejo que se tornou um dos principais alvos de atividades ilícitas e fraudulentas. Isso está exigindo pesquisas e investimentos significativos em segurança para proteger dados e as transações dos clientes.

1.1. Contextualização, tema e problema de pesquisa

Como qualquer organização, as instituições financeiras estão sujeitas a uma ampla gama de riscos durante o curso de suas atividades. Conhecer as suas características e particularidades é essencial, uma vez que os riscos desconhecidos representam os mais marcantes (Martin et al., 2004). Dentre essas ameaças, a fraude se destaca como uma das mais significativas enfrentadas globalmente pelos bancos. Ao longo do tempo, as técnicas utilizadas por fraudadores têm aumentado, evoluído e se sofisticado, dificultando progressivamente sua detecção e mitigação.

As fraudes em instituições financeiras envolvem uma variedade de esquemas que exploram vulnerabilidades tecnológicas e humanas. Entre as mais comuns estão a falsificação e adulteração de documentos; a aplicação de *phishing*, que utiliza comunicações falsas para roubar dados bancários; a clonagem de cartões; e os golpes via PIX, em que criminosos simulam urgência para obter transferências indevidas. Também se destacam fraudes digitais como o uso de softwares maliciosos para capturar sessões bancárias, além da clonagem de WhatsApp e perfis falsos em redes sociais para extorquir contatos da vítima.

Paralelamente às fraudes, as anomalias representam potencial ameaça à estabilidade e segurança dos bancos. Anomalias referem-se a irregularidades ou comportamentos que desviam do esperado ou do normal. Podem indicar algo fora do comum nos sistemas, nas transações ou nos processos da instituição. Representam uma ameaça à estabilidade e segurança das instituições porque podem indicar vulnerabilidades críticas em seus sistemas,

processos ou operações, que, se não forem tratadas, podem resultar em sérias consequências financeiras, operacionais e reputacionais.

Com o avanço da tecnologia e a crescente digitalização das transações financeiras, a necessidade de sistemas eficazes de detecção de fraudes e anomalias nunca foi tão importante. A aplicação de modelos analíticos e Inteligência Artificial (IA) tem se mostrado promissor para combater essas ameaças e garantir a segurança financeira. Tais modelos permitem processar grandes volumes de dados em tempo real, proporcionando uma resposta mais rápida a atividades suspeitas. A habilidade de identificar eficientemente anomalias financeiras, reduzindo custos operacionais e elevando a precisão de fraudes, tornou-se um divisor de águas no combate a atividades fraudulentas. A integração de *machine learning* com técnicas analíticas tradicionais, formando modelos híbridos, tem sido particularmente eficaz. Modelos analíticos baseados em *machine learning* têm desempenhado um papel central na transformação da metodologia de detecção de anomalias e prevenção de fraudes por bancos e instituições financeiras.

A prevenção de fraudes e a detecção de anomalias constituem componentes primordiais da Gestão de Segurança da Informação em instituições bancárias e integram a estrutura de gestão de riscos corporativos (Damenu & Beaumont, 2017). No setor bancário, a gestão de riscos se baseia nas regras de Basileia III, um conjunto de regulamentos internacionais que buscam fortalecer a regulamentação, supervisão e gestão de riscos nas transações bancárias. Essas regras reforçam a necessidade de uma gestão de riscos mais robusta, especialmente na gestão de capital, liquidez e exposição ao risco.

Os bancos usam técnicas de benchmarking para comparar suas práticas com as exigências de Basileia III, garantindo que as práticas adotadas por eles estejam alinhadas com os padrões internacionais (Locher, 2005). A auditoria das melhores práticas deve ser realizada regularmente com o objetivo de saber o que se pode aprender delas e de adaptar planos que definam os papeis e responsabilidades na estrutura organizacional e de governança da instituição em sintonia com suas estratégias e padrões regulatórios. Assim, os bancos que conseguem realizar benchmarking eficaz podem melhorar sua posição competitiva e manter padrões de risco compatíveis com os requisitos globais, fortalecendo sua resiliência e estabilidade no mercado.

Diante da relevância do assunto para a indústria financeira, a presente tese define

métodos e modelos de detecção de fraudes e anomalias como seu tema central. Em outras palavras, a área de interesse em que esta pesquisa está enquadrada é "aplicações de modelos analíticos e de inteligência artificial na detecção de anomalias e mitigação de fraudes bancárias.

Dentro desse tema, a questão central que a tese procura resolver é como o desenvolvimento de modelos analíticos e de inteligência artificial podem ser utilizados para aprimorar a detecção de fraudes e anomalias nas transações financeiras?

1.2. Desafios e lacunas de pesquisa sobre o tema

Historicamente, a detecção de fraudes e anomalias dependia fortemente de sistemas baseados em regras e da experiência humana. No entanto, com o aumento do número, complexidade e sofisticação das fraudes, esses métodos se tornaram insuficientes. A experiência passada pode não ser satisfatória para identificar padrões inéditos ou avançados de fraude. Nas últimas décadas, a literatura especializada tem explorado com uma maior intensidade o uso de técnicas de IA, mineração de dados e aprendizado de máquina para melhorar a detecção de fraudes. Essas técnicas podem identificar padrões complexos e sutis que podem ser difíceis para os humanos detectarem.

Estudos acadêmicos recentes comprovam a eficiência da utilização dos modelos analíticos e de inteligência artificial pelas instituições financeiras. Porém, a dinâmica constante das ameaças cibernéticas e a crescente demanda dos clientes pela eficácia na segurança de seus dados e patrimônio exigem melhoras contínuas, aprimoramentos ininterruptos dos métodos e modelos existentes. A inteligência artificial, a aprendizagem de máquina e a mineração de dados emergem como poderosas ferramentas nessa batalha contra as ameaças de fraudes e comportamentos anormais que, de outra forma, poderiam passar despercebidos pelos sistemas convencionais de monitoramento.

Além disso, apesar dos avanços, ainda persistem desafios e lacunas na literatura sobre o tema.

O primeiro desafio é que a maioria dos estudos em modelos de detecção de fraudes se concentra em técnicas de aprendizado supervisionado, que requerem grandes conjuntos de dados rotulados para treinamento. Isso é um desafio, pois os dados de fraude são tipicamente

escassos comparados as transações não fraudulentas.

Outro aspecto crítico relacionado à escassez de dados rotulados é o elevado custo e a complexidade do processo de anotação, uma vez que a validação de uma transação como fraude geralmente exige investigações manuais, auditorias e conhecimento especializado. Isso faz com que os conjuntos de dados disponíveis sejam limitados e, muitas vezes, desatualizados em relação às estratégias de fraude mais recentes, reduzindo a eficácia dos modelos supervisionados. Nesse contexto, cresce o interesse por abordagens alternativas, como métodos semi supervisionados e não supervisionados, que buscam explorar o grande volume de transações não rotuladas para extrair padrões úteis e auxiliar na identificação de atividades suspeitas.

O segundo problema diz respeito à explicabilidade dos modelos da detecção da fraude. Muitas vezes, os modelos baseados em inteligência artificial são complexos e de difícil interpretação, o que pode limitar sua aceitação nas decisões. Portanto, há uma necessidade de mais pesquisas que explorem maneiras de integrar efetivamente esses sistemas de IA nos fluxos de trabalho bancários existentes, a fim de maximizar seu impacto e garantir transparência nas decisões tomadas.

O terceiro desafio trata de que muitos dos modelos atuais ainda lutam para equilibrar a precisão com a minimização de falsos positivos devido à natureza dos dados ser extremamente desequilibrada entre as classes. Um conjunto de dados é considerado desbalanceado quando uma ou mais classes têm significativamente menos observações em comparação com outras, o que compromete a capacidade do modelo de identificar adequadamente os padrões dessa classe. Como consequência, o modelo pode aprender de forma inadequada os padrões das classes com menos dados, resultando em desempenho deficiente, métricas enganosas e risco de *overfitting* nessas classes.

Deste modo, existe a necessidade de utilizar técnicas eficazes para tratar a quantidade de dados entre as classes até atingir um equilíbrio desejado para a aprendizagem e reconhecimento de padrões. A seguir, são apresentadas algumas das técnicas mais utilizadas para lidar com o desbalanceamento de classes:

1. *Oversampling*: Consiste em aumentar a representatividade da classe minoritária, seja duplicando instâncias existentes via reamostragem ou gerando novos exemplos sintéticos por meio de algoritmos. Esta técnica auxilia os modelos a aprenderem

melhor os padrões da classe minoritária, reduzindo o viés em favor da classe majoritária e melhorando o desempenho na detecção de eventos raros.

2. *Undersampling* (subamostragem): Diminui o número de observações na classe majoritária. Essa técnica ajuda a balancear a distribuição dos dados, mas deve ser usada com cautela para não descartar informações relevantes.
3. Ajuste de Pesos (*Cost-Sensitive Learning*): Muitos algoritmos de aprendizado permitem a ponderação das classes, aumentando a penalidade para erros cometidos na classe minoritária. Dessa forma, o modelo "se importa mais" em classificar corretamente as instâncias menos representadas.
4. Uso de Algoritmos Específicos ou *Ensemble Methods*: A) algoritmos como *Random Forest* ou *Gradient Boosting* podem ser adaptados para lidar melhor com classes desbalanceadas, muitas vezes combinando a reamostragem com o ajuste de pesos. B) Técnicas híbridas: Combinar *oversampling* e *undersampling* para alcançar um equilíbrio sem perder a variabilidade dos dados.
5. Alteração das Métricas de Avaliação: Em vez de focar apenas em acurácia, é importante usar métricas que deem visibilidade para as classes minoritárias, como F1-*Score, Precision, Recall* e AUC-ROC. Essa abordagem permite monitorar e ajustar o desempenho do modelo de maneira correta e equilibrada.

Cada uma dessas estratégias pode ser aplicada isoladamente ou combinada, dependendo do contexto e da complexidade do problema. A escolha da técnica mais adequada deve considerar além do equilíbrio dos dados, a preservação da informação relevante e o impacto das modificações na generalização do modelo. Assim, este trabalho propõe métodos e modelos para atuar contra o desequilíbrio de classes com geração de dados sintéticos via utilização de inteligência artificial - IA generativa.

Nesse contexto, a presente tese justifica-se pela urgência e constante necessidade do setor financeiro de desenvolver métodos e modelos inovadores para identificar fraudes e comportamentos atípicos. Além disso, propõe novas soluções para as lacunas ainda existentes com destaque para o desafio do desbalanceamento de classes entre fraudes e não fraudes, por meio de métodos e modelos voltados à geração de dados sintéticos e á minimização deste problema.

1.3. Objetivos da Tese

A identificação de atividades irregulares ou suspeitas envolve a detecção de transações fraudulentas, risco operacionais, lavagem de dinheiro ou roubo de identidade e dados. Para isso, modelos analíticos e IA são usados para analisar grandes volumes de dados de transações para identificar padrões e comportamentos anômalos que podem indicar atividades fraudulentas. Esses sistemas têm a capacidade de aprender continuamente, adaptando-se a novas táticas de fraude e melhorando a precisão da detecção, identificando práxis sistemáticas, fazendo previsões e tomando decisões com base em dados sem a necessidade de intervenção humana direta. À medida que os sistemas são expostos a mais dados e situações, eles adquirem habilidades e melhoram suas capacidades e desempenho.

Diante desse cenário, a presente tese tem como objetivo central desenvolver aplicações e métodos de inteligência artificial para melhora de modelos de prevenção a fraudes e detecção de anomalias na indústria financeira, atuando nas lacunas e desafios identificados.

Como objetivos específicos, buscará: i) analisar a literatura acadêmica e o estado da arte sobre o tema por meio de uma revisão sistemática; ii) propor um método e modelo de geração de dados sintéticos com utilização de modelos de inteligência artificial generativa para atuar sobre a necessidade de balanceamento de classes; iii) desenvolver modelo com aplicação de técnicas de aprendizado não supervisionado para detecção de anomalias.

1.4. Estrutura da Tese

Com a finalidade de atender os objetivos propostos, esta pesquisa adota a modalidade de tese em formato de artigos, que inclui, além de esta introdução e o segundo Capítulo de referencial teórico geral, três (3) artigos e as considerações finais da tese.

Assim, o Capítulo 2, faz uma revisão teórica atrelada aos artigos, onde aborda-se os principais conceitos e referenciais teóricos sobre fraudes, anomalias, além da base conceitual sobre modelos, ciência de dados, inteligência artificial, aprendizado de máquina e mineração de dados, voltada para finanças e banking.

No Capítulo 3, apresenta-se o primeiro artigo, em que se realiza uma revisão bibliográfica sistemática, aplicando técnicas de bibliometria e análise de redes complexas. Seu objetivo é mapear conexões citacionais e identificar os principais desafios presentes na literatura como: desbalanceamento de classes, exigência de detecção imediata, clareza

interpretativa e a contínua transformação nos padrões de fraudes, com uma notável ascensão em casos vinculados à engenharia social. O artigo foi publicado na revista *Journal of Bibliometrics in Business and Management*[1].

O Capítulo 4 contém o segundo artigo, que visa desenvolver e colocar em prática alguns métodos e modelos de geração de dados sintéticos com uso de aprendizado de máquina e inteligência artificial generativa (GenIA). A proposta é equilibrar os dados e realizar *oversampling*, aumentando a quantidade de exemplos da classe minoritária, de modo a reduzir falsos positivos e melhorar a performance de modelos preditivos de fraudes. Inicialmente, foram desenvolvidos modelos tradicionais de IA, como SMOTE, GAN e VAE. Em seguida, o trabalho avançou para o desenvolvimento de aplicações de IA Generativa para criação dos dados sintéticos que respeitassem os mesmos padrões de estrutura dos dados originais. Para tal feito, foi aplicado inicialmente engenharia de prompt, depois modelos de IA Generativa com RAG e por fim, um modelo inédito de SLM, batizado de Aurora, desenvolvido via *fine-tuning* com o conhecimento intrínseco para geração de dados sintéticos.

No Capítulo 5, inclui-se o terceiro artigo, que se dedica ao desenvolvimento e avaliação de um modelo de detecção de anomalias, denominadas de negócios não sustentáveis, utilizando dados de uma grande instituição financeira brasileira, voltados a produtos de seguridade para pessoas físicas e jurídicas.

Para finalizar, há um capítulo final de conclusão e considerações finais. Nele, são discutidas a integração dos objetivos, resultados e conclusões individuais dos três artigos, bem como, as contribuições para o campo de estudo, limitações do trabalho e possível agenda de pesquisa futura.

1.5. Justificativa da Tese

A relevância deste trabalho, assim como de futuras agendas de pesquisa sobre detecção de fraudes bancárias usando modelos analíticos e IA, é variada e abrangente para a literatura acadêmica. Primeiramente, os resultados da pesquisa trazem insights inovadores que aprofundam o conhecimento sobre como melhor detectar e prevenir fraudes bancárias,

[1] Pinto, A. C., Tessmann, M. S., & Lima, A. V. (2024). Fraud and anomaly detection models in banks: a systematic analysis and literature connection. International Journal of Bibliometrics in Business and Management, 3(2), 182-205. DOI: 10.1504/IJBBM.2024.140372

contribuindo para o desenvolvimento de modelos mais precisos e eficientes, capazes de superar as limitações dos métodos atualmente empregados.

Destaca-se que este tema de pesquisa é interdisciplinar, combinando elementos de ciência da computação, finanças, estatística e ética. Portanto, os avanços nessa área podem ter implicações em várias disciplinas acadêmicas. A realização de pesquisas neste campo é de grande importância para a academia, a indústria e a sociedade como um todo.

Buscando atender os requisitos de originalidade, ineditismo, inovação e relevância, a tese contribui com o campo de pesquisa ao levantar o estado da arte dos métodos e modelos de detecção de anomalias e fraudes; identificando os principais desafios relacionados ao assunto; fornecendo propostas com sugestões de soluções para os problemas mencionados; e executando uma agenda de pesquisa que aborde esses desafios na indústria bancária e financeira. A abordagem adotada inclui o desenvolvimento de aplicações empíricas com modelos de IA Generativa sobre o tema, detalhadas nos capítulos 4 e 5, que buscam atuar nos gaps identificados:

- Balanceamento da base de dados. Dados de fraudes são escassos e altamente desequilibrados, portanto, o desenvolvimento de modelos com *oversamping* melhores tendem a reduzir falso positivos;
- Desenvolvimento de modelos não supervisionados. A criação de modelos que não dependam exclusivamente de dados rotulados é essencial para ampliar a capacidade de detecção de anomalias;
- Integração de sistemas de IA aos fluxos de trabalho bancários: Garantir que os modelos desenvolvidos possam ser integrados aos processos já existentes para maximizar seu impacto e eficiência.

Resumindo, é importante ressaltar que os resultados dessas pesquisas têm implicações diretas na indústria financeira, ajudando as instituições a protegerem-se contra fraudes e a oferecerem um serviço mais seguro aos seus clientes. Com o cenário de fraudes em constante evolução, a pesquisa contínua é necessária para abordar novos tipos de fraudes.

Capítulo 2

2. Referencial Teórico - Principais Conceitos sobre riscos financeiros e ciência de dados

Neste capítulo, apresenta-se e discute-se a base conceitual sobre riscos financeiros, fraudes e transações atípicas nas transações financeiras, bem como os métodos para enfrentá-los. Discute-se o conjunto de teorias, modelos, estudos e evidências prévias que dão base a este trabalho. Busca-se articular a estrutura do trabalho e dar coerência à tese formada por três artigos.

Na seção 2.1, é apresentada a base conceitual sobre riscos financeiros, em especial o risco operacional, bem como sobre fraudes e anomalias. A seguir, é apresentado conceitos gerais e evolução histórica sobre inteligência artificial, aprendizado de máquina e mineração de dados, técnicas estas que são aplicadas em modelos de prevenção a fraudes e anomalias.

Em sequência, salientamos as principais características e diferenciações entre modelos supervisionados e não supervisionados, bem como os principais tipos de modelos de aprendizado de máquina (classificação, regressão e clusterização). Subsequentemente, é apresentada a metodologia Crisp-DM como principal método de desenvolvimento de modelos e, a seguir, explicita-se os principais algoritmos de aprendizado de máquina e as principais medidas de desempenho destes.

Por fim, antes da conclusão, no item 2.3 destaca-se, no que se refere aos modelos e pesquisa em fraudes e anomalias, quais são os principais desafios e lacunas de pesquisa deste tema no âmbito financeiro.

2.1. Riscos financeiros, fraudes e anomalias

Todos os agentes econômicos, independentemente do ramo de atuação, estão expostos a uma multiplicidade de riscos ao longo do ciclo operacional de seus negócios. O conhecimento desses riscos é fundamental, assim como a gestão eficaz daqueles considerados mais relevantes, levando em consideração a forma de atuação, o nicho de mercado e a escala dos negócios.

Damodaran (2010) destaca que o risco é onipresente em quase todas as atividades humanas e não há uma definição única e consensual para o termo. A discussão sobre o tema, portanto, baseia-se na distinção entre o risco passível de ser quantificado de forma objetiva e o risco subjetivo.

No setor bancário, o risco está relacionado à possibilidade reais de perdas financeiras e

dificuldades que afetem a capacidade de cumprir com as obrigações financeiras. Incluem risco operacional, de crédito, de mercado, de liquidez e sistêmico e têm várias fontes, como mudanças no mercado, volatilidade dos preços de seus ativos, flutuações na taxa de juros e câmbio, desastres naturais, ataques cibernéticos, inadimplência de clientes e falhas internas. (Bessis, 2011)

Para administrar o risco, é necessário realizar uma análise detalhada dos processos e previsões, identificar potenciais falhas, suas causas e consequências, bem como remodelar permanentemente as transações de seguros e reservas para identificar e mitigar esses perigos. Nesse sentido, as grandes instituições financeiras seguem as orientações do *Bank of International Settlements* – BIS (Banco de Compensações Internacionais), especialmente aquelas estabelecidas pelos Acordos de Basileia.

Com a publicação dos Acordos de Basileia e o aumento subsequente da regulamentação nacional do setor, a gestão de riscos adquiriu maior relevância, resultando no desenvolvimento e aprimoramento de diversos procedimentos, mecanismos e modelos para a mensuração e controle de riscos. De acordo com o conceito regulatório, as tratativas de mitigação de fraudes, é referenciada dentro do arcabouço do Risco Operacional (RO), que também incluem erros internos e falhas sistêmicas. O RO É reconhecido como um componente essencial, devido aos elevados montantes de perdas efetivas e de capital alocado por essas instituições.

O Banco de Compensações Internacionais (BIS) define o risco operacional como a possibilidade de perdas em decorrência de falhas em processos, pessoas, tecnologia ou eventos externos (BIS, 2021).

Essa abrangente definição engloba tanto os riscos externos imprevistos, fora do controle direto da empresa, bem como diversos eventos internos, tais como erros humanos, falhas em sistemas de informação, defeitos na identificação de fraudes e anomalias, imprecisões no diagnóstico do contexto externo, entre outros. Por exemplo, uma falha operacional (como um erro de avaliação de um pedido de empréstimo) pode aumentar o risco de crédito ao aprovar um empréstimo para um tomador de alto risco. Assim, a má gestão de riscos operacionais pode comprometer a capacidade da instituição financeira de gerir adequadamente seu portfólio de crédito, exacerbando as perdas. Para os autores Hull (2012) e Pesaran, Schuermann, Treutler e Weiner (2006), o risco de crédito é o principal risco operacional enfrentado pelas instituições financeiras.

O BIS (2021) propõe uma gestão de risco operacional abrangente e proativa, composta por

cinco etapas: identificação, avaliação, implementação de controles, monitoramento e acompanhamento, e cultura de gestão de riscos. Essa abordagem visa minimizar o impacto de eventos adversos, fortalecer a reputação da instituição, otimizar processos e reduzir custos, além de garantir o cumprimento das exigências regulatórias e contribuir para a vantagem competitiva.

No Brasil, o risco operacional é normatizado, para o sistema bancário, pela Resolução Conselho Monetário Nacional (CMN) 4557/2017, que se refere o risco operacional, de forma similar ao BIS, como resultados de mudanças do contexto externo e da deficiência ou inadequação de processos internos, pessoas ou sistemas. Essa Resolução estabelece diretrizes para o gerenciamento de riscos, a manutenção da estrutura de capital adequada, a governança corporativa e política de segurança cibernética.

Neste contexto, o risco operacional é causado basicamente por fraudes financeiras e anomalias em transações, ao comprometer a integridade dos processos internos. A fraude é definida como qualquer ação intencional ou comportamento enganoso executado com o objetivo de obter vantagem financeira ilícita, seja através da manipulação de informações contábeis, distorção de transações financeiras, falsificação e omissões de documentos ou qualquer outra forma de conduta desonesta dentro do contexto financeiro de uma organização ou sistema. Essas atividades fraudulentas podem resultar em perdas financeiras significativas para as partes afetadas, incluindo investidores, acionistas, instituições financeiras e consumidores.

De acordo com um estudo realizado por Hilal et. al (2022), a fraude financeira é descrita como " qualquer ato intencional ou deliberado para privar outro de propriedade ou dinheiro por astúcia, decepção ou outros meios injustos". Esses crimes podem ocorrer em diversos contextos e formas, desde transações corporativas até operações de mercado, e são frequentemente perpetradas por indivíduos ou grupos que buscam explorar vulnerabilidades nos sistemas financeiros e contábeis. Dentre os tipos de fraudes financeiras apresentado por Hilal et. al (2022), destaca-se: fraude de cartão de crédito; fraudes em seguros; lavagem de dinheiro; pirâmides financeiras e manipulação de informações de mercado.

No que tange a anomalias financeiras, estas se referem aos desvios dos padrões, comportamentos ou eventos incomuns, atípicos ou inesperados nos dados financeiros de um cliente, uma empresa, mercado ou sistema financeiro. Essas irregularidades podem indicar a possibilidade de atividades fraudulentas, falhas sistêmicas, erros contábeis e operacionais, manipulação de informações ou outros problemas financeiros, bem como choques

macroeconômicos que são eventos exógenos imprevistos que causam grandes mudanças nos principais indicadores econômicos de um país ou região.

As anomalias financeiras podem se manifestar de várias maneiras, como discrepâncias significativas nos registros contábeis, padrões incomuns de transações, variações repentinas nos fluxos de caixa, inconsistências nos relatórios financeiros ou qualquer outra irregularidade que não siga o padrão esperado de funcionamento financeiro. A detecção de anomalias tem um papel significativo na detecção de fraudes financeiras e é utilizada para extrair informações atípicas e desvios em grandes quantidades de dados (Ngai et al., 2011).

Na literatura, existe uma quantidade significativa de trabalhos aplicando métodos estatísticos, bem como técnicas de inteligência artificial e aprendizado de máquina para abordar a detecção de fraudes e anomalias em registros de pedido de empréstimo, emissões de notas fiscais, declaração de impostos, cartões de crédito e seguros, sendo a maioria focada nos dois últimos. Além disso, a maioria dos novos artigos científicos direciona seu foco para técnicas não supervisionadas, muitas das quais abordando conjuntos de dados desbalanceados e incompletos Hilal et. al (2022).

Segundo a Pesquisa Global de Identidade e Fraude, apresentada em Serasa Experian (2021), no âmbito global, 8 em cada 10 empresas disseram que agora têm uma estratégia de reconhecimento do cliente, um aumento de 26% desde o início da pandemia. Isso se deve à constatação de que as perdas causadas por fraudes têm aumentado a cada ano no Brasil e no mundo. Essa pesquisa identificou que 57% das empresas relataram perdas maiores associadas a fraudes na abertura e no roubo de contas em 2020, em comparação com 55% em 2018 e 51% em 2017. Além disso, menciona ter atuado mais de 3 mil eventos fraudulentos por segundo em 2020. Contudo, o foco na atuação e mitigação da fraude, está provocando o bloqueio de transações legítimas de muitos clientes por suspeita de fraudes, identificando-se situações de falsos positivos. Apesar do principal dano ser a perda financeira, o bloqueio errado também causa perda de confiança do cliente.

A proporção de fraudes em relação ao volume total de transações na indústria financeira pode variar dependendo do país e do período considerado. No Brasil, por exemplo, estudo de Mapa da Fraude[2], divulgado pela ClearSale em 2020, apontou que mais de 902 mil

[2] https://br.clear.sale/mapa-da-fraude

das mais de 22 milhões de transações analisadas eram uma tentativa de fraude, o que corresponde a aproximadamente 4,05% do total.

Já no primeiro trimestre de 2023, a empresa de prevenção de fraudes e segurança digital CAF relata, a partir de dados do Banco Central, que 1,73% das transações digitais nos canais eletrônicos do sistema financeiro do país tiveram intenções criminosas, ou seja, 2,8 mil tentativas de fraudes financeiras por minuto (CAF, 2023).

Esses números destacam a importância de novos investimentos, contínuas pesquisas e desenvolvimento de métodos oriundos da ciência da computação para a detecção de fraudes na indústria financeira.

2.2.Métodos da ciência de dados

O ecossistema da ciência de dados refere-se ao conjunto interligado de disciplinas, tecnologias e práticas que compõem a área de estudo e aplicação da computação. Ele abrange desde os fundamentos teóricos da ciência da computação — como algoritmos e estruturas de dados — passando por técnicas de mineração de dados e aprendizado de máquina, até aplicações práticas e avançadas, incluindo inteligência artificial (IA).

O ecossistema da ciência de dados envolve uma série de componentes interconectados que transformam dados brutos em insights úteis. Seus principais componentes seguem uma sequência. Ele começa com a obtenção de dados estruturados (bancos de dados) ou não estruturados (mídias sociais, IoT). A coleta e o armazenamento são realizados por meio de ferramentas como SQL, data lakes e nuvem. Em seguida, a engenharia de dados entra em ação, realizando a limpeza e o processamento das informações. Linguagens de programação como Python e R, juntamente com ferramentas como TensorFlow e Tableau, constituem a base técnica para análise e visualização. Os modelos e algoritmos aplicados envolvem estatística, aprendizado de máquina e inteligência artificial. Para garantir eficiência e escalabilidade, são utilizadas infraestruturas computacionais robustas, como GPUs e plataformas de big data. Profissionais especializados — cientistas e engenheiros de dados — colaboram com stakeholders para gerar valor estratégico, enquanto práticas de governança e ética asseguram a privacidade e o uso responsável dos dados. Por fim, a comunicação dos resultados é essencial, sendo realizada por meio de gráficos, dashboards e técnicas de storytelling. Trata-se de um ecossistema dinâmico, multidisciplinar e em constante evolução (Géron, 2022).

Com o avanço do hardware e a crescente capacidade de processamento, novas áreas começaram a emergir dentro do ecossistema da ciência de dados. Nos anos 1980 e 1990, a internet e o desenvolvimento de redes de computadores abriram caminho para a era da informação, onde a coleta, armazenamento e análise de grandes volumes de dados tornaram-se cruciais. Neste contexto, surgiram a mineração de dados e a ciência de dados, disciplinas focadas na extração de conhecimento útil a partir de grandes conjuntos de dados.

Desta forma, na seguinte seção será apresentada a contextualização e breve descritivo das tecnologias ligadas a ciência da computação, com viés para a ciência de dados e sua evolução histórica com o surgimento da inteligência artificial, mineração de dados, aprendizado de máquina, bem como os principais algoritmos e medidas de performance associadas.

2.2.1. Ciência da Computação

A utilização da computação evoluiu aceleradamente nas últimas décadas, dando origem a tecnologias inovadoras aplicadas em diversas áreas como a medicina, economia, finanças, robótica, linguística e em diversos setores da indústria e serviços. A tendência é que no futuro essas técnicas se façam ainda mais presente nas nossas rotinas e desempenhem papéis cruciais na análise e interpretação de dados para a tomada de decisões e automação de processos.

Conforme descreve Brookshear (2013), a Ciência da Computação é uma área de conhecimento que investiga e cria métodos, ferramentas e tecnologias computacionais para automatizar processos e solucionar problemas relacionados ao processamento de informações. Essa ciência vai além do estudo de algoritmos e sua implementação em softwares, abrangendo também técnicas de organização e gerenciamento de dados, telecomunicações, protocolos de comunicação e outros campos especializados da computação.

A ciência da computação é conceituada como um campo amplo que estuda os fundamentos teóricos e práticos da computação, abrangendo desde a concepção e construção de hardware até o desenvolvimento de software e algoritmos. É a base da Inteligência Artificial (IA), que por sua vez incorporam técnicas de mineração de dados e aprendizado de máquina.

2.2.2. Inteligência Artificial

A Inteligência Artificial (IA) é um campo da ciência da computação que se concentra no desenvolvimento de sistemas computacionais capazes de realizar e replicar tarefas que

normalmente exigiriam atuação humana. Isso inclui reconhecer padrões, aprender com dados, tomar decisões, automatizar tarefas e resolver problemas complexos de forma eficiente. Segundo Russell e Norvig (2010), a Inteligência Artificial é o estudo de como fazer os computadores realizarem tarefas que, até o momento, os seres humanos fazem melhor.

Inicialmente limitada a sistemas baseados em regras e lógica simbólica, a IA também evoluiu rapidamente, especialmente com o desenvolvimento de técnicas de aprendizado de máquina, que permitem que os sistemas "aprendam" padrões a partir de dados ao invés de serem explicitamente programados para cada tarefa. Nas últimas décadas, o aprendizado de máquina, e em particular o aprendizado profundo (deep learning), tornou-se um dos pilares mais importantes do ecossistema da ciência da computação e da Inteligência Artificial, aplicável em áreas como visão computacional, processamento de linguagem natural, reconhecimento de fala, jogos e simulações.

Conforme Gartner (2023), a IA pode ser definida como a aplicação de técnicas baseadas em lógica e análises avançadas, para interpretar eventos, apoiar e automatizar decisões e realizar ações. Essa definição está em conformidade com o crescente desenvolvimento das tecnologias e recursos que utilizam técnicas de probabilidade e estatística para treinar algoritmos, permitindo identificar padrões ocultos e incertezas, fazer inferências e previsões para tomar decisões a partir de grandes volumes de dados.

Mais recentemente, os algoritmos de Inteligência Artificial têm visto avanços significativos nos últimos anos, especialmente com o surgimento e aprimoramento de modelos de linguagem natural, que imitam a comunicação humana, com grande capacidade de processamento e compreensão de texto. Entre esses modelos, destacam-se os *Large Language Models* (LLMs), que são modelos de linguagem treinados em grandes volumes de texto e capazes de gerar e compreender linguagem natural com uma alta precisão.

A origem dos LLMs pode ser traçada a partir de avanços em técnicas de aprendizado profundo e processamento de linguagem natural. Um dos marcos iniciais foi o desenvolvimento do modelo de “saco de palavras contínuo” (Continuous Bag-of-Words - CBOW) e o modelo de Skip-gram contínuo por Mikolov et al. (2013), que introduziu a ideia de representar palavras em um espaço vetorial contínuo, permitindo que as relações semânticas entre as palavras fossem capturadas de maneira mais eficaz. Ambos os modelos utilizam a *softmax* hierárquica baseada em árvore de Huffman para eficiência computacional e buscam prever a palavra atual com base

no contexto, no primeiro modelo, e tenta maximizar a classificação de uma palavra com base em outra na mesma sentença no segundo modelo.

Contudo, um importante avanço, ocorreu com a introdução do algoritmo Transformer por Vaswani et al. (2017). Este modelo eliminou a necessidade de estruturas sequenciais, como RNNs, e introduziu a atenção auto-regressiva (*self-attention*), que permitiu um paralelismo muito mais eficiente e escalabilidade para treinamento em grandes conjuntos de dados. Esta abordagem permite um maior nível de paralelização e diminuiu significativamente o tempo de treinamento, mantendo ou superando o desempenho dos modelos existentes em tarefas de tradução de linguagem.

A arquitetura Transformer é baseada em uma estrutura de encoder-decoder composta por múltiplas camadas de self-attention, seguidas por camadas feed-forward. Cada camada de self-attention no encoder conecta todas as posições de entrada, capturando dependências de longo alcance de maneira eficiente. O decoder utiliza um mecanismo semelhante, com a adição de atenção cruzada para integrar informações do encoder, além de uma atenção auto-regressiva para a geração sequencial de saídas. (Vaswani et al., 2017)

Com o desenvolvimento do algoritmo Transformer por Vaswani et al. (2017), eliminou-se a necessidade de estruturas sequenciais, introduzindo a atenção auto-regressiva (self-attention). Esta inovação propiciou um paralelismo consideravelmente mais eficiente e uma alta escalabilidade no treinamento de grandes conjuntos de dados. Tal abordagem permite uma maior paralelização, reduzindo significativamente o tempo de treinamento, enquanto mantém ou até supera o desempenho dos modelos existentes em tarefas de tradução de linguagem.

A arquitetura Transformer, de forma simplificada, é composta por um modelo codificador-decodificador que integra múltiplas camadas de atenção auto-regressiva seguidas por camadas de redes neurais feed-forward. No codificador, cada camada de atenção está ligada a todas as posições da entrada, o que permite que o modelo capture de maneira eficaz dependências de longo alcance. O decodificador funciona de maneira semelhante, porém adiciona um mecanismo de atenção cruzada que incorpora as informações provenientes do codificador, facilitando a integração dos dados processados (Vaswani et al., 2017).

A atenção auto-regressiva é o componente central, e seu cálculo é expresso da seguinte maneira:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}))V \quad (2.1)$$

onde:

Q: (Queries) são as entradas de consulta.

K: (Keys) são as entradas de chave.

V: (Values) são as entradas de valor.

d_k: é a dimensão das chaves.

Por exemplo, quando uma pessoa lê um texto e deseja compreender o significado de uma palavra, o cérebro direciona a atenção para outras partes da frase que auxiliam na construção do sentido. De forma semelhante, o modelo de atenção identifica quais partes da informação são mais relevantes para cada palavra ou elemento em processamento.

O mecanismo de atenção calcula a média ponderada dos valores V, onde os pesos são determinados pela similaridade entre a consulta Q e a chave K. A similaridade é medida usando o produto escalar, normalizado pela raiz quadrada da dimensão das chaves para estabilizar os gradientes durante o treinamento.

Dentre os usos em processamento de linguagem natural, destaca-se o BERT (*Bidirectional Encoder Representations from Transformers*), apresentado por Devlin et al. (2018), que foi desenvolvido na Google e trata-se de um modelo de linguagem pré-treinado baseado também na arquitetura Transformer possuindo a capacidade de compreender o contexto bidirecionalmente, o que o torna eficaz em tarefas de entendimento de linguagem natural, como classificação de texto, perguntas e respostas e análise de sentimentos. Outros modelos neste mesmo contexto são o XLNet (Yang et al., 2019) e o T5, *Text-To-Text Transfer Transformer* (Raffel, et al., 2020)

Entre os modelos de linguagem mais destacados está o GPT (*Generative Pre-trained Transformer*), desenvolvido pela OpenAI, que utiliza a arquitetura Transformer. O GPT foi lançado em 2018 e desde então foi sucedido por versões mais avançadas, como o GPT-2 e o GPT-3 tornando-se mundialmente famoso após 2023. Esses modelos foram treinados em grandes quantidades de dados textuais e demonstraram habilidades impressionantes em tarefas de geração de texto, tradução automática, resumo de texto etc. Eles são notáveis por sua capacidade de gerar texto naturalmente fluente, coerente e realizar uma variedade de tarefas de processamento de linguagem natural com desempenho impressionante. Isso abre

possibilidades para aplicação em áreas que vão desde suporte ao cliente, criação de conteúdo e educação, até a automação de processos mais complexos, como programação e análise de dados.

Radford et al. (2019) descreveram o desenvolvimento e o desempenho do GPT-2, destacando sua capacidade de gerar texto de alta qualidade em uma variedade de tarefas de linguagem natural. Além disso, o artigo também discute o uso de modelos de linguagem como GPT-2 como ferramentas versáteis para uma variedade de aplicações de inteligência artificial.

O artigo descreve o treinamento do modelo em um corpus de texto massivo e não rotulado, denominado WebText. O WebText foi criado a partir de rastreamentos da web e contém trilhões de palavras. Os autores avaliaram o desempenho do GPT-2 em diversas tarefas de Processamento de Linguagem Natural (PNL), incluindo resposta a perguntas (CoQA), classificação de sentimento (SST-2) e tradução automática (WMT). Apesar de não ter sido treinado especificamente para nenhuma dessas tarefas, o GPT-2 demonstrou um desempenho surpreendentemente bom em todas elas, superando ou igualando o desempenho de modelos supervisionados treinados especificamente para cada tarefa. (Radford et al., 2019)

2.2.3. Mineração de dados

A Inteligência Artificial também está intimamente relacionada com a Mineração de Dados, pois estas técnicas buscam processar, explorar e peneirar grandes conjuntos de dados brutos para descobrir regularidades, anomalias e conexões que possam ser usados para prever resultados e tomar decisões informadas. Sendo uma das áreas da IA, a Mineração de Dados fornece dados de entrada valiosos para os sistemas de IA permitindo que eles aprendam com exemplos históricos e façam previsões sobre eventos futuros. Por exemplo, em um sistema de recomendação de filmes, a Mineração de Dados pode ser usada para analisar o histórico de visualizações de um usuário e, em seguida, alimentar esses dados em um algoritmo de IA para sugerir filmes semelhantes que o usuário possa gostar. De forma similar é possível analisar o histórico e hábitos do cliente para identificar anomalias, classificá-los em função de seu perfil e sugerir novos serviços. Além disso, as técnicas de Mineração de Dados também são usadas em outras áreas, como marketing, finanças, saúde e segurança, para melhorar a tomada de decisões, identificar fraudes, prever comportamentos futuros e otimizar processos operacionais.

Segundo Han et al. (2006), a Mineração de Dados, ou Data Mining, é a extração de padrões úteis ou conhecimento implícito em grandes quantidades de dados. Este processo envolve várias etapas, incluindo a seleção e pré-processamento dos dados, a aplicação de algoritmos de mineração para identificar padrões e a interpretação dos resultados obtidos.

As técnicas de mineração de dados surgiram nos anos 1960, quando a pesquisa acadêmica passou a demandar formas de lidar com grandes volumes de dados. No entanto, só se popularizaram na década de 1990, impulsionadas pelo aumento do poder de processamento dos computadores e pelo avanço das tecnologias de armazenamento e análise de dados. Esse crescimento ocorreu em resposta à crescente quantidade de informações acumuladas nas organizações e à necessidade de extrair insights e conhecimento útil a partir desses dados. A mineração de dados utiliza uma variedade de técnicas, incluindo *clustering*, classificação, regressão e redes neurais.

Entre os pioneiros nesta área, destacam-se Rakesh Agrawal e Jiawei Han. Rakesh é bem conhecido por desenvolver conceitos e tecnologias fundamentais de mineração de dados utilizados por grandes empresas como IBM e Microsoft. Seus artigos estão entre os mais citados nas áreas de bancos de dados e mineração de dados. Han contribuiu significativamente para a mineração de dados, mineração de texto, sistemas de banco de dados, redes de informação, através de suas pesquisas em *clustering*, classificação e mineração de padrões sequenciais. Seus livros e publicações são amplamente utilizados em cursos e pesquisas sobre mineração de dados.

2.2.4. Aprendizado de Máquina

O Aprendizado De Máquina (*Machine Learning* – ML ou Aprendizado Automático) é outra das áreas da inteligência artificial (IA). O ML foca no desenvolvimento de algoritmos e modelos computacionais capazes de aprender e melhorar seus desempenhos a partir de dados, sem a necessidade de serem explicitamente programados para realizar uma tarefa específica. Mimetizando as habilidades humanas, esse processo de aprendizado baseia-se na reprodução de ações ou habilidades de pessoas, na identificação de estereótipos e tendencias nos dados, bem como na aplicação desses padrões para fazer previsões, tomar decisões ou adquirir novas habilidades e conhecimentos observados nas melhores práticas. Ou seja, o ML permite que os computadores aprendam com experiências passadas para realizar tarefas futuras de forma

mais eficiente e precisa.

Conforme descrito por Tian et al. (2012), o *machine learning* é uma extensão da ciência da computação, que, em conjunto com aplicações da estatística, permitiu que os softwares aprendessem com modelos de comportamento, sendo utilizados principalmente em problemas de classificação e predição.

Murphy (2012, p.1) define aprendizado de máquina como "um conjunto de métodos que podem detectar padrões automaticamente em dados e, em seguida, usar esses padrões descobertos para prever dados futuros ou para realizar outros tipos de tomada de decisão sob incerteza." Em outras palavras, *machine learning* é uma abordagem computacional que se baseia na construção e estudo de sistemas que podem aprender com dados, em vez de seguir explicitamente instruções programadas.

O tema de aprendizado de máquina, assim como os modelos de detecção de fraudes e anomalias, evoluiu consideravelmente ao longo das últimas décadas, passando por diferentes fases que refletem avanços tecnológicos e teóricos. Essa evolução ocorreu tanto no desenvolvimento de novas abordagens teóricas e metodológicas quanto na aplicação prática, impulsionada pela disponibilidade de grandes bancos de dados estruturados, pelo aprimoramento dos algoritmos e pela crescente demanda por soluções mais sofisticadas e escaláveis.

O campo de aprendizado de máquina teve suas origens na década de 1950 e 1960, com contribuições significativas de pioneiros como Alan Turing e Arthur Samuel. Alan Turing (2009), na década de 1950, explorou a ideia de máquinas capazes de aprender com a experiência, introduzindo o conceito de "máquina de aprendizado" e lançando as bases para o campo da inteligência artificial. Samuel (1959), em seu trabalho com o programa de xadrez de autotreinamento em 1956, desenvolveu o primeiro programa de aprendizado de máquina conhecido, que aprimorou seu desempenho à medida que jogava. Esses esforços pioneiros estabeleceram os fundamentos teóricos iniciais do aprendizado de máquina.

Nas décadas seguintes, o desenvolvimento de algoritmos e métodos de aprendizado de máquina tornou-se significativamente mais robusto e escalável. Na década de 1970 surgiram técnicas como a árvore de decisão — posteriormente formalizadas por Breiman et al. (1984) — e as redes neurais artificiais, cujas bases foram estabelecidas por McCulloch e Pitts (1943) e aprimoradas por Rosenblatt (1958) com o perceptron. Já na década de 1980, houve avanços

em algoritmos de vizinhos mais próximos (k-NN), originalmente proposto por Fix e Hodges (1951), em um artigo intitulado *"Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties".* Houve também progressos em técnicas de agrupamento, com destaque para o algoritmo k-means, introduzido por MacQueen (1967), ampliando substancialmente o escopo de aplicação das abordagens de aprendizado de máquina.

No final do século XX e o início do século XXI, com o advento e expansão do uso da internet, concomitante com o crescimento exponencial dos dados digitais, métodos de aprendizado de máquina capazes de lidar com grandes volumes de dados de forma eficiente tornaram-se uma prioridade. O aprendizado profundo emergiu como uma área promissora, com redes neurais profundas capazes de aprender representações complexas dos dados. Seu desenvolvimento foi influenciado por trabalhos como Fukushima (1980) com a rede neocognitron, Rumelhart et al. (1986) nos avanços do algoritmo de retropropagação e LeCun et al. (2015) sobre redes neurais convolucionais e o uso de GPUs para acelerar o treinamento de modelos de aprendizado profundo.

Na última década, o campo do aprendizado de máquina tem se tornado cada vez mais interdisciplinar, recebendo contribuições significativas de áreas como neurociência, psicologia e ciência da computação. Além disso, o aprendizado por reforço emergiu como uma área de pesquisa importante, especialmente em domínios como jogos, robótica e automação. Essa abordagem permite treinar agentes de inteligência artificial a interagir com o ambiente e aprender com as consequências de suas ações. Trabalhos como os de Sutton e Barto (1998) estabeleceram as bases teóricas do aprendizado por reforço, enquanto Silver et al. (2016) demonstraram seu potencial prático com o AlphaGo, um marco no desenvolvimento de sistemas autônomos capazes de superar o desempenho humano em tarefas complexas.

Assim, o aprendizado de máquina (ML) e a mineração de dados, como área da Inteligência artificial, desempenham um papel crucial na área de segurança da informação, oferecendo diversas aplicações que contribuem para a proteção de sistemas, redes e dados contra ameaças cibernéticas.

Na detecção de anomalias, seus algoritmos podem identificar padrões incomuns nos dados, ajudando a detectar atividades suspeitas que possam indicar ataques ou violações de segurança. Na prevenção de fraudes, a análise de padrões de comportamento e transações pode auxiliar na identificação de intrusos e atividades fraudulentas, protegendo organizações

e usuários contra fraudes financeiras e digitais.

Além disso, as técnicas de ML e Data Mining são capazes de identificar e classificar Malwares (softwares maliciosos projetados para extrair dados que podem ser utilizados para obter ganhos financeiros ilícitos) garantindo a autenticidade das identidades o acesso seguro a sistemas e dados. Portanto, o uso do ML e Data Mining na Segurança da Informação proporciona uma abordagem proativa e eficaz para proteger sistemas e dados contra ameaças cibernéticas, fortalecendo a resiliência das organizações em um cenário digital cada vez mais complexo e dinâmico.

2.2.5. Principais características dos modelos supervisionados e não supervisionado

Uma característica dos modelos de Aprendizado Automático (ML) é que usam tanto técnicas econométricas e estatística computacional quanto processos de otimização matemática. A forma de aprendizado dos algoritmos pode ser classificada em: supervisionado, não supervisionado e por reforço.

O conceito de aprendizado supervisionado começou a ganhar forma na década de 1950, com o surgimento de algoritmos como a regressão linear e o perceptron, desenvolvido por Rosenblatt (1958).

No aprendizado supervisionado, parte-se de um conjunto de dados rotulados, ou seja, cada observação no conjunto de dados vem acompanhada de uma resposta correta ou saída esperada (rótulo). Geralmente esse conjunto de dados é dividido aleatoriamente em dados de treinamento e dados de teste. Usando os dados de treinamento, o modelo aprende a mapear corretamente as entradas para as saídas e faz previsões para cada observação de entrada e os compara com os rótulos fornecidos. Posteriormente, com base na diferença entre a previsão e o rótulo verdadeiro, estima-se uma função de erro ou perda, utilizada para ajustar os parâmetros do modelo, de forma que as previsões futuras fiquem cada vez mais próximas dos rótulos corretos.

Após o treinamento, o modelo deve ser capaz de generalizar, ou seja, fazer boas previsões com os dados de teste que não foram vistos durante o treinamento, com base nos padrões aprendidos. Ou seja, após o treinamento do modelo, é necessário avaliar se o modelo aprendeu adequadamente e se pode generalizar adequadamente para novos dados.

Desta forma, o conjunto de dados de teste serve para avaliar essa capacidade de generalização. Assim, o objetivo de um modelo supervisionado é não apenas aprender com os

dados de treinamento, mas ser capaz de fazer boas previsões para novos dados (Gregório, 2018). Em concordância com o objetivo da pesquisa, os modelos de aprendizado supervisionado usam métodos de Regressão, Classificação K-Vizinhos Mais Próximos (KNN), Árvores de Decisão, Redes Neurais entre outros.

Em contraposição ao modelo anterior, o aprendizado não supervisionado é um tipo de aprendizado de máquina em que o modelo é treinado com um conjunto de dados sem rótulos, ou seja, sem a saída desejada associada a cada observação. Começou a ser explorado na mesma época (década de 1950), com algoritmos como o k-means *clustering*, que foi introduzido por Hugo Steinhaus em 1956 e popularizado por James MacQueen em 1967. Seu objetivo é encontrar padrões, estruturas ou regularidades nos dados por conta própria, sem supervisão explícita.

Como definido por Bishop (2006), o aprendizado não supervisionado busca identificar estruturas ocultas nos dados, como agrupamentos (*clusters*), associações ou padrões que não são imediatamente aparentes, organizando os dados de maneira coerente com base em suas características internas. Além disso, permite a redução de dimensionalidade, que consiste em simplificar os dados, mantendo as características mais importantes e eliminando redundâncias ou ruídos, o que facilita a visualização e o processamento de grandes volumes de dados. Desta forma, é possível segmentar clientes e detectar comportamentos ou padrões incomuns em conjuntos de dados, como em fraudes financeiras, defeitos em serviços fornecidos ou problemas de segurança cibernética.

Apesar dos desafios, o aprendizado não supervisionado tem uma ampla gama de aplicações em áreas como análise exploratória de dados, reconhecimento de padrões, segmentação de mercado e processamento de linguagem natural. Ao explorar a estrutura interna dos dados, o aprendizado não supervisionado pode revelar insights valiosos e padrões ocultos que, de outra forma, poderiam passar despercebidos.

A terceira forma de aprendizagem de máquinas, a aprendizagem por reforço, é aquela em que o modelo tenta aprender qual é a melhor ação a ser tomada, dependendo das circunstâncias na qual essa ação será executada. Este conceito foi formalizado por Sutton e Barto (2018) que definiram o aprendizado por reforço como o aprendizado de máquina de como agir para maximizar uma medida de recompensa.

O aprendizado por reforço trata-se, portanto, uma das formas de aprendizado de máquina em que um agente aprende a tomar ações em um ambiente para maximizar uma recompensa

cumulativa ao longo do tempo. Desta forma, é possível vincular recompensas e punições ao aprendizado do modelo, ponderando-as da forma certa.

2.2.6. Principais tipos de modelos de ML: classificação, regressão, agrupamento

No ramo da Ciência de Dados, existem diversos tipos de modelos de Aprendizado de Máquinas, cada um com características e aplicações específicas. Entre os tipos de modelos são: classificação, regressão e agrupamento.

Os modelos de classificação são uma ferramenta poderosa para prever a classe a qual um novo dado pertence. Essa classificação pode ser binária (por exemplo, um e-mail ser spam ou não spam) ou multiclasse (por exemplo, qual o tipo de flor, qual a raça de cachorro, entre outros).

Segundo Mitchell (1997), um classificador aprende a partir de um conjunto de treinamento, onde cada exemplo é um par composto de uma entrada de descrição do objeto e uma classe. Os algoritmos de classificação são treinados com exemplos rotulados e, em seguida, usam esses exemplos para prever a classe de novos dados não rotulados.

Existem diversos tipos de modelos de classificação, cada um com suas vantagens e desvantagens. Alguns exemplos incluem Regressão Logística, K-*Nearest Neighbors* (KNN), Árvores de Decisão, *Support Vector Machine* (SVM) e Redes Neurais. A escolha do modelo mais adequado para uma tarefa específica depende de diversos fatores, como o tipo de dado, o objetivo da tarefa e a quantidade de dados disponíveis. (Hastie et al., 2009)

Já os modelos de aprendizado de máquina do tipo regressão são utilizados para prever um valor numérico contínuo. Segundo Hastie et al., (2009), o objetivo da regressão é modelar a relação entre uma ou mais variáveis de entrada e uma variável de saída. Os algoritmos de regressão são treinados com dados que possuem pares de entrada e saída, e o objetivo é encontrar uma função que mapeie as entradas para as saídas de forma mais precisa possível. Essa previsão pode ser utilizada para estimar valores futuros, como salário, o preço de uma ação ou a temperatura em um determinado dia.

Além da regressão linear, árvores de decisão para regressão e redes neurais, existem vários outros modelos amplamente utilizados. Exemplos incluem: Regressão Polinomial, que lida com relações não lineares; *Random Forest* e *Gradient Boosting Machines* (GBMs), como XGBoost e LightGBM, que oferecem maior precisão e robustez; *Support Vector Regression* (SVR), adequado para problemas com poucos dados; K-*Nearest Neighbors* (KNN), que usa os valores dos vizinhos

mais próximos; e *Gaussian Process Regression*, útil para lidar com incertezas. Para dados sequenciais, modelos baseados em séries temporais, como ARIMA, Prophet ou Redes Neurais Recorrentes (RNNs), são especialmente eficazes. Também, estão disponíveis os Modelos Lineares Generalizados (MLGs) que são uma extensão poderosa dos modelos de regressão linear tradicional, desenvolvidos para lidar com situações em que a variável resposta não segue uma distribuição normal, o que é bastante comum em dados reais — como contagens, proporções ou classificações binárias e multinominais. Esses diversos modelos oferecem um arsenal poderoso para lidar com diferentes tipos de problemas e dados (Hastie et al., 2009).

O terceiro tipo de modelo de ML consiste no agrupamento, ou *clustering*. Este é um tipo de modelo de aprendizado de máquina considerado não supervisionado, pois seu objetivo é organizar instâncias de dados semelhantes em grupos, chamados clusters. Segundo descreve Bishop (2006), o agrupamento envolve a atribuição de objetos a grupos de modo que os objetos no mesmo grupo sejam mais semelhantes entre si do que com aqueles em outros grupos. Diferentemente dos modelos supervisionados, os algoritmos de agrupamento não requerem exemplos rotulados para treinamento e buscam identificar estruturas intrínsecas nos dados.

Um dos algoritmos de agrupamento mais populares é o K-Means, que particiona os dados em k clusters, onde k é um parâmetro definido pelo usuário. Por exemplo, em um conjunto de dados de clientes de uma empresa, o K-Means pode identificar grupos com padrões de compra semelhantes, permitindo à empresa personalizar suas estratégias de marketing.

Além do K-Means, há o *Hierarchical Clustering* que cria uma hierarquia de clusters. O DBSCAN identifica formas arbitrárias e lida bem com outliers. *Gaussian Mixture Models* (GMM) usa distribuições para flexibilidade. Mean-Shift e Affinity Propagation detectam clusters baseados na densidade e em mensagens entre pontos. *Spectral Clustering* e Birch são eficazes para clusters complexos e grandes conjuntos de dados. OPTICS revela mais da estrutura dos dados além do DBSCAN, e algoritmos de *Clustering Fuzzy* permitem pertencer a múltiplos clusters. Cada técnica se adapta a diferentes tamanhos, formatos e distribuições de dados.

2.2.7. Principais algoritmos de aprendizado de máquina supervisionados

A seguir, apresenta-se uma síntese das principais técnicas de aprendizado de máquina, incluindo Regressão Logística, *Naive Bayes*, *Random Forest*, *Extreme Gradient Boosting* (XGBoost), *Support Vector Machine* (SVM) linear e Redes Neurais Artificiais

2.2.7.1. Regressão Logística

A regressão logística, também chamada de análise logit, no âmbito das finanças e analise de riscos, é uma técnica frequentemente adotada por instituições financeiras e bureaus de crédito. Esse método de análise binária é ideal para situações em que a variável dependente é categórica e apresenta apenas dois possíveis resultados: Um exemplo comum na modelagem financeira é a previsão de descumprimento de obrigações, em que um indivíduo ou empresa pode ser classificado como solvente (representado pelo número zero) ou insolvente (representado pelo número um). Outro exemplo de aplicação na identificação de fraudes é o de um serviço bancário on-line capaz de determinar se uma transação em curso é fraudulenta, utilizando informações como o endereço IP do usuário, localização geográfica, o histórico de transações e outros indicadores.

O objetivo da regressão logística é gerar uma função matemática que estime a probabilidade de uma observação pertencer a um grupo previamente determinado, com base em um do conjunto de variáveis independentes. Ou seja, atuar em um problema de classificação, estimando uma probabilidade de uma observação com um determinado perfil pertencer a uma classe. Deste modo, os coeficientes estimados pelo modelo de regressão indicam a influência de cada variável independente para a ocorrência do evento.

A função matemática central na regressão logística é a função logit, que modela o log-odds (logaritmo da chances). A forma geral da função logit é:

$$logit[p] = \log\left[\frac{p}{1-p}\right] \tag{2.2}$$

onde,

p = probabilidade do evento de interesse (por exemplo, a probabilidade de pertencimento a uma classe), que é 0 ou 1.

$\frac{p}{1-p}$ é razão de probabilidades (chance).

O uso do logaritmo do odds é fundamental em regressão logística, pois transforma probabilidades restritas ao intervalo [0, 1] em valores de $-\infty$ a $+\infty$, permitindo que o modelo

aplique uma relação linear entre as variáveis independentes e o log-odds.

$$\ln\left(\frac{p}{1-p}\right) = X'\boldsymbol{\beta}$$

$$\frac{p}{1-p} = e^{X'\boldsymbol{\beta}}$$

$$p = e^{X'\boldsymbol{\beta}}(1-p) = e^{X'\boldsymbol{\beta}} - p\, e^{X'\boldsymbol{\beta}}$$

$$p\, e^{X'\boldsymbol{\beta}} + p = e^{X'\boldsymbol{\beta}} = p(1 + e^{X'\boldsymbol{\beta}})$$

$$p = \frac{e^{X'\boldsymbol{\beta}}}{(1 + e^{X'\boldsymbol{\beta}})} = \frac{\frac{e^{X'\boldsymbol{\beta}}}{e^{X'\boldsymbol{\beta}}}}{\frac{1 + e^{X'\boldsymbol{\beta}}}{e^{X'\boldsymbol{\beta}}}} = \frac{1}{(1 + e^{-X'\boldsymbol{\beta}})}$$

Assim, na regressão logística, a probabilidade p é modelada como uma função das variáveis independentes X. A equação é:

$$p = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \tag{2.3}$$

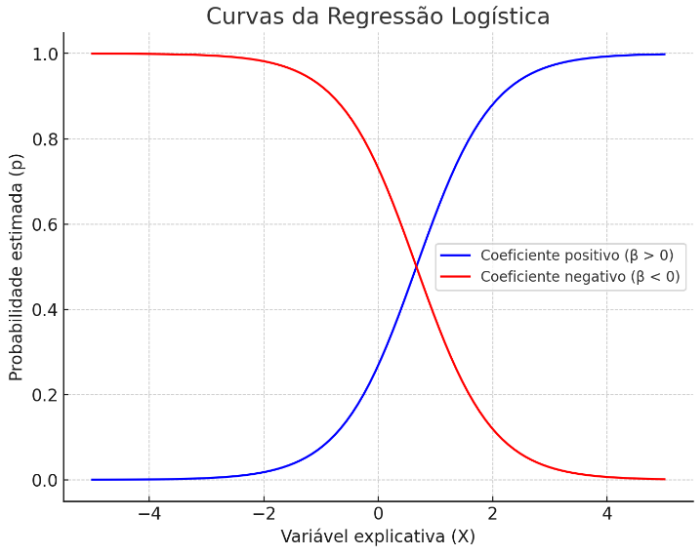
$x_1, x_2 \ldots x_n$ = são as variáveis independentes

β = são os coeficientes do modelo.

Ao interpretar os coeficientes de um modelo de classificação, cada β representa o efeito da variável explicativa sobre os log-odds do evento, sendo que o expoente de β (exp(β)) é entendido como uma razão de chances (odds ratio). Por exemplo, se exp(β) = 1,5, isso indica que a variável em questão aumenta as chances do evento ocorrer em 50%, assumindo que as demais variáveis se mantêm constantes.

Esta equação representa a função sigmoide ||mostrada na Figura x, em que a curva azul com o coeficiente positivo (β>0), indica que à medida que X aumenta, a probabilidade estimada também aumenta. A curva vermelha com coeficiente negativo (β<0) mostra que à medida que X aumenta, a probabilidade estimada diminui. Assim, mapeia-se qualquer valor real para um valor entre 0 e 1, interpretado como a probabilidade ocorrência do evento de interesse. A soma $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$ é a combinação linear das variáveis independentes, ponderada pelos seus respectivos coeficientes. Essa combinação linear é então transformada pela função logística para gerar uma probabilidade.

Figura 2.1 - Exemplos de curvas de modelo de Regressão Logísica



Fonte: Elaboração própria.

Para estimar os coeficientes β usa-se o método de máxima verossimilhança que, a partir de uma distribuição de Bernoulli de n observações independentes, é definida por:

$$l(y_1, \dots y_n, \beta) = \prod_{i=1}^{n} p^{y_i}(1-p)^{(1-y_i)}$$

2.2.7.2. Naive Bayes

O algoritmo Naive Bayes é um algoritmo de classificação probabilístico binário muito utilizado em *machine learning*. Baseado no Teorema de Bayes, este modelo é frequentemente aplicado em processamento de linguagem natural e diagnósticos médicos, entre outros. O teorema de Bayes trata da probabilidade condicional, isto é, a probabilidade de o evento A ocorrer, dado o evento B. (LEWIS, 1998)

Desta forma o Teorema de Bayes, que define a probabilidade posterior como:

$$P(C_k \mid x) = \frac{P(x \mid C_k)P(C_k)}{P(x)} \qquad (2.4)$$

Sendo:

$P(C_k \mid x)$: Probabilidade da classe C_k dado o vetor de características x. (Probabilidade posterior)

$P(x \mid C_k)$: Probabilidade do vetor de características x ser observado, dado que a classe é C_k . (Probabilidade condicional)

$P(C_k)$: Probabilidade a priori da classe C_k, obtida da distribuição dos dados.

$P(x)$: Probabilidade de observar o vetor x (Constante para todas as classes).

O algoritmo supõe que todas as características (variáveis explicativas ou features) são independentes entre si, mesmo que isso não seja verdade na prática, por isso seu nome *naive* - ingênuo. (LEWIS, 1998). Ou seja, parte-se da hipótese de que a ocorrência de um evento A em nada interfere na probabilidade de ocorrência do outro evento, B, portanto, a probabilidade de ambos ocorrerem é igual ao produto de suas probabilidades. Essa simplificação facilita o cálculo das probabilidades e torna o algoritmo eficiente usando método de máxima verossimilhança que determina os parâmetros das distribuições das features.

Assim, $P(x \mid C_k)$ pode ser decomposto como:

$$P(x \mid C_k) = \prod_{i=1}^{n} \frac{P(x_i \mid C_k) P(C_k)}{P(x)} \tag{2.5}$$

sendo:

x_i: Cada característica (ou variável) no vetor x.

n: Número total de características.

A classe predita C_{pred} é aquela que maximiza a probabilidade posterior ($P(C_k \mid x)$. Desta forma, a rede bayesiana é descrita assim:

$$C_{\text{pred}} = \arg\max_{C_k} P(C_k \mid x) \tag{2.6}$$

A classe C_k escolhida é a que maximiza a probabilidade posterior:

Substituindo $P(x_i \mid C_k)$, o Teorema de Bayes e a suposição de independência condicional:

$$C_{\text{pred}} = \arg\max_{C_k} \left(P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k) \right) \tag{2.7}$$

2.2.7.3. Random Forest

No universo do aprendizado de máquina, uma das formas de melhorar a capacidade dos algoritmos é por meio da combinação destes. Neste artigo, faz parte desta classe de modelos o *Random Forest* e Xgboost.

Estes algoritmos são conhecidos como do tipo *Ensemble,* ou seja, que combinam modelos simples e de baixo poder preditivo (*weak models*), para produzir um único forte, robusto e com maior acurácia. As principais metodologias de Ensemble são: *Bagging* e *Boosting*.

A metodologia *bagging*, proposta utilizada no *Random Forest*, foi proposta por Breiman (2001), tem por objetivo reduzir a variância das predições. Vários algoritmos são treinados separadamente em diversas reamostragens com reposição do mesmo conjunto de treinamento. De maneira geral, o método *bagging* se baseia na:

- Construção das bases de treinamento utilizando *bootstrap* na base de treinamento original. Amostragem com reposição para formação dos dados;
- Criar múltiplos algoritmos construídos para cada conjunto de dado reamostrado;
- Combinar os algoritmos: As predições são combinadas utilizando médias, moda, mediana para regressão ou voto majoritário para problemas de classificação.

O algoritmo *Random Forest* opera construindo múltiplas árvores de decisão durante o treinamento e produzindo a classe que é a moda das classes (classificação) ou a média das previsões (regressão) das árvores individuais. No caso, este modelo combina várias arvores de decisão e os valores combinados tendem a ser mais robusto que o valor gerado por um único modelo. O modelo constrói várias árvores pouco correlacionadas, onde a principal melhoria das árvores combinadas é a redução da variância. (JAMES et al, 2013)

Destaca-se como vantagem da técnica de *Random Forest* a capacidade de lidar com dados em grandes volumes e com muitas variáveis e a habilidade de identificar as variáveis mais significativas dentro de um conjunto de variáveis de entrada. Em contrapartida, como desvantagem, o modelo pode facilmente superajustar a base de dados de treino (*overfitting*), assim como dar maior importância para variáveis altamente categorizadas, mesmo que estas não possuam alto poder explicativo, além deste modelo ser de difícil interpretação. (JAMES et al, 2013)

A formulação matemática do algoritmo Random Forest é baseada na combinação de modelos (métodos ensemble) e pode ser dividida em três etapas:

1. Amostragem Bootstrap:

Dado um conjunto de treinamento D contendo N exemplos, são gerados B subconjuntos D_b (para b=1, 2, ..., B) por meio de amostragem com reposição. Cada subconjunto D_b pode conter exemplos repetidos e nem todos os exemplos de D necessariamente aparecem em cada D_b.

2. Construção das Árvores de Decisão:

Para cada subconjunto D_b, é construída uma árvore de decisão $h\_b\,(x)$ Durante a construção de cada árvore, em cada nó de decisão, é selecionado aleatoriamente um subconjunto de m atributos dentre os M atributos disponíveis ($m \ll$M). A melhor divisão no nó é escolhida apenas dentro desse subconjunto aleatório de atributos, promovendo diversidade entre as árvores.

3. Combinação das Predições:

- Classificação:

A predição final y para uma nova entrada x é determinada pelo voto majoritário das predições individuais das árvores:

$$h_t(x), para\ t\ =\ 1,2,\ldots,T \tag{2.8}$$

Onde $h_t(x)$, representa a predição da árvore t para a entrada x

$$\hat{y} = argmax_{c \in C} \sum_{t=1}^{T} 1\{\, h_t(x)\ = c\} \tag{2.9}$$

Onde:

- c é o conjunto de classes possíveis.
- $1(.)$ é a função indicadora, que vale 1 se a árvore$h_t(x)$ previu a classe c, e 0 caso contrário.

- Regressão:

A predição final é calculada pela média aritmética das predições individuais. Isso reduz a variância do modelo e melhora a capacidade de generalização:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \tag{2.10}$$

Esta formulação acima não inclui outros parâmetros que podem ser acrescentados como a função de impureza usada para os cortes, o critério de parada para o crescimento das árvores, entre outros. Além disso, a implementação exata pode variar dependendo da biblioteca de aprendizado de máquina utilizada.

2.2.7.4. XGBoost

Segundo James et al. (2013), no método *boosting,* os algoritmos são aplicados de maneira sequencial, de forma que, a cada iteração o algoritmo aplicado utiliza os resultados da iteração anterior. Ou seja, a cada iteração ajusta-se o algoritmo usando os resíduos do modelo (erros) da interação anterior como a variável dependente, no lugar da variável resposta.

Deste modo, o algoritmo XGBoost (*Extreme Gradient Boosting*), desenvolvido por Chen e Guestrin (2016) é um algoritmo de aprendizado de máquina do tipo *boosting*, e uma implementação popular e eficiente do método de aprendizado supervisionado Gradient Boosted Trees. Baseado no boosting, XGBoost se destaca pela aproximação de funções, otimizando funções de perda específicas e aplicando técnicas de regularização. Este método tem se destacado em competições de machine learning na plataforma Kaggle, muitas vezes sendo combinado com redes neurais profundas.

Aqui está uma descrição detalhada do algoritmo conforme (Chen e Guestrin, 2016). O objetivo principal do XGBoost é minimizar uma função de perda regularizada, que consiste em duas partes:

1. Função de Perda ($\mathcal{L}$): Mede a discrepância entre as previsões do modelo e os valores reais.
2. Termo de Regularização (Ω): Controla a complexidade do modelo para evitar overfitting.

Matematicamente, a função objetivo pode ser expressa como:

$$\mathcal{L}(\Phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.11}$$

Onde:

- n é o número de amostras.
- l é a função de perda.
- y_i são os valores reais.
- $\hat{y}_i$ são as previsões do modelo.
- K é o número de árvores.
- $\Omega(f_k)$ é o termo de regularização para a K-ésima árvore.

O XGBoost constrói as árvores de decisão de forma sequencial, onde cada nova árvore tenta corrigir os erros residuais das árvores anteriores. Para cada nó da árvore, o algoritmo busca a melhor divisão que minimiza a função objetivo. Isso é feito calculando os gradientes de primeira e segunda ordem da função de perda em relação às previsões.

A atualização das previsões é dada por:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot f_t(x_i) \tag{2.12}$$

onde:

- t é a iteração atual.
- η é a taxa de aprendizado.
- $f_t(x_i)$ é a previsão da t-ésima árvore para a amostra x_i.

O termo de regularização no XGBoost é definido como:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2.13}$$

onde:

- T é o número de folhas na árvore.
- w_j são os pesos das folhas.
- γ e λ são hiperparâmetros que controlam a complexidade da árvore.

O principal fator por trás do sucesso do XGBoost é sua escalabilidade em todos os cenários devido a sua otimização algorítmica. O sistema é capaz de rodar mais de dez vezes mais rápido do que outras soluções populares em uma única máquina e pode escalar para bilhões de exemplos em configurações distribuídas ou com recursos limitados de memória (Chen e Guestrin, 2016).

2.2.7.5. Suport Vector Machine - SVM

Outra técnica amplamente conhecida nos modelos de aprendizado de máquina para classificação é o *Support Vector Machine* – SVM. Desenvolvido por Boser, Guyon e Vapnik (1992), o SVM é um algoritmo de aprendizado supervisionado utilizado tanto para classificação quanto para regressão. Trata-se de um classificador linear binário não-probabilístico que classifica os dados sempre em apenas duas classes

Segundo Betancourt (2005), o SVM tem como vantagens: i) a facilidade de treinar; ii) não apresenta um ótimo local, como nas redes neurais; iii) escala relativamente bem para dados em espaços de alta dimensão; iv) a relação entre a complexidade do classificador e o erro pode ser explicitamente controlado; e v) dados não tradicionais, como caracteres, podem ser usados como entrada, em vez de vetores de recursos.

Por outro lado, a fraqueza do SVM é a necessidade de uma função "boa" do *kernel*, ou seja, são necessárias metodologias eficientes para ajustar os parâmetros de inicialização do SVM. (BETANCOURT, 2005)

O SVM opera encontrando os vetores de suporte, que são os pontos de dados mais próximos ao hiperplano de separação. Esses pontos determinam a margem, que corresponde à distância entre o hiperplano e os vetores de suporte. O objetivo do SVM é maximizar essa margem, garantindo uma melhor generalização do modelo.

A formulação matemática do algoritmo SVM é representada por (Boser, Guyon e Vapnik, 1992). A ideia central do SVM é identificar um hiperplano que separa os dados de diferentes classes de maneira ótima. Para conjuntos de dados linearmente separáveis, o SVM busca o hiperplano que maximiza a distância (margem) entre as duas classes mais próximas, conhecidas como vetores de suporte.

Matematicamente, consideremos um conjunto de treinamento com n amostras, onde cada amostra x_i pertence a uma das duas classes $y_i \in \{-1,+1\}$. O hiperplano de decisão pode ser definido pela equação:

$$w^{\top x} + b = 0 \tag{2.14}$$

onde:

w é o vetor de pesos normal ao hiperplano.

b é o termo de bias (deslocamento).

O objetivo é encontrar w e b que maximizem a margem entre as classes, sujeita às restrições de que todas as amostras sejam classificadas corretamente. Isso leva a um problema de otimização quadrática sujeito a restrições lineares.

A formulação de otimização do SVM pode ser expressa da seguinte maneira:

Maximizar a margem:

$$\text{Margem} = \frac{2}{|w|} \tag{2.15}$$

Sujeito a:

$$y_i(w^\top x_i + b) \geq 1, \quad \forall i = 1,2,\dots,n \tag{2.16}$$

Essa formulação busca maximizar a margem minimizando $\| w \|$, pois $\frac{2}{|w|}$ aumenta à medida que $|w|$diminui. Enquanto garante que todas as amostras estejam do lado correto do hiperplano com uma distância mínima de 1.

2.2.7.6. Redes Neurais Artificiais - ANN

Por fim, as Redes Neurais Artificiais (ANNs) são modelos inspirados no funcionamento do cérebro humano, descritos inicialmente por McCulloch e Pitts (1943). Eles desenvolveram o perceptron, um sistema que simula as características básicas de um neurônio biológico. As ANNs são compostas por múltiplas camadas de neurônios artificiais e são amplamente utilizadas em tarefas de classificação, regressão e outras aplicações de aprendizado de máquina.

Desta forma, as ANNs são uma técnica de processamento de informações inspirada pelo sistema nervoso humano. Conforme descrito por Haykin (2007), o cérebro humano pode ser considerado um sistema de processamento de informação extremamente complexo, não linear e paralelo, que realiza diversas atividades de maneira muito mais eficaz do que os sistemas computacionais convencionais.

Aqui está uma descrição detalhada do algoritmo:

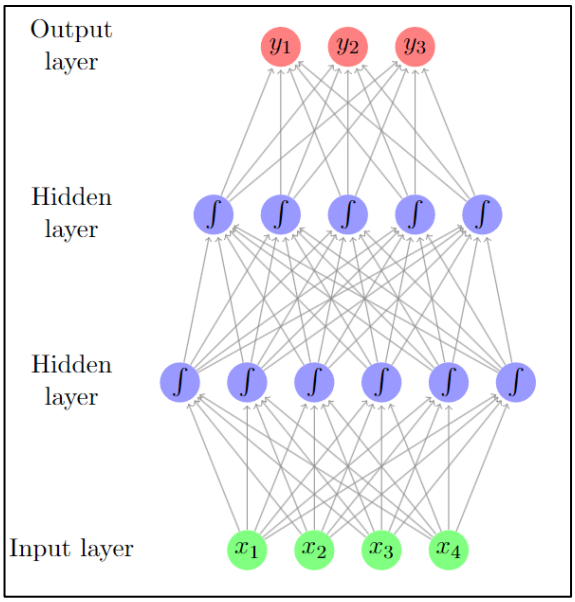
i) Neurônios: As ANNs são compostas por unidades de processamento chamadas neurônios ou nós. Cada neurônio recebe várias entradas, aplica uma função de ativação e produz uma saída.

ii) Pesos e Viés: Cada entrada de um neurônio é multiplicada por um peso e o viés é adicionado. Os pesos e o viés são os parâmetros do modelo que são aprendidos durante o treinamento.

iii) Função de Ativação: A função de ativação transforma a soma ponderada das entradas em uma saída. Exemplos comuns de funções de ativação incluem a função sigmóide, a função tangente hiperbólica e a função ReLU.

iv) Camadas: As ANNs são geralmente organizadas em camadas. Existem três tipos de camadas: a camada de entrada, as camadas ocultas e a camada de saída.

Figura 2.2 - Rede Neural feed-forward com duas camadas ocultas



Fonte: Reprodução de Goldberg (2016).

A formulação matemática de uma ANN é a seguinte:

Dado um conjunto de treinamento:

$$\{(x_i, y_i)\}_{i=1}^{n} \tag{2.17}$$

onde $x_i \in \mathbb{R}^d$ é a entrada e y_ié a saída correspondente, a saída de um neurônio é dada por:

$$f(x) = \sigma(w^T x + b) \tag{2.18}$$

Onde:

w: vetor de pesos;

b: viés;

σ: viés: função de ativação e;

T: denota transposição

Durante o treinamento, os pesos e o viés são atualizados para minimizar uma função de perda, que mede a diferença entre a saída da rede e a saída desejada.

2.2.8. Principais algoritmos de aprendizado de máquina não supervisionados

Desta forma, a seguir é apresentada uma síntese de algumas das principais técnicas de aprendizado de máquina não supervisionados, tais quais: K-means; DBSCAN; Análise de componentes principais (PCA); Isolation Forest; e Autoencoder

2.2.8.1. K-means

O algoritmo K-Means é um dos métodos mais populares e simples de aprendizado de máquina não supervisionado para a tarefa de clusterização. Ele organiza um conjunto de dados em um número pré-definido de grupos (clusters) "k". Seu principal objetivo é minimizar a variância intra-cluster, garantindo que os pontos dentro de cada grupo sejam o mais semelhantes possível entre si e o mais distintos possível dos pontos em outros grupos.

O k-means funciona através das seguintes etapas:

1. Inicialização: Escolhem-se k centroides iniciais. Isso pode ser feito de forma aleatória ou por métodos mais sofisticados, como o k-means++.
2. Atribuição: Cada ponto de dados é atribuído ao centroide mais próximo, formando k clusters.
3. Atualização: Para cada cluster, calcula-se um novo centroide, que é a média aritmética de todos os pontos que pertencem ao cluster.
4. Convergência: Repete-se os passos de atribuição e atualização até que os centroides não mudem mais ou até que o número máximo de iterações seja atingido.

O resultado final é um conjunto de clusters em que a soma das distâncias dos pontos ao seu centroide correspondente é minimizada. No entanto, como o algoritmo pode convergir para mínimos locais, a escolha inicial dos centróides pode influenciar significativamente o resultado final.

A formulação matemática do algoritmo k-means pode ser descrita como um problema de minimização, onde o objetivo é minimizar a soma das distâncias quadradas entre cada ponto de dados e o centroide do cluster ao qual ele pertence.

A formulação matemática do algoritmo k-means pode ser descrita como um problema de minimização, onde o objetivo é minimizar a soma das distâncias quadradas entre cada ponto de dados e o centroide do cluster ao qual ele pertence.

Um conjunto de dados $X = \{ x_1, x_2, \dots, x_n\}$, onde cada $x_n \in \mathbb{R}^d$ é um vetor de dados de dimensão d. O objetivo do algoritmo é particionar X em k clusters $C = \{C_1, C_2, \dots, C_n\}$, com cada cluster C_j representado por um centroide $\mu_j \in \mathbb{R}^d$.

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \| x_i - \mu_j \|^2 \tag{2.19}$$

onde:

$\| x_i - \mu_j \|^2$ é a distância euclidiana quadrada entre o ponto de dados x_i e o centroide do cluster μ_j .

J é a soma total das distâncias quadradas de todos os pontos de dados ao seu respectivo centroide.

Essa formulação matemática destaca a natureza iterativa e o objetivo de minimização do algoritmo k-means, que busca agrupar os dados em clusters de tal forma que a variabilidade dentro de cada cluster seja minimizada.

2.2.8.2. DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) é um algoritmo de *clustering* baseado em densidade que agrupa pontos em regiões de alta densidade separadas por áreas esparsas. Proposto por Ester et al. (1996), ele classifica os pontos em três categorias: núcleo, borda ou ruído, com base na quantidade de vizinhos dentro de um raio ϵ e no número mínimo de pontos necessários para formar um núcleo (min Pts).

O algoritmo inicia um cluster a partir de cada ponto núcleo, incorporando seus vizinhos diretos dentro de ϵ. O processo se expande iterativamente, adicionando pontos de borda ao cluster até que não haja mais conexões válidas. Pontos que não atendem aos critérios de núcleo e não estão próximos a nenhum núcleo são considerados ruído.

O DBSCAN depende de dois parâmetros: ϵ (eps) e min Pts. O conjunto de vizinhos diretos $N_\epsilon(p)$ de um ponto p é definido como:

$$N_\epsilon(p) = \{q \in D \mid dist(p,q) \leq \epsilon \tag{2.20}$$

onde:

D é o conjunto de dados.

$dist(p,q)$ é uma função de distância (geralmente distância Euclidiana).

Um ponto p é considerado um ponto central se:

$$|N_\epsilon(p)| \geq \min \text{Pts} \tag{2.21}$$

O DBSCAN começa verificando cada ponto no conjunto de dados para determinar se ele é um ponto central. Se for, o algoritmo cria um novo cluster, adicionando ao cluster o ponto e todos os seus vizinhos dentro da distância ϵ. Em seguida, explora recursivamente os vizinhos dos pontos adicionados, incluindo pontos de borda ao cluster. Este processo continua até que não haja mais pontos elegíveis para serem adicionados ao cluster. Após isso, o algoritmo move-se para o próximo ponto ainda não visitado e repete o procedimento, ignorando pontos classificados de ruído.

Dentre as vantagens do DBSCAN destacam-se a capacidade de identificar clusters de formas e tamanhos arbitrários e a robustez contra ruídos, já que ignora pontos considerados como ruído. Além disso, ele é determinístico, o que significa que produz o mesmo resultado em diferentes execuções com os mesmos dados. No entanto, as desvantagens incluem a escolha complexa de ϵ e $\min \text{Pts}$, sensibilidade a densidades de cluster variáveis e alta complexidade computacional, especialmente para conjuntos de dados grandes.

2.2.8.3. Isolation Forest

O algoritmo Isolation Forest (IF), introduzido por Liu et al. em 2008, é uma técnica popular para detecção de anomalias. A principal ideia por trás do algoritmo é que as anomalias são mais fáceis de serem "isoladas" das observações normais. O algoritmo constrói uma floresta de árvores de isolamento, onde cada árvore isola observações individuais por meio de cortes aleatórios nas dimensões dos dados.

Cada árvore de isolamento é construída recursivamente selecionando aleatoriamente uma característica e um valor de corte dentro do alcance dos dados dessa característica. A construção continua até que cada ponto de dados seja isolado em uma folha única ou até atingir uma profundidade máxima da árvore. O princípio chave é que, se um ponto é uma anomalia, ele será isolado em uma profundidade menor em comparação a pontos normais, pois será necessária uma menor quantidade de divisões para isolá-lo.

A profundidade de isolamento de uma instância é usada para calcular a pontuação de anomalia. A pontuação é determinada pela profundidade média de todas as árvores de isolamento. Assim, um ponto/observação com uma profundidade média baixa é mais provável de ser uma anomalia.

A pontuação de anomalia s(x,n) para uma instância x é calculada como:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{2.22}$$

onde:

$E(h(x))$ é a profundidade média da instância x nas árvores de isolamento e $c(n)$ é o custo de um nó externo para uma árvore binária completa, dado por:

$$c(n) = 2H(n-1) - \left(\frac{2(n-1)}{n}\right) \tag{2.23}$$

onde:

$H(i)$ sendo o i-ésimo número harmônico, aproximadamente igual a $\ln(i) + 0.5772156649$ (constante de Euler-Mascheroni).

O algoritmo Isolation Forest é altamente eficiente em termos de tempo e espaço, o que o torna adequado para grandes conjuntos de dados. Além disso, sua implementação é simples, pois não requer normalização prévia dos dados nem tratamento de valores ausentes. Ele também é eficaz em dados de alta dimensão e robusto a ruídos e outliers. No entanto, seu desempenho pode ser sensível à escolha de parâmetros, como a profundidade máxima das árvores e o número de árvores na floresta. A interpretação dos motivos pelos quais um ponto é considerado anômalo pode ser, ainda, menos intuitiva em comparação com outros métodos.

2.2.8.4. Auto Encoder

O Autoencoder é um tipo de rede neural projetada para aprender representações de dados não rotulados. Ele é composto por duas partes principais: o encoder e o decoder. O encoder mapeia a entrada para um espaço latente de menor dimensão, enquanto o decoder reconstrói a entrada original a partir desse espaço latente. Esse processo força a rede a aprender as características mais salientes dos dados (Legrand et al., 2018).

A estrutura e funcionamento do Autoencoder apresenta as etapas de encoder, decoder e função de custo. De forma simplificada o encoder busca reduzir a dimensionalidade dos dados e aprender uma representação latente significativa, enquanto o decoder busca reconstruir os dados de entrada a partir dessa representação. Assim temos:

a. Encoder: é a parte da rede neural que transforma a entrada de alta dimensionalidade em uma representação de baixa dimensionalidade, ou representação latente. Consiste em

uma ou mais camadas de neurônios que transformam a entrada x em uma representação latente z. A transformação pode ser descrita matematicamente como:

$$z = f(W_e x + b_e) \tag{2.24}$$

onde:

W_e são os pesos do encoder

b_e são os vieses do encoder

f é uma função de ativação não-linear, como a ReLU ou a Sigmoid

O objetivo do treinamento de um Autoencoder é minimizar o erro de reconstrução, que é a diferença entre os dados de entrada originais e a saída reconstruída. Ao forçar a rede a aprender a reconstruir os dados a partir de uma representação comprimida, o Autoencoder aprende a capturar as características mais importantes dos dados, ignorando o ruído e os detalhes irrelevantes. Durante o treinamento, o encoder aprende a captar as características mais significativas dos dados de entrada, comprimindo a informação e eliminando redundâncias.

b. Decoder: é a parte da rede neural que tenta reconstruir os dados originais a partir da representação latente. O decoder recebe a representação latente z e tenta reconstruir a entrada x através de transformações lineares e não lineares. Matematicamente, isso pode ser descrito como:

$$\hat{x} = f(W_d z + b_d) \tag{2.25}$$

onde:

W_d são os pesos do encoder

b_d são os vieses do encoder

g é uma função de ativação, muitas vezes a mesma usada no encoder

Desta forma, o objetivo do decoder é reconstruir a entrada original o mais próximo possível. Ele faz isso usando a informação comprimida pelo encoder para gerar uma saída que tenha uma baixa perda em relação à entrada original. Durante o treinamento, o decoder é ajustado para minimizar a diferença entre a entrada original e a saída reconstruída, o que força o encoder a aprender uma representação latente que retém a informação mais crítica.

c. Função de Custo: O objetivo do treinamento do autoencoder é minimizar a diferença entre a entrada x e a reconstrução $\hat{x}$. A função de custo é uma parte fundamental do treinamento de um autoencoder, pois define o objetivo de aprendizado e quantifica o

quão bem o autoencoder está realizando seu trabalho de reconstrução dos dados de entrada a partir da representação latente.

Essa diferença é medida através de uma função de custo, como o erro quadrático médio:

$$L(x, \hat{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{2.26}$$

A escolha da função de custo afeta significativamente a performance do autoencoder. Uma função de custo adequada ajuda o modelo a aprender a codificar e decodificar os dados de forma eficiente, mantendo a essência da informação enquanto descarta o ruído e as redundâncias. Durante o treinamento, o modelo ajusta seus pesos para minimizar a função de custo, melhorando a qualidade da reconstrução dos dados.

Além disso, a função de custo influencia a estabilidade e a convergência do treinamento do modelo. Uma função de custo mal escolhida pode levar a um treinamento instável ou a um modelo que não generaliza bem para dados novos.

2.2.8.5. Análise de Componentes Principais - PCA

A Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) é uma técnica de análise multivariada usada para reduzir a dimensionalidade de um conjunto de dados enquanto preserva a variância máxima possível. Ela transforma um conjunto de observações de variáveis possivelmente correlacionadas em um conjunto de valores de variáveis linearmente não correlacionadas, chamadas de componentes principais.

O PCA foi desenvolvido inicialmente por Pearson (1901), como um análogo ao método dos mínimos quadrados para problemas de regressão linear. Posteriormente, Hotelling (1933) expandiu os fundamentos matemáticos da técnica, permitindo a aplicação em uma ampla variedade de campos, especialmente na economia e na psicologia.

A técnica do PCA visa, então, reduzir a dimensionalidade, ou seja número de variáveis, em um conjunto de dados, mantendoo maior quantidade possível de informações. O PCA alcança isso através da matriz de covariância que captura as relações entre as diferentes variáveis. Se as variáveis forem padronizadas (isto é, tiverem variância unitária), a matriz de covariância se torna a matriz de correlação.

Antes de aplicar o PCA, os dados são normalmente padronizados para que cada variável tenha média zero e desvio padrão unitário. Sejam X os dados originais com n observações e p variáveis conforme abaixo:

$$X_{padronizado} = \frac{x - \mu}{\sigma} \tag{2.27}$$

onde:

μ é o vetor de médias das variáveis

σ é o vetor de desvios padrão.

Após padronizar os dados, calcula-se a matriz de covariância Σ conforme a seguir:

$$\sum = \frac{1}{n-1} X_{padronizado}^{T} X_{padronizado} \tag{2.28}$$

A matriz de covariância (ou correlação) é decomposta em seus autovalores e autovetores sendo que os autovetores representam as direções de maior variância nos dados, enquanto os autovalores quantificam a magnitude da variância em cada direção. Os autovetores são ordenados em ordem decrescente de seus autovalores correspondentes.

$$\sum v = \lambda v \tag{2.29}$$

Onde λ representa os autovalores e $v$os autovetores. Os autovalores indicam a variância explicada por cada componente principal. Os autovetores são ordenados de acordo com os autovalores em ordem decrescente. O autovetor correspondente ao maior autovalor é o primeiro componente principal.

O primeiro componente principal captura a maior quantidade de variância nos dados, o segundo componente principal captura a segunda maior quantidade de variância, e assim por diante. Os dados originais são projetados nos componentes principais selecionados, resultando em um novo conjunto de dados com dimensionalidade reduzida.

$$Z = X_{padronizado}\, x\, V \tag{2.30}$$

onde:

Z é a matriz dos componentes principais

V é a matriz dos autovetores ordenados.

Em suma o PCA apresenta a vantagem de reduzir a dimensionalidade dos dados, facilitando a visualização e análise, eliminando a redundância entre variáveis e melhorando o desempenho de modelos de aprendizado de máquina ao mitigar o *overfitting*. No entanto, a técnica pode levar à

perda de informação significativa, dificultar a interpretação devido aos componentes principais serem combinações lineares das variáveis originais, ser sensível à escala das variáveis e não capturar relações não lineares nos dados.

2.2.9. Principal metodologia para estruturar projetos de mineração de dados e aprendizado de máquina

No que se refere a criação e comparação de modelos com técnicas de *machine learning* a ciência de dados utiliza-se de alguns métodos para desenvolver e testar modelos. A metodologia mais utilizada na academia e no mercado é o *Cross Industry Standard Process for Data Mining*, conhecido como Crisp-DM. Esta metodologia reúne as algumas das melhores práticas mineração de dados, de forma que o processo de tratamento de dados e modelagem seja o mais produtivo e eficiente possível. (TUKEY, 1977).

Conforme descrita por Wirth e Hipp (2000), a metodologia Crisp-DM fornece uma visão geral do ciclo de vida de um projeto de mineração de dados. Ele contém seis fases, suas respectivas tarefas e seus resultados, assim divididos no Quadro 2.1 abaixo:

Quadro 2.1 - Etapas Metodologia Crisp-DM

Fases	Descrição
Entendimento do negócio	Essa fase inicial concentra-se no entendimento dos objetivos e requisitos do projeto de uma perspectiva comercial e, em seguida, na conversão desse conhecimento em uma definição de problema de mineração de dados e em um plano preliminar do projeto desenvolvido para atingir os objetivos.
Compreensão dos Dados	A fase de entendimento dos dados começa com uma coleta inicial de dados, familiarização e identificar dos problemas de qualidade dos dados. Também se descobre os primeiros insights e a formar as primeiras hipóteses.
Preparação dos dados	A fase de preparação de dados abrange todas as atividades para construir o conjunto de dados final a partir dos dados brutos iniciais. É provável que as tarefas de preparação de dados sejam executadas várias vezes no decorrer do processo. Esta etapa inclui seleção de tabelas, limpeza de dados, construção de novos atributos e transformação de dados para ferramentas de modelagem.

Modelagem	Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para valores ideais. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas requerem formatos de dados específicos
Avaliação	Nesta fase do projeto, avalia-se a performance preditiva do modelo e, antes de prosseguir para a implantação final do modelo, é importante avaliar mais detalhadamente o modelo e revisar as etapas executadas na construção do modelo para garantir que ele atinja adequadamente os objetivos de negócios. No final desta fase, uma decisão sobre o uso dos resultados da mineração de dados deve ser alcançada.
Implementação	O conhecimento adquirido pelo modelo precisará ser organizado e apresentado de forma que o cliente possa usá-lo. Esta fase de implantação pode ser tão simples quanto gerar um relatório ou tão complexa quanto implementar um processo repetível de mineração de dados. De qualquer forma, é importante saber de antemão quais ações precisarão ser realizadas para realmente fazer uso dos modelos criados.

Fonte: Wirth e Hipp (2000, p. 5-8). Adaptado. Elaboração própria.

2.2.10. Principais medidas de desempenho dos algoritimos

As principais medidas de desempenho dos algoritmos variam de acordo com o tipo de problema que está sendo abordado. Aqui estão algumas das medidas de desempenho mais comuns em diferentes contextos:

I. Classificação, conforme síntese apresentada por Bishop (2006):

- Matriz de Confusão: Um Quadro que mostra o número de instâncias de cada classe prevista pelo modelo em comparação com as classes reais. Isso permite uma análise mais detalhada do desempenho do modelo, incluindo taxa de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

Quadro 2.2 - Matriz de Confusão

	Previsto

		Negativo (N)	Positivo (P)
Real	Negativo	Verdadeiro Negativo (VN)	Falso Positivo (FP) Erro Tipo I
	Positivo	Falso Negativo (FN) Erro Tipo II	Verdadeiro Positivo (VP)

Fonte: elaboração própria

Como base neste quadro estima-se:

- Acurácia: A proporção de instâncias classificadas corretamente pelo modelo em relação ao total de instâncias. É uma medida geral da capacidade do modelo de fazer previsões corretas.
- Precisão: mede a proporção de instâncias classificadas como positivas que são realmente positivas, ou seja, a precisão é a razão entre o número de verdadeiros positivos e o número total de previsões positivas feitas pelo modelo (verdadeiros positivos e falsos positivos).
- Recall: mede a proporção de instâncias positivas que são corretamente identificadas pelo modelo, ou seja, a razão entre o número de verdadeiros positivos e o número total de casos positivos. Tanto precisão, quanto recall são medidas especialmente úteis em problemas com base de dados que apresentam desequilíbrio de classe.
- F1-Score: A média harmônica da precisão e recall, que fornece uma medida única do desempenho do modelo, equilibrando a precisão e a recall.

II. Regressão conforme explicitado por Deisenroth et al. (2020) e Bruce, Bruce e Gedeck (2020):

- Erro quadrático médio (*Mean Squared Error* - MSE): métrica que calcula a média de diferença ao quadrado entre o valor predito com o real. Por ter a diferença ao quadrado, penaliza-se valores que sejam muito diferentes entre o previsto e o real.

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2.31)$$

onde:

y: é o valor real

$\hat{y}$: é o valor previsto

n: é o número total de observações ou pontos de dados

y_i e $\hat{y}_i$: são o i-ésimo valor real e previsto, respectivamente.

- Raíz do Erro Médio Quadrático (*Root Mean Square Error* - RMSE): É a raiz quadrada da média dos quadrados das diferenças entre os valores previstos pelo modelo e os valores reais. É uma medida comum para avaliar a precisão das previsões em problemas de regressão.

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2.32)$$

- Erro Absoluto Médio (*Mean Absolute Error* - MAE): A média das diferenças absolutas entre os valores previstos pelo modelo e os valores reais. Por haver valores positivos e negativos, é adicionado um módulo entre a diferença dos valores. É menos sensível a outliers do que o MSE.

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (2.33)$$

- Erro percentual absoluto médio (*Mean Absolute Percentual Error* – MAPE): é a métrica que apresenta a porcentagem de erro médio em relação aos valores reais. Ou seja, demonstra o cálculo do valor da média da divisão entre a diferença entre o valor real e o predito sobre o valor real.

$$MAPE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| x100 \quad (2.34)$$

- R^2 (Coeficiente de Determinação): Uma medida que indica a proporção da variabilidade nos dados que é explicada pelo modelo. Valores mais altos de R^2

indicam um melhor ajuste do modelo aos dados. O resultado varia de 0 a 1, porém também podem ser expressos em termos percentuais.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2.35)$$

III. Agrupamento:

- Índice de Silhueta (*Silhouette Score*): Uma medida que quantifica o quão bem os clusters estão separados uns dos outros. Valores mais altos indicam que os pontos dentro de um cluster estão mais próximos uns dos outros do que dos pontos de outros clusters.
- Índice de Davies-Bouldin (Davies-Bouldin *Index*): Uma medida que avalia a dispersão entre os clusters. Valores mais baixos indicam clusters mais densos e bem separados.
- Índice de Calinski-Harabasz (Calinski-Harabasz *Score*): Uma medida que avalia a dispersão entre os clusters em relação à dispersão dentro dos clusters. Valores mais altos indicam clusters mais densos e bem separados.

2.3.Principais desafios e lacunas de pesquisa em fraudes e anomalias no setor financeiro

A aplicação de modelos analíticos de *machine learning* em bancos e instituições financeiras tem trazido benefícios significativos. Esses modelos são capazes de processar grandes volumes de dados em tempo real, permitindo a detecção mais rápida de atividades suspeitas. Além disso, contribuem para a redução de custos operacionais e aprimoram a precisão na detecção de fraudes financeiras. Os modelos analíticos de *machine learning* têm se mostrado altamente eficazes na identificação de anomalias financeiras, contribuindo para a prevenção de perdas substanciais e a manutenção da confiança dos clientes.

Em diversos trabalhos relacionados ao tema, detalhados com mais detalhes no artigo de revisão da literatura presente no capítulo 3, verifica-se a recorrência da proposição da criação de modelos analíticos híbridos para mitigar o risco operacional e reduzir perdas com fraudes e redução de falsos negativos desses modelos. Dentre os modelos híbridos, são sugeridos modelos que utilizam aprendizado de máquina em conjunto com regras de negócio ou a combinação em sequência de dois ou mais modelos analíticos para melhor precisão em casos de fraude. Também se verifica o forte uso de técnicas mais profundas como redes neurais e *deep learning*.

Essencialmente, os esforços de pesquisa recentes têm se concentrado na analisar variáveis comportamentais e cadastrais de usuários que foram vítimas de fraudes. Diversos autores também discutem os desafios e as limitações existentes na detecção de fraudes financeiras, principalmente a adaptação por se tratar de um processo em constante evolução. Destaca-se a importância de abordagens adaptativas e de técnicas de atualização contínua dos modelos de detecção de anomalias para enfrentar esses desafios. (Hilal et. al., 2022; Zamini & Hasheminejad, 2019; Zhang et al., 2022).

Assim, de forma resumida, os principais desafios identificados foram:

1. Dados desbalanceados: Os conjuntos de dados utilizados para detecção de anomalias e fraudes geralmente são altamente desbalanceados, com a maioria das transações sendo legítimas e apenas uma pequena proporção sendo fraudulentas, ou seja, há um forte desequilíbrio entre as classes de fraudes e não fraudes. Esse desequilíbrio dificulta o treinamento de modelos de aprendizado de máquina para detectar adequadamente as transações fraudulentas, uma vez que o desequilíbrio pode levar a uma baixa taxa de detecção de fraudes ou a muitos falsos positivos. Um alto número de falsos positivos pode levar a interrupções desnecessárias e incômodos para os clientes, enquanto uma baixa taxa de detecção pode permitir que fraudes passem despercebidas (Wei et al., 2013; Zhang et al., 2022; Ngai et al., 2011; West e Bhattacharya, 2016; Nami e Shajari, 2018)

2. Evolução de padrões: Os métodos de fraude estão em constante evolução, o que significa que as anomalias e fraudes podem apresentar padrões e características diferentes ao longo do tempo. Os modelos de detecção de anomalias precisam ser capazes de se adaptar a essas mudanças e detectar novos tipos de fraudes à medida que surgem. Destaca-se persuasões chamadas de engenharias sociais onde os fraudadores estão em constante evolução, desenvolvendo novas táticas e estratégias para evitar a detecção utilizando-se de atuação da própria vítima. Assim, os modelos de detecção devem ser atualizados regularmente e incorporar técnicas de detecção avançadas para acompanhar essas mudanças. (Zhang et al., 2022; West e Bhattacharya, 2016; Hilal et. al., 2022)

3. Detecção em tempo real: As instituições financeiras exigem sistemas de detecção de fraudes em tempo real para interromper atividades fraudulentas antes que causem grandes prejuízos. Isso impõe a necessidade de modelos eficientes e rápidos que possam processar grandes volumes de dados em tempo real e fornecer detecção em tempo hábil. Este desafio se desdobra em necessidade de grandes volumes de dados rotulados para treino dos modelos, no caso de modelos

supervisionados, e para os demais modelos, o uso de variáveis plausíveis de serem obtidas em tempo real, de baixa complexidade. Modelos com variáveis complexas tendem a ter melhor acurácia e precisão, porém não são disponíveis em curto espaço de tempo. (Zioviris et al., 2022; West e Bhattacharya, 2016; Bakumenko e Elragal, 2022; Zhou et al., 2021)

4. Interpretabilidade dos modelos: Modelos complexos de aprendizado de máquina e inteligência artificial podem ter dificuldades em fornecer explicações claras e interpretações dos resultados, o que pode representar um desafio em termos de confiança e aceitação dos modelos pelas empresas financeiras e reguladores.

5. Concentração de modelos supervisionados: maioria dos estudos para desenvolvimento de modelos preditivos de fraudes se concentram em técnicas de aprendizado supervisionado, que requerem grandes conjuntos de dados rotulados para treinamento. (Hilal et. al., 2022; Zioviris et al., 2022)

6. Uso de Inteligência Artificial nos processos: Integrar dados de diferentes fontes e sistemas, muitas vezes fragmentados e desorganizados, pode ser um processo complexo e demorado, exigindo expertise técnica e infraestrutura robusta. Do mesmo modo, novos modelos de IA Generativa pode atuar em partes importantes do desenvolvimento e treinamento do modelo. Contudo estas tecnologias demandam infraestrutura específica de hardware para atuação. Do mesmo modo, é necessária atenção para vieses e discriminação, assim quando a IA é treinada com dados tendenciosos, pode perpetuar vieses e discriminações existentes, levando a resultados injustos e à exclusão de grupos específicos.

Essas análises buscam decifrar os padrões e características distintas dos usuários, ajudando a detectar situações em que intrusos tentam se passar por clientes legítimos. Apesar dos avanços, a literatura também aponta diversos desafios, como os desafios apresentados acima. Estes desafios, bem como a proposição e desenvolvimento de modelo de detecção de anomalias não supervisionado, será objeto de desenvolvimento no artigo proposto no capítulo 5.

Neste contexto, para enfrentar o desafio do desbalanceamento de classes na detecção de fraudes financeiras, propomos no Capítulo 4 um modelo baseado em *Large Language Models* (LLMs) especializados em finanças, capaz de gerar dados sintéticos representativos de transações fraudulentas. A abordagem consiste na utilização de um LLM para aprender padrões de transações legítimas e fraudulentas, permitindo a geração de amostras sintéticas realistas que preservem a distribuição estatística e comportamental dos dados originais. O modelo utiliza técnicas de

engenharia de prompt e fine-tuning, garantindo que as fraudes sintéticas criadas reflitam as características mais recentes das anomalias detectadas. Dessa forma, cria-se uma solução inovadora para ser aplicada visando equilibrar e reduzir o impacto da sub-representação das transações fraudulentas no treinamento de modelos preditivos.

2.4.Conclusões

Na esfera acadêmica e no setor financeiro, a busca por modelos otimizados de detecção de fraudes assume uma relevância ímpar. Para o mundo acadêmico, o desenvolvimento e aprimoramento desses modelos não apenas alimenta a vanguarda da pesquisa em ciência de dados e finanças, mas também demonstra a aplicabilidade concreta de teorias e métodos inovadores. Paralelamente, no setor financeiro, modelos mais precisos e eficazes não só minimizam perdas financeiras significativas, mas também fortalecem a confiança e a lealdade dos clientes.

Os modelos de detecção de mitigação de fraudes são de suma importância no cenário atual, onde a digitalização está em ascensão e a busca para redução do risco operacional está cada vez maior. Modelos analíticos e de IA desempenham um papel crucial na identificação de atividades suspeitas e na prevenção de fraudes, protegendo assim as organizações e indivíduos contra perdas financeiras significativas. Além disso, esses modelos ajudam a manter a integridade dos sistemas e a confiança dos usuários, que são fundamentais para o sucesso a longo prazo de qualquer organização.

À medida que as fraudes se tornam mais sofisticadas, cresce a necessidade de modelos avançados e robustos, evidenciando a interdependência entre avanços acadêmicos e sua implementação prática para assegurar a integridade do sistema financeiro. A evolução da ciência de dados, da estatística, aprendizado de máquina e inteligência artificial tem desempenhado um papel central nesse contexto ao longo dos anos.

A complexidade crescente dos padrões de fraude, a necessidade de detecção em tempo real e a escassez de bases de dados rotuladas, somada ao alto desbalanceamento de classes, comum na análise de fraudes, continuam a representar desafios tanto para a academia quanto para a indústria financeira. Diferentes técnicas têm sido exploradas para mitigar essas dificuldades e aprimorar os modelos de detecção.

A lacuna de pesquisa identificada, detalhada no artigo do capítulo 3, evidencia a

necessidade de aprofundamento na literatura sobre modelos analíticos que não sejam estritamente supervisionados e aperfeiçoamento de técnicas para tratar o desbalanceamento de classes. Isso se deve à escassez de grandes conjuntos de dados rotulados para treinamento, bem como à necessidade de desenvolver e aprimorar novas abordagens para geração de dados sintéticos de forma a lidar com o desbalanceamento de classes. Desta forma, é objetivo deste trabalho oferecer novas soluções para geração de dados sintéticos baseadas em IA generativa e propostas de modelos não supervisionados que desviam da necessidade de dados rotulados para treino.

Nesse tipo de pesquisa, as variáveis e modelos adotados devem ser diferentes dos tradicionais. Essa abordagem contribuiria para a literatura vigente sobre o tema, além de trazer novas proposições tanto para a academia quanto para os agentes de mercado impactados pela necessidade de geração de dados.

Capítulo 3

3. Modelos de detecção de fraudes e anomalias em bancos: análise sistemática e conexão com a literatura

Alex Cerqueira Pinto

PPGA - Universidade de Brasília - UNB

Resumo

Este trabalho busca analisar e verificar conexões existentes na literatura sobre detecção de fraudes em bancos. Para isso, são analisados e classificados 227 artigos publicados até dezembro de 2022 na Web of Knowledge por meio do protocolo PRISMA. Os trabalhos foram identificados por meio das palavras-chave "Fraude", "Modelo", "Detecção", "Banca" e "Risco" e classificados em 12 categorias, como tipo de estudo, abordagem, corte, desenho, natureza, objetivo do estudo, método, abrangência espacial, período de estudo, foco, dados utilizados e resultados. Com base na classificação, estatísticas de redes complexas também são usadas para identificar as conexões de citação existentes entre elas. Os resultados mostram que há uma disseminação do uso de técnicas de aprendizado de máquina juntamente com regras de negócios para detectar possíveis casos de fraude e um aumento crescente de casos de fraude com engenharia social. Esses achados são úteis para a literatura científica que investiga o risco operacional como bem como para os profissionais responsáveis pela detecção de fraudes.

Palavras-chave: A detecção de fraude; detecção de anomalias; revisão sistemática da literatura; bibliométrica; aprendizado de máquina; bancos.

3.2. Introdução

Como qualquer empresa, uma instituição financeira está sujeita a uma ampla gama de riscos durante a condução de seus negócios. Conhecer as suas características e particularidades é essencial, uma vez que os riscos a que está exposto e que se desconhece são os mais marcantes (Martin et al., 2004). O risco está onipresente em quase todas as atividades e não há unanimidade quanto à definição do seu termo. Assim, toda a discussão se baseia na distinção entre risco mensurável e risco subjetivo.

Em geral, a capacidade de um banco fazer frente a perdas inesperadas é muito dependente dos modelos de risco que a instituição possui e do montante de capital em seu patrimônio. Devido às crises econômicas ocorridas no final do século XX e início do século XXI, houve uma tendência mundial para o desenvolvimento de regras cada vez mais complexas e rígidas sobre os modelos de gestão de risco e capital que os bancos devem manter

(Hull, 2012). Segundo Damodaran (2012), em finanças, o risco é definido como uma função da variabilidade dos retornos obtidos em um investimento em comparação com o retorno esperado do investimento, enquanto Jorion (2000) descreve o risco como a volatilidade de resultados inesperados.

As boas práticas relacionadas à gestão de riscos e capital devem ser perenes e incluir atividades relacionadas a definições estratégicas, controles e incluir a definição de papéis e responsabilidades na estrutura de governança, além de buscar o atendimento aos padrões regulatórios. Ter uma gestão de riscos de qualidade para uma instituição financeira é essencial para sua confiança e sustentabilidade diante de perdas esperadas e inesperadas.

A gestão de riscos de segurança da informação e a prevenção de fraudes são os principais componentes da Gestão de Segurança da Informação em Bancos e devem ser vistos como componentes da gestão de riscos corporativos (Damenu e Beaumont, 2017). O setor bancário, por sua vez, se baseia nas regras de Basileia III para gerenciar riscos operacionais corporativos, que estão diretamente relacionados ao gerenciamento de riscos de segurança da informação (Locher , 2005). Munir e Manarvi (2010) defendem que esses gerenciamentos de segurança da informação, juntamente com o combate à fraude, devem ser combinados com o gerenciamento de riscos operacionais.

Damenu e Beaumont (2017) apresentam a necessidade de pesquisas que tenham como foco a avaliação dos aspectos sociotécnicos de segurança e prevenção de fraudes que são cada vez mais importantes no ambiente corporativo. Além disso, as abordagens predominantes para avaliações de segurança geralmente seguem abordagens automatizadas e mecânicas, concentrando-se em componentes e possivelmente omitindo questões holísticas e envolvendo funcionários. Para capturar esses aspectos, os autores recomendam o uso de modelos analíticos de aprendizado de máquina combinados com regras de negócios.

O objetivo deste trabalho é fornecer uma avaliação abrangente, transparente e replicável de toda a literatura relevante sobre modelos de fraude e detecção de anomalias em bancos, por meio de uma revisão sistemática da literatura. Isso inclui examinar as conexões existentes entre os estudos-chave realizados nesta área.

Da mesma forma, este estudo busca abordar a questão de pesquisa de identificar os principais trabalhos atuais em modelos de detecção de fraudes bancárias. Ele visa identificar tendências predominantes, os métodos e técnicas mais comumente utilizados, as fontes de

dados empregadas, o foco da pesquisa e os resultados alcançados. Além disso, o estudo se esforça para mapear as lacunas existentes no campo de estudo e analisar colaborações e influências entre autores, instituições e países neste domínio.

Para isso, a metodologia é composta por duas técnicas: a revisão sistemática da literatura que classifica as obras em doze características como tipo de estudo, abordagem, corte, desenho, natureza, objetivo do estudo, método, abrangência espacial, período do estudo, foco, dados utilizados e resultados e, posteriormente, a análise de redes complexas que identificam as conexões existentes entre essas obras por meio de suas citações.

Certas análises bibliométricas são cruciais em uma revisão de literatura, pois permitem a identificação de tendências, padrões e lacunas no campo de estudo, além de mapear colaborações e influências entre autores, instituições e países. Essa abordagem também é essencial para avaliar a importância e o impacto de diferentes autores, artigos e tópicos, por meio da análise de citações e coautoria. Facilita a compreensão das interconexões entre várias disciplinas, revelando como diferentes áreas do conhecimento estão interligadas.

As principais contribuições deste trabalho incluem fornecer uma síntese abrangente do conhecimento atual, identificar lacunas que necessitam de investigação adicional e, assim, moldar uma agenda de pesquisa futura sobre o assunto, além de oferecer uma base baseada em problemas do mundo real para facilitar uma abordagem baseada em evidências e possibilitar o desenvolvimento de novas técnicas para a detecção, previsão e análise de fraudes. As contribuições são alcançadas por meio da análise do trabalho explicado e categorizado na seção de metodologia e analisado na seção de resultados.

Nossos resultados mostram que há um uso generalizado de algoritmos de aprendizado de máquina em conjunto com regras de negócios para detecção de fraudes e um aumento crescente de casos de fraude com engenharia social. Esses achados são úteis para a literatura científica que investiga os riscos operacionais e para os agentes econômicos que buscam detectar fraudes nas organizações.

Além desta introdução, este trabalho é composto por mais quatro seções, nas quais a segunda traz um breve referencial teórico, a terceira explica a metodologia e as técnicas empregadas, a quarta apresenta os resultados e, por fim, a seção cinco conclui.

3.3. **Referencial Teórico**

Silva e cols. (2017) apresentam uma revisão sistemática da literatura sobre risco financeiro sistêmico. Para tanto, os autores analisaram e classificaram 266 artigos publicados até setembro de 2016 nas bases de dados Scopus e Web of Knowledge; esses artigos foram identificados por meio das palavras-chave "risco sistêmico", "estabilidade financeira", "financeiro", "medição", "indicador" e "índice". Eles foram avaliados com base em 10 categorias, a saber, tipo de estudo, tipo de abordagem, objeto de estudo, método, abrangência espacial, abrangência temporal, contexto, foco, tipo de dados utilizados e resultados. Em relação aos artigos mais importantes na visão dos próprios pesquisadores, foi construída uma rede de 102 artigos considerados importantes para o avanço do assunto. Foi realizada uma análise aprofundada dos 27 artigos que descreviam a principal trajetória desse campo de pesquisa. Em conclusão, a análise e classificação dessa literatura permitiram identificar as lacunas remanescentes na literatura sobre risco sistêmico; isso contribuiu para uma futura agenda de pesquisa sobre o assunto. Além disso, foram identificados os artigos mais influentes neste campo de pesquisa e os artigos que compõem a pesquisa *mainstream* sobre risco financeiro sistêmico. Segundo as suas conclusões, o bom funcionamento do sistema financeiro depende fundamentalmente da confiança dos agentes muito mais do que noutros setores da economia. Quanto aos artigos mais importantes, na visão dos próprios pesquisadores, há uma diversidade de objetos e abordagens para propor formas inovadoras ou medidas de risco úteis para mensurar o risco financeiro sistêmico.

No mesmo sentido, Fahim e Sillitti (2019) apresentam os resultados de uma revisão sistemática da literatura sobre técnicas de detecção de anomalias, exceto nos domínios de segurança e análise de risco. Os autores usaram estudos publicados de 2000 a 2018 nas áreas de aplicação de ambientes de vida inteligentes, sistemas de transporte, sistemas de saúde, objetos inteligentes e sistemas industriais. A principal fonte de dados utilizada foi o monitoramento de sensores, soluções de baixo custo e alto impacto em diversos domínios de aplicação. Os sensores geram uma enorme quantidade de dados que podem ser analisados para identificar comportamentos não saudáveis. Foram identificadas várias lacunas de pesquisa relacionadas à coleta de dados, análise de grandes conjuntos de dados desequilibrados, limitações de métodos estatísticos para processar os enormes dados sensoriais e poucos artigos de pesquisa sobre a previsão de comportamento anormal em cenários reais. Com base na análise deste trabalho, pesquisadores e profissionais podem se

familiarizar com as abordagens existentes, usá-las para resolver problemas reais e/ou contribuir ainda mais para o desenvolvimento de novas técnicas de detecção, previsão e análise de anomalias.

Em outra revisão sistemática, Pourhabibi et al. (2020) desenvolveu uma estrutura para sintetizar a literatura existente sobre a aplicação de métodos de detecção de anomalias baseados em gráficos (GBAD) na detecção de fraudes publicada entre 2007 e 2018. Este estudo visa investigar as tendências atuais e identificar os principais desafios que exigem esforços de pesquisa significativos para aumentar a credibilidade da técnica. Usando oito perguntas que investigam aspectos específicos da pesquisa de detecção de fraude baseada em GBAD, os autores desenvolveram uma estrutura de classificação para analisar sistematicamente 39 trabalhos acadêmicos identificados por meio de uma pesquisa sistemática na literatura. Algumas aplicações mais recentes das técnicas GBAD mostraram que é possível atribuir um nível de confiança a usuários individuais para indicar a probabilidade de cada indivíduo, por exemplo, pagar um empréstimo ou se esses indivíduos são quem dizem ser, com base em cada dados de indivíduos descobertos na Internet.

Em outra revisão publicada mais recentemente, Hilal et al. (2022) apresentam pesquisa que buscou investigar e apresentar uma revisão completa das técnicas de detecção de anomalias mais populares e eficazes aplicadas para detectar fraudes financeiras, com foco em destacar os avanços recentes nas áreas de aprendizado semi-supervisionado e não supervisionado. A metodologia por trás desta pesquisa foi impulsionada pela missão de que uma revisão abrangente das técnicas de detecção de anomalias facilite ao leitor a compreensão de suas vantagens e limitações quando aplicadas a uma área específica de fraude. As direções mais promissoras para pesquisas futuras, na opinião dos autores, envolvem a investigação do desempenho de modelos de detecção que incorporam tanto o poder de sobreamostragem quanto o poder discriminativo de modelos generativos, como LSTMs e CNNs, que capturam relacionamentos temporais de longo e curto prazo nos dados. e, finalmente, resultar em um sistema mais robusto e eficiente. Além disso, pode valer a pena explorar as áreas de fraude menos pesquisadas, como fraude de valores mobiliários e commodities, fraude hipotecária, informações privilegiadas e outras.

Por fim, Nonnenmacher e Gómez (2021) realizaram uma revisão sistemática dos estudos existentes que aplicam a detecção não supervisionada de anomalias em um contexto

de auditoria. Uma abordagem para resolver a crescente quantidade de dados causada pela transformação digital é aplicar regras aos dados. Uma desvantagem disso é que as regras provavelmente encontrarão apenas erros, enganos ou desvios que já foram antecipados pelo auditor. A detecção de anomalias não supervisionada pode ir além desses recursos e detectar novos desvios de processo ou novas tentativas de fraude. Depois de analisar profundamente 16 trabalhos de 2006 a 2019, os resultados revelam que a maioria dos estudos desenvolve uma abordagem apenas para um conjunto de dados específico e não aborda a integração no processo de auditoria ou como os resultados devem ser melhor apresentados ao auditor. Portanto, os autores desenvolveram uma agenda de pesquisa abordando tanto a generalização da detecção de anomalias de auditoria não supervisionada quanto a preparação de resultados para auditores.

Em resumo, esses autores identificaram lacunas de pesquisa em vários espectros, tais como: lacunas relacionadas à coleta de dados, análise de conjuntos de dados grandes e desequilibrados, limitações dos métodos estatísticos para processamento de dados em massa (Fahim e Sillitti, 2019); o poder discriminativo de modelos generativos como LSTMs e CNNs, bem como outros tipos de fraudes, incluindo fraudes em títulos mobiliários, fraudes hipotecárias e outros (Hilal et al., 2022); e a generalização da detecção de anomalias em auditoria não supervisionada quanto à preparação de resultados para auditores (Nonnenmacher e Gómez, 2021).

Diante desses desenvolvimentos, um estudo bibliométrico renovado se faz necessário devido à progressão temporal do tema, ao aumento exponencial recente de publicações e à introdução de novas técnicas envolvendo aprendizado de máquina, inteligência artificial e mudanças no comportamento de agentes. A literatura recente apresenta perspectivas variadas, destacando tendências de pesquisa emergentes e a lacuna notável no estudo de fraudes envolvendo engenharia social.

3.4. **Metodologia**

3.4.1. Revisão Sistemática da Literatura

Ao se aprofundar no estudo de um tema, os resultados na literatura costumam ser contraditórios e, para contornar esse problema, os pesquisadores podem se valer de uma

revisão sistemática da literatura. É um tipo de investigação que visa identificar, selecionar, avaliar e sintetizar as principais evidências disponíveis na literatura, deve ser abrangente e não tendenciosa para que os critérios adotados sejam divulgados e outros pesquisadores possam repetir o procedimento e, desta forma, são considerados o melhor nível de evidência para a tomada de decisão (Galvão e Pereira, 2014).

Neste trabalho, utiliza-se a pesquisa meta-síntese, apropriada quando uma revisão sistemática visa integrar a pesquisa qualitativa para sintetizar estudos qualitativos sobre um tema, localizar temas, conceitos ou teorias-chave que forneçam explicações novas ou mais poderosas para o fenômeno sob revisão. (Galvão e Ricarte, 2019; Siddaway et al., 2019).

O critério adotado para a possível seleção dos artigos baseou-se em buscas em bases de dados de acesso livre. Assim, realizou-se pesquisa bibliográfica por meio de publicações de artigos científicos obtidos em meio eletrônico na base de dados Web of Science. Na pesquisa foram utilizadas as palavras-chave "Fraude", "Modelo", "Detecção" e posteriormente foram analisados artigos mais aderentes ao tema do setor financeiro e bancário, utilizando filtros como "banking" ou "banco", "risco", "detecção de anomalias" e "financeiro". Não foram utilizadas restrições na construção do banco de dados quanto aos anos de publicação, sendo o idioma inglês a única restrição para a inclusão dos estudos. Assim, foram selecionados 227 artigos.

Como limitação do estudo, está a utilização apenas do fluxograma no formato do protocolo PRISMA aplicado de forma específica, contudo, outras premissas do referido protocolo de revisão sistemática não são utilizadas.

A Tabela 3.1 apresenta os critérios para categorização dos artigos definidos neste trabalho. Esta categorização foi adaptada de Silva et al. (2017).

Tabela 3.1 - Critérios de Categorização dos Artigos

Ordem	Características	Código
1	Tipo de estudo	A – Teórico B – Empírico C – Ambos
2	Abordagem	A – Quantitativo B – Qualitativo C – Quantitativo e qualitativo D – Não aplicável

3	Recorte	A – Transversal B – Longitudinal C – Não aplicável
4	Projeto	A – Experimentar B – Quase-Experiência C – Correlacional D – Preditivo E – Observacional
5	Natureza	A – Exploratória B – Descritiva C – Explicativa
6	Propósito do estudo	A – Regulação B – Risco Operacional/Fraude C – Risco de Crédito D – Segurança E – Auditoria F – Outros
7	Método	A – Econométrico/Estatístico B – Machine Learning C – Otimização Matemática D – Não aplicável
8	Escopo Espacial	A – Um país B – Mais de um país C – Global D - Não especificado
9	Período do Estudo	A – Até 2 anos B – De 2 a 5 anos C – De 5 a 10 anos D – Mais de 10 anos E – Não aplicável
10	Foco	A – Bancos tradicionais B – Instituições financeiras não bancárias C – Fundos de investimento D – Empresas não financeiras E – Setor público
11	Dados utilizados	A – Público/mercado B – Balanço das companhias abertas C – Privado D – Reguladores E – Outros F – Não informado

12	Resultados	A – Consistente com o que foi verificado na literatura B – Nova perspectiva C – Replicação com resultados divergentes D – Estudo comparativo

Fonte: Elaborado pelo autor.

Com isso, o presente trabalho busca contribuir com essa literatura ao considerar 227 trabalhos que investigaram a detecção de fraudes em bancos e instituições financeiras e classificando-os segundo doze características como tipo de estudo, abordagem, corte, desenho, natureza, objetivo do estudo, método, abrangência espacial, período de estudo, foco, dados utilizados e resultados, além de identificar por meio das métricas de redes complexas as conexões existentes entre esses trabalhos, a fim de quantificar sua relevância na literatura. Essas doze características ajudam a responder à pergunta de pesquisa, ao objetivo proposto e também foram validadas em outras bibliografias em finanças, como Silva et al. (2017).

3.4.2. Métricas de Redes Complexas

Buscando analisar conexões entre os trabalhos selecionados e, assim, compreender melhor seus achados identificando sua influência na literatura, são utilizadas métricas de redes complexas. Como pode ser visto em Passos et. al (2022), a análise da rede pode ocorrer no nível dos agentes (nós) ou no nível da rede (como um todo).

A análise de dados por meio de redes complexas tem emergido como uma ferramenta poderosa para investigar sistemas complexos encontrados em uma variedade de disciplinas acadêmicas e aplicações do mundo real. Essa abordagem, fundamentada na teoria dos grafos, permite a representação e análise de interações entre elementos individuais, proporcionando insights valiosos sobre a estrutura e dinâmica desses sistemas interconectados.

Em essência, conforme Passos et. al (2022) uma rede complexa é composta por nós (ou vértices) que representam entidades individuais, e arestas (ou conexões) que representam as interações entre essas entidades. Essa abordagem transcende a organização tradicional de dados, revelando padrões e insights ocultos em conjuntos de dados interconectados. Através da modelagem matemática de redes, podemos explorar a dinâmica subjacente a sistemas complexos, desde redes sociais e biológicas até sistemas de transporte e financeiros.

Dentre as métricas das redes complexas podemos destacar:

1. Grau (Degree): O grau de um vértice é o número de conexões que ele possui. Em uma rede de transações bancárias, o grau de um cliente representa quantas transações ele realizou. Um cliente com alto grau pode ser um influenciador na rede.

2. Centralidade de Proximidade (Closeness Centrality): Mede a distância média entre um vértice e todos os outros vértices na rede. Em uma rede de investidores em um mercado de ações, a centralidade de proximidade de um investidor indica quão rapidamente ele pode acessar informações de outros investidores. Um investidor central pode ter vantagem competitiva.

3. Centralidade de Intermediação (Betweenness Centrality): Avalia a importância de um vértice como intermediário nas comunicações entre outros vértices. Em uma rede de empresas e seus fornecedores, a centralidade de intermediação de uma empresa indica se ela atua como intermediária nas transações entre outras empresas.

4. Coeficiente de Aglomeração (Clustering Coefficient): Mede a tendência de vértices vizinhos formarem grupos densamente conectados. Em uma rede de colaboração entre empresas, o coeficiente de aglomeração de uma empresa indica se ela está inserida em um grupo coeso de parceiros. Empresas com alto coeficiente de aglomeração podem formar alianças estratégicas.

O foco, neste caso, é a análise no nível dos agentes, especialmente as métricas de centralidade de grau (ou valência) e o algoritmo PageRank, cuja base teórica é a centralidade do autovetor.

O algoritmo PageRank é um algoritmo utilizado para avaliar a importância relativa dos nós em uma rede de grafos, especialmente em motores de busca na internet. Ele foi desenvolvido pelos fundadores do Google, Larry Page e Sergey Brin, enquanto eram estudantes na Universidade de Stanford, e foi um dos principais componentes do algoritmo de classificação de páginas usado pelo Google em seu mecanismo de busca.

O princípio subjacente ao PageRank é bastante intuitivo: um site é considerado importante se for vinculado por outros sites importantes. Assim, o algoritmo atribui uma pontuação de importância a cada página da web com base nas páginas que a vinculam e na importância dessas páginas. A ideia fundamental é que páginas importantes tendem a ser vinculadas por outras páginas importantes.

Matematicamente, o algoritmo PageRank pode ser representado por um sistema de

equações lineares. Seja $PR(u)$ a pontuação de PageRank atribuída a uma página u, então a pontuação de PageRank de uma página u é calculada pela fórmula:

$$\mathrm{PR(u)} = \sum_{(\mathrm{v} \in \mathrm{B_u})\mathrm{x}} \mathrm{PR(v)/L(v)} \tag{3.1}$$

onde:

$PR(u)$ é a pontuação de PageRank da página u.

Bu é o conjunto de páginas que possuem um link para a página u.

$L(v)$ é o número de links na página v.

Essencialmente, o PageRank de uma página é a soma dos PageRanks de todas as páginas que a vinculam, dividido pelo número de links nessas páginas. Isso reflete a ideia de que uma página importante é aquela que é vinculada por outras páginas importantes, mas também leva em conta o número de links nessas páginas.

O algoritmo PageRank é iterativo e converge para uma distribuição de PageRank estável em que as pontuações de PageRank refletem a importância relativa das páginas na rede. Ele tem sido fundamental para melhorar a qualidade dos resultados de pesquisa na internet, ajudando os usuários a encontrar informações relevantes de forma mais eficaz.

As medidas de nível de agente que usamos são centralidade de grau ou valência; e centralidade do autovetor. A centralidade de grau (ou valência) é definida pela seguinte expressão:

$$k_i = \sum_{j=1}^{n} a_{ij}, \quad 0 < k_i < n \tag{3.2}$$

e

$$k_v = \left|N_v\right| \quad 0 < k_v < n \tag{3.3}$$

onde:

um $_{ij}$ é a entrada da i-ésima linha e j-ésima coluna da matriz de adjacência A.

N $_v$ é a vizinhança do agente (nó ou vértice) V.

Para redes direcionadas temos:

k_i^+ = In-degree (número de agentes de entrada, ou seja, número de arestas ou relações iniciando no agente v).

k_i^- = grau de saída (número de agentes de saída, ou seja, número de arestas ou relações que terminam em agente v).

$$k_i^+ = \sum_{j=1}^{n} a_{ij} \tag{3.4}$$

$$k_i^- = \sum_{j=1}^{n} a_{ij} \tag{3.5}$$

A medida de grau nas redes segmentadas também é conhecida como prestígio, expressão muito utilizada em ARS (análise de redes sociais).

Existem dois tipos de prestígio: (i) apoio; e (ii) influência. O de apoio é o grau de entrada e o de influência é o grau de saída. Em redes pesadas (ou ponderadas), a força é equivalente ao grau. É igual à soma dos pesos das arestas adjacentes a um dado agente (ou das relações ligadas a este agente). Como em (3.6):

$$k_i^w = \sum_{j=1}^{n} a_{ij}^w \tag{3.6}$$

A equação (3.7) fornece a centralidade do autovetor, que é a métrica que serviu de base para o desenvolvimento do PageRank ™:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} a_{ij} x_j \tag{3.7}$$

onde:

xi/ xj representa a centralidade do agente i / j;

aij denota a matriz de adjacência A (aij = 1 se os agentes iej são conectados por uma aresta e aij = 0 se não são); e

λindica o maior autovetor da matriz A.

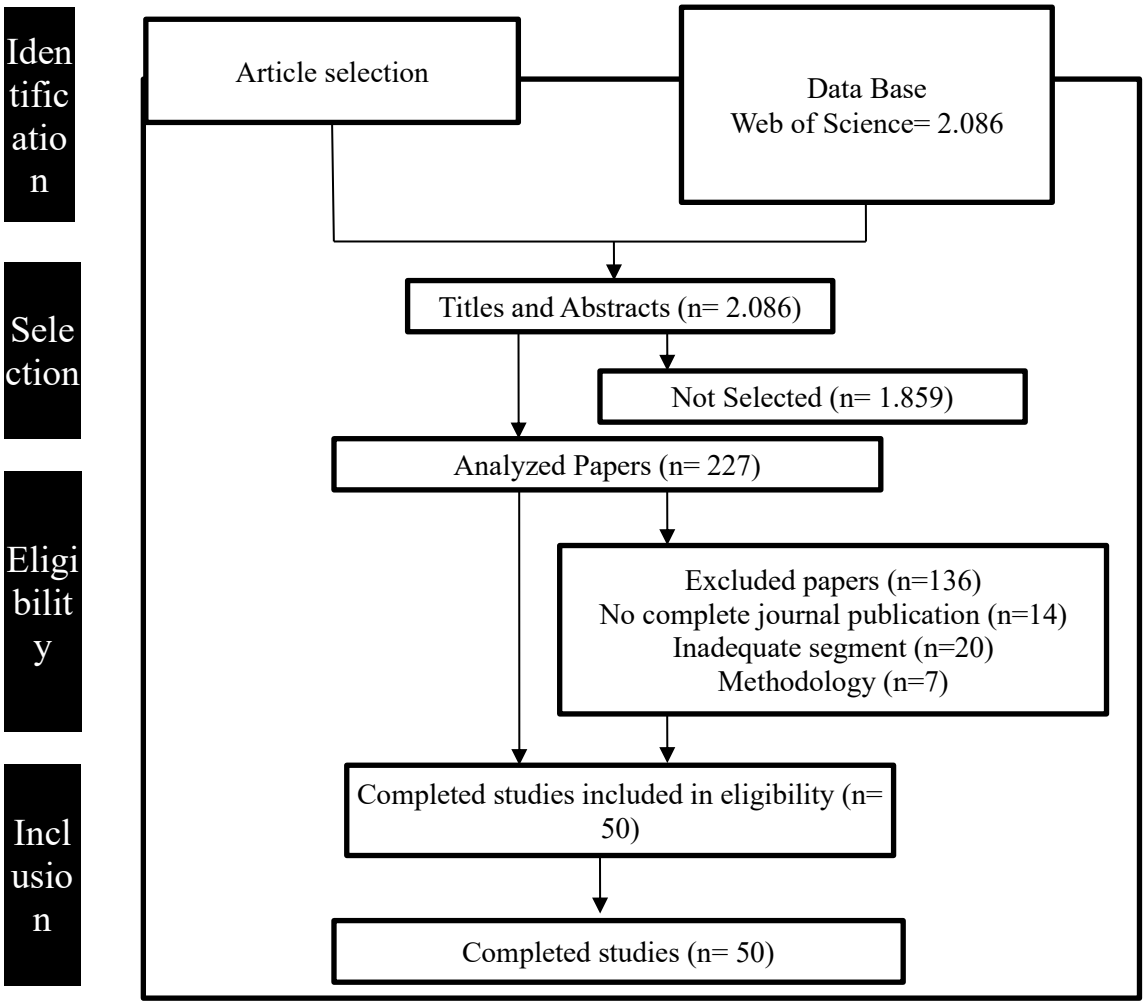
A centralidade do autovetor é uma medida proposta por Bonacich (1987) e se baseia na noção de que a centralidade de um agente é definida pela centralidade dos agentes com os quais ele se relaciona (via trocas, transações, etc.). Assim, o poder ou situação econômico-financeira de um agente é definido pelo poder ou situação econômico-financeira de seus alteres. Os alters são os agentes diretamente relacionados ao agente central (também chamados de ponto focal ou ego). A centralidade do autovetor é a combinação linear das centralidades de seus vizinhos de primeira ordem.

3.5. **Resultados**

3.5.1. Seleção e Análise Exploratória

Os dados utilizados são descritos e apresentados na Figura 3.1 do Fluxograma PRISMA, conforme descrito por Liberati et al. (2009).

Figura 3.1 - Fluxograma de seleção de artigos na revisão sistemática – PRISMA

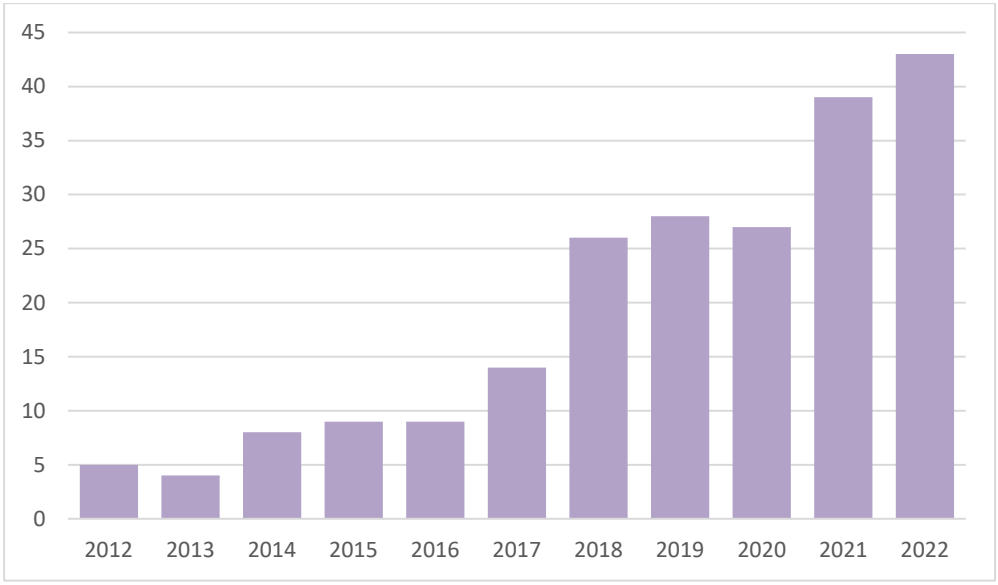


Fonte: Elaborado pelo autor.

Ao analisarmos a base de dados considerando os anos de publicação, observamos que, em média, havia aproximadamente seis publicações por ano com o tema e filtro de busca

utilizados nos primeiros cinco anos. No entanto, essa média cresce fortemente a partir de 2017 e se mantém até 2022, com níveis médios quatro vezes superiores ao período de 2006 a 2016, conforme podemos observar na Figura 3.2.

Figura 3.2 - Número de artigos por ano

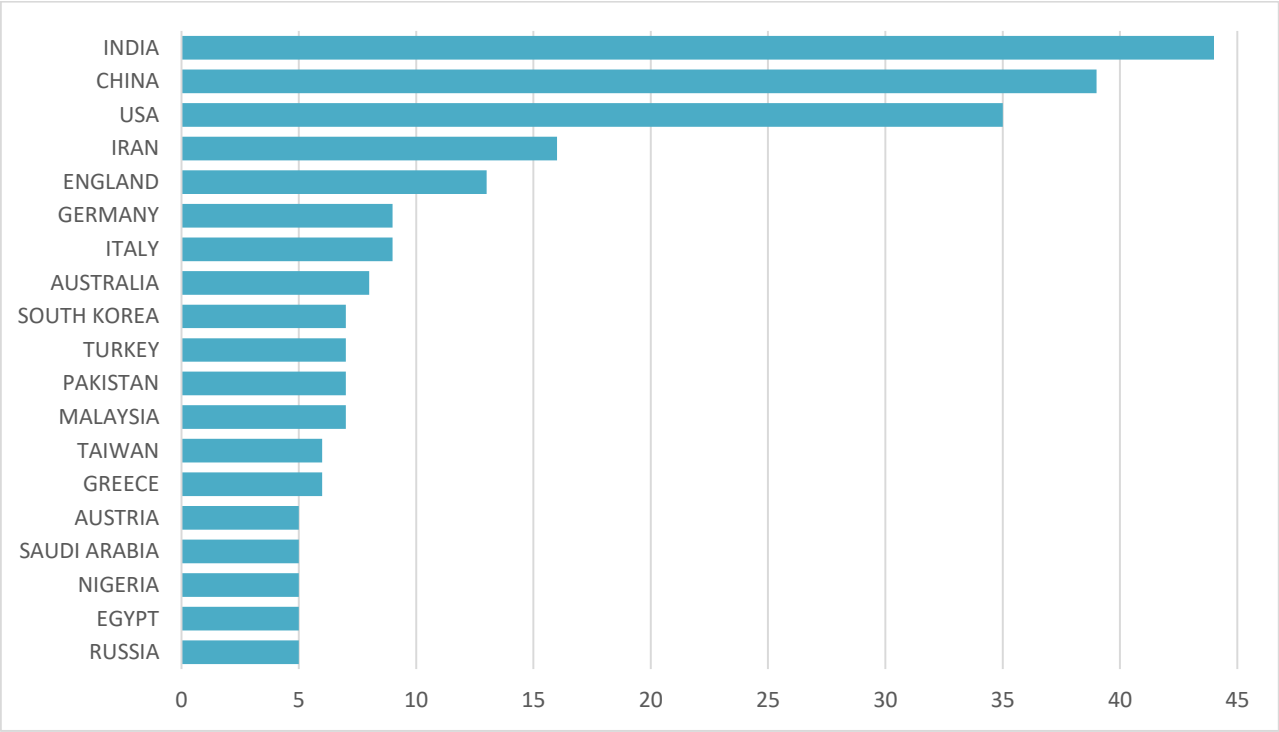


Fonte: Elaborado pelo autor.

A apresentação quantitativa de publicações categorizadas por países e regiões não apenas aprimora a compreensão da distribuição e impacto da pesquisa, mas também lança luz sobre as dinâmicas globais e locais que influenciam a produção científica. Essa categorização é essencial por várias razões. Em primeiro lugar, ela oferece insights sobre a diversidade geográfica e cultural da pesquisa, auxiliando na avaliação do impacto global versus local dos estudos e oferecendo perspectivas sobre sua relevância e aplicabilidade em diferentes contextos geográficos. Por fim, a categorização por país e região pode revelar padrões de colaboração internacional, ilustrando como o conhecimento é compartilhado e disseminado entre diferentes partes do mundo, o que é crucial para a compreensão das dinâmicas e interconexões da pesquisa científica global.

Analisando as publicações por região, verificamos que elas estão espalhadas por vários países, porém, 24 delas possuem apenas uma publicação. Vale citar a Índia e a China com seus pesquisadores destacados frente ao número de publicações, com 38 e 26 publicações, respectivamente. A Figura *3.3* mostra a distribuição das publicações por países e regiões.

Figura 3.3 - Publicações por países e regiões



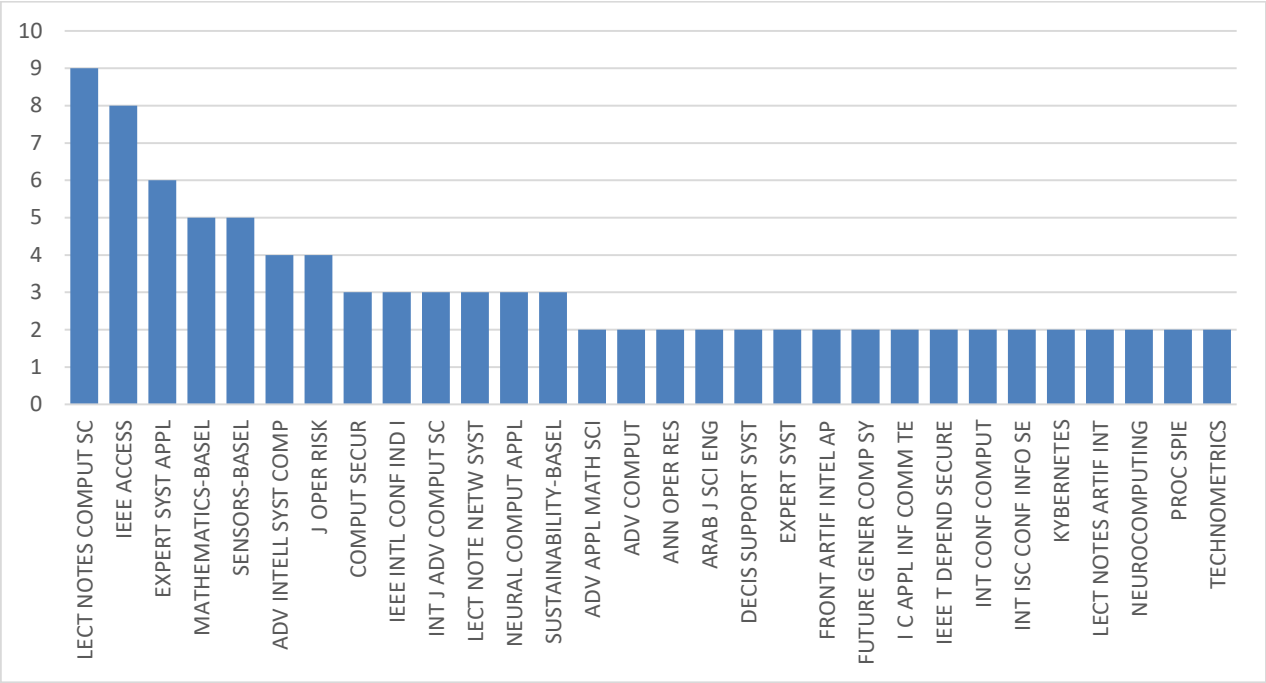
Fonte: Elaborado pelo autor.

A apresentação quantitativa de publicações por origens, fontes e periódicos em uma revisão sistemática contribui para uma compreensão mais profunda e contextualizada do campo de estudo e da ideologia dessa origem, além da diversidade, perspectivas e credibilidade da fonte.

Dessa forma, ao considerarmos a fonte científica na qual os trabalhos foram publicados, observa-se que as 10 principais fontes concentraram apenas 16% dos artigos publicados - 36 de 227. Uma grande dispersão de periódicos foi observada para os 116 artigos restantes, com a grande maioria das fontes - 109 - publicando apenas um artigo e 7 fontes publicando dois trabalhos, como mostrado na Figura 3.4.

Análises bibliométricas específicas baseadas em redes desempenham um papel vital em revisões de literatura. Elas possibilitam a descoberta de tendências, padrões e áreas inexploradas dentro do campo de estudo, além de auxiliarem no mapeamento das redes colaborativas e influentes entre autores, instituições e países. Tal abordagem é fundamental para avaliar a importância e influência de diversos autores, artigos e temas por meio de análises de citação e coautoria. Isso auxilia na compreensão das conexões entre múltiplas disciplinas, revelando as inter-relações entre diferentes campos do conhecimento.

Figura 3.4 - Papéis por diários



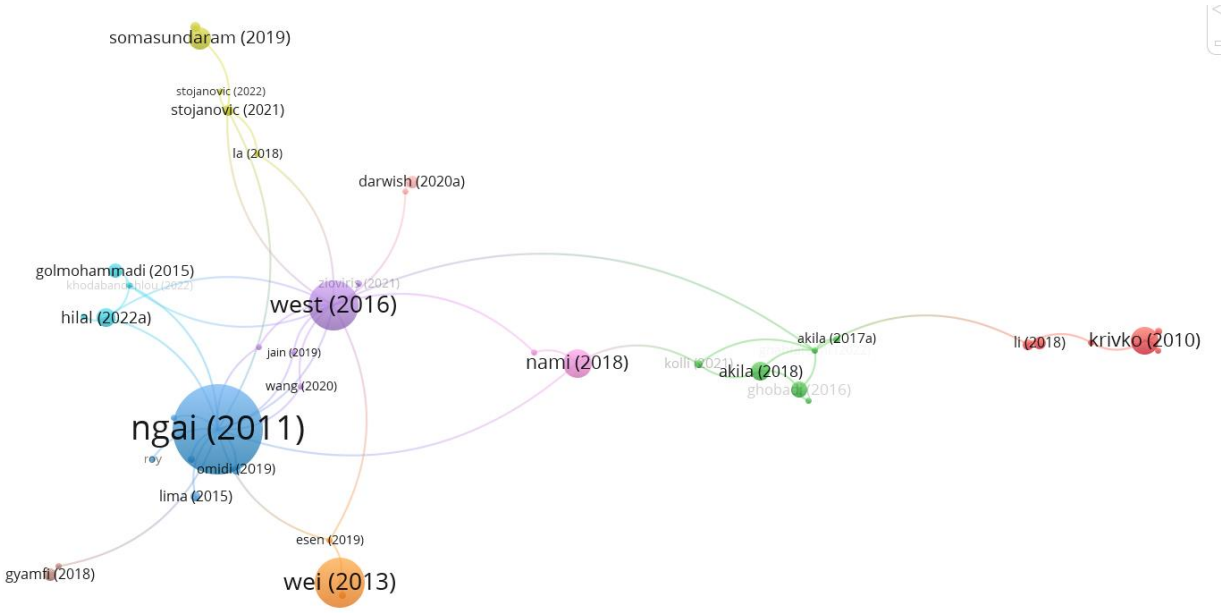
Fonte: Elaborado pelo autor.

Dessa forma, em relação às medidas de centralidade apresentadas a seguir, a rede de interação por citação mensura a influência de artigos ou autores com base no número de vezes em que são citados, auxiliando na identificação das obras mais influentes. A rede de interação por coautoria em pesquisa mapeia as colaborações entre pesquisadores, revelando os principais colaboradores e tendências colaborativas no campo. Por fim, os links entre palavras-chave analisam as conexões entre diferentes tópicos e conceitos, auxiliando na identificação de temas dominantes e na compreensão de como estão interconectados, orientando a estrutura temática da revisão.

Ao analisar a literatura, é importante verificar as conexões existentes entre os trabalhos publicados para entender como ela foi desenvolvida. Verificar a dinâmica da citação é uma dessas maneiras. No que diz respeito à rede de citações entre os trabalhos mais influentes na literatura que investiga fraudes bancárias, verifica-se que a rede de citações também é esparsa com pluralidade de trabalhos. Observe que o artigo mais citado é Ngai et al. (2011) que discute a aplicação de técnicas de mineração de dados para detecção de fraudes e apresenta uma revisão da literatura acadêmica sobre o assunto. A Figura 3.5 apresenta as conexões entre as

obras.

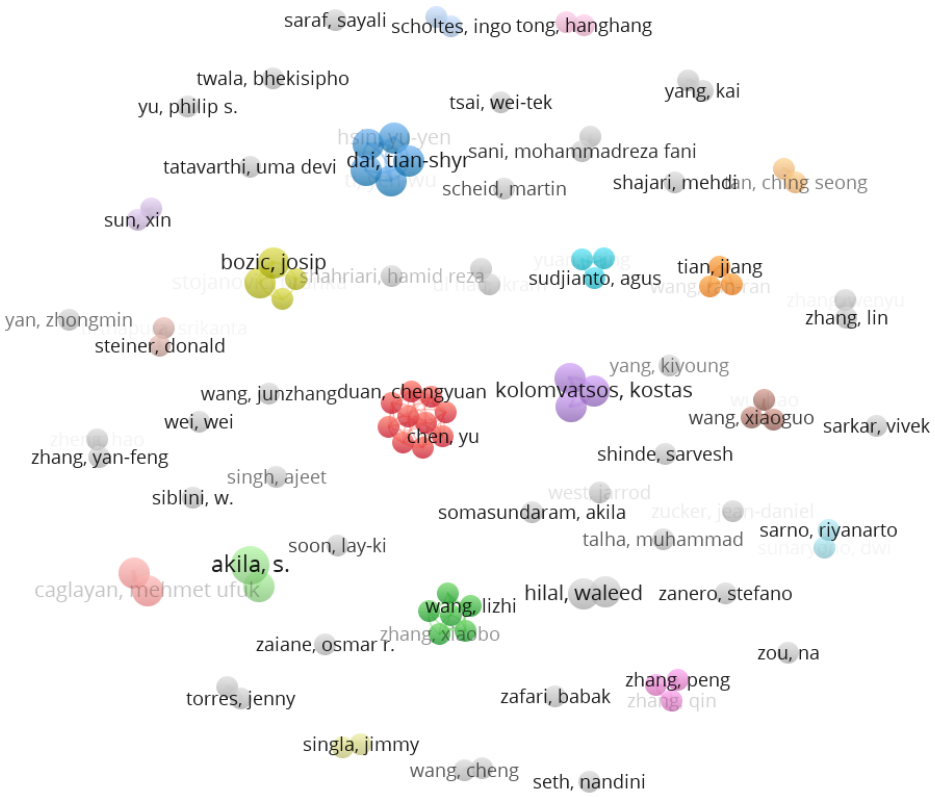
Figura 3.5 - Rede de interações de citações



Fonte: Elaborado pelo autor.

Ao analisar as redes de interações entre os autores da literatura sobre detecção de fraudes em bancos, também é possível melhorar seu entendimento. Neste caso, redes de influência não ocorrem em pesquisas para este campo de análise, uma possível motivação para isso pode ser devido a diferentes formas de abordar o assunto em diferentes campos de pesquisa como riscos, modelos computacionais, auditorias, ciência da computação, finanças e etc. A Figura 3.6 demonstra que não há conexões entre os autores das obras.

Figura 3.6 - Rede de interação co-autor de pesquisa

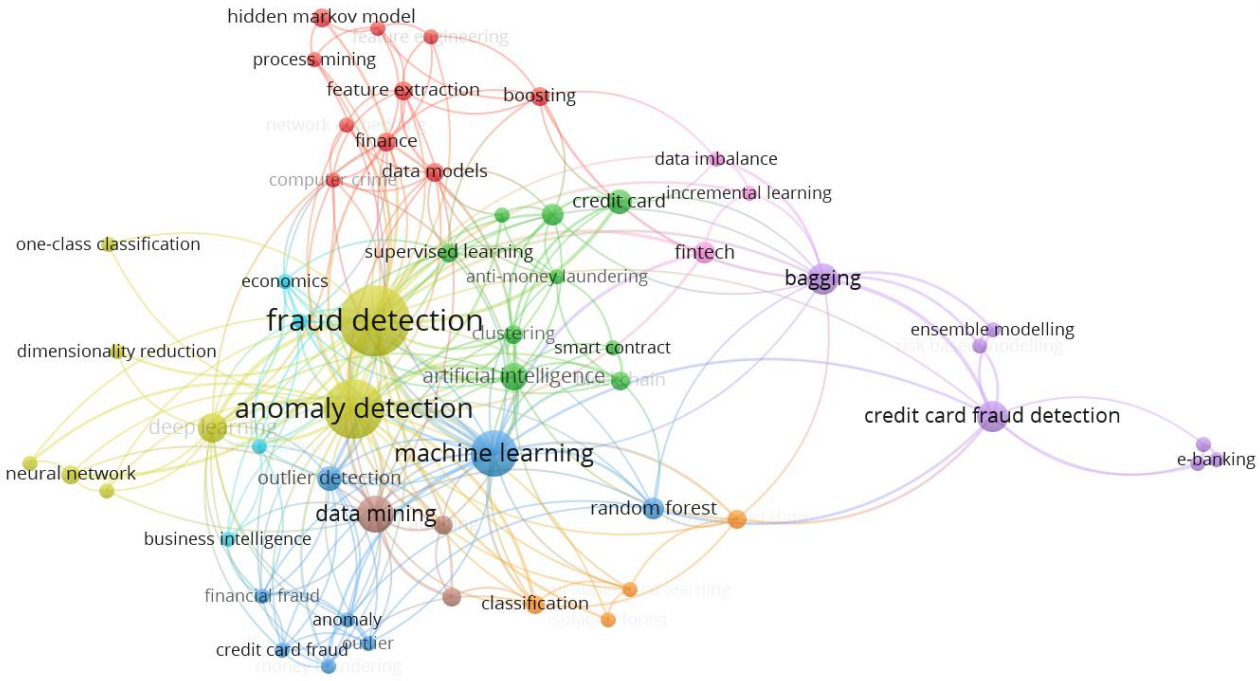


Fonte: Elaborado pelo autor.

A análise bibliométrica de redes de palavras-chave é importante para identificar temas centrais em um campo de estudo, mapear tendências ao longo do tempo, descobrir relações temáticas entre diferentes áreas e inspirar a formulação de novas hipóteses de pesquisa. Observa-se que, nos temas de pesquisa representados pelas palavras-chave, os temas de detecção de fraude estão intimamente alinhados com algoritmos de detecção de anomalias, pois as técnicas são aplicadas em ambas as situações devido ao desequilíbrio de classe comumente encontrado em tais casos. Da mesma forma, o uso de aprendizado de máquina e mineração de dados exibe uma conexão intrínseca com o tema no desenvolvimento desses modelos.

Da mesma forma, considerando as palavras-chave dos trabalhos, há uma forte ligação com 'detecção de fraude', 'detecção de fraude de cartão de crédito', 'classificação' e 'modelo'. A rede completa pode ser analisada na Figura 3.7.

Figura 3.7 - Links entre as palavras-chave



Fonte: Elaborado pelo autor.

3.5.2. Classificação por Categorias

Dos 227 trabalhos obtidos, os resumos foram lidos a priori e também foi iniciada uma avaliação com base nos critérios de inclusão e exclusão, com o intuito de selecionar os estudos a serem analisados na íntegra. Para a seleção final, o objetivo do artigo deve coincidir com a questão problema e envolver a análise sob a ótica da gestão de riscos. Dessa forma, foram retirados os artigos que não atendiam a esses critérios e aqueles que não possuíam publicações completas em periódicos revisados por pares. Assim, a seleção final concentrou-se em 50 artigos, conforme aplicação do protocolo PRISMA apresentado na metodologia.

Após a identificação dos artigos, eles foram classificados, conforme descrito na seção de metodologia. A Tabela 3.2 apresenta a classificação.

Tabela 3.2 - Classificação dos Artigos

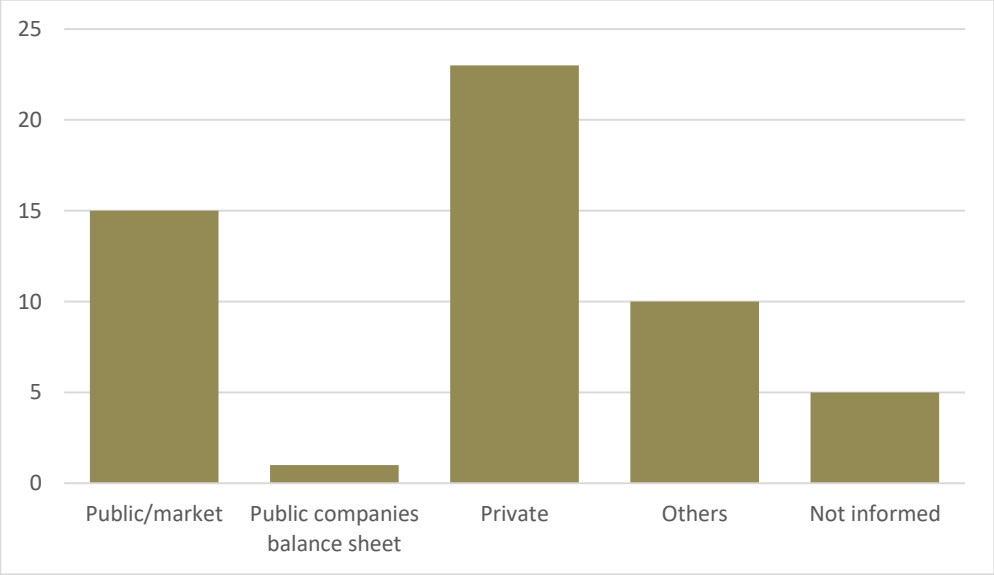
Artigo	Tipo de estudo	Abordagem	Recorte	Projeto	Natureza	Objeto de estudo	Método	Escopo Espacial	Período de estudo	Foco	Dados	Resultados
Zioviris et ai. (2022)	1B	2A	3A	4D	5B	6B; 6D	7B; 7C	8C	9E	10A	11A	12B
Bernardo e cols. 2022)	1B	2C	3A	4E	5B	6E	7C	8D	9E	10C	11E	12C
Wei e outros. (2013)	1B	2C	3B	4D	5B	6B	7B	8A	9A	10A	11C	12B; 12D
Akila e Reddy (2017)	1B	2A	3B	4D	5B	6B	7C	8A	9A	10A	11C	12B
Nami e Shajari (2018)	1B	2A	3A	4C; 4D	5C	6B; 6D	7A; 7B	8A	9A	10A	11C	12B
Darwish (2020)	1C	2A	3C	4D; 4E	5B	6B	7A; 7B	8D	9E	10A; 10B	11F	12D
Kumar et ai. (2019)	1C	2A	3B	4B; 4D	5B	6B; 6D; 6E	7A; 7B	8D	9A	10A	11C	12A
Nanduri et ai. (2020)	1C	2C	3B	4B	5C	6B; 6D; 6F	7B; 7C	8D	9A	10D	11C	12C; 12D
Karthik et ai. (2022)	1C	2C	3B	4B; 4D	5B	6A; 6E	7A; 7B	8D	9B	10B; 10D	11C; 11E	12B; 12D
Pandey (2010)	1C	2C	3A	4C	5A	6A	7A; 7B	8A	9E	10A	11C	12D
Labanca et ai. (2022)	1B	2A	3A	4D	5B	6B	7B	8A	9E	10A	11C; 11E	12B
Can et ai. (2020)	1C	2C	3B	4A; 4D	5C	6B; 6D; 6E	7A; 7B	8A	9B	10A	11C	12B; 12D
Babu e Vasavi (2017)	1B	2A	3B	4B	5B	6A; 6B; 6E	7A; 7B	8A	9D	10E	11C	12C
Wu e outros. (2014)	1B	2A	3A	4D	5A; 5B	6C	7A	8A	9D	10A; 10D	11C	12A
Saha et al. (2016)	1B	2C	3A	4C; 4D	5B	6E	7A	8D	9E	10A	11C	12A
Ravi (2021) See More	1B	2C	3A	4B; 4D	5B	6F	7A	8D	9E	10B	11C	12A
Teng e Lee (2019)	1B	2A	3A	4D	5B	6C	7B	8A	9C	10A	11A	12A
Rahman et ai. (2021)	1C	2C	3A	4B	5A	6F	7A	8A	9E	10A; 10B	11F	12A; 12D
Boyle et ai. (2015)	1C	2C	3A	4C	5B	6E	7D	8A	9E	10D	11F	12A; 12D
Mu e Carroll (2016)	1B	2C	3A	4C	5A	6B; 6E	7D	8A	9E	10D	11C	12B; 12D
Lokanan et ai. (2019)	1B	2C	3A	4B	5B	6E	7A	8A	9D	10A; 10D	11B	12A
Ashfaq et ai. (2021)	1C	2A	3B	4A; 4D	5A; 5B	6B	7A; 7B	8D	9E	10A; 10B	11E	12A
Bakumenko e Elragal (2022)	1C	2A	3A	4A; 4D	5A; 5C	6B	7B	8A	9E	10A	11A	12A; 12D
Bose et ai. (2017)	1A	2C	3C	4B	5B	6B	7D	8D	9E	10A	11E	12A
Carminati et ai. (2015)	1C	2A	3B	4A; 4D	5A; 5C	6B	7A; 7B	8A	9A	10A	11C	12B
Carminati et ai. (2018)	1B	2A	3A	4B	5B	6B; 6D	7B	8A	9A	10A	11C	12A
Cui et al. (2021)	1C	2A	3A	4B	5B	6B	7B	8A	9A	10A	11C	12B
Esen et al. (2019)	1B	2A	3A	4B	5B	6B	7B	8A	9C	10A	11C	12B
Hewapatirana (2019)	1A	2C	3C	4E	5A; 5B	6B	7B	8D	9E	10A	11A	12A
Hsin et al. (2022)	1B	2A	3A	4A; 4D	5B	6B	7B	8D	9E	10A	11A	12A
Khodabandehlou e Golpayegani (2022)	1A	2B	3C	4C	5A	6B	7B	8D	9E	10A	11A	12A

Kim e outros. (2014)	1A	2B	3C	4C; 4E	5C	6B; 6D	7D	8D	9E	10A	11F	12B
Li et ai. (2021)	1C	2A	3C	4D	5A; 5B	6B; 6C	7B; 7C	8A	9E	10A	11C; 11E	12B
Mosavi et al. (2020)	1C	2A	3A	4E	5A	6B; 6D; 6F	7B	8D	9E	10A; 10B; 10C	11A	12A; 12D
Nesvijevskaia et ai. (2021)	1B	2A	3A	4B	5C	6B	7A	8A	9A	10A	11A	12A
Nicholls et ai. (2021)	1A	2B	3C	4E	5B	6D	7B	8D	9E	10B	11A	12A
Oliveira e cols. (2006)	1C	2C	3A	4B	5B	6B; 6F	7A; 7B	8D	9E	10A; 10B	11A	12B
Omidi et al. (2019)	1B	2A	3A	4B	5B	6B	7B	8A	9E	10A	11A	12D
Prabha et al. (2021)	1A	2B	3C	4E	5B	6B	7D	8D	9B	10A	11E	12D
Sair e outros. (2019)	1C	2A	3A	4B	5A; 5C	6B	7A; 7B	8B	9A	10B; 10D	11A	12B
Stojanovic et ai. (2021)	1B	2A	3A	4B	5A	6B	7B	8B	9A	10B	11A	12B
Sudjianto et al. (2010)	1A	2A	3C	4E	5C	6B	7A; 7B	8D	9E	10A	11A	12A
Ti et al. (2022)	1C	2A	3A	4B	5B	6B	7B	8A	9E	10A	11C	12A
Oeste e Bhattacharya (2016)	1B	2A	3A	4B	4B	6B	7B	8B	9B	10A; 10B	11C	12A
Zafari et ai. (2022)	1C	2A	3A	4B; 4D	5B	6B	7A	8A	9C	10D	11E	12B
Zamini e Hasheminejad (2019)	1A	2B	3C	4E	5B	6B; 6D	7D	8B	9C	10A	11E	12D
Zhang et ai. (2021)	1C	2A	3A	4D	5C	6B; 6D; 6F	7A; 7B	8D	9E	10A; 10B	11C; 11E	12B; 12D
Zhang et ai. (2022)	1B	2A	3A	4B	5B	6B	7B	8D	9E	10A	11F	12D
Zhou et ai. (2021)	1C	2A	3A	4D	5C	6B; 6F	7B; 7C	8D	9E	10A	11C	12B; 12D
Zhu e Yang (2019)	1C	2A	3A	4B	5B	6B; 6F	7B	8D	9E	10A	11A	12B; 12D

Com os resultados apresentados na Tabela 3.2 , é possível verificar um grande número de trabalhos que analisaram fraudes com cartões de crédito, possivelmente motivados pelo tema estar atrelado às maiores necessidades da atualidade dada a preocupação da indústria de meios de pagamento e fintechs em evitar perdas operacionais. Da mesma forma, verifica-se que todos os estudos utilizam uma abordagem quantitativa e o tipo de estudo é classificado como empírico, muitas vezes juntamente com o tipo classificado como teórico.

Em relação ao período do estudo, a maioria dos trabalhos, 68% dos quais com série histórica, utilizou dados de até cinco anos para a construção dos modelos e esses dados foram obtidos, em sua maioria, por meio de instituições privadas e seus próprios históricos séries, seguidas das bases de dados públicas disponíveis no mercado, conforme a Figura 3.8.

Figura 3.8 - Banco de dados utilizado nos artigos

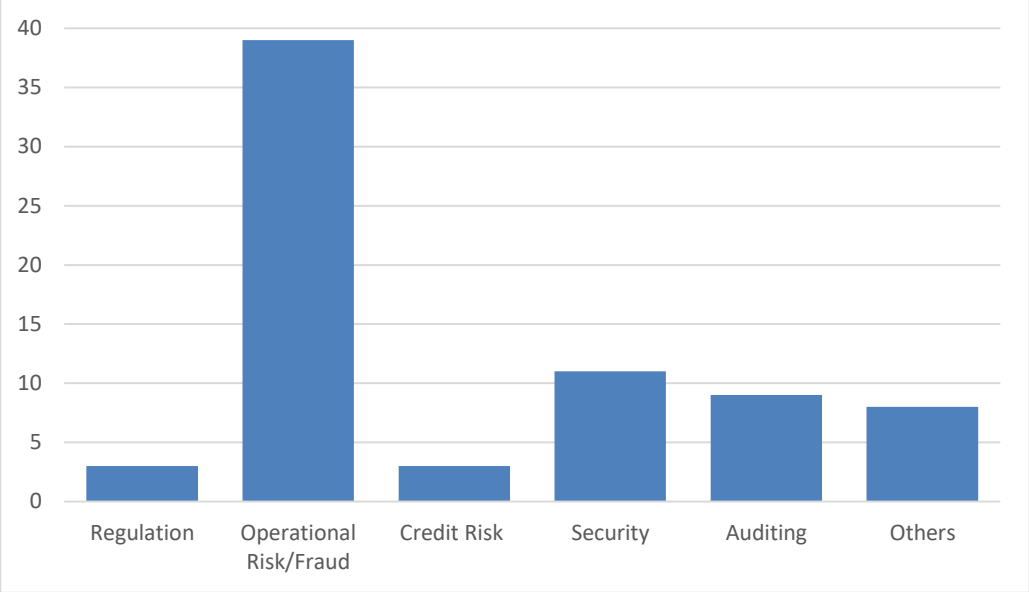


Fonte: Elaborado pelo autor.

Quanto ao principal objetivo do trabalho, a ampla maioria dos objetos de pesquisa é referente a risco operacional e fraudes, seguido por modelos de segurança, representando 53% e 15%, respectivamente. Considerando os temas risco operacional e detecção à fraudes e anomalias, o método de pesquisa mais utilizado faz uso de *aprendizado de máquina.* Nos últimos anos, a utilização de modelos analíticos não é mais um diferencial, mas tornou-se padrão no mercado devido a enorme quantidade de dados a ser processado e a disseminação global de empresas de tecnologiafuncionava no ramo de *banking* e meios de pagamento. Os principais objetivos dos trabalhos e a distribuição dos métodos podem ser verificados na Figura

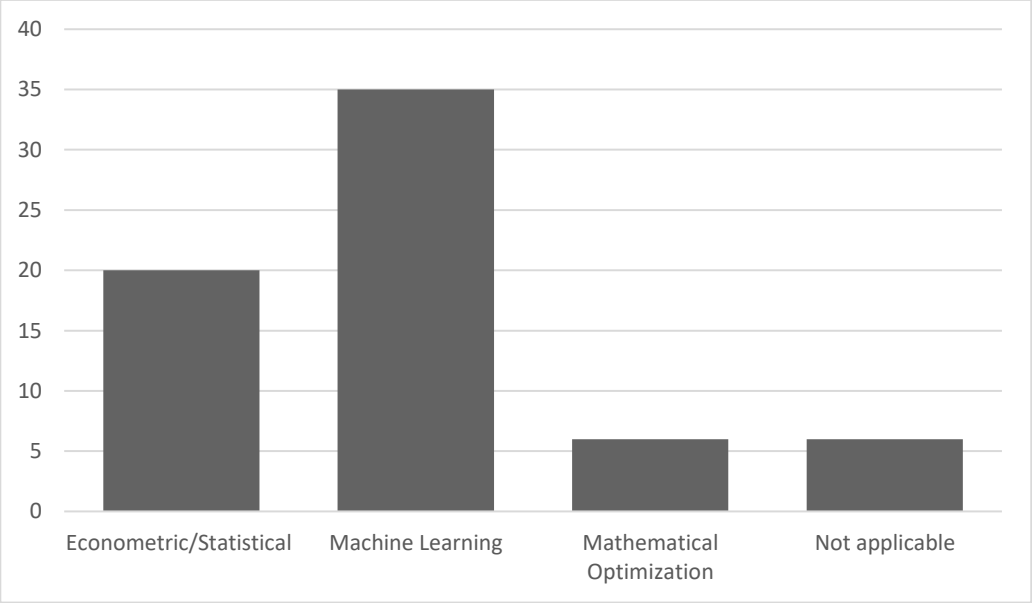
3.9 e Figura 3.10, respectivamente.

Figura 3.9 - Principais objetivos dos artigos



Fonte: Elaborado pelo autor.

Figura 3.10 - Principais métodos dos artigos



Fonte: Elaborado pelo autor.

O principal setor econômico que utiliza informações e modelos de detecção de fraudes é o bancário tradicional, seguido por outras instituições financeiras não bancárias. São consideradas instituições financeiras não bancárias as instituições que são apenas meios de pagamento, como as processadoras de cartões de crédito, e que são o principal setor alvo dos modelos analíticos. Ao

analisar as conclusões dos trabalhos, verifica-se a recorrência da proposição da criação de modelos analíticos híbridos para mitigar o risco operacional e reduzir perdas com fraudes e redução de falsos negativos desses modelos. Dentre os modelos híbridos, são sugeridos modelos que utilizam aprendizado de máquina em conjunto com regras de negócio ou a combinação em sequência de dois ou mais modelos analíticos para melhor precisão em casos de fraude.

É importante ressaltar que os modelos desenvolvidos nos trabalhos utilizam variáveis comportamentais e cadastrais de usuários lesados por fraude, a fim de identificar características e padrões de usuários para detectar momentos em que outra pessoa tenta se passar pelo usuário/cliente. Assim, os resultados dos artigos são - em sua maioria - comparativos com os resultados já descritos na literatura e que, por alguns motivos específicos, podem ter melhor aplicação prática conforme cada um desses estudos apresenta.

3.6. **Conclusão**

Este trabalho buscou compreender a produção científica sobre detecção de fraudes em bancos e identificar as conexões existentes entre os trabalhos, visto que o tema é relevante para o setor bancário devido aos potenciais prejuízos decorrentes de falhas na detecção, que podem inviabilizar o funcionamento da organização, fato que acarreta altos volumes de investimentos em métodos para reduzir riscos e perdas decorrentes de fraudes.

Para isso, a metodologia utilizada foi composta por uma combinação de métodos: a revisão sistemática da literatura e as métricas de redes complexas. Além disso, foram considerados 227 trabalhos científicos sobre o tema, o que possibilitou classificá-los segundo doze características como tipo de estudo, abordagem, corte, desenho, natureza, finalidade do estudo, método, abrangência espacial, período do estudo, foco, dados utilizados e resultados e medir a conectividade existente entre eles por meio de suas citações, coautorias e recorrência de palavras-chave.

Através do estudo apresentado, buscamos alcançar várias contribuições significativas. Em primeiro lugar, foi fornecida uma síntese abrangente do conhecimento atual, refletindo a profundidade e amplitude da pesquisa realizada. Em segundo lugar, identificamos lacunas críticas no conhecimento existente, estabelecendo assim uma agenda de pesquisa futura sobre o tema, que pode orientar investigações subsequentes. Por fim, com base em problemas do mundo real, desenvolvemos e fortalecemos um agrupamento de casos e modelos reais para possibilitar o avanço e refinamento de novas técnicas para detecção, previsão e análise de

fraudes, destacando o impacto prático e teórico de nosso trabalho. A análise de rede bibliométrica foi crucial principalmente porque facilita a visualização das relações entre autores, obras e conceitos, auxilia na identificação de autores-chave e trabalhos fundamentais e ajuda a mapear a estrutura e evolução de um campo de estudo.

A investigação revela que temas relacionados à detecção de fraudes estão significativamente correlacionados com algoritmos de detecção de anomalias. Esse alinhamento ocorre porque as técnicas empregadas em ambos os contextos são semelhantes, especialmente na abordagem do desequilíbrio de classe frequentemente encontrado em tais cenários. A maioria dos estudos fornece análises descritivas e introduz modelos preditivos para detecção de fraudes. O foco principal desses trabalhos está nos bancos tradicionais, e os resultados estão alinhados com os verificados anteriormente na literatura, embora com novas perspectivas.

Os resultados mostram que o uso de técnicas de aprendizado de máquina para detecção de fraudes são os métodos mais utilizados, em que muitos trabalhos buscam criar diferentes modelos para identificar padrões de comportamento e registro que indiquem que outra pessoa está tentando usar os acessos do usuário do site. instituição. Muitos desses modelos utilizam combinações de técnicas de aprendizado de máquina ou econometria/estatística com regras de negócios e dois tipos de modelos de aprendizado de máquina em sequência – métodos híbridos – buscando reduzir o número de falsos positivos e ser mais precisos na detecção de fraudes.

Mesmo com o crescimento de aplicações de outras formas de fraude como a engenharia social, não foram encontradas publicações de modelos analíticos sendo desenvolvidos para prevenir fraudes desta natureza. Ressalte-se que tais fraudes envolvendo engenharia social utilizam senhas, cartões físicos originais e dispositivos normalmente já utilizados pelo usuário da instituição, burlando alguns dos principais controles de mitigação de fraudes. Esses achados são úteis para a literatura científica que investiga os riscos operacionais quando apresentados, bem como para os agentes econômicos responsáveis pela identificação de fraudes nas organizações.

Comparado aos estudos de revisão existentes identificados na seção de revisão de literatura, há uma notável similaridade nas obras quanto à adoção generalizada de vários algoritmos de aprendizado de máquina combinados entre si na busca de modelos mais precisos. Além disso, um desafio comum no desenvolvimento desses modelos é a necessidade de bancos de dados extensos e a presença de dados desbalanceados. Com relação às lacunas na literatura

e às potenciais agendas de pesquisa, também é necessária uma exploração mais aprofundada em modelos generativos, conforme descrito por Hilal et al. (2022). Deste modo, vislumbra-se a possibilidade de desenvolvimento de pesquisa aprofundando detecção e anomalias com combinações de técnicas e o desenvolvimento de modelos generativos (LLMs) no auxílio aos desafios dos dados desbalanceados.

Também neste diagnóstico, constatou-se que não há publicações sobre modelos diretamente ligados à detecção de fraudes quando as operações são realizadas pelos próprios usuários em processos conhecidos como engenharia social. Nesse tipo de pesquisa, as variáveis e modelos devem ser diferentes dos tradicionais e a contraparte da operação deve ser observada. O aprofundamento desse tema poderia ser objeto de uma relevante agenda de pesquisa em finanças bancárias e meios de pagamento, com vistas a ampliar o escopo da prevenção de fraudes.

O desenvolvimento de trabalhos sobre modelos de detecção de fraudes e anomalias para bancos é fundamental para a consolidação e síntese do conhecimento teórico atual, possibilitando a identificação de métodos e técnicas comprovadamente eficazes na detecção de fraudes. Além disso, inspira o desenvolvimento de novas abordagens devido às lacunas na pesquisa existente e oferece insights valiosos sobre a aplicação de avanços tecnológicos, como aprendizado de máquina e inteligência artificial, na melhoria contínua dos sistemas de segurança bancária.

Capítulo 4

4. Aplicações de IA Generativa para Geração de Dados Sintéticos – O uso no Balanceamento de Classes em Predição de Fraudes

Alex Cerqueira Pinto

Universidade de Brasília - UNB

Resumo

O desequilíbrio de classes é um problema comum em aprendizado de máquina que pode prejudicar o desempenho de modelos preditivos. Este estudo apresenta uma abordagem inovadora para lidar com dados desbalanceados por meio da criação de dados sintéticos utilizando Inteligência Artificial Generativa (GenIA). A técnica de *data augmentation* via geração de dados sintéticos proposta emprega modelos generativos e Large Language Models (LLMs), para gerar amostras sintéticas das classes minoritárias, equilibrando assim a distribuição das classes no conjunto de dados. Desta forma, foram desenvolvidos tanto modelos tradicionais de *oversamplig* (SMOTE, GAN e VAE) quanto modelos de LLM com aplicação de RAG e Fine-tuning treinados para gerarem sobreamostragem. O novo LLM desenvolvido, via Fine-Tuning, com habilidade para gerar dados sintéticos foi batizado de Aurora. A metodologia e modelos foram aplicados e testados em base de dados de um grande banco nacional e sua eficiência comparada com os modelos atuais preditivos de fraudes. Os resultados indicaram que o modelo LLM Aurora desenvolvido *para data augmentation*, bem como engenharia de prompt e RAG, são viáveis e eficazes. Os testes de similaridade e correlação entre os dados originais e sintéticos demonstrou eficácia dos modelos de LLM para este fim. No entanto, identificou-se como limitação a alta dependência dos LLMs ao processamento do hardware (GPU), à quantidade de tokens e ao ambiente de plataforma em que estão hospedados.

Palavras-chave: Dados sintéticos, IA Generativa, Oversampling, Desequilíbrio de classes, Data Augmentation, LLM, SMOTE**.**

4.1. Introdução

Hoje, as transações financeiras são mais frequentes e diversas devido à globalização e digitalização. Infelizmente, este ambiente também proporcionou oportunidades para a proliferação de fraudes e golpes, levando a perdas significativas para bancos e seus clientes.

No cenário financeiro contemporâneo, a integridade e segurança das transações tornaram-se componentes cruciais na garantia da confiança dos clientes e na preservação da imagem das instituições bancárias. Paralelamente, a crescente sofisticação e frequência dos ataques cibernéticos, particularmente em fraudes, têm se intensificado. Portanto, é fundamental que os bancos se aprofundem no desenvolvimento e implementação de modelos avançados de detecção de fraudes, priorizando tanto a preservação do patrimônio dos clientes quanto a sustentabilidade operacional da própria instituição.

De forma sintética, modelos de detecção de fraudes são sistemas algorítmicos desenvolvidos para identificar atividades anômalas que podem indicar a ocorrência de uma fraude. Estes modelos, geralmente baseados em técnicas de aprendizado de máquina, analisam padrões históricos de transações e comportamentos dos usuários para prever e identificar transações potencialmente fraudulentas em tempo real. Empregam desde métodos estatísticos clássicos até abordagens mais avançadas, como redes neurais e florestas aleatórias. A escolha do algoritmo adequado depende da natureza dos dados, do tipo de fraude que se pretende detectar e dos recursos disponíveis para a implementação.

Além disso, a indústria tem investido intensivamente no uso de modelos analíticos, tanto puros quanto híbridos. A combinação de características provenientes de diferentes fontes de informação — como dados transacionais, cadastro de contas, comportamento do usuário e características temporais — contribui para aumentar a precisão na detecção de fraudes (Wei et al., 2013).

Essa tendencia atende também á crescente regulamentação e supervisão dos setores financeiros em muitas jurisdições. Bancos que não adotam medidas rigorosas de detecção de fraudes podem enfrentar penalidades regulatórias e litígios.

No entanto, um dos problemas recorrentes nesse campo é o desequilíbrio de classes nos conjuntos de dados, onde uma ou mais classes estão sub-representadas em relação às demais. Esse desbalanceamento pode levar a modelos de aprendizado de máquina tendenciosos e menos eficazes, especialmente em tarefas de classificação.

Para mitigar esse problema, técnicas de aumento de dados (*Data Augmentation*) e *oversampling* têm sido amplamente exploradas. O aumento de dados consiste em complementar uma coleta de observações com dados similares gerados a partir da informação já existente. Essa abordagem é frequentemente utilizada para superar a limitação conjuntos de dados pequenos, evidenciada em diversos estudos anteriores (Bej et al., 2021; Chang et al., 2013; Mohammed et al., 2020; Mujahid et al., 2024; Yang et al., 2024).

Tradicionalmente, métodos como o SMOTE (Synthetic Minority Over-sampling Technique) têm sido utilizados para gerar novas amostras da classe minoritária. No entanto, avanços recentes em Inteligência Artificial Generativa (GenIA), especialmente com o uso de *Large Language Models* (LLM), oferecem novas possibilidades para a criação de dados sintéticos de alta qualidade.

Os LLMs são modelos de linguagem treinados em grandes quantidades de texto, capazes de

gerar e compreender linguagem natural com um nível de sofisticação impressionante. Esses modelos têm mostrado um desempenho excepcional em várias tarefas de processamento de linguagem natural (NLP) e vêm sendo adaptados para aplicações em outras áreas, como a geração de dados sintéticos para aprendizado de máquina.

Os benefícios dos LLMs incluem a capacidade de capturar complexidades e nuances dos dados, criando amostras sintéticas que são mais realistas e diversificadas. Isso pode levar a uma melhor generalização dos modelos de aprendizado de máquina, reduzindo o risco de overfitting e aumentando a robustez contra dados desbalanceados.

Assim, o presente artigo tem como problema de pesquisa enfrentar o desbalanceamento de classes presente nas bases de dados em modelos de prevenção a fraudes. Este estudo investiga a eficácia do uso de LLMs para gerar dados sintéticos em conjuntos de dados desbalanceados.

Desta forma, o objetivo deste trabalho é desenvolver modelo de IA Generativa LLM próprio capaz de aplicar técnicas de *oversampling* e geração de dados sintéticos que equilibrem as classes de uma base de dados.

Objetivos específicos: i) desenvolver e mensurar a performance de modelos tradicionais de dados sintéticos, como SMOTE, GAN e VAE; ii) aplicar modelos de LLM para geração de dados sintéticos, via engenharia de prompt, e mensurar sua performance; iii) desenvolver e aplicar modelo de LLM com direcionamento para geração de dados sintéticos via RAG Geração Aumentada de Recuperação (*Retrieval-Augmented Generation*); iv) desenvolver um novo modelo inédito de LLM próprio com conhecimento intrínseco para geração de dados sintéticos.

Com base na questão de pesquisa, formulam-se as seguintes hipóteses: i) O uso de modelos LLM para geração de dados sintéticos não influencia significativamente a performance e a eficiência de modelos destinados à mitigação e prevenção de fraudes; ii) os modelo LLM geradores de dados sintéticos são eficientes na geração de dados similares aos reais; e iii) os modelos de geração de dados sintéticos com LLM (IA generativa) não apresentam diferenças significativas em relação a outros modelos de geração de dados sintéticos para mitigar e prevenir fraudes.

Os experimentos foram conduzidos em um conjunto de dados real de modelagem para fraudes desbalanceados. A performance dos modelos de geração de dados sintéticos foi comparada com base em dois critérios: primeiro, a similaridade entre as distribuições reais e sintéticas e, em segundo lugar, a melhora na performance de modelos preditivos de classificação de fraudes após a aplicação do *oversampling* sintético.

Os resultados demonstraram uma melhora significativa no poder preditivo de modelos que

usam data *augmentation*, bem como na capacidade de generalização dos modelos, comprovando a eficácia da abordagem proposta. Os testes de similaridade e correlação entre os dados originais e sintéticos demonstraram a eficácia dos modelos de LLM para este fim. Além disso, a análise das métricas de desempenho, como precisão, recall e F1-score, reforçou a similaridade do uso de dados sintéticos gerados por LLM em comparação com métodos tradicionais de *oversampling*, como SMOTE. Os resultados obtidos reforçam a promessa desta abordagem, destacando sua relevância e aplicabilidade em diversos cenários práticos.

Concluímos que a utilização de LLM para a criação de dados sintéticos representa uma solução promissora para problemas de desequilíbrio de classes, proporcionando uma maneira eficaz de melhorar a robustez e a performance de modelos de aprendizado de máquina em cenários realistas. Além disso, a geração de dados sintéticos com características similares às reais que podem ser aplicadas para fins de privacidade de dados.

A análise demonstra que a utilização de dados sintéticos melhora a precisão dos modelos preditivos, abordando problemas de escassez e desbalanceamento de dados. Os resultados indicam que a abordagem é eficaz e propõem futuras pesquisas para otimizar o uso de dados sintéticos na detecção de fraudes. Este estudo contribui significativamente para o campo ao mostrar a viabilidade e os benefícios do uso de LLM na geração de dados sintéticos.

A geração de dados sintéticos utilizando LLM para *oversampling* em problemas de aprendizado de máquina é de relevante para o campo de pesquisa acadêmico pois aborda e propõe uma solução a desafio crítico na modelagem de dados: o desequilíbrio de classes. Isso ajuda a evitar modelos tendenciosos e melhora a precisão e generalização das previsões.

Este trabalho se justifica ao explorar o uso de técnicas avançadas de IA Generativa, como os LLMs, em uma abordagem inovadora para criar amostras sintéticas realistas e diversificadas, potencialmente superando métodos tradicionais como o SMOTE. Além disso, a aplicação de LLMs e SLMs (*Small Language Models*) podem trazer insights valiosos sobre a capacidade desses modelos de capturar e replicar complexidades intrínsecas dos dados, contribuindo para o desenvolvimento de soluções mais robustas e precisas em aprendizado de máquina. Assim, este estudo não só enriquece o corpo teórico da ciência de dados e IA, mas também tem implicações práticas significativas para diversas áreas que dependem de previsões precisas a partir de dados desbalanceados.

4.2. **Revisão da Literatura**

As principais técnicas utilizadas em modelos analíticos de detecção de fraudes incluem regras e heurísticas, que estabelecem critérios pré-definidos para identificar transações suspeitas com base em limites de valor, padrões de comportamento do cliente, e localização geográfica. Além disso, técnicas estatísticas e de aprendizado de máquina são amplamente empregadas, utilizando modelos preditivos como árvores de decisão, regressão logística e redes neurais, aplicados por meio de algoritmos de classificação.

A detecção de fraudes é um campo crítico no contexto de segurança financeira e tecnológica, exigindo a aplicação de modelos de aprendizado de máquina para identificar transações suspeitas. Uma característica marcante desse domínio é o desbalanceamento de classes, em que as instâncias de fraudes representam uma pequena fração em comparação com as transações legítimas. Esse desequilíbrio impõe desafios significativos ao desenvolvimento de modelos eficazes, pois algoritmos convencionais tendem a ser enviesados em favor da classe majoritária, resultando em uma alta taxa de falsos negativos. Consequentemente, a revisão da literatura sobre métodos de detecção de fraudes frequentemente aborda estratégias específicas para lidar com esse desbalanceamento, incluindo técnicas de reamostragem, ajustes de penalização e a aplicação de algoritmos especializados, visando aumentar a sensibilidade e a precisão na identificação de fraudes.

Neste sentido Ngai et al. (2011) oferecem uma visão abrangente sobre a aplicação de técnicas de mineração de dados na detecção de fraudes financeiras, apresentando um framework de classificação para facilitar a compreensão e o desenvolvimento de abordagens eficientes. Eles discutem os desafios e limitações, como a falta de conjuntos de dados reais e a necessidade de lidar com desequilíbrios de classe que representam obstáculos significativos para a criação de modelos de prevenção a fraudes. Ngai et al. (2011) classificam as fraudes financeiras em quatro grupos: fraude bancária, fraude em seguros, fraude em títulos e commodities, e outras fraudes financeiras. Eles destacam a necessidade de uma abordagem integrada que combine diferentes métodos e técnicas para resultados mais eficazes.

West e Bhattacharya (2016) abordam a detecção de fraudes financeiras como um campo em constante evolução, necessitando de abordagens avançadas para enfrentar a crescente sofisticação das fraudes. Eles revisam técnicas como análise de padrões, mineração de dados, redes neurais, algoritmos genéticos, lógica difusa e sistemas especialistas, enfatizando os desafios do desequilíbrio de classes e a importância da interpretabilidade dos modelos. Além disso, defendem uma abordagem multidisciplinar que combine conhecimentos financeiros com

técnicas avançadas de análise de dados em tempo real.

Em um dos primeiros trabalhos a utilizarem de novas técnicas para geração de dados sintéticos aplicados a modelos, Wei et al. (2013) propõem uma estrutura eficaz para a detecção de fraudes bancárias online em dados desbalanceados, utilizando a técnica SMOTE para gerar um conjunto sintético de fraudes e melhorar o desempenho dos modelos. Eles aplicam algoritmos como Random Forest, utilizando características transacionais, informações de contas e comportamentos de usuários para melhorar a precisão da detecção. Os resultados mostram uma melhoria significativa na detecção de fraudes e redução de falsos positivos.

Com o mesmo objetivo, Nami e Shajari (2018) apresentam uma abordagem de duas etapas para a detecção de fraudes em cartão de crédito, considerando tanto o desequilíbrio de classe quanto os custos de erros de classificação. A metodologia combina o algoritmo Dynamic Random Forest (DRF) com o método k-Nearest Neighbors (k-NN), adaptando o DRF para lidar com o desequilíbrio e utilizando k-NN para ajustar as classificações finais. O resultado da proposta é uma nova medida de similaridade baseada no tempo de transação que atribui maior peso às transações recentes, mostrando que o comportamento recente dos portadores de cartão é crucial na avaliação de transações fraudulentas.

Zioviris et al. (2022) exploram a detecção de fraudes em cartões de crédito por meio de um modelo de aprendizado profundo em múltiplos estágios. A abordagem utiliza autoencoders para seleção de recursos e redes neurais convolucionais (CNNs) e recorrentes (RNNs) ~~para~~ que capturam características transacionais complexas. O estudo destaca a importância da detecção eficaz de fraudes em tempo real e da interpretabilidade do modelo para obter a confiança dos usuários, discutindo métodos para tornar o modelo mais explicável e transparente, sustentando sua aplicabilidade prática.

No que se refere ao tema de fraudes com engenharia social, é importante notar a existência de várias referências acadêmicas significativas na incluindo trabalhos de Mitnick e Simon (2002), Hadnagy (2011), Sheng et al. (2009), Stajano e Wilson (2010), e Hijji e Alam (2021). Em suma, estes autores descrevem a engenharia social como um conjunto de técnicas usadas por fraudadores para enganar pessoas e obter informações confidenciais ou acesso indevido a sistemas, frequentemente para atividades fraudulentas. Para eles, as interações entre a engenharia social e a detecção de fraudes e anomalias acontecem de diversas formas, como:

i) Phishing e spear phishing, onde os fraudadores se passam por entidades legítimas usando e-mails, mensagens de texto ou chamadas telefônicas falsas para solicitar informações

confidenciais;

ii) Engenharia direcionada, que envolve pesquisa detalhada sobre os alvos e coleta de informações pessoais disponíveis publicamente;

iii) Pretexting, que usa histórias fictícias ou desculpas para obter informações;

iv) Engenharia social reversa, onde fraudadores se passam por clientes ou superiores em instituições financeiras para obter acesso a sistemas ou informações confidenciais.

Estes estudos exploram diferentes aspectos da engenharia social, oferecendo percepções valiosas sobre as técnicas dos fraudadores, o perfil das vítimas e estratégias de prevenção. Destacam-se as abordagens comportamentais de clientes e fraudadores, indicando a necessidade de expandi-las por meio de modelos analíticos baseados em aprendizado de máquina. Este campo representa uma área promissora para pesquisa adicional e proposição de soluções inovadoras tanto para a academia quanto para a indústria.

No que se refere a trabalhos focados na geração de dados sintéticos como forma *oversamping* - ou seja, visando de aumentar artificialmente o número de exemplos da classe minoritária - diversos autores aplicam diferentes técnicas para equilibrar as classes minoritárias e majoritárias da base de dados. Entre as técnicas mais recentes e eficientes destacam-se: i) *Synthetic Minority Oversampling Technique* - SMOTE; ii) *Generative Adversarial Networks* - GAN, ou Redes Adversárias Generativas, em português, e iii) Autoencoders Variacionais (VAE).

O algoritmo SMOTE foi desenvolvido por Chawla et. al. (2002) com o objetivo de realizar a sobreamostragem da classe minoritária envolvendo a criação de exemplos sintéticos com base nas características dos ~~exemplos~~ observações reais dessa classe. A ideia principal é sobre-amostrar a classe minoritária e criar exemplos com base nos vizinhos mais próximos a essa classe. Em outras palavras, a classe minoritária é sobreramostrada ao gerar novos exemplos dados sintéticos similares a seus k-vizinhos mais próximos. Essa abordagem aumenta a diversidade dos dados da classe minoritária, evitando a simples duplicação e, consequentemente, reduzindo o risco de overfitting. Essa técnica ajuda a melhorar o desempenho dos modelos ao balancear o conjunto de dados, proporcionando métricas de avaliação mais robustas, como precisão, recall e F1-score. (Chawla et. al., 2002).

Embora o SMOTE seja eficaz para lidar com o desequilíbrio de classes, ele também tem algumas limitações. A técnica pode gerar exemplos que não são realistas se a distribuição da classe minoritária for muito complexa. Além disso, o SMOTE pode introduzir ruído se os exemplos gerados não forem representativos da distribuição real dos dados (Chawla et. al., 2002).

No que se refere a modelos generativos, as duas técnicas amplamente utilizadas para a geração de dados sintéticos baseados em redes neurais são as redes generativas adversárias (GAN) e os Autoencoders Variacionais (VAE). Essas técnicas têm se tornado extremamente populares devido à sua capacidade de capturar distribuições de dados complexas. A utilização de GANs para a geração de dados está ganhando crescente popularidade na comunidade de aprendizado de máquina, embora exija o treinamento de vários modelos, o que acarreta desafios e sobrecarga computacional na busca pelos parâmetros ótimos do modelo. Por outro lado, o método VAE faz suposições fortes sobre a distribuição dos dados, o que pode prejudicar o desempenho do modelo gerativo (Marco et al., 2023).

As Redes Adversárias Generativas (*Generative Adversarial Networks* - GANs) apresentadas por Goodfellow et al. (2014) são uma técnica poderosa para a geração de dados sintéticos, com ampla aplicação em diversas aplicações, desde a criação de imagens realistas até a síntese de dados em áreas como detecção de fraudes.

As GANs utilizam dois modelos principais que competem entre si: o Gerador e o Discriminador. O Gerador cria dados sintéticos a partir de vetores de ruído, tentando fazer com que esses dados sejam indistinguíveis dos dados reais. O Discriminador, por sua vez, tenta distinguir entre os dados reais e os gerados pelo Gerador. Durante o treinamento, o Gerador é otimizado para "enganar" o Discriminador, enquanto o Discriminador é otimizado para melhorar sua capacidade de distinguir entre dados reais e sintéticos. Esse processo adversarial resulta em um jogo na qual ambos os modelos se aprimoram continuamente (Goodfellow et al., 2014).

Outro modelo amplamente utilizado para a geração de dados sintéticos, e que vem ganhando popularidade nessa área, são os Autoencoders Variacionais (VAEs). Estes modelos, também baseados em redes neurais, foram desenvolvidos e apresentados por Kingma e Welling (2013), que combinou conceitos de autoencoders e variáveis latentes estocásticas para criar uma técnica ~~que pode~~ capaz de gerar novos dados a partir da distribuição aprendida durante o treinamento. Eles são uma extensão dos autoencoders tradicionais, incorporando princípios da estatística bayesiana para fornecer uma representação latente mais rica e probabilística.

Desta forma, os Autoencoders Variacionais (VAEs) são redes neurais projetadas para aprendizado não supervisionado que transformam dados de entrada em uma representação latente probabilística. O modelo consiste em um codificador que mapeia a entrada para uma distribuição latente, e um decodificador que reconstrói a entrada a partir de amostras dessa distribuição. Durante o treinamento, os VAEs maximizam a evidência inferior variacional (ELBO), que inclui um termo

de reconstrução para garantir a fidelidade da reconstrução e uma divergência de Kullback-Leibler (KL) para regularizar a distribuição latente. A técnica de reparametrização é usada para permitir a propagação do gradiente através do processo de amostragem, possibilitando a otimização eficiente dos parâmetros do modelo (Kingma e Welling, 2013).

Uma descrição mais detalhada dos três métodos citados será apresentada na seção seguinte de metodologia.

No que se refere ao estado da arte de algoritmos desenvolvidos para criação de dados sintéticos para equilíbrio de classes, destaca-se a utilização de modelos híbridos que combinam duas ou mais técnicas acima citadas.

Cheah, Yang e Lee (2023) abordam o problema do desequilíbrio de classes em conjuntos de dados de fraudes financeiras, que comumente resulta em previsões tendenciosas para a classe não fraudulenta e, consequentemente, em um desempenho insatisfatório na detecção de fraudes. Para mitigar esse problema, os autores exploram e comparam a eficácia de técnicas híbridas de geração de dados, combinando a Técnica de Sobreamostragem de Minoria Sintética (SMOTE) com Redes Generativas Adversariais (GAN).

A pesquisa de Cheah, Yang e Lee emprega diferentes arquiteturas de redes neurais, incluindo Redes Neurais Feed-forward (FNN), Redes Neurais Convolucionais (CNN) e uma combinação das duas (FNN+CNN), para avaliar o impacto das técnicas híbridas sobre o desempenho de detecção de fraudes. Os resultados indicam que as técnicas híbridas, especialmente SMOTified-GAN e GANified-SMOTE, superam ou igualam o desempenho das técnicas SMOTE e GAN isoladamente, demonstrando eficácia superior na detecção de fraudes financeiras. Independentemente do tamanho da amostra de fraudes geradas, as técnicas híbridas apresentaram um desempenho consistente, destacando-se como abordagens promissoras para melhorar a precisão e recall na detecção de fraudes, contribuindo significativamente para a mitigação do problema do desequilíbrio de classes.

No contexto da medicina, o artigo de Eom e Byeon (2023) aborda o desafio do desequilíbrio de classes em conjuntos de dados estruturados, com ênfase em dados clínicos. Compara métodos tradicionais de sobreamostramento com técnicas baseadas em redes adversárias generativas condicionais (CGAN) e redes adversárias generativas tabulares condicionais (CTGAN). Utilizando dados epidemiológicos de pacientes com demência de Parkinson do Biobanco Nacional da Coreia, os autores ajustaram a razão de desequilíbrio para diferentes valores e analisaram o desempenho das técnicas de sobre amostragem.

Os resultados mostraram que CGAN e CTGAN superaram significativamente os métodos tradicionais, como ROS, SMOTE, B-SMOTE e ADASYN, em termos de AUC e F1-score, melhorando a classificação das classes minoritárias. Este estudo não só amplia a aplicação de GAN para dados estruturados, mas também oferece uma solução eficaz para o problema de desequilíbrio de dados, sugerindo direções futuras para pesquisa nessa área. Conclui-se que CGAN e CTGAN são técnicas promissoras para lidar com o desequilíbrio de classes, apresentando um desempenho superior na classificação de dados médicos desequilibrados.

Kruschwitz e Schmidhuber (2024) investigam o uso de dados sintéticos gerados por modelos de linguagem de grande escala (LLMs) para melhorar a detecção de linguagem tóxica. O estudo avalia a eficácia desses dados sintéticos em comparação com dados reais na tarefa de classificação de toxicidade, especialmente em contextos em que há escassez de dados. O objetivo central é determinar se é possível utilizar exclusivamente dados sintéticos para treinar modelos de detecção de toxicidade com desempenho comparável ao de modelos treinados com dados reais.

A metodologia adotada envolve a geração de textos sintéticos usando o GPT-3 Curie, seguida por uma etapa de filtragem com um classificador treinado em dados reais para garantir a relevância e a qualidade dos textos gerados. Os autores realizam experimentos combinando dados originais e sintéticos, além de comparar desempenhos entre modelos treinados exclusivamente com dados sintéticos e com dados reais (Kruschwitz e Schmidhuber, 2024).

Kruschwitz e Schmidhuber (2024) utilizaram prompts cuidadosamente elaborados para guiar a saída de um modelo de linguagem de grande escala (LLM), como o GPT-3 Curie, com o objetivo de garantir a geração de dados relevantes e de alta qualidade. As métricas de avaliação incluem precisão, recall e F1-score, aplicadas em diferentes cenários de detecção de toxicidade, destacando melhorias notáveis na detecção de linguagem tóxica não-odiável e condescendente quando se utilizam dados sintéticos em combinação com dados reais.

Smith et al. (2024) propõem a utilização de modelos de linguagem de larga escala (LLMs) para a geração de dados sintéticos, com o objetivo de melhorar a performance de classificadores em conjuntos de dados desbalanceados. A metodologia OPAL, desenvolvida pelos autores, envolve a utilização de prompts específicos que guiam os LLMs na produção de novos registros sintéticos, preservando as propriedades estatísticas dos dados reais, como correlações entre variáveis e distribuições marginais. Esta abordagem foi comparada com métodos tradicionais de *oversampling*, como SMOTE e duplicação de dados.

Os resultados demonstraram que o OPAL supera os métodos tradicionais em todas as

métricas avaliadas, reduzindo significativamente a taxa de erro dos classificadores em cenários de dados desbalanceados. Por exemplo, na base de dados Diabetes, a taxa de erro com OPAL foi significativamente menor em comparação com o SMOTE e sem *oversampling*. As conclusões indicam que o uso de LLMs para *oversampling* sintético é uma estratégia eficaz e promissora, proporcionando uma melhor generalização dos modelos de aprendizado de máquina (Smith et al., 2024).

Zhou et al. (2024) fornecem uma visão abrangente dos métodos de aumento de dados orientados por grandes modelos, categorizando os estudos relevantes e explorando suas aplicações, técnicas de pós-processamento, sucessos e limitações. O artigo busca contribuir para a geração de dados suficientes e diversos para treinar modelos grandes mais sofisticados, oferecendo insights críticos para pesquisadores. Os autores abordam métodos de aumento de dados orientados por grandes modelos, como modelos de linguagem e de difusão, categorizando-os em aumento de imagens, aumento de texto e aumento de dados pareados. Além disso, explora técnicas de pós-processamento de dados e discute suas aplicações em processamento de linguagem natural, visão computacional e processamento de sinais de áudio. O objetivo é fornecer uma visão abrangente dessas técnicas, avaliando seus sucessos e limitações, e sugerindo desafios e direções futuras, visando contribuir para a geração de dados suficientes e diversos para treinar modelos grandes mais sofisticados.

Os resultados indicam que os métodos de aumento de dados orientados por grandes modelos superam as abordagens tradicionais em diversas aplicações. Em visão computacional, os modelos geraram imagens sintéticas de alta qualidade, melhorando a acurácia dos sistemas de reconhecimento de imagem. No processamento de linguagem natural, os métodos aumentaram a diversidade e a qualidade dos textos gerados, resultando em melhorias nas tarefas de tradução e compreensão de texto. As conclusões destacam a necessidade de desenvolver técnicas mais robustas e escaláveis, sugerindo que futuras pesquisas devem focar em melhorar a eficácia e a diversidade dos dados gerados por grandes modelos. No que se refere ao aumento de dados pareados, que envolve a combinação de dados de diferentes modalidades, os autores concluem que, embora esse tipo de aumento ofereça melhorias substanciais na qualidade e quantidade de dados de treinamento, ainda há desafios relacionados à manutenção da coerência e relevância dos dados gerados (Zhou et al., 2024).

Neunzig et al. (2023) apresentaram um estudo sobre a geração de dados sintéticos para melhorar o desempenho preditivo de características de teste hidráulico. A metodologia envolveu o

uso de Redes Adversariais Generativas e Autoencoders Variacionais. Os resultados indicaram que a combinação de regressão da altura de gap e classificação não fornece uma abordagem adequada para um conjunto de dados de produção com uma divisão de séries temporais. No entanto, a adição de dados sintéticos ao conjunto de dados original melhorou o modelo de classificação. A Regressão por Floresta Aleatória foi usada para a regressão de altura de gap, produzindo erros médios absolutos mais baixos e resultando em um R^2 R2 mais alto em comparação com outros métodos de regressão.

Marco et al. (2023) desenvolveram um modelo de autoencoder variacional condicional (CVAE) combinado com uma função de relevância para reamostragem, visando a geração de dados sintéticos para a tarefa de regressão no contexto da estimativa de esforço de software (SEE). A principal inovação deste estudo reside na adaptação do CVAE, originalmente utilizado para classificação, para no contexto de regressão, abordando a limitação de dados e melhorando a precisão das previsões. O desempenho do modelo foi avaliado e comparado com sete métodos populares de geração de dados sintéticos, incluindo SMOTER, GAN, CTGAN, entre outros, utilizando métricas como Erro Médio Absoluto (MAE), Raiz do Erro Quadrático Médio (RMSE), Erro Absoluto Relativo (RAE) e Coeficiente de Determinação (R^2). Os resultados demonstraram que o CVAE proposto supera os métodos existentes nas bases de dados China e Desharnais. As conclusões do estudo sugerem que o uso do CVAE com reamostragem baseada em relevância é capaz de gerar dados sintéticos que são altamente semelhantes aos reais, melhorando a qualidade das previsões de SEE em comparação com modelos que utilizam apenas os dados originais e outros métodos de geração de dados sintéticos.

Nas finanças, Ding et al. (2023) investigaram a aplicação de uma Rede Generativa Adversarial Variacional (VAEGAN) aprimorada na detecção de fraudes com cartões de crédito, enfrentando o desafio do desequilíbrio de dados. A metodologia envolveu a geração de dados sintéticos da classe minoritária para aumentar o conjunto de treinamento, utilizando um VAEGAN aprimorado. Os resultados mostraram que este método de *oversampling* supera técnicas tradicionais como GAN, VAE e SMOTE, melhorando significativamente a precisão e o F1 score dos modelos de classificação. Conclui-se que a VAEGAN aprimorada é eficaz para resolver problemas de classificação em conjuntos de dados desequilibrados, oferecendo uma abordagem robusta para a detecção de fraudes em cartões de crédito.

Veigas et al. (2021) propuseram um modelo de detecção de fraudes em transações com cartões de crédito, utilizando um Conjunto Empilhado Otimizado (OSE) que incorpora técnicas de

sobreamostragem como SMOTE e GAN para gerar dados sintéticos. O estudo demonstrou que, após o balanceamento do conjunto de dados, os classificadores individuais (MLP, kNN e SVM) mostraram melhorias na pontuação F1. O modelo final, que combina esses classificadores em um meta-classificador, alcançou uma precisão de 99,86% e um aumento de 16% na pontuação F1, destacando-se como uma solução eficaz para a detecção de fraudes em cenários reais.

Patki et al. (2016) apresentam o Synthetic Data Vault (SDV), um sistema que gera dados sintéticos para viabilizar projetos de ciência de dados, preservando a privacidade dos dados originais. O SDV utiliza uma técnica de modelagem recursiva chamada "agregação de parâmetros condicionais" para criar um modelo generativo de bancos de dados relacionais. A técnica usa a cópula Gaussiana para capturar a distribuição multivariada dos dados. Testado em cinco conjuntos de dados públicos, o sistema demonstrou que os dados sintéticos produzidos podem substituir efetivamente os dados reais na construção de modelos preditivos, sem perda significativa de acurácia. Os resultados mostraram que não houve diferença estatisticamente significativa na performance dos modelos preditivos desenvolvidos com dados sintéticos comparados aos desenvolvidos com dados reais. Conclui-se que o SDV é uma solução viável e eficiente para a geração de dados sintéticos, capaz de atender às necessidades de diversas aplicações em ciência de dados, promovendo a segurança e a privacidade dos dados originais.

Do mesmo modo, segundo Tanaka e Aranha (2019), a utilização de Redes Adversariais Generativas (GANs) para a geração de dados sintéticos pode ser uma técnica eficaz em situações de conjuntos de dados desequilibrados e onde a privacidade dos dados é uma preocupação. Em seus experimentos, um classificador de Árvore de Decisão treinado com dados sintéticos alcançou resultados comparáveis, e em alguns casos superiores, aos obtidos com dados reais. Além disso, ao comparar o desempenho de diferentes métodos de aumento de dados para classes minoritárias, verificou-se que as GANs proporcionaram melhorias significativas, ainda que não superiores ao SMOTE e ADASYN em todos os cenários avaliados. Esses resultados indicam que o uso de GANs é promissor para evitar overfitting e melhorar a representatividade dos dados em tarefas de aprendizado de máquina.

Cai et al. (2023) propuseram uma abordagem para resolver o problema do desequilíbrio de dados na inferência de tópicos disciplinares hierárquicos, um problema de processamento de linguagem natural, utilizando aumento de dados baseado em grandes modelos de linguagem (LLMs). O estudo foca em propostas de pesquisa submetidas para financiamento, que apresentam desequilíbrios de dados entre diferentes disciplinas. A metodologia envolve a amostragem de

classes minoritárias e a construção de prompts para gerar dados adicionais com o modelo de linguagem Llama V1. As propostas de pesquisa geradas, usando os prompts desenvolvidos, são capazes de abordar os problemas de desequilíbrio e gerar dados científicos de alta qualidade. Os resultados indicam uma melhoria na precisão dos modelos de inferência de tópicos e na equidade da atribuição de revisores especialistas.

De modo a avaliar os riscos e viéses que os modelos de oversampling também podem apresentar Van den Goorbergh et al. (2022) apresentam uma análise detalhada dos efeitos das correções de desequilíbrio de classes em modelos de regressão logística, tanto padrão quanto penalizada. O principal achado é que métodos como *random undersampling* (RUS), *random oversampling* (ROS) e SMOTE não melhoram a discriminação dos modelos, medida pelo AUROC, mas causam uma superestimação sistemática das probabilidades para a classe minoritária, resultando em forte descalibração dos modelos. Essa distorção nas estimativas de probabilidade compromete a utilidade clínica dos modelos, podendo levar a decisões equivocadas, como o encaminhamento excessivo de pacientes para tratamentos especializados. Além disso, o estudo evidencia que o uso de RUS reduz artificialmente o tamanho da amostra, aumentando o risco de *overfitting* e a variância dos modelos. Mesmo após a aplicação de procedimentos de recalibração, os modelos corrigidos por desequilíbrio continuaram apresentando desempenho inferior em termos de calibração e discriminação. Os autores concluem que o desequilíbrio de classes não é um problema intrínseco para modelos de predição e alertam para os riscos de aplicar correções de desequilíbrio sem evidência clara de benefício, especialmente em contextos clínicos onde a calibração precisa das probabilidades é essencial para a tomada de decisão.

Neste mesmo sentido, Chen et al (2024) oferece uma revisão abrangente das abordagens recentes para lidar com dados desbalanceados, incluindo técnicas de pré-processamento, métodos algorítmicos e estratégias de aprendizado em conjunto. A principal conclusão é que, embora existam diversas técnicas promissoras para melhorar o desempenho de modelos em cenários de desequilíbrio, cada uma apresenta limitações significativas que devem ser cuidadosamente consideradas. O *oversampling*, por exemplo, pode levar ao *overfitting*, aumentar a complexidade computacional e introduzir ruído nos dados, o que compromete a capacidade de generalização dos modelos. Técnicas como SMOTE e suas variantes podem gerar amostras duplicadas ou falsas, sendo inadequadas para dados de alta dimensionalidade e sensíveis ao ruído. Métodos algorítmicos como o aprendizado sensível a custo enfrentam dificuldades na estimativa precisa dos custos de erro e podem sofrer com viés de distribuição. Estratégias de *ensemble learning*, como *boosting* e

bagging, embora eficazes, exigem alto custo computacional e são sensíveis a dados ruidosos ou mal rotulados. O trabalho destaca que a escolha inadequada de técnicas pode comprometer a performance em dados não vistos e reduzir a robustez dos modelos, sendo necessário um equilíbrio entre complexidade, custo computacional e capacidade de generalização para lidar com o problema de forma eficaz.

4.3. **Metodologia**

Os modelos analíticos de detecção de fraudes são abordagens avançadas que utilizam técnicas estatísticas, de aprendizado de máquina (*machine learning*) e inteligência artificial para identificar padrões e comportamentos suspeitos em transações financeiras. Por meio de algoritmos e análise de dados, esses modelos processam grandes volumes de informações para detectar atividades fraudulentas ou anomalias que indiquem possíveis fraudes.

Esses modelos analíticos podem incorporar diferentes variáveis e indicadores, como padrões de comportamento do cliente, histórico de transações, informações demográficas, entre outros. Ao aplicar técnicas estatísticas e de aprendizado de máquina, os modelos analíticos são capazes de detectar comportamentos atípicos ou suspeitos, permitindo uma intervenção rápida e eficaz para prevenir fraudes. No entanto, é importante ressaltar que nenhum modelo analítico é infalível. Novas técnicas de fraude estão constantemente surgindo, exigindo uma atualização contínua dos modelos e a combinação de abordagens analíticas com a expertise humana.

Desta forma, neste trabalho explora-se a aplicação de técnicas de geração de dados sintéticos em quatro frentes conforme Ding et al. (2023) e Smith et al. (2024):

i. Utilizando métodos de aprendizado de máquinas via Redes Adversárias Generativas, Redes Variacionais e SMOTE.
ii. Metodologias baseadas em *Large Language Models* (LLM) com engenharia de prompt,
iii. Aplicação de LLM com método RAG Geração Aumentada por Recuperação).
iv. Por fim, desenvolve-se do modelo inédito de SLM (*small language model*), a partir de um LLM com *fine-tuning,* com habilidade de geração de dados sintéticos.

Desta forma, primeiramente foi realizado revisão da literatura adjacente sobre o tema, com uma análise para identificar tendências, desafios e soluções em detecção de fraudes em bancos e em seguida, o desenvolvimento de modelos de aprendizado de máquina e de Inteligência Artificial

para atuar sobre um dos principais desafios encontrados – desbalanceamento de classes.

Conforme abordado na revisão de literatura, um dos principais métodos será o SMOTE de Chawla et. al. (2002). O SMOTE atua criando exemplos sintéticos da classe minoritária por meio da interpolação entre amostras de classes minoritárias e seus K vizinhos mais próximos, em vez de simplesmente duplicar os exemplos existentes. A metodologia básica pode ser descrita nas seguintes etapas:

1. Seleção de exemplos minoritários: Para cada exemplo x da classe minoritária, seleciona-se um conjunto de k vizinhos mais próximos (geralmente utilizando a distância Euclidiana).
2. Geração de exemplos sintéticos: Para cada exemplo x da classe minoritária, um ou mais dos k vizinhos são selecionados aleatoriamente. Um novo exemplo sintético é gerado ao interpolar linear entre o exemplo x e um de seus vizinhos $x\_vizinho$

Desta forma, a formulação matemática do SMOTE pode ser descrita como: seja x um exemplo da classe minoritária, e $x\_vizinho$ um dos k vizinhos mais próximos de x. O novo exemplo sintético $x_sintético$é gerado pela seguinte fórmula:

$$x\_sintético = x + \lambda(x\_vizinho - x) \tag{4.1}$$

onde:

- λ é um número aleatório no intervalo [0,1].
- x é o vetor de características do exemplo minoritário,
- $x\_vizinho$ é o vetor de características de um dos k vizinhos mais próximos de x.

Essa interpolação gera um ponto ao longo da linha que conecta x e $x\_vizinho$, distribuído aleatoriamente no espaço entre eles.

O segundo modelo a ser combinado com os demais de forma híbrida para geração dos dados sintéticos são as Redes Adversárias Generativas (GANs) conforme apresentado por Goodfellow et al., (2014). As GANs consistem em dois modelos neurais que competem entre si: o Gerador e o Discriminador. O Gerador produz dados sintéticos a partir de vetores de ruído, buscando criar dados que se pareçam o máximo possível com os dados reais. Já o Discriminador tem a tarefa de diferenciar entre os dados reais e os dados gerados pelo Gerador. Durante o treinamento, o Gerador é ajustado para enganar o Discriminador, enquanto o Discriminador é ajustado para melhorar sua capacidade de distinguir entre dados reais e gerados. Esse processo adversarial cria um ciclo de aprimoramento contínuo entre os dois modelos, resultando em um jogo de soma zero.

A formulação matemática das GANs pode ser descrita como um problema de otimização em duas partes, onde G representa o Gerador e D representa o Discriminador. A função de perda

$V(D, G)$ é definida como:

$$min_G max_D V(D, G) = E_{x \sim p_{data}(x)}[logD(x)] + E_{z \sim p_z(z)}\left[log\left(1 - D\big(G(z)\big)\right)\right] \quad (4.2)$$

onde:

$E_{x \sim p_{data}(x)}$ é a expectativa da probabilidade logarítmica que o Discriminador D atribui a dados reais x.

$E_{z \sim p_z(z)}$é a expectativa da probabilidade logarítmica que o Discriminador atribui a dados sintéticos $\mathrm{G(z)}$ gerados a partir do vetor de ruído z.

O procedimento de treinamento das Redes Adversárias Generativas (GANs) começa com a definição da arquitetura da rede, seguida pela inicialização dos pesos do Gerador e do Discriminador. Posteriormente, o processo alterna entre a atualização dos parâmetros do Discriminador (D) para maximizar a função de perda e a atualização dos parâmetros do Gerador (G) para minimizá-la.

Na etapa seguinte, o Discriminador é treinado para maximizar sua capacidade de distinguir entre dados reais e sintéticos. Isso é feito amostrando um minibatch de dados reais e um minibatch de vetores de ruído. O Gerador utiliza esses vetores para criar dados sintéticos, e os parâmetros do Discriminador são então ajustados para aumentar a probabilidade de classificação correta.

Após isso, o Gerador é treinado para minimizar a função de perda, criando dados sintéticos que o Discriminador não consiga diferenciar dos reais. Para isso, um novo minibatch de vetores de ruído é amostrado, gerando novos dados sintéticos, e os parâmetros do Gerador são ajustados para reduzir a probabilidade de que o Discriminador detecte esses dados como falsos. Esse ciclo de treinamento, alternando entre o ajuste do Discriminador e do Gerador, é repetido até que o modelo convirja, idealmente quando o Discriminador não consegue distinguir entre dados reais e sintéticos.

Autoencoders Variacionais (VAEs), conforme apresentado por Kingman e Welling (2013), representam uma classe de redes neurais probabilísticas que se consolidaram como ferramentas poderosas para aprendizado de representações, geração de dados e visualização de informações complexas. Sua capacidade de modelar distribuições de probabilidade sobre dados de alta dimensão os torna particularmente úteis em diversos domínios, incluindo visão computacional, processamento de linguagem natural e análise de dados.

A estrutura fundamental de um VAE consiste em dois componentes principais: um codificador e um decodificador. O codificador recebe um dado de entrada e o mapeia para um espaço latente probabilístico, representado por uma distribuição de probabilidade. Essa distribuição

captura as características essenciais do dado original de forma compacta e eficiente. Já o decodificador recebe amostras do espaço latente e as utiliza para reconstruir dados semelhantes aos dados de entrada.

A chave para a natureza probabilística dos VAEs reside na função de perda, que incorpora tanto o erro de reconstrução do decodificador quanto a divergência entre a distribuição do espaço latente e uma distribuição de referência pré-definida, geralmente uma distribuição normal ou uniforme. Essa regularização probabilística incentiva o VAE a aprender representações latentes que capturam não apenas as informações dos dados de treinamento, mas também a variabilidade inerente à classe de dados.

A abordagem proposta baseia-se na reparametrização da fronteira (limite) inferior variacional para obter um estimador que pode ser diferenciado, que pode ser otimizado utilizando técnicas de gradiente estocástico. A inferência variacional envolve a otimização de uma aproximação à distribuição posterior, geralmente intratável. O objetivo do VAE é maximizar a evidência inferior variacional (ELBO), que é dada por:

$$L(\theta, \phi;\, x) \;=\; E_{z\sim q\phi(z|x)}[log\; p_\theta(x|z)] \;-\; D_{KL}(q_\phi(z|x)\;||\;p(z)) \tag{4.3}$$

onde:

$E_{z\sim q\phi(z|x)}[log\; p_\theta(x|z)]$: é o termo de reconstrução, que mede quão bem o decodificador pode reconstruir a entrada x a partir da amostra latente z.

$D_{KL}(q_\phi(z|x)\;||\;p(z))$: é a divergência de Kullback-Leibler, que regulariza a distribuição latente q_ϕ(z | x) para que esteja próxima do prior p(z).

Durante o treinamento, o codificador produz os parâmetros da distribuição latente q_ϕ (z | x), geralmente uma média μ e uma variância σ^2. Uma amostra z é obtida usando a reparametrização:

$$\mathrm{z} = \mu + \sigma \odot \epsilon \quad \text{com} \quad \epsilon \sim \mathrm{N}(0, \mathrm{I}) \tag{4.4}$$

Isso permite que o gradiente seja propagado através do processo de amostragem, permitindo o uso de métodos de otimização padrão, como o gradiente descendente.

O segundo método aplicado de geração de dados sintéticos será por meio da utilização de modelos de LLM, via engenharia de prompt. Esta abrange um conjunto diversificado de técnicas e estratégias destinadas a aprimorar o desempenho dos modelos de IA. Entre essas práticas, destaca-se a utilização de prompts contextuais que fornecem (incorporam) informações adicionais para direcionar as respostas, a experimentação com diferentes formatos e estruturas de indagações, bem

como a implementação de métodos iterativos de teste e refinamento. A personalização dos prompts para tarefas específicas adapta os modelos a diversos contextos. Dessa forma, a engenharia de prompt não apenas eleva a qualidade das respostas geradas, mas também expande a versatilidade e a aplicabilidade dos sistemas de IA generativa em múltiplos cenários.

Além da necessidade de construir e testar prompts para o modelo, emprega-se a metodologia, *In Context Learning,* também conhecida como *Few-Shot Learing.* Conforme descrito Brown et al. (2020) a ICL constitui uma abordagem avançada na utilização de modelos de inteligência artificial generativa (GenAI), permitindo que esses sistemas aprendam e adaptem-se a tarefas específicas com base apenas nas informações contextuais fornecidas durante a interação, sem a necessidade de re-treinamento extensivo. As vantagens do ICL incluem a flexibilidade para lidar com uma ampla variedade de tarefas, a eficiência na utilização de dados contextuais para melhorar a precisão das respostas e a capacidade de personalização conforme as necessidades específicas do usuário.

No setor financeiro, a aplicação do ICL pode ser particularmente benéfica em áreas como análise de risco, previsão de mercado, atendimento ao cliente personalizado e detecção de fraudes. Além disso, a capacidade do ICL de interpretar e contextualizar grandes volumes de dados financeiros possibilita decisões mais informadas e estratégicas, aumentando a eficiência operacional e a competitividade das instituições financeiras.

Neste trabalho, exploramos o *In-Context Learning* (ICL) por meio da construção e teste de diversos prompts estruturados com seguinte formato: Contexto, Instrução, Objetivo, Relevância, Expectativa, Restrições, Demonstração e Indicador de Saída

Em sequência, o terceiro método de geração de dados a ser testado para o objetivo deste trabalho será a aplicação da técnica de *Retrieval Augmented Generation* (RAG) em alguns modelos LLM de código aberto para verificar a performance.

Considera-se RAG uma abordagem que busca enriquecer a capacidade de geração de texto dos modelos de linguagem de grande porte (LLMs) por meio de recursos de recuperação de informação. Em termos gerais, o RAG combina técnicas de busca em bases de conhecimento externas (como bancos de dados, wikis ou coleções de documentos) com a capacidade de geração de texto de modelos como GPT ou BERT. Dessa forma, quando se faz uma consulta ou pergunta, o sistema extrai trechos relevantes do conjunto de documentos disponíveis e integra essas informações no processo de geração de respostas, resultando em textos mais precisos e contextualizados. A principal vantagem desse método é que ele permite ao modelo incorporar

conhecimento externo, evitando, por exemplo, limitações decorrentes de treinamento estático ou alucinações de conteúdo, pois o processo de busca se apoia em fontes confiáveis e específicas.

A formulação matemática básica do RAG pode ser entendida em duas etapas. Na primeira, dada uma consulta Q (query), busca-se um conjunto de documentos relevantes $\{\mathrm{d}_1, \mathrm{d}_2, \ldots, \mathrm{d}_k\}$ através de uma função de similaridade, frequentemente modelada por uma probabilidade condicional $p(\mathrm{d}_i|q)$. Na segunda etapa, gera-se o texto com base nos documentos recuperados, combinando a distribuição do modelo de linguagem $\mathrm{p}_\theta(y|q, \mathrm{d}_i)$ para cada documento. Uma forma resumida dessa abordagem pode ser descrita como:

$$p(y \mid q) \approx \sum_{i=1}^{k} p_\theta(y \mid d_i, q) p(d_i \mid q), \tag{4.5}$$

onde, y representa a sequência de texto a ser gerada, e θ são os parâmetros do modelo treinado. Essa formulação possibilita ao sistema integrar fontes externas de informação, incrementando sua capacidade de resposta baseada em evidências concretas.

No presente estudo, descrevemos as etapas de desenvolvimento de um sistema RAG nesta seção, tendo como base o modelo DeepSeek V3 e utilizando a biblioteca LangChain para integração entre repositórios vetoriais e o LLM. Destaca-se a importância de um processo metodológico sistemático, abrangendo desde a preparação de dados até a manutenção contínua do modelo em produção.

A fim de construir um sistema RAG eficaz, propõe-se uma metodologia estruturada em cinco etapas principais, baseado em Shen et al. (2024) e (Lewis et al., 2020), descritas a seguir. Cada fase será descrita, enfatizando os aspectos práticos e os cuidados necessários para permitir qualidade e maior precisão do sistema.

1. Preparação e Análise de Dados

1.1. Coleta e Organização

A construção de um conjunto de dados robusto inicia-se pela seleção e aquisição de fontes adequadas (p. ex., bases públicas, repositórios corporativos, APIs especializadas). Para melhor gerenciamento, recomenda-se armazenar as informações em estruturas padronizadas, como arquivos CSV, JSON ou bancos de dados relacionais.

1.2. Limpeza e Pré-Processamento

A etapa a seguir envolve a remoção de inconsistências, duplicações e ruídos, bem como a padronização do texto. Práticas comuns incluem a normalização de *encoding*, e a remoção de caracteres especiais. A segmentação do texto em sentenças ou parágrafos é frequentemente

utilizada para facilitar etapas futuras de geração de *embeddings*.

1.3. Validação dos Dados

Antes de proceder ao treinamento, realiza-se uma validação minuciosa para assegurar a integridade e a completude do conjunto de dados (Kandpal et al., 2023). Essa verificação pode ser feita por meio de inspeção manual ou ferramentas de *data validation*, visando identificar lacunas ou problemas que comprometam a fase de modelagem.

2. Geração de Embeddings

2.1. Seleção do Modelo

A definição de um modelo de *embedding* pré-treinado deve levar em conta o idioma, o tamanho do vocabulário e a natureza dos textos (Reimers & Gurevych, 2019). SentenceTransformers e Hugging Face Transformers são exemplos de ferramentas que oferecem modelos consolidados e atualizados para a geração de vetores.

2.2. Conversão Vetorial e Indexação em Banco de Dados Vetorial

Com o modelo escolhido, cada unidade textual (documento, parágrafo ou sentença) é convertida em um vetor de alta dimensão. Essas representações vetoriais codificam aspectos semânticos, permitindo a comparação de similaridade entre diferentes textos de forma eficiente. Em seguida, os vetores são armazenados em bancos de dados especializados, tais como VectorDB, Mongo, que viabilizam buscas por similaridade de maneira escalável. A adoção de técnicas de indexação otimiza a recuperação de documentos relevantes, reduzindo a latência do sistema.

3. Integração com a Plataforma de RAG

3.1. Carregamento do LLM Base

O primeiro passo consiste no carregamento do modelo de linguagem base para aplicação do RAG. Neste projeto, optou-se pela utilização do Qwen 2.5, Llama 3.3, O3 e DeepSeek V3, de forma a testar a performance e viabilidade com um leque de modelos.

3.2. Configuração do Pipeline de Recuperação

Em seguida, define-se o fluxo de trabalho que une o LLM ao banco de dados vetorial. A biblioteca LangChain fornece componentes para orquestrar o processo de consulta, de modo que o *prompt* do LLM seja enriquecido com os textos mais relevantes antes da geração de resposta.

3.3. Integração Recuperação-Geração

A etapa chave consiste em unir os documentos recuperados aos mecanismos de geração do LLM. Nessa fase, configura-se se o texto externo será concatenado, sumarizado ou transformado de outras formas, garantindo que a saída final seja contextualizada e coerente (Shen et al., 2024).

4. Testes e Avaliação do Sistema

4.1. Definição de Métricas

A mensuração de desempenho do RAG pode envolver indicadores como *Precision*, *Recall*, *F1 Score* ou métricas de similaridade para avaliar a relevância dos resultados (caso aplicado neste trabalho). Em aplicações específicas, avaliações humanas também podem ser empregadas para verificar a adequação das respostas (Liang et al., 2022).

4.2. Ajuste de Parâmetros

Com base na análise dos resultados, parâmetros como modelo de *embedding*, tamanho do *embedding*, *hyperparameters* do LLM (p. ex., *temperature*) e algoritmos de recuperação podem ser ajustados. Esse processo iterativo de refinamento visa incrementar a exatidão e a utilidade prática do sistema.

5. Manutenção e Atualização

5.1. Atualização de Dados

Em cenários onde novas informações surgem continuamente, a atualização periódica do repositório é imprescindível. Nessa etapa, adicionam-se documentos recentes, removem-se dados obsoletos e, quando necessário, são corrigidas eventuais inconsistências.

5.2. Reindexação e Retreinamento

Sempre que mudanças substanciais ocorrem na base de dados, recomenda-se a geração de novos *embeddings* e a reindexação do banco vetorial, além de possíveis ajustes no modelo de linguagem, especialmente se a distribuição dos dados sofrer alterações significativas.

6. Monitoramento Contínuo

Finalmente, acompanha-se o desempenho em produção por meio de *logs* e indicadores de qualidade. Uma queda de desempenho ou o aumento de erros de recuperação podem sinalizar a necessidade de intervenções pontuais, seja na base de dados ou na configuração do sistema.

Por último, será também desenvolvido um modelo de SLM inédito, chamado Aurora, gerado por fine-tuning. Esta técnica consiste na adaptação de um modelo já treinado em grandes quantidades de dados de texto (um “modelo base”) para tarefas específicas, por meio do treinamento adicional em um conjunto de dados de tamanho menor, porém mais especializado. Esse processo permite ao modelo refinar seus parâmetros, ajustando-se aos padrões e ao vocabulário específicos de uma determinada aplicação (como análise de sentimentos, detecção de fraude ou geração de texto acadêmico). O fine-tuning aproveita o conhecimento geral do modelo base, tornando o processo de treinamento mais eficiente, tanto em termos de tempo quanto de

recursos computacionais, quando comparado ao treinamento a partir do zero.

Em termos de especialização, o fine-tuning (ajuste fino) surge como uma estratégia fundamental para adaptar modelos de linguagem às demandas de domínios ou tarefas específicas. Ao expor o modelo a um conjunto de dados especializado, ajustam-se seus parâmetros de modo a potencializar a acurácia das respostas em contextos determinados. Assim, o ajuste fino assegura que os modelos, previamente treinados de forma geral, tornem-se efetivos em cenários particulares. Conforme demonstrado por Howard & Ruder (2018), essa metodologia tem se mostrado particularmente eficaz em tarefas de processamento de linguagem natural, proporcionando ganhos significativos na acurácia e na capacidade de generalização dos modelos.

A matemática do fine-tuning envolve o ajuste e a reotimização do conjunto de parâmetros θ do modelo. Para uma tarefa T com um conjunto de dados anotados $D = \{(x_i, y_i)\}$, onde x_i representa a entrada e y_i o rótulo ou saída desejada, minimiza-se uma função de perda L(θ) específica para a tarefa. Assim, busca-se:

$$\theta^* = argmin\, L(\theta; D), \tag{4.6}$$

sujeito às restrições e regularizações adequadas. Como o modelo base já contém parâmetros que capturam representações linguísticas gerais, o processo de fine-tuning tende a convergir mais rapidamente e a exigir menos dados especializados para atingir desempenho satisfatório na tarefa de destino.

O fine-tuning em um modelo de LLM oferece inúmeras vantagens, especialmente quando se busca personalização e precisão em tarefas específicas. Esse processo permite ajustar um modelo pré-treinado com dados específicos para o contexto desejado, refinando sua capacidade de compreender e responder a demandas únicas. Como resultado, o fine-tuning pode melhorar substancialmente a qualidade das saídas do modelo, reduzindo inconsistências e alinhando suas respostas aos objetivos de aplicação.

A metodologia de fine-tuning envolve o ajuste dos parâmetros de um modelo, com o objetivo de adaptar-se a um novo conjunto de dados sem comprometer o conhecimento previamente adquirido durante o pré-treinamento. Durante esse processo, é comum optar por congelar as camadas iniciais – responsáveis pela extração de características gerais – e atualizar apenas as camadas superiores, ou mesmo ajustar todos os parâmetros com uma taxa de aprendizado reduzida, minimizando o risco de overfitting.

Para avaliação dos modelos, além da performance na aplicação em modelos de classificação de fraudes, calculou-se diversos indicadores de similaridade entre os dados reais e sintéticos. A

verificação das correlações foi realizada por meio de métricas como distância de Jensen-Shannon, divergência Kullback-Leibler (KL), *Mean Absolute Difference* MAD, índice de Similaridade e teste de Mantel. Essa série de indicadores, é a base comparativa de qualificação para validar se os dados sintéticos gerados são similares aos dados reais da classe minoritária.

Para calcular a JS, foi necessário medir a distância da distribuição de probabilidade de cada variável e calcular a média entre elas. A interpretação dos valores de JS é a seguinte:

- Distribuições são consideradas muito similares quando o JS é menor que 0.1.
- Distribuições são relativamente similares quando o JS está entre 0.1 e 0.5.
- Distribuições são consideravelmente diferentes quando o JS é maior que 0.5.

Valores mais altos de divergência KL indicam maior diferença entre as distribuições. Entende-se que KL é uma medida de diferença entre duas distribuições de probabilidade. Especificamente, ela quantifica o quanto uma distribuição de probabilidade (Q) diverge de uma distribuição de referência (P).

Em termos práticos, a divergência KL é usada em várias áreas, como aprendizado de máquina e teoria da informação, para medir a eficiência de um modelo probabilístico em relação a um modelo de referência. Uma divergência de Kullback-Leibler igual a 0 indica que as funções e distribuições P e Q são muito parecidas, enquanto uma divergência de 1 indica que se comportam de maneira diferente.

Para avaliar o quão próximo um *dataset* sintético (gerado a partir de um *dataset* original) está em termos de correlações, é necessário, basicamente, comparar as duas matrizes de correlação — a do conjunto original e a do conjunto sintético — e resumir essa comparação em um único número de avalie quanto das correlações estão próximas. Assim, é analisada as correlações das diferentes técnicas de geração de dados sintéticos comparando as matrizes de correlação aplicando a métrica de Diferença Média Absoluta (*Mean Absolute Difference*, MAD). Quanto menor este valor, mais similares são as correlações.

$$MAD = \frac{1}{n(n-1)/2} \sum_{i<j} | r_{ij}(O) - r_{ij}(S)| \tag{4.7}$$

Ro: matriz de correlação (n×n) do dataset original.

Rs: matriz de correlação (n×n) do dataset sintético

A seguir, foi aplicado a normalização baseada na fórmula abaixo, para obter um Índice de Similaridade de 0 a 1. Onde quanto mais próximo de 1, mais similar são as matrizes de correlação.

$$Similaridade = 1 - \frac{1}{\binom{n}{2}} \sum_{i<j} | r_{ij}(O) - r_{ij}(S)| \quad (4.8)$$

Em determinadas situações, quando o objetivo é verificar se as matrizes são estatisticamente diferentes, existem testes específicos, como os baseados na distribuição de Wishart ou nos desvios assintóticos da matriz de correlação. No entanto, esses testes produzem um valor-p em vez de um índice contínuo de similaridade. Assim, a pergunta a ser respondida é se há evidência de diferença significativa entre a matriz original e a matriz sintética.

O teste de Mantel (Derivado da Distribuição Wishart) avalia a igualdade das matrizes de covariância entre dois ou mais grupos. Neste contexto, consideramos dois grupos: dados originais e dados sintéticos. Se o p-valor for pequeno (por exemplo, < 0,05), rejeitamos a hipótese de que as matrizes são iguais, indicando evidências de diferença nas correlações. Caso contrário, não há evidência estatística significativa de diferença.

Cabe destacar, que muitos modelos de LLM são acessados por meio de plataformas e agentes de IA. Entende-se por agentes de Inteligência Artificial (IA) sistemas computacionais projetados para realizar tarefas que normalmente requerem inteligência humana. Eles percebem o ambiente ao seu redor, processam informações, tomam decisões e executam ações para alcançar objetivos específicos, utilizando algoritmos avançados e técnicas de aprendizado de máquina para analisar dados, reconhecer padrões e adaptar seu comportamento com base em novas informações. Em um contexto acadêmico, os agentes de IA são frequentemente estudados em termos de sua capacidade de raciocínio, aprendizado, planejamento e interação com humanos e outros agentes.

4.4. **Resultados dos Estudos**

4.4.1. Base de Dados

As informações empregadas neste estudo provêm de um conjunto de dados privado, contendo registros reais referentes a um modelo de detecção de fraudes em operações de crédito espúrio, disponibilizado por uma grande instituição financeira nacional. O referido *dataset* é composto por 168.596 observações e 37 atributos, os quais englobam variáveis cadastrais e comportamentais. A distribuição entre fraudes e não fraudes é de 0,94% e 99,06%, respectivamente, evidenciando um forte desbalanceamento na base de dados.

O Quadro 4.1 Quadro 1 apresenta, de forma consolidada para a confidencialidade do processo interno da instituição, as características das 25 variáveis do *dataset*.

Quadro 4.1 - Análise descritiva das variáveis do dataset

Grupo Temático das Variáveis	Quantidade de Variáveis	Descrição
Perfil do Cliente	6	Contempla atributos pessoais e características que ajudam a entender o contexto individual de cada pessoa.
Dados Financeiros	7	Envolve aspectos relacionados à capacidade financeira e movimentação econômica no tempo.
Histórico e Atividades Bancárias	6	Refere-se a eventos que indicam o uso de recursos e a relação do cliente com transações anteriores.
Relacionamento com Produtos e Serviços	5	Abrange interações relacionadas ao acesso e posse de soluções oferecidas no ambiente financeiro.
Interações e Uso de Tecnologia	5	Observa como o cliente se engaja com ferramentas digitais e acessa diferentes canais disponíveis.
Equipamentos e Vínculos	3	Considera elementos que demonstram ligações entre dispositivos e usuários em múltiplas situações.
Risco e Análises Cadastrais	5	Relaciona fatores que ajudam a compreender padrões de comportamento e medidas de segurança aplicáveis.

Fonte: Elaboração do autor.

Conforme apresentado no Quadro 4.1, a estrutura agrupa variáveis de um modelo de fraude de forma discreta e organizada, destacando diferentes aspectos do comportamento financeiro e interação do cliente com o sistema bancário. Entre os principais achados, observa-se uma diversidade de fatores que influenciam a análise de risco, desde características pessoais até padrões de transações e acessos digitais. Notavelmente, o grupo relacionado ao histórico bancário revela informações sobre movimentações e uso de recursos ao longo do tempo, o que pode indicar mudanças na relação do cliente com o banco. Além disso, fatores vinculados a equipamentos e acessos digitais demonstram a relevância da tecnologia na identificação de perfis e possíveis inconsistências. A abordagem permite uma visão estratégica para a avaliação de padrões, fortalecendo medidas de segurança e gestão de fraudes.

A Tabela 4.1 apresenta a análise descritiva e estatística das variáveis.

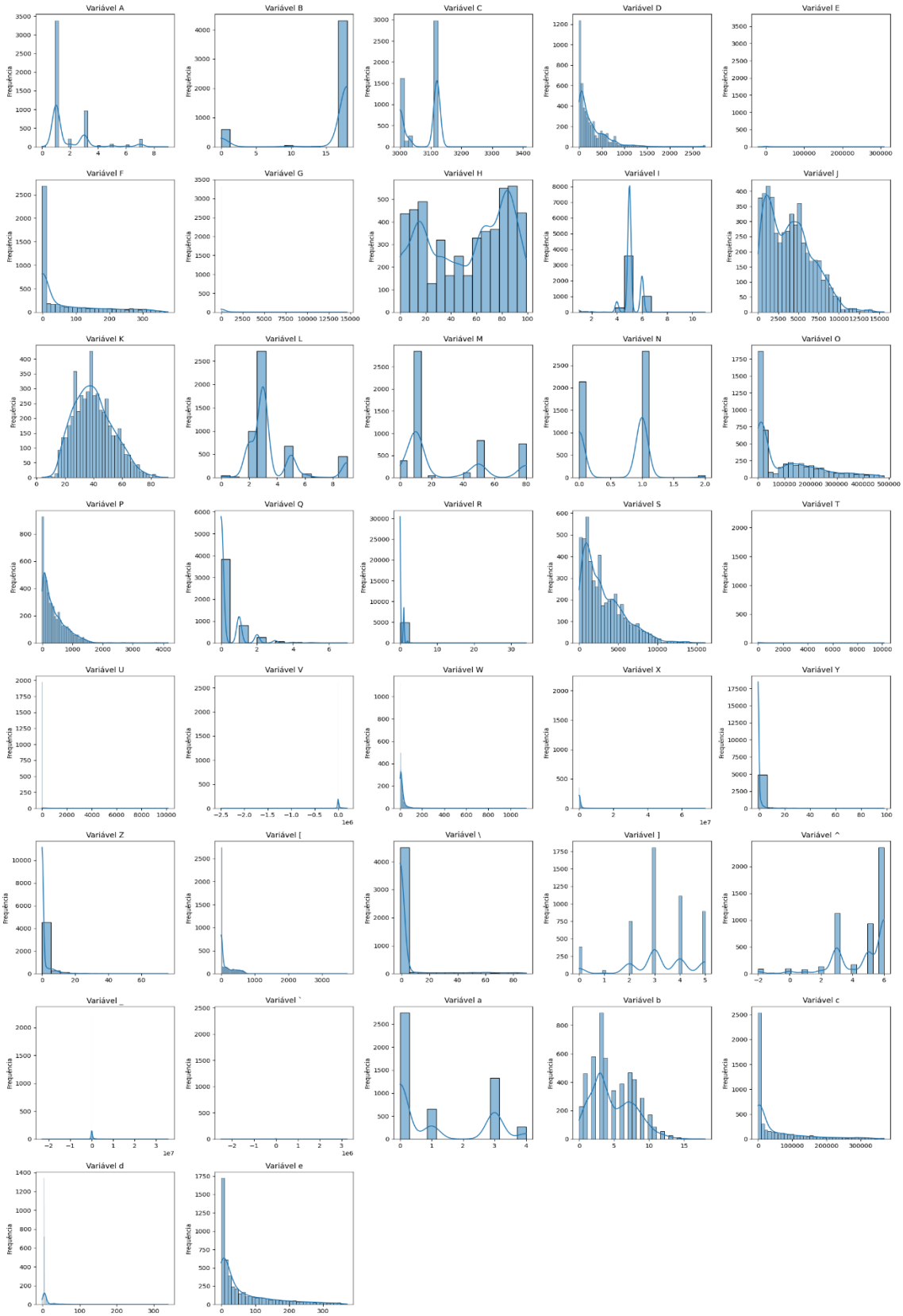
Tabela 4.1- Análise descritivas das variáveis – Base de dados

	mean	std	min	25%	50%	75%	max	IQR
1	1,75	1,61	0	1	1	3	99	2
2	15,28	6,35	0	18	18	18	18	0
3	3072,77	68,33	0	3000	3120	3120	3410	120
4	311,42	372,77	0	55	190	458	3953	403
5	3836,47	18513,95	-70398	0	0	0	1206459	0
6	65,15	99,12	-1	-1	-1	110	375	111
7	305,28	861,17	0	0	0,01	51,14	5547,75	51,14
8	51,38	32,48	0	19	61	83	99	64
9	5,02	0,76	1	5	5	5	11	0
10	3637,37	2638,04	-1	1369	3318	5345,25	17181	3976,25
11	38,50	13,50	0	28	37	47	124	19
12	3,64	2,05	0	3	3	3	9	0
13	27,68	27,58	0	10	10	50	80	40
14	0,55	0,52	0	0	1	1	2	1
15	84937,65	98282,30	0	10120	20680	153199	363241	143079
16	382,46	373,29	-1	91	268	568	5060	477
17	0,35	0,76	0	0	0	0	25	0
18	0,25	0,48	0	0	0	0	34	0
19	2849,47	2458,14	-1	949	2135	4307	17181	3358
20	467,87	1428,50	-1	1	2	8	10379	7
21	472,58	1436,34	-1	1	3	10	10379	9
22	6704,93	34563,90	-418761	0	88	7281	101814	7281
23	17,89	35,63	-1	2	9	21	3700	19
24	303309,04	433124,06	-10	5904,75	88775,5	444906	2178443	439001,3
25	-0,34	3,67	-1	-1	-1	-1	124	0
26	1,39	4,31	0	0	0	0	200	0
27	133,64	204,98	-1	-1	-1	224	3977	225
28	3,72	15,99	-1	-1	-1	-1	89	0
29	3,21	1,38	0	3	3	4	5	1
30	4,56	1,79	-2	3	5	6	6	3
31	183333,84	910227,52	-316690	0	9883	159618,75	137492463	159618,8
32	-21,06	53954,37	-418761	0	0	347	7912262	347
33	1,18	1,45	0	0	0	3	4	3
34	4,39	2,90	0	2	4	7	19	5
35	40928,81	63660,08	0	0	5355,14	61076,028	260529,8	61076,03
36	11,75	22,69	-1	4	6	8	375	4
37	65,13	86,34	-1	5	25	96	377	91

Fonte: Elaborado pelo autor

A distribuição com os histogramas das variáveis é apresentada na Figura 4.1. Os nomes das variáveis foram omitidos e substituídas por uma sequência de letras de acordo com as regras de sigilo estabelecida com a fornecedora das informações. A

Figura 4.1 - Histograma das nove primeiras variáveis da base de dados



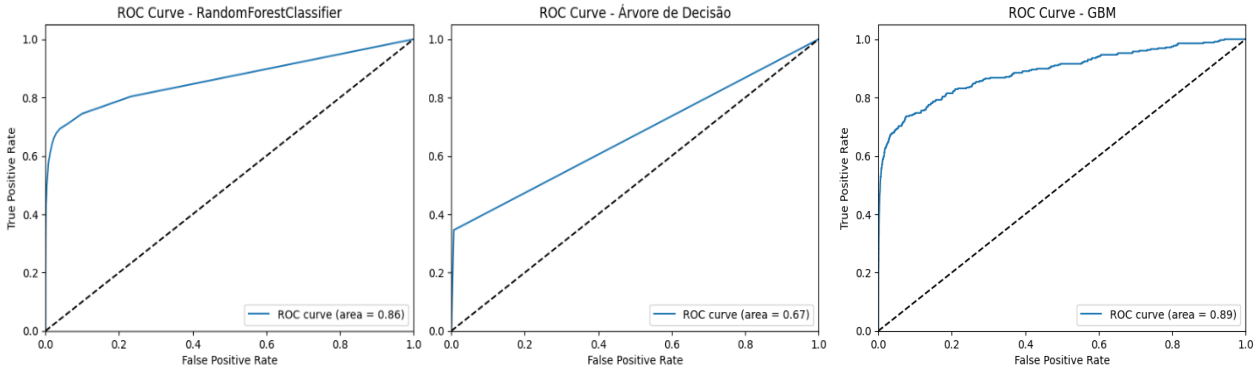
Fonte: Elaboração do autor

4.4.2. Modelo de classificação de fraudes – Base original

Para comparar a performance das diferentes técnicas de geração de dados sintéticos, foram realizados treinamentos com três modelos preditivos de classificação de fraudes e para cada modelo, apresentam-se indicadores de performance: acurácia, precisão, recall, F1 score e AUC.

Este procedimento utilizou-se como referência primária para comparar e avaliar o ganho, ou eventual ausência de melhoria, na performance dos modelos treinados com as diferentes técnicas de geração de dados sintéticos, implementados nas próximas etapas.

Figura 4.2 - Gráficos da Curva ROC dos modelos gerados – Dados desbalanceados



Fonte: Elaboração do autor.

Tabela 4.2 - Comparativo de Performance dos modelos - Sem balanceamento de classes

Modelo	Acurácia	Precisão	Recall	F1-Score	AUC
Decision Tree	0,988113	0,30434783	0,353933	0,3272727	0,67363466
Random Forest	0,993735	0,88785047	0,266854	0,4103672	0,63328815
Gradient Boosting	0,993024	0,66666667	0,292135	0,40625	0,64546587

Fonte: Elaboração do autor.

Conforme Tabela 4.2, verifica-se que o modelo Decision Tree apresentou o melhor AUC com valor de 0,67, enquanto o modelo Random Forest obteve a melhor precisão, atingindo 0,887. Estes indicadores de performance foram utilizados por serem os que minimizam os falsos negativos, objetivo destes modelos.

4.5.1. **Dados Sintéticos com SMOTE, GAN e CVAE**

Para comparar a performance dos modelos preditivos de classificação para prevenção de fraudes apresentados na seção anterior, foi realizado um balanceamento dos dados, de forma que a

classe minoritária atingisse uma proporção de 50% dos dados de treino, garantindo equilíbrio na aferição dos indicadores de performance dos modelos. Antes do balanceamento, o conjunto de dados passou por pré-tratamentos, como a remoção de outliers e a correção de valores nulos, preparando-o para a aplicação dos modelos.

4.4.1.1. Performance modelos e análise da qualidade dos dados sintéticos

Os resultados da geração de dados sintéticos, utilizando as técnicas apresentadas na seção de Metodologia – SMOTE, GAN e CVAE, são apresentados nas Tabela 4.3 a Tabela *4.5*.

Tabela 4.3 - Comparativo de Performance dos modelos - Balanceamento SMOTE

Modelo	Acurácia	Precisão	Recall	F1-Score	AUC
Decision Tree	0,97865896	0,15	0,34550562	0,20918367	0,6646898
Random Forest	0,99235853	0,54339623	0,40449438	0,46376812	0,7008474
Gradient Boosting	0,97948506	0,2076087	0,53651685	0,29937304	0,7598252

Fonte: Elaboração do autor.

Tabela 4.4 - Comparativo de Performance dos modelos - Balanceamento GAN

Modelo	Acurácia	Precisão	Recall	F1-Score	AUC
Decision Tree	0,987425	0,2767442	0,33427	0,302799	0,6635371
Random Forest	0,993483	0,8529412	0,244382	0,379913	0,6220175
Gradient Boosting	0,993299	0,7580645	0,264045	0,391667	0,6316754

Fonte: Elaboração do autor.

Tabela 4.5 - Comparativo de Performance dos modelos - Balanceamento VAE

Modelo	Acurácia	Precisão	Recall	F1-Score	AUC
Decision Tree	0,988732847	0,3180593	0,33146067	0,3246217	0,662803587
Random Forest	0,99357474	0,87254902	0,25	0,3886463	0,624849614
Gradient Boosting	0,993069898	0,67763158	0,28932584	0,4055118	0,64409608

Fonte: Elaboração do autor.

Observa-se que as técnicas utilizadas de geração de dados sintéticos, apesar de suas particularidades, melhoram a performance dos modelos de classificação, em comparação aos modelos com dados desbalanceados apresentados na Tabela 4.2. A técnica SMOTE obteve resultados de AUC de até 0,75, utilizando Gradiente Boosting enquanto o GAN e o CVAE alcançaram até 0,66 no Decision Tree.

Em comparação com os modelos de classificação treinados sem dados desbalanceados, os modelos Random Forest e Gradient Boosting com dados sintéticos para balanceamento de classes

tiveram evolução do AUC, principalmente utilizando a técnica SMOTE.

Nos dados gerados com SMOTE, observou-se aumento no Recall, ou seja, todos os modelos exibem um aumento significativo no recall, indicando melhor detecção da classe minoritária. Contudo, a precisão diminui, refletindo um aumento nos falsos positivos. Neste mesmo modelo, houve melhoria no AUC, especialmente no *Gradient Boosting*, o mesmo aumenta para 0,7598252, sugerindo melhor capacidade de classificação geral.

Os dados gerados com GAN e VAE apresentaram resultados semelhantes aos dados desbalanceados. As métricas não indicaram melhorias significativas em relação ao cenário original, mostrando apenas pequenas variações, tanto positivas quanto negativas, sem tendências claras de aprimoramento.

De modo geral, o balanceamento de classes usando SMOTE foi o mais eficaz para melhorar o desempenho dos modelos, especialmente em termos de Recall e AUC. O aumento no recall veio acompanhado de uma redução na precisão, o que mostra a existência um trade-off. Esse trade-off é comum em problemas de desbalanceamento de classes e deve ser considerado conforme o contexto do problema, principalmente nas estratégias de minimização de fraudes. Em detecção de fraudes, é geralmente mais crítico minimizar os falsos negativos, ou seja, evitar que transações fraudulentas passem despercebidas. Portanto, priorizar o aumento do recall, mesmo que isso venha com um aumento nos falsos positivos.

A seguir apresentamos a distribuição dos histogramas por cada método de geração de dados sintéticos utilizados:

Figura 4.3 - Distribuição Variáveis Dados Sintéticos vs Reais – Método SMOTE

Fonte: Elaborado pelo autor.

Figura 4.4 - Distribuição Variáveis Dados Sintéticos vs Reais – Método GAN

Fonte: Elaborado pelo autor.

Figura 4.5 - Distribuição Variáveis Dados Sintéticos vs Reais – Método VAE

Fonte: Elaborado pelo autor.

Podemos verificar com a análise das distribuições a consistência visual dos dados sintéticos gerados por três métodos diferentes: SMOTE, GAN e CVAE; ou seja, os dados gerados pelos esses métodos apresentam formas ou padrões semelhantes aos dados originais.

A Figura 4.3 mostra a distribuição das variáveis dos dados sintéticos gerados pelo método SMOTE em comparação com os dados reais. Observa-se que o SMOTE consegue replicar bem a distribuição dos dados originais, mantendo a forma geral dos histogramas. No entanto, pode haver algumas discrepâncias nas extremidades das distribuições, onde os dados sintéticos podem apresentar uma leve superestimação ou subestimação em relação aos dados reais.

Neste mesmo sentido, a Figura 4.4 apresenta a distribuição das variáveis dos dados sintéticos gerados pelo método GAN em comparação com os dados reais. Os histogramas gerados pelo GAN tendem a capturar melhor as nuances dos dados reais, resultando em uma correspondência mais próxima entre as distribuições sintéticas e reais. No entanto, a complexidade do modelo GAN pode introduzir algumas variações que não estão presentes nos dados originais, especialmente em variáveis com distribuições mais complexas.

Do mesmo modo, a Figura 4.5 ilustra a distribuição das variáveis dos dados sintéticos gerados pelo método VAE em comparação com os dados reais. O VAE também mostra uma boa capacidade de replicar a distribuição dos dados originais, com histogramas que se assemelham bastante aos dos dados reais. No entanto, assim como o SMOTE, pode haver pequenas discrepâncias nas extremidades das distribuições, onde os dados sintéticos podem não capturar perfeitamente a variabilidade dos dados reais.

A seguir, a tendência observada foi validada por meio da aplicação do indicador de Distância de Jensen-Shannon (JS), que quantifica a similaridade entre distribuições originais e sintéticas. A Distância de Jensen-Shannon é uma medida de divergência entre duas distribuições de probabilidade, sendo uma versão suavizada da divergência de Kullback-Leibler. No Quadro 4.2, são apresentados os valores de JS para três técnicas de geração de dados sintéticos: SMOTE, GAN e VAE.

Quadro 4.2 - Comparativo Distância de Jensen-Shannon (JS) das técnicas

	SMOTE	GAN	VAE
Distância de Jensen-Shannon	0.16982	0.32384	0.39743
Divergência de Kullback-Leibler	0.18834	0.78501	0.56996

Fonte: Elaboração do autor.

A análise dos resultados indica que o método SMOTE apresenta a menor distância de JS (0.16982), sugerindo que os dados sintéticos gerados por este método são relativamente similares aos dados originais. Embora não sejam considerados muito similares (JS < 0.1), a similaridade é aceitável. O método GAN apresenta uma distância de JS intermediária (0.32384), indicando que os dados sintéticos gerados por este método são relativamente similares aos dados originais, mas com uma maior divergência em comparação ao SMOTE. O método VAE apresenta a maior distância de JS (0.39743), sugerindo que os dados sintéticos gerados por este método são os menos similares aos dados originais entre os três métodos analisados. A divergência é significativa, mas ainda dentro da faixa de similaridade relativa.

No Quadro 4.3 a seguir, aplicamos a análise por divergência de Kullback-Leibler (KL). Para analisar-lo, é importante entender que ela sempre será não-negativa e será zero apenas quando as duas distribuições forem idênticas.

Quadro 4.3 - Comparativo divergência de Kullback-Leibler (KL) das técnicas

	SMOTE	GAN	VAE
Divergência de Kullback-Leibler	0.18834	0.78501	0.56996

Fonte: Elaboração do autor.

A análise dos resultados indica que o método SMOTE apresenta a menor divergência de KL (0.18834), sugerindo que os dados sintéticos gerados por este método são os mais próximos dos dados originais. Embora não sejam idênticos (KL = 0), a diferença é relativamente pequena.

Por outro lado, o método GAN apresenta a maior divergência de KL (0.78501), indicando que os dados sintéticos gerados por este método são os mais distantes dos dados originais entre os três métodos analisados. A magnitude da divergência é significativa, sugerindo que o GAN pode introduzir variações que não estão presentes nos dados originais.

Por fim, o método VAE apresenta uma divergência de KL intermediária (0.56996), revelando maior proximidade com os dados originais em comparação ao GAN, embora ainda com diferenças relevantes.

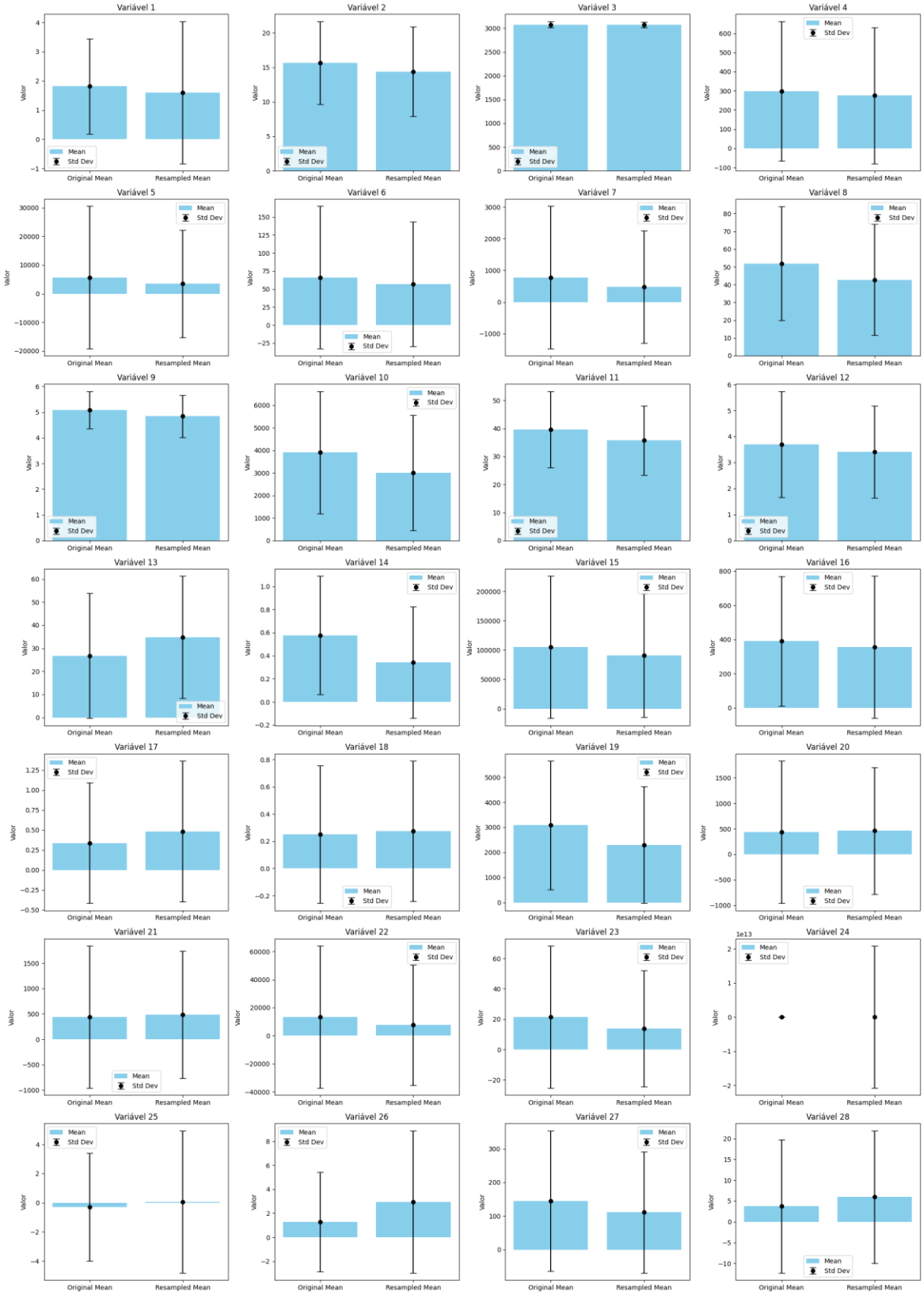
Como próximo indicador, a Figura 4.6, Figura 4.7 e Figura 4.8 apresentam uma comparação detalhada das estatísticas descritivas das variáveis originais com aquelas geradas pelos métodos sintéticos SMOTE, GAN e VAE, respectivamente. A análise visual dessas figuras indica que, em geral, os três métodos reproduzem com razoável precisão as médias e os desvios padrão das variáveis originais. Observa-se, contudo, que o método SMOTE (Figura 4.6) apresenta uma aderência

bastante satisfatória às médias originais, com pequenas variações nos desvios padrão em algumas variáveis, destacando sua robustez na geração de dados sintéticos para variáveis com diferentes escalas e distribuições.

Por outro lado, o modelo GAN (Figura 4.7) apresenta resultados igualmente consistentes, com diferenças pouco expressivas na maioria das variáveis. Contudo, percebe-se uma leve ampliação na variabilidade dos desvios padrão de algumas variáveis específicas, sugerindo que este método pode introduzir alguma dispersão adicional em certas características dos dados sintéticos gerados. Em geral, o método GAN se mostra adequado na preservação da estrutura estatística dos dados originais, ainda que apresente esta leve tendência ao aumento na variabilidade em comparação ao método SMOTE.

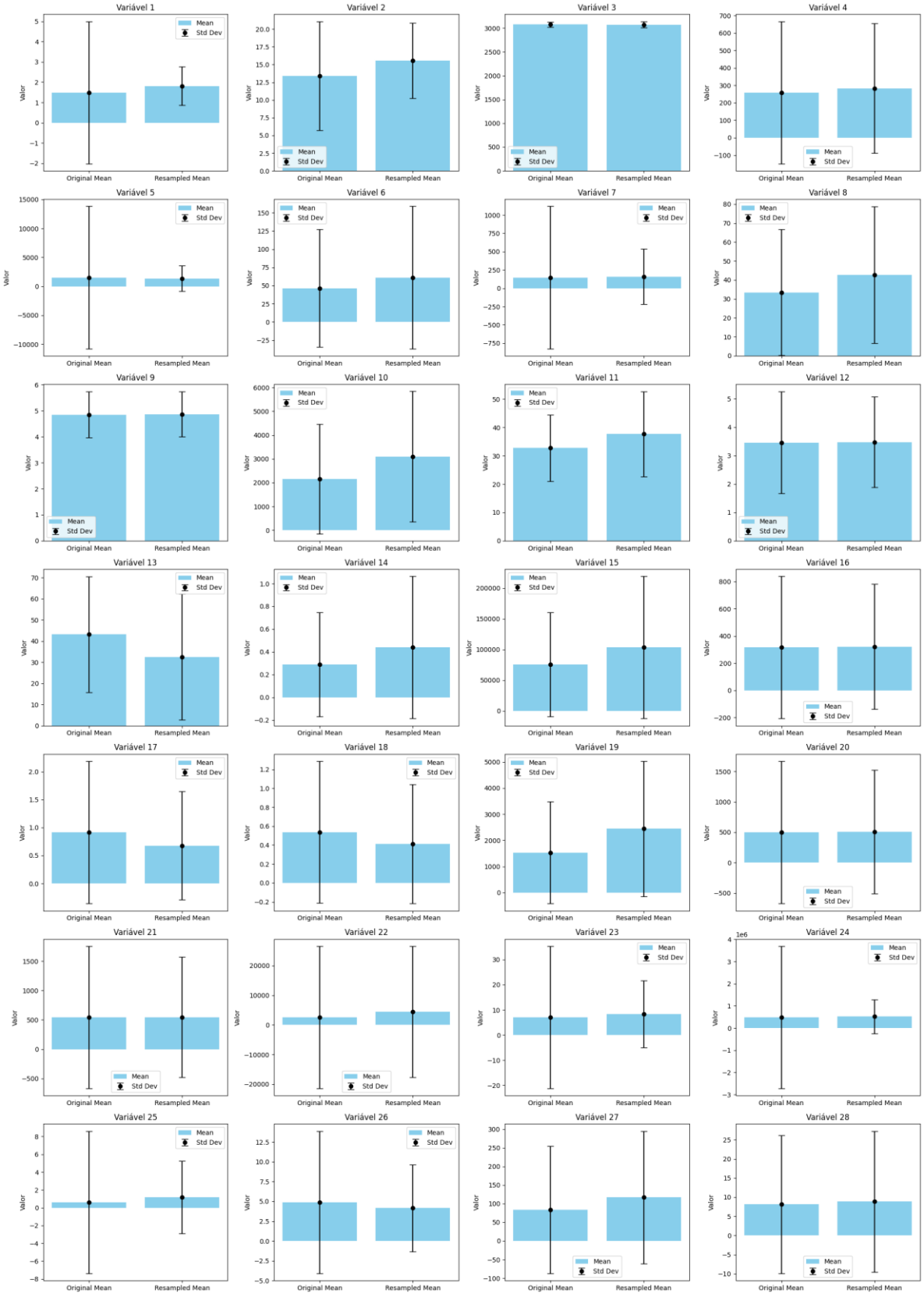
Finalmente, o método VAE, ilustrado na Figura 4.8, revela um padrão semelhante aos métodos anteriores na manutenção das médias originais, embora apresente diferenças ligeiramente mais pronunciadas nos desvios padrão em algumas variáveis específicas. Essas variações indicam que o modelo VAE pode, em determinados contextos, modificar a dispersão original dos dados, aspecto que deve ser considerado conforme a finalidade de uso dos dados sintéticos. Em suma, os três métodos demonstraram eficácia satisfatória na reprodução das estatísticas essenciais das variáveis originais, com pequenas diferenças que podem orientar a escolha específica de um método dependendo das necessidades analíticas e do contexto dos dados analisados.

Figura 4.6 - Análise descritiva dados gerados com originais - Smote



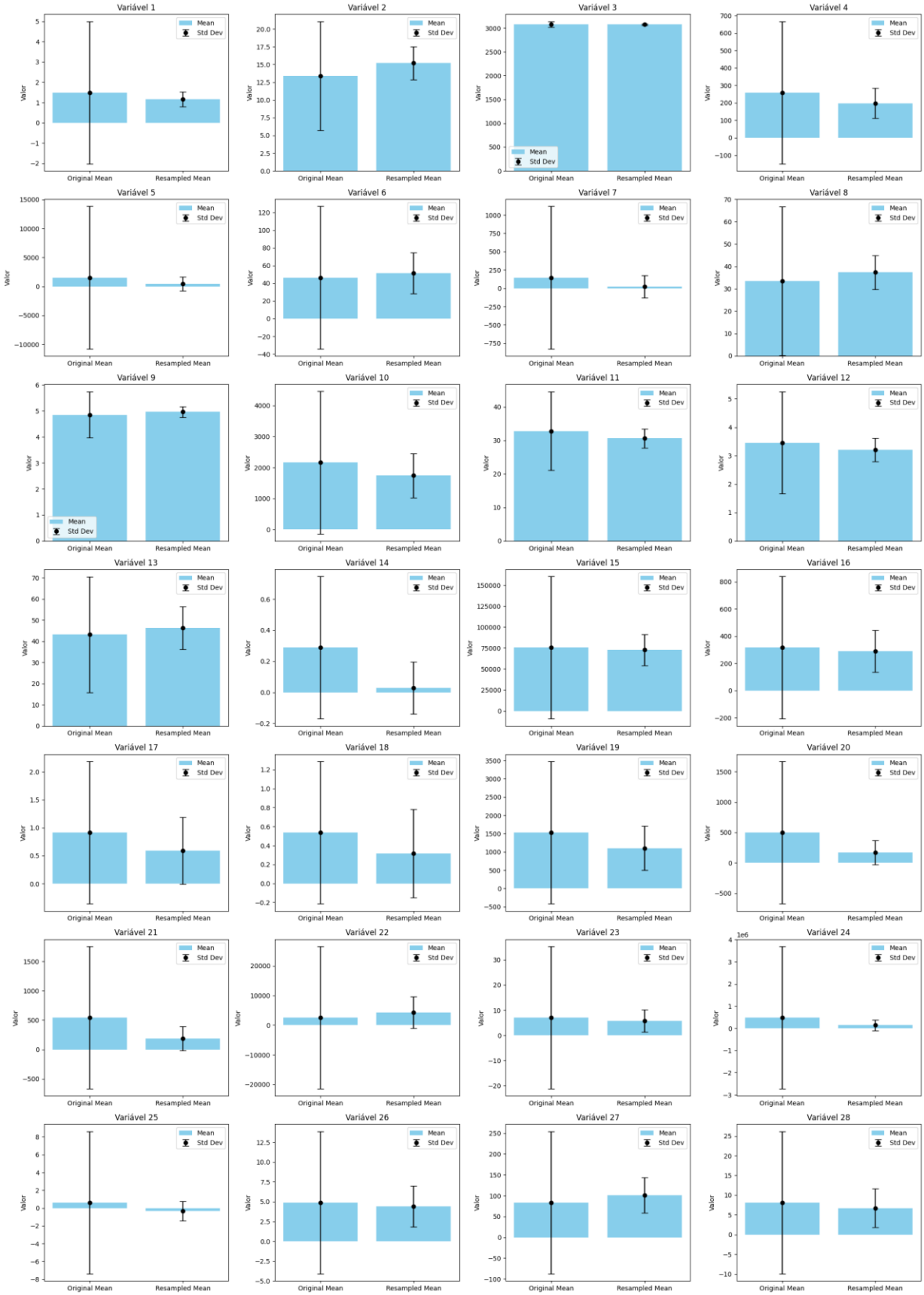
Fonte: Elaborado pelo autor.

Figura 4.7 – Análise descritiva dados gerados com originais - GAN



Fonte: Elaborado pelo autor.

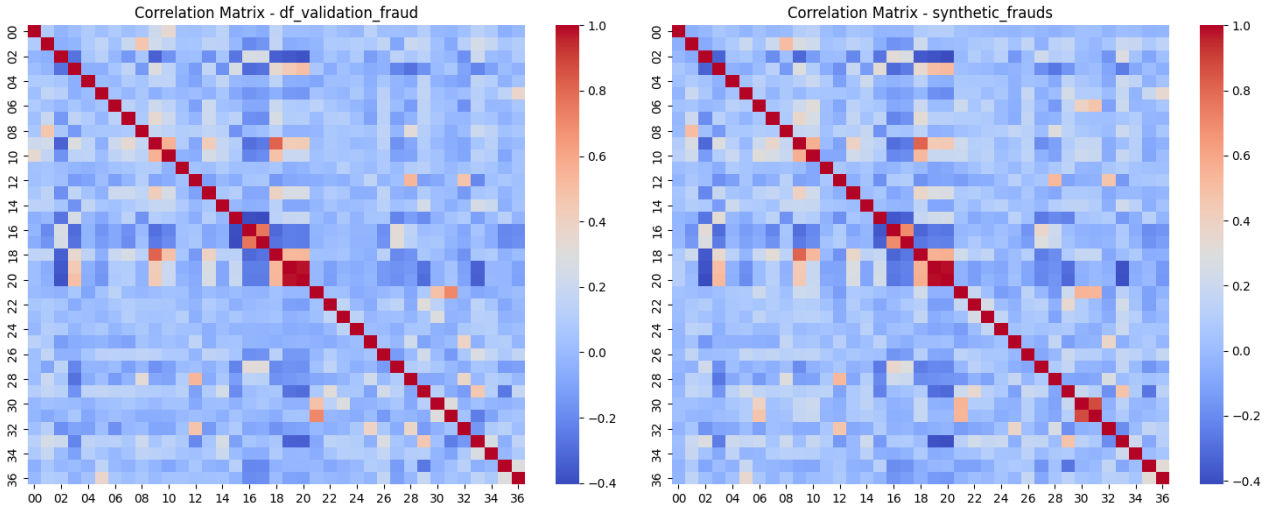
Figura 4.8 – Análise descritiva dados gerados com originais - VAE



Fonte: Elaborado pelo autor.

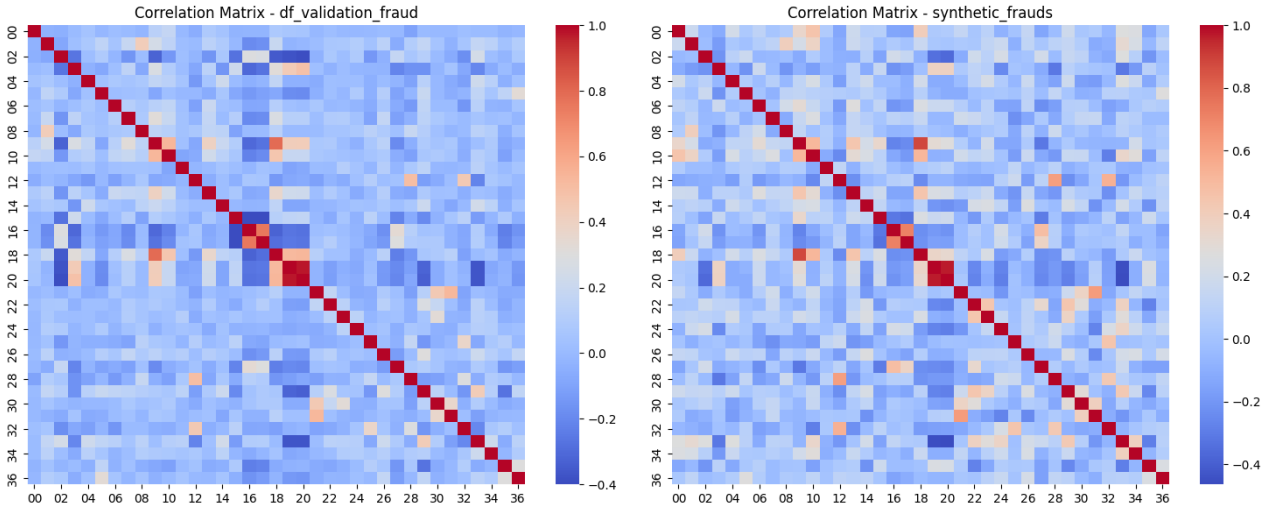
Por fim, as Figuras Figura *4.9*,Figura *4.10* eFigura *4.11*, apresentam os gráficos de correlação entre as variáveis reais e sintéticas, , respectivamente, por SMOTE, GAN e VAE, de modo a verificar se as correlações significativas são preservadas.

Figura 4.9 – Correlação entre as variáveis reais e geradas - SMOTE



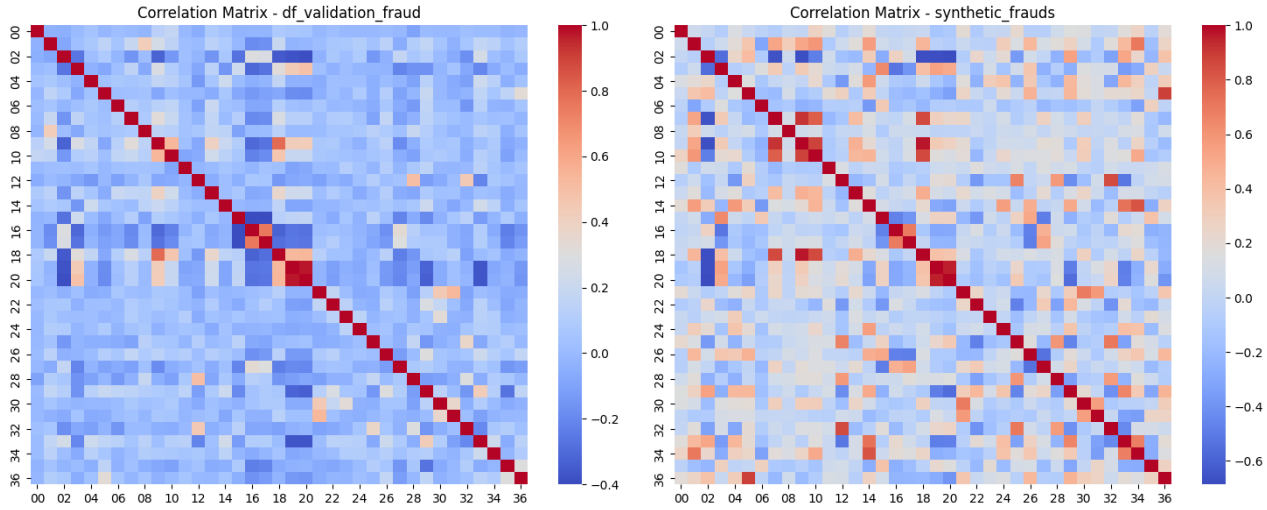
Fonte: Elaborado pelo autor.

Figura 4.10 – Correlação entre as variáveis reais e geradas - GAN



Fonte: Elaborado pelo autor.

Figura 4.11 – Correlação entre as variáveis reais e geradas - VAE



Fonte: Elaborado pelo autor.

Na Figura 4.9, referente ao método SMOTE, é possível observar que as estruturas gerais das correlações são mantidas de maneira consistente, embora algumas variações sutis nas intensidades das correlações possam ser identificadas. As associações mais fortes e significativas das variáveis originais tendem a permanecer bem representadas nos dados gerados, indicando que o SMOTE é eficaz na preservação das estruturas relacionais essenciais presentes no conjunto original.

No mesmo modo, a Figura 4.10, correspondente ao método GAN, exibe um padrão geral de preservação das correlações nas matrizes dos dados originais e sintéticos. Contudo, há uma leve variação das correlaciones entre variáveis. Isso sugere que o modelo GAN introduz pequenas alterações nas estruturas correlacionais dos dados sintéticos, apesar de ainda mantenha um bom nível de fidelidade em relação às correlações mais robustas e relevantes.

Por fim, a Figura 4.11, relativa ao método VAE, evidencia uma maior discrepância nas correlações em comparação aos dois métodos anteriores. Observa-se que algumas correlações, especialmente entre determinadas variáveis, apresentam diferenças notáveis na intensidade, o que pode implicar em mudanças mais pronunciadas nas relações lineares representadas. Portanto, o método VAE, embora capaz de gerar dados sintéticos coerentes, pode exigir cautela em contextos em que a preservação exata da estrutura correlacional seja essencial.

A seguir, é apresentado a aplicação os indicadores MAD, similaridade e Teste de Mantel para as bases geradas, em comparação as originais, vinculados a semelhanças por correlação similares.

Tabela 4.6 - Comparativo entre as semelhanças por correlação

Técnica	MAD	Similaridade	Mantel Test
SMOTE	0.0397	0.9603	0.9002 p-value: 0.0010
GAN	0.0902	0.9098	0.6679 p-value: 0.0010
VAE	0.5828	0.4172	-0.0599 p-value: 0.3430

Fonte: Elaborado pelo autor.

Como observamos na Tabela 4.6 - Comparativo entre as semelhanças por correlação, a análise dos resultados indica que, dentre as técnicas avaliadas, o método SMOTE obteve o melhor desempenho em termos de similaridade entre os dados sintéticos e originais. Isso é demonstrado pelo menor valor do erro absoluto médio (MAD = 0,0397) e pela alta similaridade correspondente (0,9603), destacando uma reprodução bastante fiel dos padrões originais. Além disso, o Teste de Mantel corroborou essa semelhança, apontando uma correlação alta (0,9002) e estatisticamente significativa (p-valor = 0,0010).

Por outro lado, a técnica GAN apresentou desempenho intermediário, com um valor de MAD um pouco mais elevado (0,0902), resultando em uma similaridade ainda aceitável (0,9098), embora inferior à do SMOTE. Esse desempenho também se reflete na correlação obtida pelo teste de Mantel (0,6679), que, embora estatisticamente significativa (p-valor = 0,0010), indica menor capacidade do modelo em preservar as relações originais entre as variáveis.

Já a técnica VAE demonstrou resultados consideravelmente inferiores às anteriores, com um valor de MAD substancialmente mais elevado (0,5828) e uma similaridade bastante reduzida (0,4172). Além disso, o Teste Mantel indicou ausência de correlação entre os dados sintéticos e os originais (correlação= -0,0599), sendo este resultado não significativo (p-valor = 0,3430). Tal desempenho sugere que, neste caso específico, os dados gerados por VAE não refletem adequadamente a estrutura original da base de dados analisada.

4.5.3.1. Conclusão dados sintéticos SMOTE, GAN e VAE

De forma geral, verifica-se que estes métodos estatísticos e de aprendizado de máquina tradicional avaliados são eficientes e adequados para aplicações em *data augmentation* na geração de dados sintéticos, sendo adequados para aplicações em *data augmentation* e *oversampling*, com o objetivo de aprimorar o treinamento de modelos de classificação.

No contexto da geração de dados sintéticos, o modelo foi modificado (adaptado) para incluir uma inovação com a adição de um termo de regularização baseado na correlação, visando aproximar a correlação dos dados sintéticos aos dados originais. Dessa forma, foi introduzida uma penalização de correlação (*correlation penalty*), dada a divergência da correlação dos dados sintéticos em relação aos originais. A função responsável pela criação e treinamento dos modelos também foi personalizada para permitir o controle sobre a perda de reconstrução e a perda K.

Uma segunda inovação foi a aplicação de uma camada customizada de reamostragem nos modelos GAN e VAE, permitindo a retropropagação correta do gradiente pela amostra estocástica. Verificou-se que os modelos que utilizam redes neurais (GAN e VAE) geram dados sintéticos com maior correlação entre as variáveis, em comparação aos métodos tradicionais.

A análise da Distância de Jensen-Shannon revela que o método SMOTE é o mais eficaz em gerar dados sintéticos que se assemelham aos dados originais, seguido pelo GAN e, por último, pelo VAE. Esses resultados são importantes para a escolha do método de geração de dados sintéticos em aplicações onde a similaridade com os dados originais é crucial.

De modo complementar, a análise da divergência de Kullback-Leibler revela que o método SMOTE é o mais eficaz em gerar dados sintéticos que preservam a distribuição dos dados originais, seguido pelo VAE e, por último, pelo GAN. Esses resultados são importantes para a escolha do método de geração de dados sintéticos em aplicações onde a similaridade com os dados originais é crucial.

No que se refere a análise estatística comparativa entre média e desvio padrão de cada variável gerada, todos os três métodos (SMOTE, GAN e VAE) demonstram uma capacidade razoável de gerar dados sintéticos que replicam as estatísticas descritivas dos dados originais. O GAN parece oferecer uma correspondência mais próxima, enquanto o SMOTE e o VAE apresentam pequenas discrepâncias nas extremidades das distribuições.

Em resumo, todos os três métodos (SMOTE, GAN e CVAE) demonstram uma capacidade razoável de gerar dados sintéticos que replicam a distribuição dos dados originais. O GAN parece oferecer uma correspondência mais próxima, enquanto o SMOTE e o CVAE apresentam pequenas

discrepâncias nas extremidades das distribuições. Cabe alertar, que se o objetivo é manter padrões de correlação, o método VAE mostrou discrepâncias mais pronunciadas nas correlações entre algumas variáveis, sugerindo que seu uso deve ser avaliado com cuidado em aplicações nas quais a preservação exata das relações lineares originais seja essencial.

Considerando as evidências apresentadas, conclui-se que a técnica SMOTE é a mais eficaz na preservação da similaridade estrutural dos dados originais, seguida pelo GAN, que apresentou resultados razoáveis. Por outro lado, a abordagem baseada em VAE mostrou-se inadequada, comprometendo significativamente a fidelidade dos dados gerados. Dessa forma, recomenda-se o uso preferencial do método SMOTE em aplicações que demandam alta fidelidade na reprodução das características estatísticas dos dados originais, especialmente em contextos de bases desbalanceadas.

4.5.2. **Dados Sintéticos com LLM**

No que se refere a geração de dados sintéticos com uso de IA Generativa na modalidade LLM (*Large Language Model*) foram realizados testes e cenários para que o modelo gere os dados sintéticos da classe minoritária apontada. Neste tipo de modelo, a engenharia de prompt tornou-se de suma importância para o atingimento dos objetivos.

A engenharia de prompt constitui um processo sistemático de formulação, desenvolvimento e aprimoramento de comandos ou instruções (prompts) direcionados a modelos de linguagem baseados em inteligência artificial generativa, como o GPT-4. Este campo emergente é crucial para maximizar a eficiência das interações entre usuários e sistemas de IA assegurando que as respostas produzidas sejam precisas, pertinentes e alinhadas com as expectativas e requisitos específicos dos usuários. A eficácia da engenharia de prompt reside na capacidade de elaborar solicitações que exploram de maneira otimizada as capacidades inerentes aos modelos, reduzindo ambiguidades e orientando a geração de conteúdo de forma controlada, eficiente e alinhada aos contextos de aplicação.

A qualidade das instruções do Prompt impacta diretamente o desempenho do modelo, pois prompts precisos e minuciosamente estruturados fornecem contexto, exemplos e orientações claras. Nesse cenário, a formulação de bons prompts torna-se indispensável para maximizar a pertinência dos resultados, bem como para alinhar as saídas do modelo aos objetivos específicos de cada aplicação. Conforme descrito na seção de metodologia,

utilizaremos nesta etapa do trabalho a aplicação do prompt no formato do *In-Context Learning* (ICL).

Além das plataformas próprias dos modelos (OpenAI, Gemini, etc) foram utilizadas as seguintes ferramentas e plataformas locais e versões webs: POE, Anything LLM, LM Studio, Novita, Playground LLM, Perplexity.AI.

Como limitação deste trabalho, não está no escopo a geração de dados sintéticos com dados em séries temporais.

4.5.3.1.1. Performance dos modelos e qualidade dados gerados

Nesta seção são apresentados os resultados quantitativos da geração de dados sintéticos aplicada via metodologia de engenharia de prompt para diversos modelos de LLM em suas plataformas.

Diante das limitações de tokens e desempenho de alguns modelos de linguagem (LLMs), a utilização do dataset completo tornou-se inviável em certos cenários de teste. Em resposta, empregou-se um subconjunto reduzido, composto exclusivamente por dados da classe minoritária, visando otimizar a alocação de tokens, especialmente em aplicações locais onde grandes datasets geravam erros.

Assim, foram realizados testes com o *dataset* completo e com dados minoritários para a geração de dados sintéticos. Adicionalmente, o *dataset* completo foi utilizado para verificar a capacidade do modelo GenAI em discernir padrões entre transações fraudulentas e não fraudulentas. Constatou-se que a performance da geração de conteúdo depende do tamanho do modelo (parâmetros), limites de tokens no input/output, presença de mecanismos de *reasoning*, ambiente de execução (GPU) e plataformas de ambiente (Docker, etc.).

Para otimizar o desempenho, foram testadas variações de prompts, aplicando técnicas de ICL (In-Context Learning) e de engenharia de prompts baseadas em IA, conforme detalhado na seção anterior. Neste teste foram avaliados os seguintes modelos:

- OpenAI GPT: 4o, o1, o3-mini-high, ADA.
- Claude: 3.5 Haiku, 3.5 Sonnet, 3 Opus.
- Gemini: 1.5 Pro, 2.0 Advanced, 1.5 Flash, 2.0 Flash.
- Llama: 3.2 8B, 3.1-70B, 3.1 405B
- Qwen: 2.5 Math 7B
- DeepSeek: V3, R1

- Gemma: 2 7B
- Mistral: Large2
- Grok: 2

Assim, a aplicação de LLMs via prompts para a geração de dados sintéticos envolveu diversas abordagens, com foco em modelos de inteligência artificial e as respostas dos modelos indicavam métodos estatísticos. O output dos modelos foram apresentados em arquivos .csv, quando a plataforma permitia, ou no formato texto no próprio chat do modelo.

De forma geral, o modelo GPT (em versões, como GPT-4o e GPT-o1), tentou aplicar técnicas como SMOTE e VAE, mas encontrou erros de execução ao tentar rodar. Esse modelo, em conjunto com o3, utilizaram abordagens estatísticas baseadas nas distribuições observadas com reamostragem de variáveis. Para variáveis numéricas, a geração de dados sintéticos buscou preservar a distribuição original, reamostrando valores com base na média e desvio padrão, além de manter as faixas mínima e máxima e o comportamento estatístico geral. Para variáveis categóricas, a amostragem foi realizada de forma a preservar a frequência das categorias existentes.

O modelo Claude em suas versões 3.5 Haiku, 3.5 Sonnet e 3 Opus, também tentou utilizar o SMOTE sem sucesso. Esses modelos aplicaram técnicas de amostragem aleatória e geração de valores respeitando as distribuições originais das variáveis. No entanto, enfrentaram dificuldades devido ao tamanho da base de dados e limitações na execução dos comandos. Embora a metodologia declarada pelo 3.5 Haiku fosse o SMOTE, a confirmação dessa aplicação não foi possível. Após novos requisitos, foi possível gerar novos dados. A técnica usada para geração de dados sintéticos foi de amostragem aleatória, de cada variável, que conste dentro do intervalo mínimo e máximo da variável (calculado primeiramente).

Os modelos Gemini, em suas versões 1.5 Pro, 2.0 Advanced e 1.5 Flash, também foram testados. Esses modelos utilizaram técnicas de interpolação aleatória e amostragem com reposição para gerar dados sintéticos. Apesar de algumas limitações na execução de código Python e na geração de arquivos para download, os modelos conseguiram gerar dados sintéticos seguindo as distribuições originais das variáveis. A técnica alegada foi uma combinação de amostragem aleatória com reposição (para variáveis categóricas) e interpolação linear (para variáveis numéricas), aplicada individualmente a cada variável

Os modelos Llama, em suas versões 3.2, 3.1 405G e 3.1 70B, foi executado no LLM Studio e na interface POE. Esses modelos aplicaram técnicas como RAG (Retrieved 3 relevant citations for

user query) e amostragem por sobrevivência, mas enfrentaram dificuldades devido ao tamanho da base de dados e limitações na geração de arquivos para download. A geração de dados sintéticos foi realizada utilizando técnicas estatísticas e preservação de correlações entre as variáveis. O modelo sugere que utiliza o SMOTE, porém não é possível confirmar. Esta plataforma gerou dados impressos no corpo do chat.

Por fim, outros modelos como DeepSeek, Qwen2.5, Gemma, Mistral e Grok, também foram testados para a geração de dados sintéticos. Esses modelos aplicaram técnicas variadas, como amostragem estatística e interpolação linear. No entanto, enfrentaram limitações das ferramentas, principalmente por serem modelos menores e de aplicação local, na execução de código, geração de arquivos e preservação das distribuições originais das variáveis. A geração de dados sintéticos foi realizada utilizando uma técnica baseada em amostragem estatística e preservação de correlações entre as variáveis. Destaca-se para o modelo da Qwen 2.5 Math, onde a técnica utilizada foi amostragem aleatória com reposição incluindo variações de acordo com suas estatísticas.

Após a aplicação e testes nos diferentes modelos, foram realizados testes de similaridade entre os dados sintéticos gerados e os dados originais, e suas correlações. Desta forma, a seguir, apresentamos o quadro com as estatísticas por modelos no Quadro 4.4 a seguir.

Quadro 4.4 - Modelos de LLM testados e sua performance

Modelo	**Versão**	**Distância de Jensen-Shannon**	**Divergência Kullback-Leibler**	**Mean Absolute Difference MAD**	**Índice de Similaridade**	**Mantel Test Correlação observada**
GPT	4o	0.04260	0.03082	0.0279	0.9721	0.4635 p-value: 0.0010
GPT	o1	0.39841	0.74062	0.2750	0.7250	0.0512 p-value: 0.3900
GPT	o3 mini-high	0.38265	0.4899	0.2606	0.7394	0.0429 p-value: 0.5160
GPT	ADA[3]	0.302876	0.43987	0.1240	0.8760	0.0835 p-value: 0.3240
Claude	3.5 Haiku	0.13139	0.17972	0.1241	0.8759	0.4170

[3] Advanced Data Analysis is a feature within ChatGPT's GPT-4 that allows users to upload data directly to ChatGPT to write and test code.

						p-value: 0.0010
Claude	3.5 Sonnet	0.13172	0.16884	0.1242	0.8758	0.4170 p-value: 0.0010
Claude	3 Opus	0.31459	0.39253	0.1753	0.8247	0.1535 p-value: 0.0130
Gemini	1.5 Pro	0.04251	0.04305	0.1235	0.8765	0.0940 p-value: 0.2210
Gemini	2.0 Advanced	0.24145	0.38300	0.1635	0.8365	0.2979 p-value: 0.0010
Llama	3.1- 70B	0.179443	0.18779	0.1228	0.8772	0.3821 p-value: 0.0010
Llama	3.1 – 405 B	0.32278	0.23006	0.2371	0.7629	-0.0859 p-value: 0.1900
DeepSeek	Math / V3	0.14844	0.20160	0.1131	0.8869	0.3465 p-value: 0.0010
DeepSeek	R1	0.30335	0.33086	0.2028	0.7972	0.1535 p-value: 0.0130
Mistral	Large 2	0.14118	0.19123	0.1191	0.8809	0.3507 p-value: 0.0010
Qwen2.5	Math	0.08733	0.11283	0.0716	0.9284	0.4364 p-value: 0.0010
Gemini	1.5 Flash	Modelo apresentou a lógica, porém não gerou saída com dados gerados.				
Gemini	2.0 Flash	Modelo apresentou a lógica, porém não gerou saída com dados gerados.				
Gemma	7b	Modelo apresentou a lógica, porém não gerou saída com dados gerados.				
Grok	2	Modelo apresentou a lógica, porém não gerou saída com dados gerados.				

Fonte: elaboração própria

Os modelos GPT 4o, GPT ADA, Claude 3 Opus e Gemini 1.5 PRO enfrentaram dificuldades em utilizar o SMOTE para gerar dados sintéticos da classe minoritária. Como alternativa, todos os modelos recorreram a distribuições estatísticas para criar os registros. Eles utilizaram estatísticas descritivas e frequências de cada variável na classe minoritária, realizando reamostragens seguindo a média e o desvio padrão (distribuição normal). Para variáveis categóricas, os valores foram gerados por amostragem aleatória com reposição, garantindo a manutenção das proporções

originais.

Além disso, cada modelo adotou abordagens específicas para gerar os dados sintéticos. O GPT 4o e o GPT ADA utilizaram distribuições normais (gaussianas) com média e desvio padrão iguais aos da variável original. O modelo GPT ADA, após tentativa de criação de dados com SMOTE, ele realizou a criação por meio de amostragem aleatória com reposição, para variáveis numéricas e utilizou a moda para as variáveis categóricas. O Claude 3 Opus criou uma função para gerar valores aleatórios a partir da distribuição de frequência da variável, enquanto o Gemini 1.5 PRO utilizou uma técnica de interpolação aleatória entre as amostras existentes.

Os modelos Claude intentam aplicar técnicas mais robustas; contudo, a plataforma de hospedagem não conseguiu gerar mais de 500 casos completos. A execução foi interrompida prematuramente, possivelmente devido a limitações de processamento ou tokens, e o modelo sugeria a aplicação do código em Python pelo usuário. Adicionalmente, a plataforma carece (no momento em que a pesquisa foi realizada) da funcionalidade de download em formato CSV.

A comparação dos modelos testados revela desempenhos significativamente variados em relação aos indicadores avaliados, destacando diferenças importantes em termos de proximidade aos dados originais e correlação observada pelo teste de Mantel.

Entre os modelos analisados, o GPT versão 4o apresentou uma performance notoriamente superior, evidenciada pelos menores valores da distância de Jensen-Shannon (0,04260), divergência Kullback-Leibler (0,03082) e média das diferenças absolutas (MAD de 0,0279), além de um elevado índice de similaridade (0,9721) e uma correlação Mantel significativa (0,4635, p-value = 0,0010).

O modelo Qwen2.5 Math também demonstrou desempenho robusto, com baixa distância de Jensen-Shannon (0,08733), pequena divergência Kullback-Leibler (0,11283) e o menor MAD (0,0716) entre os demais modelos, bem como uma forte correlação Mantel observada (0,4364, p-value = 0,0010).

Os modelos Claude 3.5 Haiku e Claude 3.5 Sonnet exibiram desempenhos semelhantes, com distâncias Jensen-Shannon e divergências Kullback-Leibler próximas (aproximadamente 0,131 e 0,17 respectivamente), MAD quase idêntico (em torno de 0,124) e correlação Mantel significativa (0,4170, p-value = 0,0010).

Por sua vez, o modelo Gemini 1.5 Pro, apesar de apresentar bons valores de distância Jensen-Shannon (0,04251) e divergência Kullback-Leibler (0,04305), exibiu correlação Mantel fraca e não significativa (0,0940, p-value = 0,2210).

Em contraste, modelos como GPT o1 e GPT o3 mini-high apresentaram desempenho

relativamente fraco em todos os indicadores avaliados, com altas distâncias Jensen-Shannon (aproximadamente 0,39) e divergências Kullback-Leibler elevadas, acompanhadas por baixas correlações Mantel e p-valores não significativos.

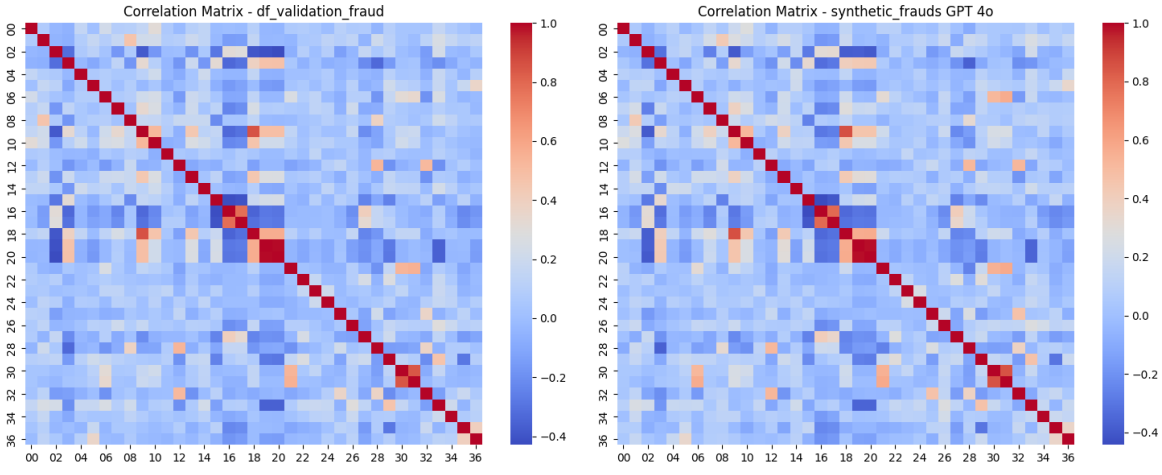
Cabe destacar ainda que modelos como Gemini 1.5 Flash, Gemini 2.0 Flash, Gemma 7b e Grok 2 não forneceram resultados quantitativos, tendo apenas demonstrado lógica, impossibilitando uma avaliação numérica precisa.

Em resumo, modelos como GPT 4o, Qwen2.5 Math e Claude 3.5 (Haiku e Sonnet) demonstraram superioridade significativa nos indicadores avaliados, enquanto os demais modelos exibiram performances inferiores ou inconsistentes.

A Figura 4.12, de modo complementar ao Quadro 4.4, apresentada um comparativo das matrizes de correlações entre os modelos com dados reais e os dados sintéticos gerados para o exemplo que utilizou o modelo GPT 4o da OpenAI. Verifica-se, desta forma, que após a aplicação da engenharia de prompt de geração de dados sintéticos as variáveis dos dados sintéticos possem distribuição semelhante as variáveis reais utilizando muitos dos modelos aplicados.

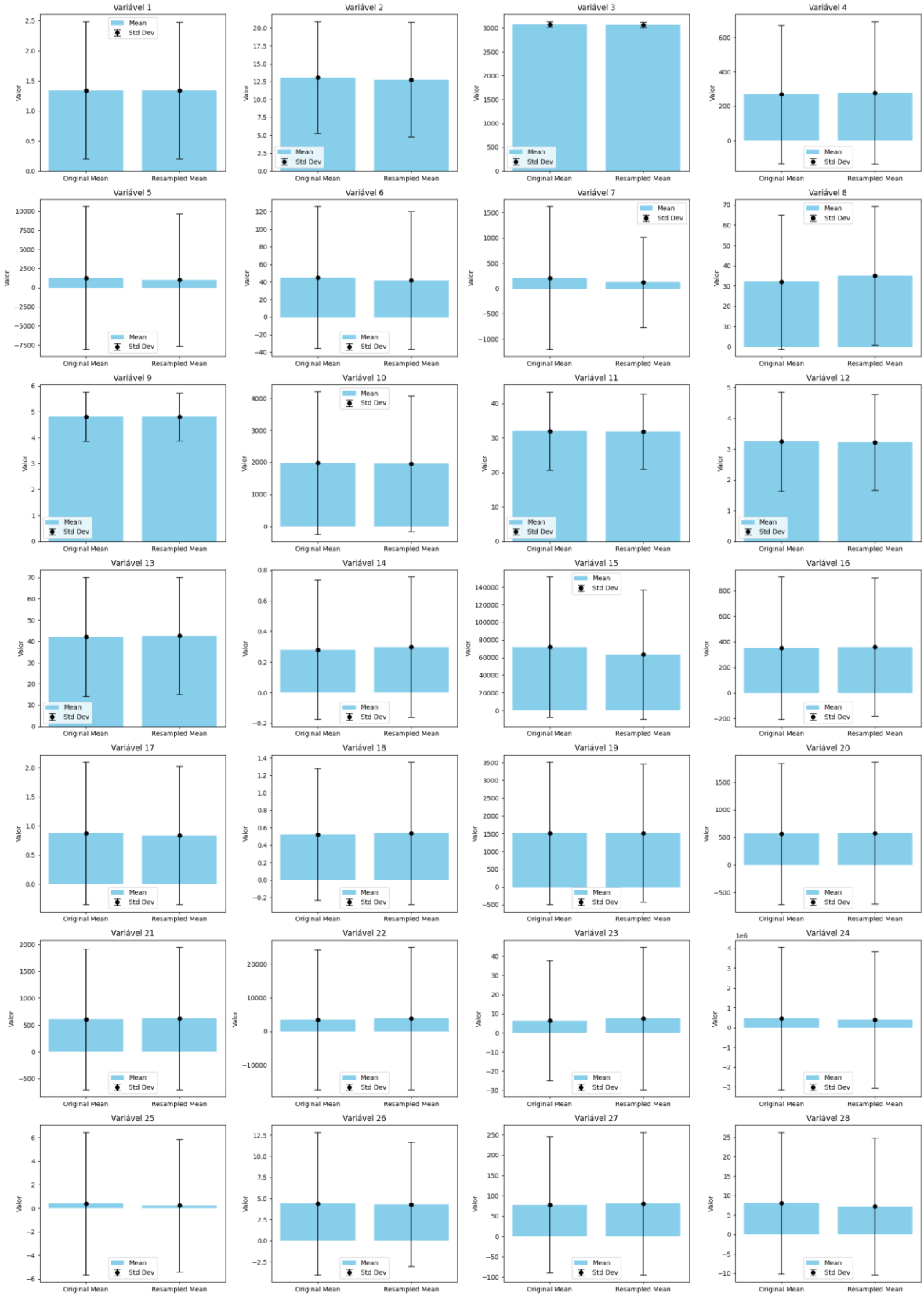
Do mesmo modo, as FigurasFigura *4.13* e Figura *4.14* apresentam os gráficos de distribuição (histograma) das variáveis reais versus os dados sintéticos gerados, e as estatísticas descritivas (Box Plot) para o modelo GPT 4o, respectivamente. Os gráficos para os demais modelos também se encontram no Anexo 1.

Figura 4.12 – Correlação entre as variáveis reais e geradas – LLM GPT4o



Fonte: elaboração própria

Figura 4.13– Estatísticas descritivas – Box Plot – Modelo GPT 4o



Fonte: elaboração própria

Figura 4.14 – Distribuição das variáveis - Modelo GPT 4o

Fonte: elaboração própria

4.5.3.2. Performance modelos LLM nos modelos preditivos de fraudes

Nesta seção, foi verificado se os novos dados sintéticos melhoram a performance dos modelos preditivos aplicados na sessão 4.5.1. Foi desenvolvido um modelo preditivo de classificação, cujo desempenho foi mensurado por meio das métricas de precisão, recall e AUC.

Para este teste, com base nos critérios apresentados no Quadro 4.4- principalmente menor Distância de Jensen-Shannon, menor divergência Kullback-Leibler, menor *Mean Absolute Difference* (MAD), maior Índice de similaridade, e maior correlação observada com significância estatística (p-value ≤ 0.05), foram selecionados os cinco melhores em performance: GPT 4o; Gemini 1.5 Pro; Qwen2.5 Math; Claude 3.5 Haiku e Claude 3.5 Sonnet.

O resultado da aplicação dos modelos preditivos é apresentado no Quadro 4.5. A comparação desses valores demonstra claramente os benefícios de aplicar técnicas de geração de dados sintéticos para *oversampling* em bases desbalanceadas. Embora os modelos da Claude (Sonnet e Haiku) geraram poucos dados sintéticos, seus resultados continuam sendo vantajosos comparada a base original desbalanceada. Destacam-se ainda que os modelos 4o e Gemini 1.5 Pro apresentaram resultados excelentes comparados a base original.

Quadro 4.5 - Comparativo dos modelos utilizando dados sintéticos gerados por LLM

Modelo e Métrica	**Base Original**	**OpenIA 4o**	**Gemini 1.5 Pro**	**Qwen2.5 Math**	**Claude Haiku**	**Claude Sonnet**
LR - Precision	1,00000	0,96203	0,70430	0,77778	0,90000	0,85714
LR - Recall	0,03125	0,85876	0,74011	0,14894	0,17308	0,27907
LR -AUC	0,51563	0,90938	0,68672	0,56748	0,58297	0,63254
DT -Precision	0,66667	0,93296	0,91124	0,77778	0,88889	0,85294
DT - Recall	0,62500	0,94350	0,87006	0,74468	0,61538	0,67442
DT -AUC	0,77802	0,93175	0,88503	0,83738	0,79341	0,81973
GBM -Precision	0,86364	0,98235	0,94857	0,85714	0,94444	0,97059
GBM -Recall	0,59375	0,94350	0,93785	0,76596	0,65385	0,76744
GBM -AUC	0,78653	0,96175	0,93893	0,86200	0,81978	0,88022

Fonte: Elaborado pelo autor.

Para o modelo de regressão logística (LR), a base original (ou seja sem *oversampling*) apresentou alta precisão (1,00), mas recall extremamente baixo (0,03125) e AUC próximo à

aleatoriedade (0,51563), indicando dificuldades significativas na identificação da classe minoritária. Ao aplicar técnicas de geração sintética, destacaram-se especialmente o modelo OpenAI 4o e Gemini 1.5 Pro, com melhoras expressivas no recall (0,85876 e 0,74011, respectivamente) e AUC (0,90938 e 0,68672), sugerindo eficácia na redução do impacto negativo do desbalanceamento. Qwen2.5, Claude Haiku e Claude Sonnet tiveram desempenho inferior, especialmente em recall e AUC, indicando limitações em gerar dados sintéticos efetivos para o modelo LR.

No caso do modelo de árvore de decisão (DT), os resultados evidenciam um ganho consistente em todas as técnicas de *oversampling,* quando comparadas à base original. Destacaram-se novamente OpenAI 4o e Gemini 1.5 Pro que apresentaram elevadas taxas de precisão (0,93296 e 0,91124) e recall (0,94350 e 0,87006), resultando em AUC muito superiores (0,93175 e 0,88503) em relação à base original (AUC=0,77802). Embora as outras técnicas (Qwen2.5, Claude Haiku e Claude Sonnet) também tenham gerado ganhos expressivos, ficaram atrás em termos de recall e AUC em comparação com os dois primeiros métodos mencionados.

Por fim, para o modelo *Gradient Boosting* (GBM), todas as técnicas avaliadas apresentaram melhorias substanciais em comparação com à base original, especialmente no recall e no AUC. Destacam-se novamente os métodos OpenAI 4o e Gemini 1.5 Pro, ambos com altos valores de recall (0,94350 e 0,93785) e AUC (0,96175 e 0,93893). O modelo Claude Sonnet também apresentou resultados robustos com alta precisão (0,97059) e um bom equilíbrio em recall (0,76744) e AUC (0,88022). Qwen2.5 e Claude Haiku, embora superiores à base original, ficaram atrás em métricas-chave como recall e AUC, indicando limitações relativas nestes métodos.

4.5.3.3. Conclusão dados sintéticos LLM

Verifica-se que, seguindo a metodologia proposta, com a aplicação adequada de prompt e do pré-processamento de dados, é possível criar bons dados sintéticos. A solicitação de geração de dados sintéticos a modelos LLM, por meio de engenharia de prompt mostra-se viável, porém possui limitações associadas à quantidade de tokens, às restrições das plataformas, às características dos dados utilizados e, principalmente, ao tipo de modelo aplicado.

Ao aplicar a metodologia, verificou-se que os modelos GPT 4o, GPT ADA, Claude 3 Opus e Gemini 1.5 Pro não obtiveram sucesso com a técnica SMOTE para geração sintética da classe minoritária e optaram por métodos estatísticos alternativos. GPT 4o e GPT ADA adotaram

distribuições normais (gaussianas) baseadas nas médias e desvios padrões originais. Claude 3 Opus utilizou uma função para gerar valores conforme as frequências originais das variáveis, enquanto Gemini 1.5 Pro aplicou interpolação aleatória entre amostras já existentes. Variáveis categóricas foram tratadas por amostragem aleatória com reposição para manter as proporções originais.

Foi realizada uma avaliação comparativa entre diferentes modelos de linguagem (LLMs), analisando suas performances por meio de indicadores estatísticos como Distância de Jensen-Shannon, Divergência de Kullback-Leibler, Mean Absolute Difference (MAD), Índice de Similaridade e Correlação obtida no teste de Mantel. Os resultados destacaram o GPT-4o e o Gemini 1.5 Pro como modelos de alto desempenho, apresentando valores particularmente baixos em divergências e elevados índices de similaridade, com destaque para o GPT-4o (Distância Jensen-Shannon = 0,04260; Índice de Similaridade = 0,9721), que também obteve correlação significativa ($r = 0,4635$; $p = 0,0010$). Em contrapartida, os modelos como GPT o1 e GPT o3 Mini-High tiveram performance significativamente inferior, evidenciada pelos maiores valores de divergência e menor similaridade, além de correlações não significativas ($p > 0,05$).

Alguns modelos como Claude 3.5 Haiku, Claude 3.5 Sonnet, Llama 3.1-70B, DeepSeek Math V3, Mistral Large 2 e Qwen2.5 Math também demonstraram boas performances, com valores satisfatórios nos indicadores e correlações significativas ($p = 0,0010$). Outros, como Gemini 1.5 Flash, Gemini 2.0 Flash, Gemma 7b e Grok 2, apesar de demonstrarem compreensão lógica das solicitações, não conseguiram produzir resultados efetivos com dados gerados. Esses achados apontam diferenças relevantes na capacidade preditiva e na geração de respostas coerentes entre os modelos, destacando a importância da escolha adequada do LLM com base em critérios quantitativos claros para diferentes aplicações práticas.

A capacidade de processamento de tokens em modelos de linguagem (LLMs) desempenha papel crucial na análise e geração de dados sintéticos. Modelos menores, como o Gemma 2, apresentam limitações consideráveis devido à sua capacidade reduzida de leitura e processamento de tokens, impactando negativamente na capacidade de manipular grandes *datasets*. Muitos modelos, mesmo aqueles disponíveis comercialmente em versões pagas, enfrentam dificuldades significativas em processar *datasets* extensos devido a restrições intrínsecas à sua arquitetura de tokenização e capacidade de entrada.

Os testes realizados evidenciaram dificuldades práticas para aplicação direta das técnicas tradicionais de geração de dados sintéticos, como SMOTE, GAN e ADASYN, nos ambientes atuais dos LLMs. Observou-se que modelos como GPT-4o, mesmo após análises detalhadas, não

conseguiram aplicar efetivamente estas técnicas devido à limitação operacional e à complexidade dos dados. Em contrapartida, foi possível utilizar uma abordagem alternativa baseada em amostragem aleatória com perturbações, obtendo resultados limitados.

Outro ponto importante identificado foi a sensibilidade dos modelos ao tamanho dos *datasets* utilizados. Modelos locais com menor capacidade, em torno de 8 bilhões de parâmetros, apresentaram tempos prolongados de execução e desempenho inferior devido ao tamanho significativo das bases de dados. Observou-se também que modelos dependentes exclusivamente da informação contida em prompts apresentaram resultados inferiores em comparação aqueles capazes de processar integralmente arquivos no formato CSV. Nesse contexto, fica evidente que tanto a estrutura da plataforma utilizada quanto o método de fornecimento dos dados influenciam significativamente os resultados.

Por fim, constatou-se que a quantidade de dados sintéticos gerados pela maioria dos modelos ainda é insuficiente, frequentemente gerando apenas cerca de 50 exemplos, mesmo diante da necessidade prática mínima de aproximadamente 500 exemplos. Nesse sentido, ressalta-se a importância de agentes executores integrados nas plataformas de inteligência artificial para a realização eficiente e escalável de tarefas relacionadas à geração de dados sintéticos, considerando-se as atuais limitações técnicas e a rápida evolução dessas ferramentas.

No que tange a aplicação para *oversampling*, o estudo revelou que o uso de técnicas de geração sintética de dados pode melhorar significativamente o desempenho dos modelos preditivos, especialmente em cenários altamente desbalanceados. Os métodos baseados em OpenAI 4o e Gemini 1.5 Pro foram consistentemente superiores, sugerindo forte eficácia na geração de dados sintéticos úteis e representativos para treinar modelos de classificação mais robustos e confiáveis.

4.5.3. **Dados Sintéticos via modelo LLM com RAG**

A utilização de LLMs pré-treinados oferece grande potencial para diversas aplicações. Contudo, a necessidade de acessar informações específicas e privadas exige uma abordagem mais refinada. Nesse contexto, o método RAG (*Retrieval-Augmented Generation*) surge como solução promissora, ao combinar a capacidade de geração de texto fluente dos LLMs com a precisão da

recuperação de informação.

A técnica de Geração Aumentada por Recuperação (RAG) combina modelos de linguagem de grande porte (LLMs) com sistemas de recuperação de informações. Seu objetivo é enriquecer a geração de texto dos LLMs com informações relevantes extraídas de uma base de dados ou documentos externos. Isso permite que o modelo produza respostas mais precisas, atualizadas e contextualmente relevantes, superando limitações de conhecimento inerentes ao modelo pré-treinado.

4.6.5.1. Contextualização de ferramentas e processos de IA Generativa

A construção de um modelo LLM com RAG requer um conjunto de ferramentas essenciais: um LLM pré-treinado como base (ex: GPT-4o), um banco de dados vetorial (ex: ChromaDB, Pinecone) para armazenar e pesquisar informações, um framework de incorporação (ex: SentenceTransformers) para transformar texto em vetores numéricos, e uma biblioteca RAG (ex: LangChain) para integrar o LLM com o banco de dados vetorial.

Destaca-se a relevância de plataformas como Anything LLM, LM Studio e iniciativas colaborativas como CrewAI, destacando sua importância para a consolidação de um ecossistema robusto de Inteligência Artificial.

Um dos elementos importantes para a orquestração do RAG é a ligação do mesmo com banco de dados vetorial que armazenam as informações e codificam dados em embeddings, o que permite executar buscas por similaridade semântica. Tal característica é especialmente útil para tarefas de recomendação, categorização de conteúdos em larga escala e aprimoramento de aplicações que utilizam RAG, pois a eficiência na recuperação de informações adequadas a cada prompt melhora a precisão dos resultados apresentados.

Nesse contexto, destaca-se o LangChain, uma estrutura de desenvolvimento projetada para viabilizar a criação de aplicações que integram grandes modelos de linguagem (LLMs) com diversas fontes de dados e ferramentas externas. Seu objetivo principal é facilitar a construção de sistemas que vão além da simples geração de texto, permitindo interações mais complexas e robustas. Essa integração é alcançada por meio de componentes modulares que suportam o gerenciamento de prompts, a orquestração de fluxos de trabalho e a implementação de memórias para armazenamento de estados entre interações. O LangChain também permite a conexão com APIs, bancos de dados e outros recursos externos, o que possibilita a criação de soluções personalizadas em domínios como atendimento ao cliente, análise de dados e geração

de conteúdo.

Um fluxo de trabalho (workflow) relevante que aplica o encadeamento descrito com as funções do LangChain é o LangFlow. Esta é uma ferramenta de low-code que simplifica a criação de aplicações e fluxos de trabalho baseados em modelos de linguagem. Com uma interface gráfica intuitiva, LangFlow permite que desenvolvedores e pesquisadores conectem módulos como agentes de linguagem, templates de prompt, memórias de conversação e integrações com APIs, facilitando a prototipagem rápida e a experimentação em projetos de inteligência artificial. Essa plataforma modular promove a customização e a integração de processos complexos sem a necessidade de escrever códigos extensos, viabilizando, por exemplo, a implementação de estratégias de fine-tuning para ajustar modelos a tarefas específicas, bem como a construção de sistemas baseados em RAG.

Essa dualidade de uso – combinando a eficiência da recuperação de dados com a capacidade de adaptação dos modelos – é ressaltada em comparativos recentes na literatura, como na análise de RAG versus fine-tuning além dos fundamentos teóricos apresentados por Lewis et al. (2020), que embasam o conceito de RAG.

Nesse sentido, ferramentas como Anything LLM e LM Studio facilitam a adoção prática de modelos de linguagem em diferentes setores, ao oferecer interfaces intuitivas, recursos de ajuste fino e integração com bases de dados. Essas plataformas permitem a análise e o gerenciamento de modelos em larga escala, possibilitando um desenvolvimento mais célere de aplicações que empregam metodologias de engenharia de prompt, RAG e outras técnicas correlatas. O objetivo central consiste em simplificar o processo de implantação e manutenção de soluções de Inteligência Artificial, alcançando resultados mais eficazes na interação homem-máquina.

Desta forma, nesta seção, será desenvolvido um modelo LLM com aplicação do protocolo e metodologia proposta de RAG para que as respostas do modelo LLM sejam efetivas para situações de geração de dados, *oversampling, data augmentation* e similares.

4.6.5.2. Performance Modelos LLM com RAG

Neste estudo de caso, a base de dados material utilizada na etapa 1 da metodologia RAG (Retrieval-Augmented Generation) será composta por informações qualificadas e organizadas no formato JSON. Essa base reunirá dados estruturados sobre os diferentes métodos de geração de dados sintéticos e técnicas de *oversampling*, conforme detalhado a seguir.

- Funções do Código do Smote original em python.
- Geração de Dados Sintéticos por Interpolação Aleatória
- Geração de Dados Sintéticos Amostra aleatória e Distribuição Normal e Perturbações
- *Oversampling* Simples
- Geração de Dados Sintéticos Usando Quantis e Perturbação
- *Data Augmentation* com *Bootstrap* e Ruído Gaussiano
- Mistura Aleatória de Atributos (*Feature Blending*)
- Geração de Dados Sintéticos com K-Means e Perturbação dos Centróides
- Geração Iterativa de Dados Sintéticos com Ruído Gaussiano
- Geração de Dados Sintéticos com SMOTE
- *Oversampling* com SMOTE para Balanceamento de Classes

A implementação do modelo, com a aplicação da metodologia RAG foi feita por meio das ferramentas, LM Studio, Anything LLM, Perplexity e POE. A escolha do banco de dados vetorial para *embedding* variou de acordo com a plataforma utilizando por exemplo, LanceDB e MongoDB. Para a geração dos vetores, adotou-se o modelo de embedding text-embedding-3-large. Assim, foi possível aplicar e testar o modelo desenvolvido.

Como métricas de avaliação, foram utilizadas as mesmas da seção anterior, a saber: Distância de Jensen-Shannon; Divergência Kullback-Leibler; *Mean Absolute Difference*; MAD - Índice de Similaridade; e Mantel Test - correlação observada. O resultado consta no Quadro 4.6 a seguir:

Quadro 4.6 - Modelos de LLM com RAG testados e sua performance de similaridade

Modelo	**Versão**	**Distância de Jensen-Shannon**	**Divergência Kullback-Leibler**	**Mean Absolute Difference MAD**	**Índice de Similaridade**	**Mantel Test Correlação observada**
GPT	4o	0.07305	0.03020	0.0466	0.9534	0.9580 p-value: 0.0010
GPT	o3	0.11306	0.05875	0.1167	0.8833	0.8295 p-value: 0.0010
Llama	3.1 8B	0.16010	0.22311	0.1133	0.8867	0.3480

						p-value: 0.0010
Llama	3.3 70B	0.15139	0.09867	0.2061	0.7939	0.6668 p-value: 0.0010
Qwen	2.5 7B	0.09734	0.05463	0.0955	0.9045	0.8973 p-value: 0.0010
DeepSeek	R1	0.142572	0.04242	0.1447	0.8553	0.8116 p-value: 0.0010

Fonte: Elaboração própria

Os experimentos com diferentes modelos de linguagem revelaram avanços e desafios notáveis na geração de dados sintéticos, na velocidade de resposta e na precisão de métodos implementados. Modelos da OpenAI, como o OpenAI o1, aplicado via ferramenta AnythingLLM, conseguiram criar códigos robustos utilizando técnicas como a amostragem aleatória com perturbações, mas enfrentaram falhas ao gerar dados sintéticos de forma consistente. Já o modelo 4o, utilizou interpolação aleatória entre amostras reais, demonstrou a necessidade de estudos adicionais devido a erros persistentes. Esses resultados destacam a importância de refinar abordagens para garantir a confiabilidade e aplicabilidade dos modelos.

A análise dos modelos da família Llama indicou uma ampla variação de desempenho. Enquanto o Llama 3.1 8B apresentou respostas completas e exemplos em Python, ele foi prejudicado por lentidão nos testes RAG, porém demonstrou desempenho moderado em código e casos sintéticos. O modelo 11B, integrante do Llama 3.2, apresentou limitações que comprometeram sua capacidade de geração de dados. Por sua vez, o modelo 3B não conseguiu apresentar uma resposta com dados sintéticos gerados. Apesar do maior porte, o Llama 3.3 70B também ficou aquém das expectativas no tocante à lentidão, impactado pelo ambiente de produção do modelo.

Por fim, outras abordagens, como os modelos Gemma2, Phi, Qwen e DeepSeek R1, também apresentaram resultados diversificados. O Gemma2 foi rápido, mas limitado em técnicas RAG. Os modelos Phi enfrentaram instabilidades com respostas inconsistentes e erros nos testes RAG, particularmente o Phi3.1-mini-128k e o Phi4-14B, enquanto o Phi3.5 3.8B mostrou dificuldades similares. O modelo Qwen 2.5 apresentou equilíbrio, rapidez e completude. O DeepSeek R1, após ajustes em sua plataforma, demonstrou potencial em técnicas específicas, como a interpolação controlada com perturbações. Essa diversidade de resultados reforça a necessidade de abordagens

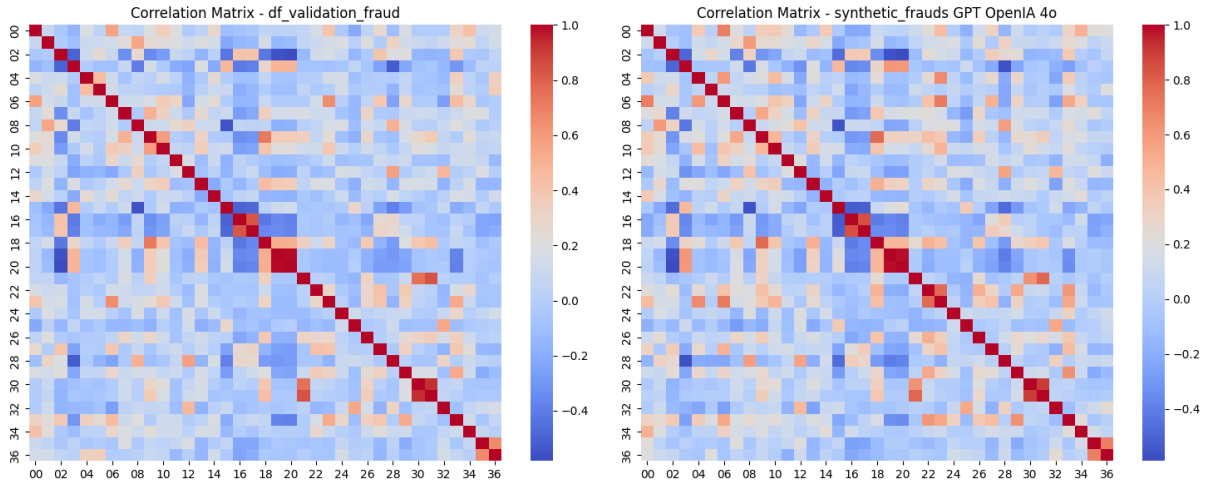
customizadas e avanços tecnológicos para atender às demandas variadas de aplicação.

A análise dos modelos de LLM com RAG testados evidencia um desempenho satisfatório e coerente com as expectativas estabelecidas. O GPT-4o destacou-se por sua alta similaridade (0.9534) e correlação observada (0.9580), além dos menores índices de divergência de Jensen-Shannon (0.07305) e Kullback-Leibler (0.03020). Em contraste, o GPT-o3 apresentou índices ligeiramente inferiores, com similaridade de 0.8833 e correlação de 0.8295, mostrando maior divergência entre as métricas (0.11306 e 0.05875, respectivamente). Os modelos da família Llama, como o 3.1 8B, alcançaram uma similaridade de 0.8867, embora apresentassem uma baixa correlação de 0.3480 e um índice de Jensen-Shannon relativamente alto (0.16010), enquanto o Llama 3.3 70B obteve similaridade de 0.7939 e correlação de 0.6668, com diferenças absolutas médias mais elevadas (0.2061).

Outros modelos demonstraram performances equilibradas, com destaque para o Qwen 2.5 7B, que obteve uma similaridade de 0.9045, correlação de 0.8973 e divergência de Jensen-Shannon de 0.09734, sendo mais consistente que o DeepSeek R1, que alcançou similaridade de 0.8553 e correlação de 0.8116, com um índice de divergência de 0.142572. Apesar de limitações pontuais, os resultados gerais reforçam a eficiência desses modelos nas tarefas de similaridade, com variações notáveis conforme suas capacidades específicas e abordagens técnicas.

Como verificado na Figura 4.15 (para LLM GPT 4o) e Anexo 2 (para demais modelos testados) as análises de correlações mantiveram padrões muito semelhantes, provando o sucesso dos modelos com RAG.

Figura 4.15 – Correlação entre as variáveis reais e geradas – LLM GPT4o



Fonte: elaboração própria

Desta forma, verifica-se que todos os modelos onde foi possível aplicar a RAG, foram

gerados dados sintéticos de forma aceitável e com características similares aos dados originais. Nos testes foram verificados que algumas vezes o modelo gera dados idênticos aos originais. Portanto, cabe ao cientista de dados excluir estes dados da análise e ou aplicação que de deslumbra aplicar.

4.6.5.3. Performance do modelo LLM - RAG com modelo preditivo

Desta forma, na Tabela 4.7 comparativa a seguir, são analisados os resultados de diferentes modelos de classificação aplicados a um conjunto de dados sintético, avaliados por meio Árvore de Decisão que foram medidos em termos de Precisão, Recall e AUC. A Precisão reflete a capacidade do modelo em minimizar falsos positivos, indicando a proporção de predições corretas entre as classificações positivas realizadas. O Recall, por sua vez, avalia a sensibilidade do modelo ao identificar corretamente os casos positivos existentes, enquanto a AUC (Área sob a Curva ROC) fornece uma métrica agregada que captura a habilidade discriminativa do classificador independentemente do limiar de decisão adotado.

Os modelos comparados foram ajustados de forma a permitir uma análise mais direta entre as abordagens, com o modelo "original" estabelecendo a base de comparação. Essa estrutura comparativa possibilita identificar os trade-offs entre as técnicas de classificação, evidenciando, por exemplo, se um aumento na Precisão pode ocorrer em detrimento do Recall ou se ambos os aspectos podem ser otimizados simultaneamente sem comprometer a capacidade discriminativa global (AUC).

Tabela 4.7 - Comparativo dos modelos utilizando dados sintéticos gerados por LLM

Modelo	**Dado Original**	**Llama 3.1-8B**	**Qwen 2.5-7b**	**OpenIA o3**	**OpenIA 4o**	**Llama 3.3 70B**	**Deepseek R1**
DT Precision	0,6429	0,8125	0,7778	0,9024	0,9452	0,7000	0,8140
DT Recall	0,6207	0,9070	0,8000	0,8605	0,9583	0,7778	0,7778
DT AUC	0,7241	0,8772	0,8310	0,8951	0,9407	0,7889	0,8199
LR Precision	0,7273	0,6818	0,6000	0,6667	0,7536	0,5429	0,6471
LR Recall	0,5517	0,6977	0,6000	0,6047	0,7222	0,5278	0,7333
LR AUC	0,7241	0,7302	0,6793	0,6883	0,6976	0,6306	0,7115
GBM Precision	0,8400	0,9500	0,8438	0,8864	0,9855	0,8378	0,9149
GBM Recall	0,7241	0,8837	0,7714	0,9070	0,9444	0,8611	0,9556
GBM AUC	0,8276	0,9249	0,8426	0,9096	0,9626	0,8806	0,9433

Fonte: Elaborado pelo autor.

A aplicação da geração de dados sintéticos com RAG, utilizando diferentes modelos de grandes linguagens (LLM) para *oversampling,* demonstrou impactos positivos significativos nos modelos preditivos avaliados - Regressão Logística (LR), Árvore de Decisão (DT) e Gradient Boosting Machine (GBM) - os resultados apresentados na análise comparativa da Tabela 4.7.

Observando os resultados, percebe-se que todos os modelos, em geral, apresentaram uma melhoria substancial ao incorporar dados sintéticos, com destaque para as métricas de precisão, *recall* e Área Sob a Curva ROC (AUC).

Na análise detalhada do desempenho por algoritmos, a Árvore de Decisão mostrou notável melhoria em todas as métricas quando alimentada com dados sintéticos gerados pelos modelos OpenAI-4o e OpenAI-o3, destacando-se com valores elevados de precisão (0,9452) e recall (0,9583) no OpenAI-4o, bem acima dos valores observados na base original (0,6429 e 0,6207, respectivamente). Similarmente, o *Gradient Boosting* obteve incrementos expressivos com o uso de dados sintéticos, especialmente no modelo OpenAI-4o, atingindo precisão quase perfeita (0,9855), recall alto (0,9444) e uma excepcional AUC de 0,9626. Esses resultados indicam que a qualidade da geração sintética de dados proporcionada pelos modelos da OpenAI impactou positivamente e de forma consistente o desempenho preditivo desses classificadores.

Contudo, para esta base de dados analisada, a Regressão Logística apresentou resultados mais variados e menos consistentes em comparação aos outros algoritmos. Embora tenha havido ganhos em recall com Deepseek-R1 (0,7333) e OpenAI-4o (0,7222), observou-se queda na precisão na maioria dos modelos avaliados, particularmente notável no Llama 3.3-70B (0,5429) e Qwen 2.5-7b (0,6000), ambos abaixo da precisão obtida na base original (0,7273). Além disso, os ganhos na métrica AUC para a Regressão Logística foram modestos, com valores próximos ao original (0,7241), variando pouco entre 0,6306 e 0,7302.

4.6.5.4. Conclusão e resultado geração dados sintéticos com RAG

Verifica-se que é possível e eficiente gerar dados sintéticos por meio de modelos de LLM com RAG conforme modelo proposto neste trabalho.

Diversos fatores influenciam a geração de dados sintéticos e podem impactar diretamente a performance de um modelo de agente de IA. Primeiramente, a capacidade de tokens do modelo é essencial, pois muitos modelos dependem de uma quantidade máxima de tokens para funcionar de maneira eficiente. Modelos menores tendem a apresentar limitações nesse aspecto, afetando negativamente a performance e a qualidade dos dados gerados. Além disso, a plataforma utilizada

para rodar o modelo também desempenha um papel importante; por exemplo, é necessário que a arquitetura suporte a execução de Python em Docker no *backend*, garantindo estabilidade e compatibilidade.

Outro aspecto crítico é o hardware utilizado durante a execução. Quando rodado localmente, o desempenho do modelo está diretamente relacionado à memória disponível e ao uso de GPU no ambiente, sendo ambos determinantes para otimizar a performance. Por fim, em sistemas de chat local, respostas anteriores são continuamente carregadas no prompt a cada nova interação, o que pode acumular tokens de contexto e resposta, resultando em lentidão progressiva. Diante dessas variáveis, o desenvolvimento de um modelo de agente de IA especializado na geração de dados sintéticos poderia abordar essas limitações e oferecer resultados mais robustos e consistentes.

A implementação de um sistema RAG, conforme descrita, evidencia a relevância de uma abordagem meticulosa para cada etapa do processo, desde a preparação de dados até a manutenção contínua do modelo. A adoção dos diferentes modelos de LLM com aplicação da biblioteca LangChain para orquestração entre o banco de dados vetorial e o modelo de geração, bem como uso de plataformas com estrutura de RAG (LM Studio, POE.AI e AnythingLLM) se mostraram vantajosas, pois ambas as ferramentas oferecem flexibilidade para lidar com diversos cenários e exigências de domínio. Além disso, a constante reavaliação das métricas e o refinamento de parâmetros asseguram a qualidade e a pertinência das respostas produzidas, evidenciando o potencial do RAG para aplicações que demandem informações atualizadas e contextualizadas.

Os modelos analisados cumpriram seus objetivos, com destaque para o GPT-4º, que alcançou os menores índices de divergência (Jensen-Shannon: 0.07305) e a maior similaridade (0.9534). O Qwen 2.5 7B também obteve resultados equilibrados, com uma similaridade de 0.9045 e baixa divergência (0.09734), superando o desempenho do DeepSeek R1 (similaridade de 0.8553). Modelos como o Llama apresentaram variações, sendo o 3.1 8B mais consistente (similaridade de 0.8867) em relação ao 3.3 70B, que teve menor eficiência (similaridade de 0.7939).

No geral, os resultados destacam a eficiência dos modelos em tarefas de similaridade, apesar de limitações específicas. Ajustes técnicos podem maximizar seu desempenho para atender às demandas futuras com maior precisão e confiabilidade.

Nos testes referentes a aplicação dos dados sintéticos gerados pelos modelos utilizando RAG, demostram que o uso de *oversampling* com dados sintéticos gerados por modelos de grandes linguagens (LLMs) melhora significativamente o desempenho dos classificadores DT (Árvore de

Decisão) e GBM (*Gradient Boosting Machine*), especialmente com o modelo OpenAI-4o, que obteve os maiores valores em precisão (DT: 0,9452, GBM: 0,9855), recall (DT: 0,9583, GBM: 0,9444) e AUC (DT: 0,9407, GBM: 0,9626). A Regressão Logística, porém, apresentou ganhos menos expressivos e resultados mais instáveis entre os modelos avaliados, destacando-se apenas no recall com Deepseek-R1 (0,7333).

De maneira geral, os testes evidenciam que a eficácia na geração de dados sintéticos e na execução de código Python varia significativamente entre os modelos, dependendo de fatores como o limite de tokens, a técnica de amostragem empregada (por exemplo, *random sampling with perturbations* ou *synthetic data generation with random interpolation between samples*) e a plataforma utilizada (como AnythingLLM, LM Studio ou GPT4all). Enquanto alguns modelos se destacam pela rapidez na resposta ou pela capacidade de gerar código, muitos apresentam desafios na consistência e integridade dos dados sintéticos gerados, indicando a necessidade de novos testes e ajustes metodológicos para aprimorar seus desempenhos em cenários práticos.

4.5.4. **Dados Sintéticos Modelo SLM Aurora com Fine-Tuning**

Nesta seção, conforme apresentado na metodologia, foi evoluída a abordagem e aplicada a estratégia de fine-tuning para treinar um modelo SLM (*Small Language Model*) com conhecimento de criação de dados sintéticos. O modelo desenvolvido foi batizado de Aurora. Foi utilizado como base o modelo de código aberto Qwen 2.5 7B pois este apresenta boa performance matemática e de geração de código conforme verificado na seção anterior. A aplicação deste novo modelo criado será para uso em máquinas locais, tal fato é um limitador de desempenho comparado a outros modelos LLM que são hospedados em servidores de GPU em nuvem.

Essa abordagem é especialmente relevante, pois a geração de dados sintéticos pode contribuir significativamente para a superação de desafios relacionados à escassez de dados e à proteção da privacidade, conforme discutido por Radford et al. (2019). Ao ajustar um LLM pré-treinado para capturar padrões específicos do domínio de interesse, deve-se obter um modelo capaz de sintetizar dados realísticos. Estes dadospoderão ser utilizados para aprimorar a validação e o desenvolvimento de aplicações avançadas em ciência de dados.

4.7.5.1. Performance de Modelo com Fine-Tuning - Aurora

Da mesma forma aos testes realizados com modelos abertos de IA generativa para a geração de dados sintéticos apresentados anteriormente, nesta sessão, modelos de LLM específicos foram desenvolvidos via fine-tuning, incorporando o conhecimento de data augmentation.

Para desenvolvimento do modelo próprio Aurora, o fine-tuning eficiente (ajuste fino ou treinamento adicional) do modelo de LLM base foi realizado usando a biblioteca Unsloth, integrada à técnica LoRA (*Low-Rank Adaptation*), que permite treinar grandes modelos de linguagem com pouca memória, rapidez e eficiência.

O fine-tuning foi executado no Google Collab, usando GPU A100 e T4 com aplicação do SFTTrainer, especializado para treinamento supervisionado com a biblioteca TRL (*Transformers Reinforcement Learning*). O modelo final é aplicado localmente na máquina do usuário via ferramenta LMSudio, desta forma, possui uma arquitetura limitada de operacionalização.

O conteúdo para o Fine-tuning foi baseado na coleta de dados com diferentes métodos e técnicas de geração de dados sintéticos, *oversampling* e *data augmentation* validadas na sessão anterior, além disso, códigos Python específicos para a aplicação desses métodos foram acrescentados e validados. Desta forma, o modelo Aurora recebe a habilidade de geração destes dados.

A avaliação do modelo Aurora, fine-tuned a partir do Qwen 2.5 7B, demonstra uma capacidade notável na geração de dados sintéticos, conforme as métricas apresentadas no Quadro 4.7 . As baixas Distância de Jensen-Shannon (0.1310) e Divergência Kullback-Leibler (0.1114) indicam que a distribuição estatística dos dados gerados pelo Aurora é muito próxima à dos dados originais, sugerindo uma captura eficaz das características gerais. Isso é complementado por um alto Índice de Similaridade (0.8644), que reforça a elevada semelhança global entre os conjuntos de dados sintético e de referência.

Quadro 4.7 - Modelos de LLM Aurora – Desenvolvido com Fine-Tuning

Modelo Base	**Versão**	**Distância de Jensen-Shannon**	**Divergência Kullback-Leibler**	**Mean Absolute Difference MAD**	**Índice de Similaridade**	**Mantel Test Correlação observada**
Qwen	2.5 7B	0.1310	0.1114	0.1356	0.8644	0.8372

						p-value: 0.0010

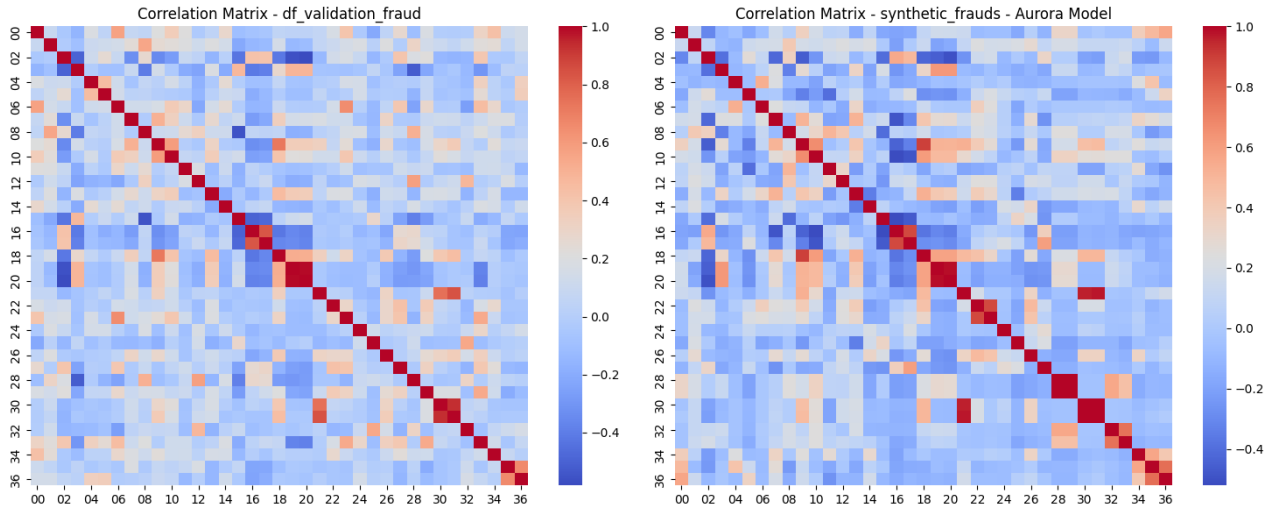
Fonte: elaboração própria

Além da similaridade distributiva e geral, a qualidade dos dados sintéticos é evidenciada pela precisão em níveis mais granulares e estruturais. O baixo *Mean Absolute Difference* (0.1356) sugere que as diferenças médias entre valores ou estatísticas pontuais são mínimas. Mais importante ainda, o Teste de Mantel apresentou uma correlação forte e estatisticamente significativa (0.8372, p=0.001), indicando que o Aurora conseguiu preservar eficazmente as relações e a estrutura interna presentes nos dados originais. Esse é um aspecto crucial para a utilidade dos dados sintéticos em tarefas analíticas complexas.

No que tange a análise de correlação, a Figura 4.16 apresenta duas matrizes de correlação de Pearson, comparando os dados reais (df_validation_fraud) e os dados sintéticos gerados pelo modelo Aurora (synthetic_frauds). A análise visual revela que as duas matrizes exibem padrões de correlação muito semelhantes, indicando que o modelo Aurora foi bem-sucedido em replicar as relações entre as variáveis presentes nos dados reais. Demostra que gerou dados sintéticos que preservam com precisão as características de correlação dos dados reais.

Essa preservação é evidente na semelhança dos padrões de correlação, na distribuição das correlações positivas e negativas e na presença de uma diagonal principal vermelha intensa em ambas as matrizes. Esses resultados comprovam o sucesso do modelo na replicação da estrutura de dependência entre as variáveis, indicando que os dados sintéticos podem ser utilizados de forma confiável para análises e modelagem.

Figura 4.16 - Correlação de Pearson - dados originais vs gerado modelo Aurora



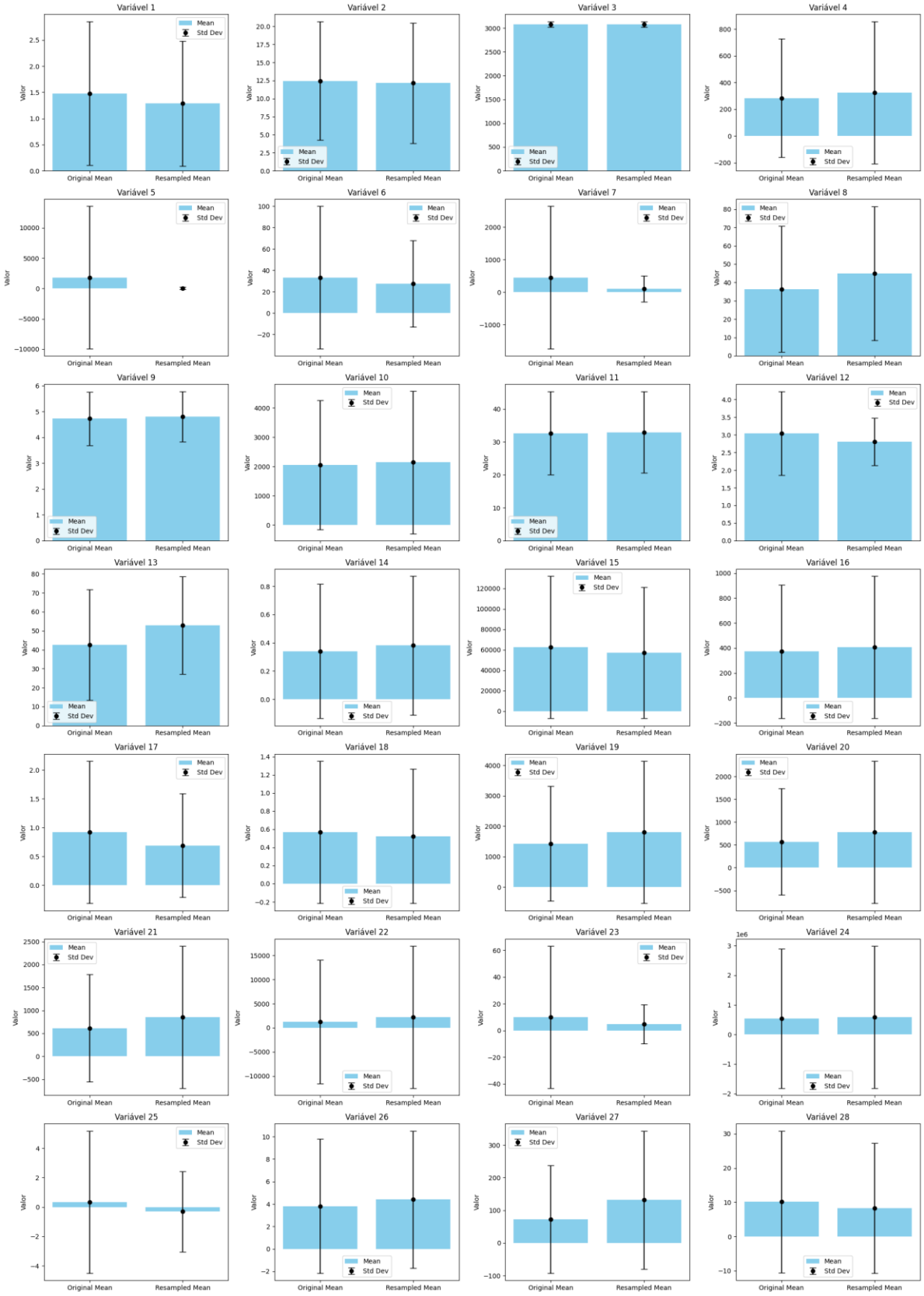
Fonte: elaboração própria

Os box-plots na Figura 4.17 mostram que a média e o desvio padrão são quase idênticos entre os dados reais e sintéticos. Isso indica que o modelo gerador de dados sintéticos conseguiu capturar com precisão as características estatísticas centrais dos dados reais. A semelhança na forma e na extensão dos box-plots sugere que a distribuição geral das variáveis foi bem preservada no processo de geração de dados sintéticos.

Destaca-se o fato inédito de ter um modelo de IA Generativa SLM para ser utilizado de forma local para geração de dados sintéticos. Tal fato por si só apresenta limitação intrínseca de ambiente de aplicação do modelo pelos limitadores de hardware e arquitetura em comparação aos grandes modelos disponibilizados online nas plataformas da OpenAI, Google e etc.

No que se refere a análise das distribuições das variáveis, conforme apresentado na Figura 4.18, observa-se uma semelhança entre os conjuntos real e sintético. Em quase todos os gráficos, os histogramas dos dados sintéticos seguem de perto o padrão dos histogramas dos dados reais, indicando que o modelo gerador de dados sintéticos capturou com precisão as características de distribuição das variáveis originais. Essa semelhança é observada tanto na forma geral das distribuições quanto nas frequências relativas dos valores, sugerindo que o modelo replicou com sucesso as nuances estatísticas dos dados reais.

Figura 4.17 – Estatísticas descritivas – Box Plot – Modelo Aurora



Fonte: Elaborado pelo autor

Figura 4.18 – Distribuição das variáveis – Modelo Aurora

Fonte: Elaborado pelo autor

O Quadro 4.8 apresenta uma comparação do desempenho de três modelos preditivos (Regressão Logística, Árvore de Decisão e GBM) usando tanto dados sintéticos gerados pelo modelo Aurora quanto dados originais. O modelo GBM, treinado com dados sintéticos do Aurora, obteve a maior precisão de 90,32%, recall de 77,78% e AUC de 0,89751, superando o modelo treinado com dados originais em todas as métricas. A Árvore de Decisão com dados sintéticos alcançou 73,53% de precisão e 69,44% de recall, enquanto a Regressão Logística apresentou 63,33% de precisão e 52,78% de recall.

Quadro 4.8 - Comparativo de modelos preditivos utilizando dados sintéticos gerados pelo modelo Aurora de SLM

Modelo	Logistic Regression	Decision Tree	GBM
Precisão (Aurora)	0,63333	0,73529	0,90323
Recall (Aurora)	0,52778	0,69444	0,77778
AUC (Aurora)	0,76676	0,76964	0,89751
Precisão (Original)	0,72222	0,83333	0,80000
Recall (Original)	0,44828	0,68966	0,68966
AUC (Original)	0,78537	0,81034	0,89180

Fonte: Elaborado pelo autor.

Em geral, os modelos treinados com dados sintéticos do Aurora demonstraram desempenho competitivo em relação aos modelos treinados com dados originais. O GBM, em particular, apresentou um aumento significativo no desempenho, com um aumento de 10,32% na precisão, 8,81% no recall e 0,57% no AUC em comparação com o modelo treinado com dados originais. Isso sugere que o modelo Aurora pode gerar dados sintéticos eficazes para treinar modelos preditivos, com potencial para superar o desempenho obtido com dados originais em alguns casos.

Cabe salientar que o modelo desenvolvido foi aplicado em máquina local, não possuindo os recursos e ecossistema disponíveis das grandes plataformas, bem como utilização de grandes GPUs. Tal fato limitou a capacidade do modelo neste teste, porém apresenta condições e expectativas dos resultados serem ampliados com a utilização de equipamentos e infraestruturas mais robustas.

4.7.5.2. Conclusão e resultado modelo Aurora Fine Tuning

Num cenário de constante evolução da inteligência artificial, a geração de dados sintéticos desponta como uma abordagem promissora, capaz de mitigar desafios relacionados à privacidade, à escassez e ao desbalanceamento de dados, além de atender à necessidade de conjuntos de treinamento mais robustos e representativos. É nesse contexto que se destaca o modelo Aurora, uma inovação focada no desenvolvimento de uma inteligência artificial generativa especializada, destinada à produção de dados sintéticos de alta fidelidade.

Diferentemente de modelos de propósito geral, o Aurora foi especificamente concebido e refinado (através de fine-tuning sobre bases como o Qwen 2.5 7B) para dominar a complexa tarefa de gerar conjuntos de dados artificiais, sintéticos. Seu diferencial reside na capacidade de replicar não apenas as características estatísticas, mas também a estrutura intrínseca de dados reais, o que representa um avanço focado na capacidade das IAs generativas.

A análise conjunta das métricas quantitativas valida de forma robusta o processo de fine-tuning e confirma a proficiência do modelo Aurora na geração de dados sintéticos de alta fidelidade, a partir da base Qwen 2.5 7B. Os resultados são convincentes e consistentes quanto à similaridade distributiva dos dados gerados, marcadas pela baixas Distância de Jensen-Shannon (0.1310) e Divergência Kullback-Leibler (0.1114). Estes valores atestam que a distribuição estatística geral dos dados sintéticos gerados pelo Aurora se alinha estreitamente com a dos dados originais.

A proximidade de valores encontrados, mensurado pelo baixo Mean Absolute Difference (0.1356) sugere que, em média, os valores numéricos ou estatísticas descritivas dos dados sintéticos diferem pouco dos seus correspondentes originais. Do mesmo modo, o elevado Índice de Similaridade (0.8644) quantifica uma forte semelhança global entre os conjuntos de dados sintético e de referência.

Crucialmente, o Teste de Mantel revelou uma correlação alta (0.8372) e estatisticamente muito significativa (p=0.001), demonstrando que as relações e a estrutura interna (distâncias relativas entre pontos de dados) foram eficazmente preservadas no conjunto de dados sintético.

A análise dos box-plots confirma a eficácia do modelo em gerar dados sintéticos que se assemelham estatisticamente aos dados reais. A proximidade das médias e a similaridade na dispersão dos dados indicam que o modelo conseguiu replicar as características importantes das variáveis originais. Essa semelhança estatística é crucial para garantir que os dados sintéticos

possam ser usados de forma confiável em análises e modelagens subsequentes.

A análise visual dos histogramas revela também uma forte semelhança entre as distribuições dos dados reais e sintéticos. O modelo de geração de dados sintéticos foi bem-sucedido em replicar as características de distribuição das variáveis originais, tanto na forma geral quanto nas frequências relativas dos valores.

A análise comparativa do desempenho de modelos preditivos treinados com dados sintéticos gerados pelo modelo Aurora e dados originais revela resultados promissores. O modelo GBM, em particular, demonstrou um desempenho superior ao modelo treinado com dados originais, alcançando uma precisão de 90,32%, um recall de 77,78% e um AUC de 0,89751. Esse resultado indica que o modelo Aurora foi capaz de gerar dados sintéticos de alta qualidade, permitindo que o modelo GBM aprendesse padrões mais eficazes e superasse o desempenho obtido com dados reais. Em geral, os modelos treinados com dados sintéticos do Aurora apresentaram um desempenho competitivo, demonstrando o potencial do modelo para gerar dados sintéticos úteis em diversas aplicações de aprendizado de máquina.

Cabe destacar que o uso de um modelo de IA Generativa SLM de forma local para a criação de dados sintéticos apresentou resultados satisfatórios mesmo com as limitações inerentes ao ambiente em que é aplicada, principalmente devido às restrições de hardware e arquitetura. O desempenho desse modelo pode ser comparável ao de soluções de mercado disponíveis em plataformas online, mesmo não dispondo dos mesmos recursos avançados e da infraestrutura robusta oferecida por grandes provedores, como OpenAI e Google.

Essa convergência de resultados — baixas divergências, baixa diferença média, alta similaridade geral e, notavelmente, a preservação da estrutura interna validada estatisticamente — fornece forte evidência do sucesso da abordagem de fine-tuning. Assim, fica estabelecida a capacidade do modelo Aurora de produzir dados sintéticos de alta qualidade, que mimetizam de perto as características essenciais e relacionais dos dados de referência utilizados.

4.6. **Conclusão Geração dados Sintéticos com GenAI**

No gerenciamento de modelos de detecção de fraudes, as instituições financeiras aplicam critérios rigorosos para lidar com falsos positivos. Frequentemente, priorizam modelos com alto *recall* para garantir a detecção da maioria das fraudes, enquanto mantêm a precisão em um nível aceitável. Isso é crucial para evitar a sobrecarga dos sistemas de verificação e minimizar o impacto nos clientes. Do mesmo modo, devem implementar sistemas de pontuação de risco ou etapas

adicionais de verificação para transações sinalizadas, ajudando a reduzir o impacto dos falsos positivos. É essencial, portanto, realizar uma análise custo-benefício que pondere os prejuízos associados a falsos positivos e falsos negativos, ajustando o modelo conforme necessário para otimizar o retorno geral.

Conforme verificado na seção de geração de dados sintéticos com modelos tradicionais de inteligência artificial, os resultados evidenciam que os três métodos avaliados (SMOTE, GAN e VAE) são eficazes na geração de dados sintéticos, cada um com características distintas. O modelo GAN apresentou a melhor correspondência estatística com os dados originais, enquanto o SMOTE se destacou na preservação da correlação entre variáveis. A introdução de um termo de regularização baseado na correlação melhorou a aproximação entre os dados sintéticos e originais, e uma camada customizada de reamostragem foi essencial para otimizar a retropropagação do gradiente nos modelos baseados em redes neurais. A análise da Distância de Jensen-Shannon indicou que o SMOTE gera os dados sintéticos mais similares aos originais, seguido pelo GAN e pelo VAE. Contudo a divergência de Kullback-Leibler apontou uma inversão parcial, com SMOTE mantendo a liderança, mas com o VAE superando o GAN nesse critério.

Em termos de desempenho, os modelos de redes neurais foram mais eficazes em cenários onde a normalização não afeta a interpretação dos resultados, enquanto o SMOTE mostrou superioridade ao garantir que os dados pudessem ser retornados à escala original. O VAE apresentou discrepâncias mais pronunciadas na preservação das correlações, tornando-o menos adequado para aplicações onde essa característica é essencial. Dessa forma, para cenários que exigem alta fidelidade na preservação da estrutura dos dados, recomenda-se o uso do SMOTE, seguido pelo GAN, com o VAE sendo utilizado com cautela devido à sua menor capacidade de replicar padrões de correlação.

A geração de dados sintéticos por meio de engenharia de prompt em modelos de LLM demonstrou ser viável, embora tenha apresentado limitações importantes. Modelos como GPT-4o e Gemini 1.5 Pro apresentaram alto desempenho, manifestado pela baixa Distância de Jensen-Shannon (0,04260 para GPT-4o) e alto Índice de Similaridade (0,9721 para GPT-4o), além de correlações significativas. No entanto, modelos como GPT o1 e GPT o3 mini-high tiveram desempenho inferior, e a aplicação de técnicas tradicionais como SMOTE mostrou-se problemática, exigindo a adoção demétodos alternativos como distribuições gaussianas e interpolação. Além disso, a capacidade limitada de processamento de tokens e a sensibilidade ao tamanho dos datasets impactaram negativamente a geração de dados sintéticos.

Apesar dos desafios, o uso de dados sintéticos gerados por LLMs mostrou potencial para melhorar o desempenho de modelos preditivos, especialmente em cenários desbalanceados. Modelos como GPT-4o e Gemini 1.5 Pro demonstraram eficácia na geração de dados representativos, superando as limitações técnicas e a insuficiência na quantidade de dados gerados (cerca de 50 exemplos, em média). A necessidade de agentes executores integrados nas plataformas de IA para otimizar a geração de dados sintéticos foi destacada, visando a superação das limitações atuais e aprimoramento da escalabilidade.

Na sequência, ao aplicar o LLM com RAG para a geração de dados sintéticos, você aproveita a capacidade do modelo de linguagem em gerar texto coerente e contextualmente rico, enquanto o mecanismo de recuperação assegura que as informações relevantes e específicas das classes sub-representadas sejam incorporadas, resultando em um dataset mais balanceado e um modelo preditivo mais robusto.

O estudo dos modelos LLM com RAG destacam que a eficiência dessa abordagem é influenciada por diversos fatores, incluindo a capacidade de tokens dos modelos, a plataforma de execução (que deve suportar Python em Docker no backend) e o hardware disponível (memória e GPU). A implementação de um sistema RAG, utilizando a biblioteca LangChain para orquestração entre o banco de dados vetorial e o modelo de geração, junto a plataformas como o LMStudio, demonstraram ser vantajosa, permitindo flexibilidade e adaptação a diferentes cenários. Modelos como GPT-4o e Qwen 2.5 7B apresentaram resultados promissores, com baixos índices de divergência (Jensen-Shannon de 0.07305 para GPT-4o) e alta similaridade (0.9534 para GPT-4o), evidenciando a eficácia da abordagem RAG na geração de dados sintéticos de alta qualidade.

De forma complementar, a aplicação dos dados sintéticos gerados via RAG em tarefas de *oversampling* melhorou significativamente o desempenho de modelos de classificação, especialmente Árvores de Decisão (DT) e Gradient Boosting Machine (GBM), sendo o modelo OpenAI-4o o que se destacou. No entanto, a regressão logística (LR) apresentou resultados mais instáveis. A pesquisa também apontou variações na eficácia dos modelos em relação à execução de código Python e à consistência dos dados gerados, influenciadas por fatores como o limite de tokens e a técnica de amostragem utilizada. Verificou-se também a necessidade de aprimorar e manter o ambiente de execução dos modelos para otimizar o desempenho dos LLMs e SLM locais em cenários práticos de geração de dados sintéticos com RAG.

Nesse contexto, o modelo Aurora se destaca como uma inovação significativa, um projeto focado no desenvolvimento de uma IA generativa especializada na criação de dados sintéticos de

alta fidelidade.

reais. A análise conjunta de métricas quantitativas valida robustamente o processo de fine-tuning e confirma a proficiência do modelo Aurora na geração de dados sintéticos de alta fidelidade a partir da base Qwen 2.5 7B. Os resultados são convincentes e consistentes, com baixa Distância de Jensen-Shannon (0.1310) e Divergência Kullback-Leibler (0.1114), atestando a similaridade distributiva dos dados gerados.

Os resultados obtidos demonstram a capacidade do modelo Aurora na produção de dados sintéticos de alta qualidade, que mimetizam de perto as características essenciais e relacionais dos dados de referência utilizados. A convergência de resultados - manifestada pelas baixas divergências, baixa diferença média (Mean Absolute Difference de 0.1356), alta similaridade geral (Índice de Similaridade de 0.8644) e a preservação da estrutura interna (Teste de Mantel com correlação de 0.8372 e p=0.001) - valida estatisticamente o sucesso da abordagem de fine-tuning. Além disso, a análise comparativa do desempenho de modelos preditivos treinados com dados sintéticos gerados pelo modelo Aurora e dados originais revela resultados promissores, com o modelo GBM alcançando uma precisão de 90,32%, recall de 77,78% e AUC de 0,89751, demonstrando o potencial do modelo para gerar dados sintéticos úteis em diversas aplicações de aprendizado de máquina.

Capítulo 5

5. Modelo de Detecção de Anomalias em Produtos de Seguridade: Aplicação Empírica em uma Instituição financeira Nacional

Alex Cerqueira Pinto

PPGA - Universidade de Brasília - UNB

Resumo

Este estudo apresenta a aplicação de modelos não supervisionados de detecção de anomalias em produtos de seguridade no setor financeiro, com o objetivo de superar limitações das abordagens tradicionais de monitoramento interno. A ausência de dados rotulados impôs desafios significativos à validação dos modelos, exigindo abordagens alternativas, como pseudo-supervisão e amostragem especializada, para viabilizar a análise de desempenho. Através do treinamento do modelo com algoritmos de aprendizado de máquina, como *autoencoder* e *isolation forest*, busca-se alcançar uma performance superiores na detecção de irregularidades as técnicas tradicionais de amostra estratificada de controles internos. A interpretabilidade dos resultados foi analisada com a aplicação do SHAP value, os quais revelaram contribuições negativas significativas de variáveis operacionais críticas: atrasos em etapas processuais (SHAP ≈ -1,03), altas taxas de desistência recentes (SHAP ≈ -2,42) e perdas financeiras acumuladas (SHAP ≈ -1,94), todas indicando padrões consistentes com riscos operacionais imediatos de acordo com os especialistas e fazem sentido negocial. A análise de interpretação via SHAP Waterfall complementou ao demonstrar a discrepância entre o valor esperado da saída do modelo (E[f(x)] = 14,843) e a predição observada (f(x) = 8,332), explicada principalmente pelas variáveis críticas mencionadas. Em outro exemplo, perdas contratuais acumuladas (SHAP = -1,94) e inadimplência recorrente (SHAP = -1,38) foram os principais determinantes do desvio, enquanto atributos de alto valor nominal, como transações de grande montante, exerceram pouco impacto. Esses resultados reforçam a coerência negocial da modelagem, pois os fatores de risco identificados alinham-se com expectativas práticas do domínio financeiro. Apesar da precisão satisfatória (0,3135), o modelo apresentou baixo recall (0,0922), refletindo a dificuldade em capturar a totalidade das perdas operacionais, mas mantendo alta especificidade (0,8918) e acurácia global de 61,27%. Os achados indicam ganhos operacionais relevantes, com redução de falsos positivos e fortalecimento do processo decisório por meio de insights acionáveis. Recomenda-se o uso de estratégias complementares, como ajuste de limiares, aprendizado semissupervisionado e fine-tuning com feedback especializado, para aprimorar o desempenho e adaptabilidade dos sistemas em ambientes complexos e dinâmicos. A pesquisa contribui para a evolução de práticas de monitoramento de riscos no setor financeiro, demonstrando o potencial da inteligência artificial na detecção proativa de condutas irregulares, promovendo maior segurança, eficiência e confiabilidade institucional.

Palavras -chave: anomalias; perdas operacionais; não supervisionado; relacionamento com clientes; prevenção;

5.1. Introdução

O mundo financeiro moderno opera em uma escala e complexidade sem precedentes, com

uma imensidão de transações ocorrendo a cada segundo. Em um ambiente tão dinâmico e interconectado, bancos e instituições financeiras enfrentam o constante desafio de garantir a integridade e autenticidade de cada operação, detectando anomalias para prevenir fraudes, manter a confiança do cliente e evitar perdas financeiras.

A identificação de anomalias, ou padrões que se desviam da norma, surge como um mecanismo vital para detectar e prevenir atividades fraudulentas. Esta prática não só serve como uma barreira contra possíveis ameaças, mas também como uma ferramenta para melhorar a confiabilidade e eficiência das operações financeiras. No âmbito das instituições financeiras, a detecção de anomalias surge como um elemento crítico na salvaguarda da integridade das operações e na preservação da confiança dos stakeholders. À medida que o ambiente bancário se torna cada vez mais digitalizado, interconectado e complexo, a capacidade de identificar e responder prontamente a atividades atípicas assume uma importância sem precedentes.

A detecção de anomalias refere-se ao processo de identificar padrões em um conjunto de dados que não se conformam ao comportamento esperado. Estas "anomalias" ou "outliers" podem ser indicativas de uma variedade de eventos, desde erros de dados até atividades fraudulentas. No contexto bancário, onde a precisão das transações é fundamental, a identificação precoce de tais irregularidades pode ser a diferença entre uma operação segura e um potencial desastre financeiro.

Torna-se evidente que os modelos de detecção de anomalias não são mais um luxo, mas uma necessidade estratégica. As abordagens tradicionais, que muitas vezes dependem de regras rígidas ou limiares predefinidos, estão rapidamente se tornando obsoletas. Em contraste, métodos modernos, muitas vezes ancorados em aprendizado de máquina e análise de big data, oferecem flexibilidade, adaptabilidade e precisão sem paralelo.

A natureza intrincada das transações financeiras, aliada à complexidade das atividades fraudulentas, impõe desafios específicos (Hilal et al., 2022; Zamini & Hasheminejad, 2019; Zhang et al., 2022). Um dos principais obstáculos é o desbalanceamento dos dados, uma vez que a grande maioria das transações é legítima, enquanto as fraudes representam apenas uma pequena fração. Esse desequilíbrio pode comprometer a capacidade dos modelos de detecção de anomalias, resultando em falsos positivos que, além de gerar incômodo aos clientes, exigem revisões manuais desnecessárias e elevam os custos operacionais.

Outro desafio relevante é o a interpretabilidade dos modelos: para ganhar a confiança dos tomadores de decisão, os modelos precisam ser transparentes e facilmente interpretáveis. Técnicas complexas de aprendizado de máquina, como redes neurais profundas, muitas vezes são vistas como

"caixas-pretas", tornando difícil entender suas decisões. Soma-se a isso, o desafio da baixa latência (entendida como o atraso de tempo entre o momento em que uma ação é iniciada (ou um dado é enviado) e o momento em que essa ação é concluída (ou o dado é recebido/processado): em ambientes bancários que operam em tempo real, os modelos precisam processar informações e tomar decisões em milissegundos. Garantir essa velocidade sem comprometer a precisão impõe um elevado custo de implementação, visto que o desenvolvimento, treinamento e manutenção de modelos sofisticados, especialmente aqueles que empregam tecnologias emergentes, podem ser onerosos. emergentes.

Neste contexto, o presente artigo tem como objetivo, desenvolver modelos analítico de detecção de anomalias em produtos de bancários utilizando técnicas de aprendizado de máquina, e verificar se esses modelos superam as abordagens tradicionais de monitoramento de controles internos. Para tanto, serão utilizados dados de um banco nacional, com aplicação prática dos modelos e mensuração dos resultados obtidos.

Este estudo busca avançar o conhecimento acadêmico ao aprofundar e expandir investigações anteriores sobre detecção de anomalias em bancos e instituições financeiras, tema de crescente relevância diante dos desafios contemporâneos de segurança e conformidade regulatória. A pesquisa justifica-se pela lacuna existente na literatura quanto à aplicação prática e à eficácia das principais técnicas de detecção em cenários com ausência de dados rotulados, uma realidade comum no setor financeiro. Como contribuição, o trabalho propõe o aprimoramento de métodos já consolidados, realiza testes empíricos com modelos desenvolvidos, analisa criticamente os resultados obtidos e apresenta soluções viáveis e fundamentadas para superar limitações atuais. Ao enfatizar o potencial dos modelos de detecção de anomalias, o artigo reforça sua utilidade prática e seu papel estratégico no fortalecimento dos mecanismos de monitoramento e prevenção de irregularidades, oferecendo subsídios teóricos e aplicados para futuras pesquisas e implementações no campo.

Com a implementação dessas abordagens analíticas, espera-se ampliar a abrangência do monitoramento de atipicidades, tanto aquelas discutidas na literatura acadêmica quanto as observadas no sistema financeiro. O uso de técnicas modernas de *analytics* para detecção de anomalias permite uma atuação preventiva, com potencial para redução de perdas operacionais, mitigação do risco de conduta e levantamento de informações para retroalimentação dos processos de gestão de risco e de controles internos.

A implementação bem-sucedida desses modelos não só fortalece a segurança contra fraudes,

mas também impulsiona a confiança do cliente e eficiência operacional. Trata-se, portanto, de um fator crítico para garantir a integridade das operações nas instituições financeiras. Tal monitoramento busca encontrar operações financeiras incomuns que possam indicar ilícitos e negócios não sustentáveis como: venda casada, operações feitas apenas para bater meta que depois são canceladas/estornadas; operações sem o conhecimento do cliente; entre outros. Tais operações podem indicar riscos operacionais ou legais e a perdas operacionais para os clientes e para o banco.

O desenvolvimento e aplicação de modelos analíticos para atipicidades em negócios não sustentáveis está de acordo com a Resolução CMN Nº 4.949, de 30 de setembro de 2021, que dispõe sobre princípios e procedimentos a serem adotados no relacionamento com clientes e usuários de produtos e de serviços. Nesta Resolução, fica estabelecido que as instituições devem instituir mecanismos de acompanhamento, de controle e de mitigação de riscos relacionados ao cumprimento da política de relacionamento, onde cada cliente deve ter ofertado seu produto e serviço de acordo com seu perfil.

Dessa forma, há atuação preventiva, com potencial para redução de perdas operacionais, mitigação do risco de conduta e levantamento de informações para retroalimentação dos processos de gestão de risco e de controles internos.

5.2. Revisão da Literatura

5.2.1. Anomalias

A detecção de anomalias em instituições financeiras desempenha um papel crucial na identificação de atividades fraudulentas e na garantia da segurança e integridade dos dados. Para esse fim, utilizam-se modelos analíticos voltados à verificação de atividades incomuns ou comportamentos anômalos em conjuntos de dados. Esses modelos aplicam algoritmos sofisticados baseados em técnicas estatísticas e de aprendizado de máquina para fornecer insights valiosos no combate a fraudes e proteção das operações financeiras

Os modelos analíticos com uso de técnicas de *machine learning* fornecem uma abordagem sistemática e automatizada para detectar essas anomalias. Segundo Chalapathy e Chawla (2019), essas técnicas podem ser combinadas para aumentar a sensibilidade na detecção de anomalias e reduzir falsos positivos.

As principais técnicas de análise de anomalias buscam identificar transações que diferem significativamente do padrão normal de comportamento. Entre os algoritmos mais utilizados nesse contexto, destacam-se o One-Class Support Vector Machine (OC-SVM), Autoencoders, Support

Vector Machines (SVM) e o DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

De modo mais abrangente, Hilal et al. (2022) apresentam estudo que revisa uma variedade de técnicas de detecção de anomalias aplicadas à detecção de fraudes financeiras. Isso inclui métodos estatísticos, aprendizado de máquina, mineração de dados e abordagens baseadas em inteligência artificial. Cada técnica é discutida em termos de sua aplicabilidade, vantagens e desvantagens na detecção de fraudes financeiras. Além disso, o artigo apresenta avanços recentes na detecção de fraudes financeiras, como o uso de técnicas de detecção de anomalias baseadas em redes neurais, algoritmos genéticos e aprendizado de máquina profundo. Essas abordagens avançadas facilitam a identificação de padrões complexos e sutis associados a fraudes financeiras, aprimorando a capacidade de detecção e prevenção de atividades fraudulentas.

Dando suporte ao tema, destaca-se o recente trabalho de Bakumenko e Elragal (2022) que aborda a importância da detecção de anomalias em dados financeiros e apresenta uma revisão abrangente de algoritmos de aprendizado de máquina aplicados a essa tarefa. Os autores argumentam que as técnicas tradicionais podem ser insuficientes diante do volume e da velocidade com que os dados financeiros são gerados atualmente. O estudo ressalta a importância da preparação adequada dos dados, do ajuste de hiperparâmetros e do tratamento de conjuntos de dados desequilibrados no contexto da detecção de anomalias. Deste modo, os autores exploram algoritmos como k-means, Isolation Forest, Autoencoders e One-Class Support Vector Machines (OC-SVM), cujas vantagens e limitações são discutidas em diferentes contextos financeiros.

No que se refere a modelos de detecção de anomalias não supervisionados, objeto deste trabalho, dois modelos foram treinados no estudoi de Bakumenko e Elragal (2022): Isolation Forests e autoencoders. O Isolation Forest foi avaliado com base em *scores* de anomalia e para decidir quais pontos de dados seriam considerados anomalias, um *threshold* (limiar estatístico foi estabelecido a partir desses mesmos scores. Os pontos de dados com *scores* acima desse threshold foram rotulados como anomalias, enquanto aqueles com *scores* abaixo foram considerados normais. Já os autoencoders, utilizam redes neurais para redução de dimensionalidade e reconstrução de dados. O modelo é treinado para reconstruir os dados normais com alta precisão, de modo que desvios na reconstrução de dados pode indicar uma anomalia. A diferença entre os dados originais e os dados reconstruídos foi calculada como um erro de reconstrução para cada ponto de dados no conjunto de teste. Quanto maior esse erro, maior a suspeita de anomalia (Bakumenko e Elragal, 2022).

O trabalho de Zhang et al. (2021) apresenta um framework unificado e abrangente para a

detecção de anomalias em séries temporais multivariadas. O framework proposto, denominado AURORA, integra várias técnicas de detecção e propõe uma solução robusta que combina diferentes técnicas e abordagens. Os resultados mostram sua eficácia na identificação de anomalias, proporcionando insights valiosos para a detecção precoce de eventos anômalos sendo projetado para lidar com os desafios específicos das séries temporais multivariadas, como a dependência entre as variáveis e a heterogeneidade dos padrões de anomalias. Além disso, os experimentos indicam que AURORA supera outras abordagens existentes na identificação precisa de eventos anômalos (Zhang et al., 2021).

O estudo realizado por Zhang et al. (2022) apresenta uma abordagem otimizada para lidar com conjuntos de dados desequilibrados na detecção de fraudes em cartões de crédito. Como as fraudes são relativamente raras em comparação com as transações normais, os autores empregam técnicas de reamostragem, como *oversampling* e *undersampling*, para equilibrar as classes de fraude e não fraude. Isso permite que os modelos de detecção de anomalias aprendam adequadamente com ambos os tipos de instâncias. Os resultados demonstram uma melhora significativa na precisão e na taxa de detecção de fraudes, confirmando a eficácia da abordagem proposta. A precisão e a taxa de detecção de fraudes são significativamente aprimoradas, demonstrando a eficácia da abordagem proposta (Zhang et al., 2022).

Neste contexto, Zhou et al. (2021) aborda o problema da detecção de anomalias em redes dinâmicas atribuídas. Redes dinâmicas atribuídas são redes que evoluem ao longo do tempo e possuem atributos associados a seus nós e arestas. A detecção de anomalias nesse contexto é importante para identificar comportamentos anômalos ou atividades suspeitas em diferentes aspectos das redes.

Os autores propõem um novo método para detecção de anomalias em redes dinâmicas atribuídas, que combina a análise de atributos e a análise de topologia da rede. A abordagem utiliza aprendizado de máquina para modelar e identificar padrões normais de comportamento na rede, permitindo a detecção de anomalias quando ocorrem desvios significativos desses padrões. Além disso, para lidar com a natureza dinâmica da rede, onde os padrões evoluem ao longo do tempo, os autores apresentam uma estratégia baseada em janelas deslizantes, que captura variações no comportamento da rede e ajusta continuamente o modelo de detecção de anomalias. Os resultados experimentais mostram a eficácia do método em diferentes cenários de aplicação (Zhou et al., 2021).

No que se refere a modelos de detecção de anomalias não supervisionado, Zong et al. (2018) apresenta proposta de um modelo numa arquitetura denominada "DAGMM" (*Deep Autoencoding*

Gaussian Mixture Model). A metodologia aplicada envolve a utilização de redes autoencoder para reduzir a dimensionalidade dos dados, seguida pela modelagem da distribuição gaussiana nos espaços de características de baixa dimensão. A principal inovação do DAGMM é seu treinamento de ponta a ponta, que permite que a rede de estimativa e a rede de compressão melhorem mutuamente seu desempenho.

Para avaliar o desempenho do modelo DAGMM, Zong et al. (2018) utilizaram métricas, como a precisão, a recall e o F1-score. Os resultados indicam que que o modelo supera outras técnicas de detecção de anomalias, em benchmarks públicos, alcançando até 14% de melhoria no F1-score. A arquitetura DAGMM demonstrou superioridade na detecção de anomalias em diversos conjuntos de dados, incluindo KDDCUP, Thyroid, Arrhythmia e KDDCUP-Rev. Em particular, a DAGMM mostrou uma alta capacidade de detectar anomalias em situações em que outras técnicas tiveram dificuldade. Como conclusão, os autores afirmam que o modelo DAGMM é uma abordagem eficaz para a detecção de anomalias em dados de alta dimensão. Além disso, destacam a importância de considerar a relação entre os hiperparâmetros na função objetivo do DAGMM para otimizar o desempenho (Zong et al., 2018).

Em um trabalho relevante fora da área de finanças, Schlegl et al. (2017) aborda a detecção de anomalias em imagens médicas utilizando Redes Adversariais Generativas Profundas (*Deep Generative Adversarial Networks* - GANs). O foco principal foi desenvolver um modelo capaz de identificar anomalias em dados não vistos anteriormente, treinando-o exclusivamente em imagens saudáveis. Os autores propõem um modelo, batizado de AnoGAN, que utiliza dados normais para aprender a representação de imagens saudáveis. Com base nesse processo, eles introduzem uma métrica de pontuação de anomalia, que quantifica a diferença entre as imagens reais e as imagens geradas, permitindo a identificação de padrões anômalos.

O desempenho do modelo proposto foi avaliado usando-se a curva ROC (*Receiver Operating Characteristic*) e a Área sob a Curva ROC (AUC) e métricas tradicionais, como precisão, recall, sensibilidade e especificidade. Além disso, os autores destacam a "Perda Residual" que mede a dissimilaridade visual entre a imagem real de consulta e a imagem gerada pelo modelo. Os autores concluem, que o modelo foi capaz de detectar diferentes anomalias conhecidas, como fluido retiniano e HRF (*Hyperreflective Foci*), mesmo sem tê-las visto durante o treinamento.

Além disso, o método utiliza AnoGAN (Anomaly GAN), um modelo baseado em Redes Adversariais Generativas (GANs) para detecção de anomalias em imagens médicas, que se mostrou capaz de gerar imagens médicas realistas e identificar diferenças notáveis entre imagens normais e

anômalas (Schlegl et al., 2017).

Por fim, Goldstein e Uchida (2016), realizaram uma avaliação abrangente de 19 diferentes algoritmos de detecção de anomalias não supervisionados em 10 conjuntos de dados oriundos de diversos domínios de aplicação. A análise desses algoritmos é mais complexa do que na classificação supervisionada tradicional. O estudo abrange diversas abordagens, incluindo aqueles baseados em vizinhos mais próximos, *clustering*, estatísticas, densidade e outros métodos. Os algoritmos foram avaliados segundo métricas como precisão, sensibilidade, estabilidade e velocidade de processamento.

Os autores indicam para avaliação de algoritmos de detecção de anomalias não supervisionados classificar os resultados de acordo com a pontuação da anomalia e depois aplicar iterativamente um limite do primeiro ao último rank. Isso resulta em N valores de tupla (taxa de verdadeiros positivos e taxa de falsos positivos), que formam a curva ROC (Receiver Operating Characteristic). A área sob a curva (AUC), a integral da ROC, é utilizada como métrica de desempenho da detecção. Muitas vezes, um parâmetro k precisa ser ajustado. Goldstein e Uchida (2016) sugerem testar múltiplos valores de k e reportar a média e o desvio padrão da AUC como estratégia para otimização.

Os resultados indicaram variação significativa no desempenho dos algoritmos, dependendo do objetivo do modelo. Os pesquisadores recomendaram o uso de métodos baseados em vizinhos mais próximos, como o k-NN, para tarefas de detecção de anomalias globais, e o LOF para detecção de anomalias locais. No entanto, observou-se que o desempenho dos algoritmos também dependia da natureza do conjunto de dados e das características do problema em questão. Portanto, as recomendações foram adaptadas às necessidades específicas de cada aplicação (Goldstein e Uchida, 2016).

5.2.2. Desafios e Considerações

Ao analisar a bibliografia apresentada, observa-se a recorrente proposta de criação de modelos analíticos híbridos para mitigar o risco operacional, reduzir perdas com fraudes e minimizar a ocorrência de falsos negativos. Dentre esses modelos híbridos, destacam-se aqueles que combinam aprendizado de máquina com regras de negócio, ou que integram dois ou mais modelos analíticos em sequência, visando maior precisão na detecção de fraudes. Também se verifica o forte uso de técnicas avançadas como redes neurais e *deep learning*.

Os principais desafios na detecção de fraudes incluem o desbalanceamento dos dados, a

evolução constante dos métodos fraudulentos, necessidade de detecção em tempo real, interpretabilidade dos modelos e a predominância de abordagens supervisionadas.

O desbalanceamento ocorre porque a maioria das transações é legítima, o que dificulta o treinamento de modelos que consigam identificar fraudes sem gerar muitos falsos positivos ou negligenciar casos suspeitos. Além disso, as fraudes evoluem constantemente, exigindo que os modelos sejam atualizados regularmente para detectar novos padrões e evitar ataques que exploram vulnerabilidades humanas, como a engenharia social. Já a necessidade de detecção em tempo real impõe desafios técnicos, pois os modelos precisam ser rápidos e eficientes sem comprometer a precisão.

A interpretabilidade dos modelos também é um obstáculo, especialmente quando se utilizam técnicas avançadas de inteligência artificial, que muitas vezes funcionam como "caixas-pretas". Isso pode dificultar a aceitação por parte de instituições financeiras e reguladores, que precisam confiar nos modelos para a tomada de decisões.

Outro ponto crítico é a predominância de modelos supervisionados, que exigem grandes volumes de dados rotulados para o treinamento. No entanto, a obtenção desses rótulos é cara, demorada e, em alguns casos, impraticável, visto que fraudes são eventos raros e em constante mutação. Essa dependência de dados rotulados limita a aplicabilidade dos modelos supervisionados em cenários reais, tornando necessário o desenvolvimento de alternativas mais flexíveis.

Nesse contexto, modelos não supervisionados surgem como uma solução promissora, pois conseguem identificar anomalias sem a necessidade de rótulos. Técnicas como *clustering* e autoencoders permitem que os modelos aprendam padrões normais dos dados e detectem desvios que possam indicar atividades fraudulentas. Além disso, esses modelos são mais adaptáveis a novos tipos de fraudes e podem operar em tempo real, tornando-os mais viáveis para cenários dinâmicos.

5.3. **Metodologia**

Paralelamente às fraudes, as anomalias representam potencial ameaça à estabilidade e segurança dos bancos. Anomalias podem surgir em diversas formas, desde transações atípicas até comportamentos de clientes que se desviam dos padrões habituais. Identificar tais anomalias é essencial para prevenir atividades suspeitas, como corrupção, lavagem de dinheiro, financiamento ao terrorismo e outros delitos financeiros. Portanto, a capacidade de detectar anomalias em tempo real e adotar medidas proativas para mitigar os riscos associados é crucial para a estabilidade e a credibilidade de qualquer instituição financeira.

As abordagens de detecção de anomalias são aplicadas para identificar comportamentos atípicos ou não conformes nas transações financeiras. Esses modelos utilizam técnicas como a distância de Mahalanobis, *Isolation Forest*, SVM, *Auto Encoders* e densidade local para identificar padrões incomuns nos dados. Embora sejam eficazes na detecção de anomalias, esses modelos podem gerar um número significativo de falsos positivos devido à dificuldade em definir limites claros entre comportamentos normais e anormais (Bakumenko & Elragal, 2022; Hilal et. al., 2022; Zhang et al., 2022)

No trabalho deste capítulo, aplicaremos técnicas de aprendizado de máquina com abordagens de detecção de anomalias seguindo os modelos não supervisionados.

Após revisar a literatura, foram desenvolvidos modelos analíticos não supervisionados focando em produtos de capitalização. Para tal, foi aplicado três modelos analíticos de detecção de anomalias, apresentados a seguir, ao qual sua descrição foi apresentada no capítulo dois desta tese:

- *Isolation Forest*: A metodologia Isolation Forest fundamenta-se na construção de múltiplas árvores de decisão para identificar anomalias. Nela, as observações atípicas são aquelas que requerem um número significativamente menor de partições para serem isoladas em um nó da árvore (Liu et al., 2008).
- COPOD: é uma metodologia para identificação de anomalias que utiliza cópulas para gerar um score de atipicidade. A utilização de cópulas nos permite utilizar distribuições de probabilidade marginais de cada variável separadamente da estrutura de dependência presente entre as variáveis presentes no conjunto de dados (Li et al, 2020).
- *Fully connected Autoencoder*: É um tipo especial de rede neural que busca aprender uma representação eficiente do conjunto de dados, mesmo em cenários de alta dimensionalidade. Nesse contexto, estudos anteriores, como os de Legrand et al. (2018) e Xu et al. (2018), destacam métodos promissores para lidar com tais desafios.

Por se tratar de modelos de *machine learning* não supervisionados, a avaliação de performance dos modelos será feita sob o aspecto negocial comparativamente a técnicas e processo atualmente estabelecidos na instituição financeira. Esta avaliação, dentro deste arcabouço experimental, será em três etapas destacadas a seguir:

1. Interpretabilidade SHAP do modelo: Esta etapa, aplicada ao algoritmo de Isolation Forest, explica a contribuição de cada variável na saída interpretativa de cada uma

das variáveis utilizadas no modelo. Representa uma primeira verificação do alinhamento negocial das descobertas.

2. Consistência Negocial: Envolve a avaliação de especialistas em fraudes e controles internos. Eles analisarão os casos mais anômalos utilizando os gráficos SHAP value e SHAP Waterfall, com o objetivo de validar o sentido negocial das variáveis apontadas como mais importantes, contrastando-as com os métodos tradicionais de controle.
3. Avaliação da consistência com as práticas de monitoramento de movimentações não convencionais: Consiste no cruzamento dos casos considerados anômalos de cada modelo com a regra de monitoramento atualmente existentes nas equipes de fiscalização e controles internos específica do produto aplicado. Este processo visa verificar o ganho de eficiência e possibilitou a geração de uma 'base-teste' com dados rotulados para análise estatística. Tal etapa é considerada crítica e importante para o processo, visto que atualmente demanda processamento manual e limitação de recursos, não sendo possível fiscalizar cem porcento das propostas. Assim o atual método baseia-se em amostragem estratificada por conveniência, ou seja, por regras que possui hoje taxa de acerto que varia de 15 a 20% das amostras. A ótica apresentada é de identificar perda operacional para o cliente e/ou banco.

Nesta última etapa, após avaliação manual dos especialistas, será realizada avaliação de performance estatística dos modelos por meio das métricas de acurácia, precisão e recall, conforme aplicado por Zhang et al. (2022) e Zong et al. (2018) e com base na área sob a curva ROC (AUC) conforme Schlegl et al. (2017) e Goldstein e Uchida (2016). A realização desta avaliação é essencial, visto que, apesar de envolver técnicas de modelos não supervisionados, estamos diante de uma tarefa de classificação. O resultado do monitoramento será uma proxy do rótulo de classificação aplicada para inferir os indicadores de performance estatísticos típicos de modelos supervisionados.

Deste modo, serão realizadas comparações das variáveis mais importantes e análise do sentido negocial. O objetivo é confrontar os resultados obtidos pelos modelos desenvolvidos com os dos sistemas convencionais de detecção de fraudes da instituição financeira, visando discutir a aplicabilidade e eficácia desses modelos em ambientes reais.

Os modelos de detecção de anomalias terão como objetivos a mitigação de perdas operacionais com o uso de inteligência analítica visando atuar como foco no monitoramento de padrões atípicos na comercialização de produtos de seguridade.

Tal monitoramento busca encontrar operações financeiras incomuns que possam indicar ilícitos ou serem inviáveis, tais como: venda casada, operações feitas apenas para bater meta que depois são canceladas/estornadas; operações sem o conhecimento do cliente; entre outros. Tais operações podem indicar riscos operacionais ou legais, levando a perdas para os clientes e para o banco. Dessa forma, a atuação do monitoramento é preventiva, com potencial para redução de perdas operacionais, mitigação do risco de conduta e o levantamento de informações para retroalimentação dos processos de gestão de risco e de controles internos.

Tendo como premissa que a maior parte das contratações realizadas nas dependências da instituição financeira buscam alinhamento com os padrões éticos relacionados à Política de Relacionamento preconizada pelo Bacen conforme Resolução CMN Nº 4.949, foram criados escores de anomalia para quantificação do distanciamento do padrão usual para cada uma das contratações.

5.3.1. Base de dados

A amostra utilizada para desenvolvimento e treino dos modelos deste trabalho é composta por operações de capitalização contratadas de clientes pessoa física, disponibilizadas anonimizadas por uma intuição financeira nacional. Por se tratar de abordagem não supervisionada, ou seja, sem uma variável resposta, todas as operações são analisadas com o objetivo de se encontrar aquelas que são mais propensas a serem anomalias.

O período de coleta de dados para treino dos modelos foi de janeiro/2022 a julho/2022. A contagem total de 417 mil operações no período. Para cada um dos contratos, foram levantadas informações sobre as partes envolvidas na comercialização dos produtos.

Após as análises iniciais, processamentos dos dados e testes de consistência, foram identificadas 25 variáveis, utilizadas para aplicação dos modelos. Elas estão apresentadas no Quadro 5.1 de forma consolidada para a confidencialidade do processo interno da instituição.

Quadro 5.1 - Variáveis de Modelagem

Tema das variáveis	Quantidade	Descrição Geral do Tema
Métricas de Cancelamento (Churn)	8	Variáveis que medem taxas, percentuais ou indicadores de cancelamento de contratos/serviços por clientes e dependentes.
Métricas Operacionais e Internas	8	Variáveis relacionadas a cancelamentos, captação e perdas financeiras associadas a funcionários ou gerentes.
Comportamento e Engajamento do Cliente	3	Variáveis que descrevem interações do cliente com a instituição, como tempo de relacionamento e histórico de reclamações.

Dados Demográficos do Cliente	3	Variáveis que caracterizam o perfil socioeconômico e etário dos clientes
Métricas Financeiras e de Perda (Foco Cliente)	3	Variáveis que quantificam perdas financeiras acumuladas ou eventos de resgate associados ao cliente.

Fonte: Elaboração do autor.

Tabela 5.1 - Análise descritivas das variáveis – Base de treino dos modelos

Variable	mean	std	min	25%	50%	75%	max	IQR
1	3,4203	6,71685	0	0	0	3	30	3
2	0,00374	0,069614	0	0	0	0	5	0
3	0,546133	0,907992	0	0	0	0,86	12	0,86
4	0,004892	0,024868	0	0	0	0	1	0
5	0,006001	0,012434	0	0	0	0,008621	0,269231	0,008621
6	0,003556	0,021556	0	0	0	0	1	0
7	0,004479	0,010548	0	0	0	0,006283	0,263393	0,006283
8	0,003735	0,015071	0	0	0	0	0,613095	0
9	0,184902	1,23274	0	0	0	0	12	0
10	30,6359	50,68056	0	0	11	43	731	43
11	0,29524	1,460998	0	0	0	0	135	0
12	0,005341	0,025516	0	0	0	0	1	0
13	0,005271	0,020064	0	0	0	0,00578	1	0,00578
14	10317,08	40764,72	0	2098,005	4596,405	9375,538	3333276	7277,533
15	58,80535	15,92075	16	48	61	71	103	23
16	3,688894	1,552705	0	3	3	5	9	2
17	0,004617	0,021732	0	0	0	0	1	0
18	0,003391	0,018454	0	0	0	0	1	0
19	0,001925	0,043832	0	0	0	0	1	0
20	36,32097	723,4677	0	0	0	0	74810,56	0
21	0,001339	0,032177	0	0	0	0	5,24292	0
22	360,712	34,78419	0	365	365	365	365	0
23	998,3492	2766,951	0	0	200	970,85	95616,67	970,85
24	0,005179	0,021506	0	0	0,001168	0,00537	3,42647	0,00537
25	0,003898	0,009918	0	0	0,00039	0,004305	0,759246	0,004305

Fonte: Elaborado pelo autor

O Quadro 5.1 fornece uma visão estruturada das variáveis selecionadas para a modelagem, categorizando-as em cinco temas principais. Observa-se uma concentração de variáveis nas categorias "Métricas de Cancelamento (Churn)" e "Métricas Operacionais e Internas", ambas com 8 variáveis, indicando um foco significativo na mensuração direta do fenômeno de churn e nos fatores internos da organização que podem influenciá-lo. As categorias subsequentes, "Comportamento e Engajamento do Cliente", "Dados Demográficos do Cliente" e "Métricas Financeiras e de Perda

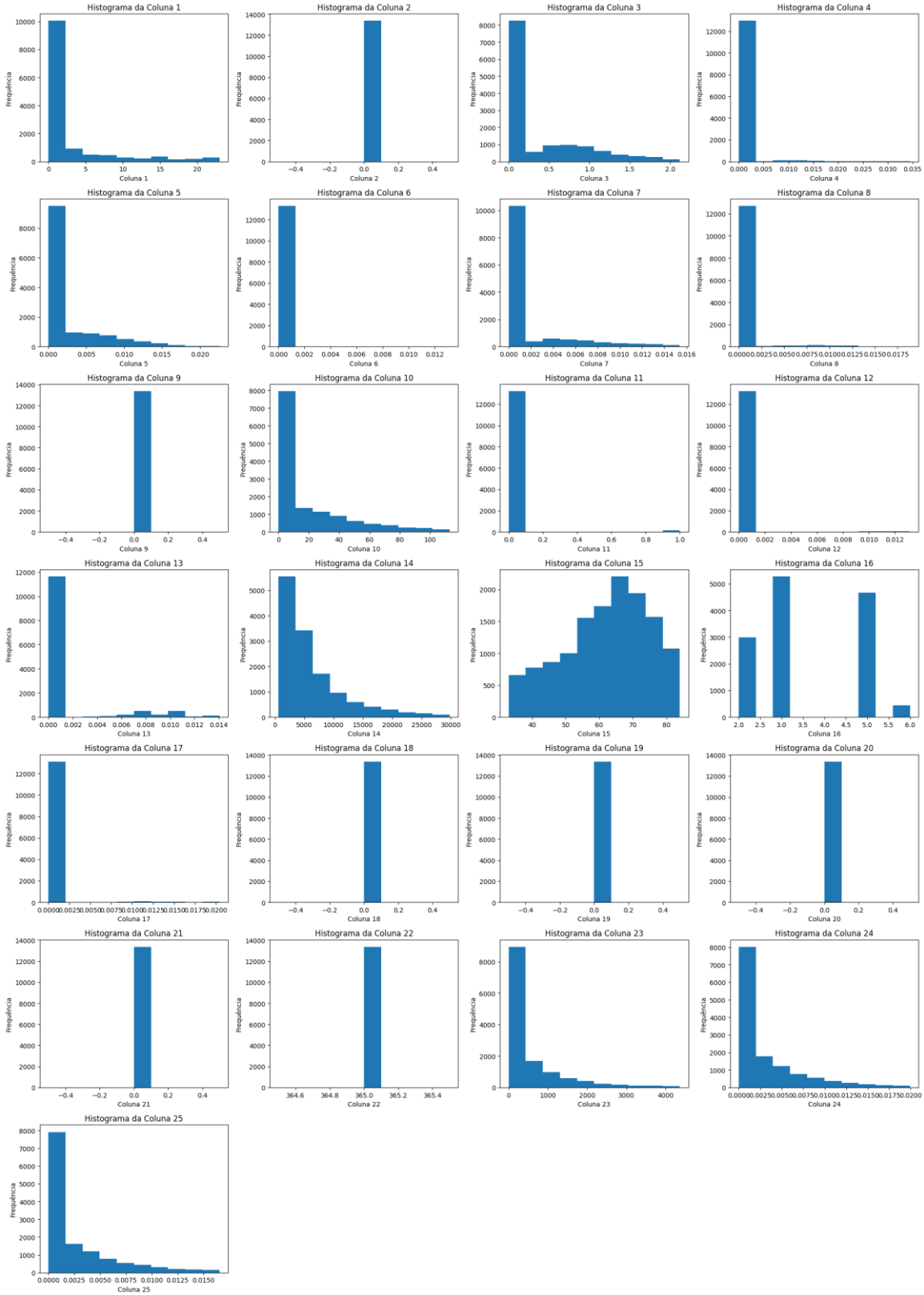
(Foco Cliente)", contêm 3 variáveis cada, sugerindo a incorporação de aspectos relacionados às interações do cliente, seu perfil socioeconômico e o impacto financeiro associado, totalizando 25 variáveis. Essa distribuição temática evidencia uma abordagem multifacetada para compreender os determinantes do cancelamento, abrangendo desde indicadores diretos até características contextuais e operacionais.

A análise descritiva exposta na Tabela 5.1 revela características cruciais da base de treino para a modelagem. Evidencia-se uma acentuada heterogeneidade nos dados, marcada por expressivas disparidades de escala entre variáveis e uma forte predominância de assimetria positiva – muitas com mediana de valor zerado e valores médios/máximos consideravelmente superiores. Essa configuração sugere a presença de outliers e distribuições não normais, tornando imperativas etapas de pré-processamento, como escalonamento e transformações adequadas, para garantir a validade e otimizar o desempenho dos modelos subsequentes.

Com base na análise visual dos histogramas fornecidos na Figura 5.1, observa-se uma confirmação gráfica robusta das características identificadas na análise descritiva numérica. A maioria expressiva das variáveis demonstra uma severa assimetria à direita, visualizada pela concentração maciça de dados na primeira classe (usualmente em torno de zero) e uma cauda longa que se estende para valores elevados, sugerindo a presença de outliers e a não aderência à distribuição normal. As distintas escalas nos eixos horizontais entre os gráficos também são evidentes, reforçando a heterogeneidade nas ordens de grandeza.

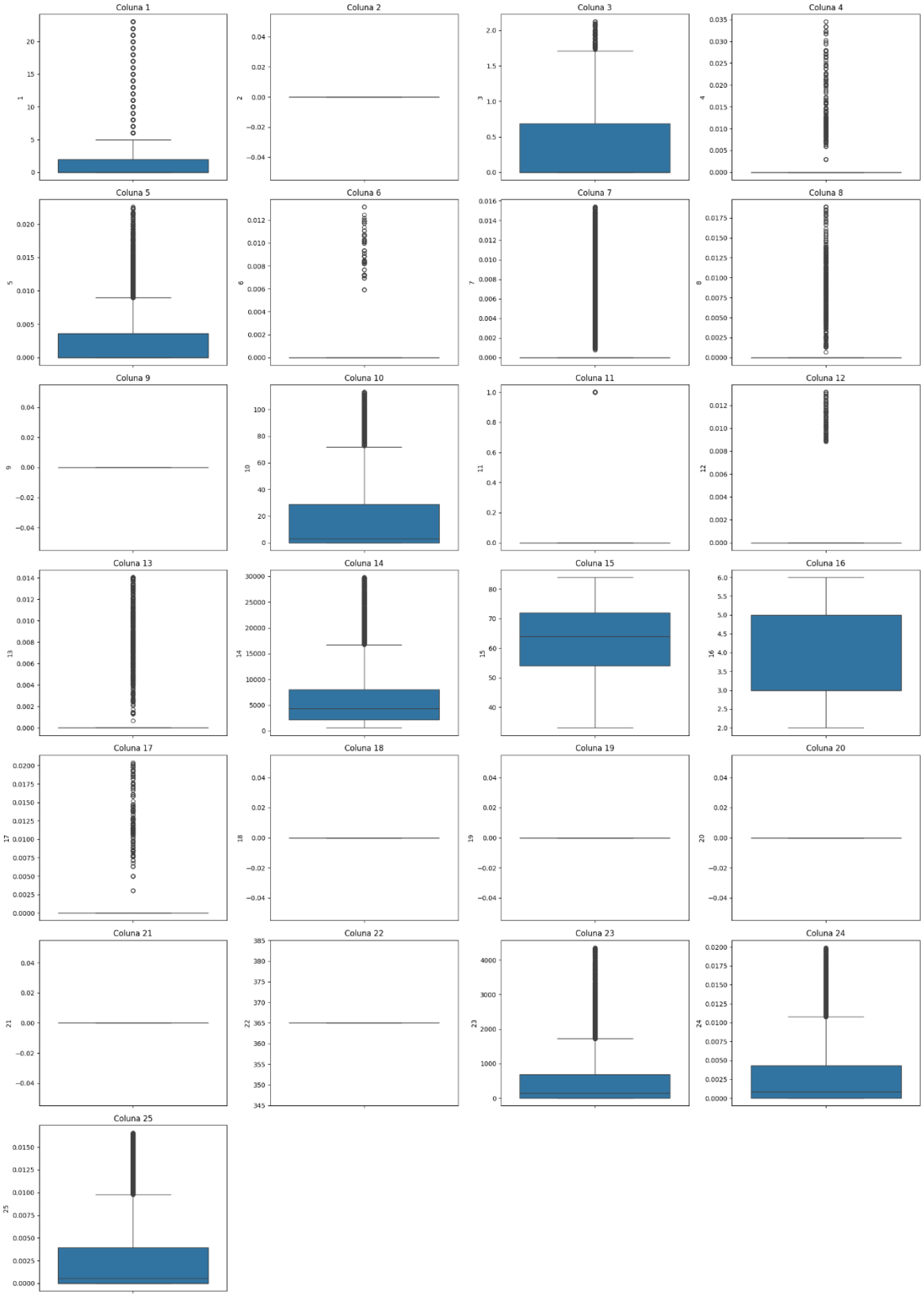
Esta representação gráfica sublinha a criticidade da etapa de pré-processamento, indicando a necessidade de transformações para corrigir a assimetria e de escalonamento para normalizar as diferentes magnitudes das variáveis antes da aplicação em algoritmos de modelagem.

Figura 5.1 – Histograma distribuição das variáveis de modelagem



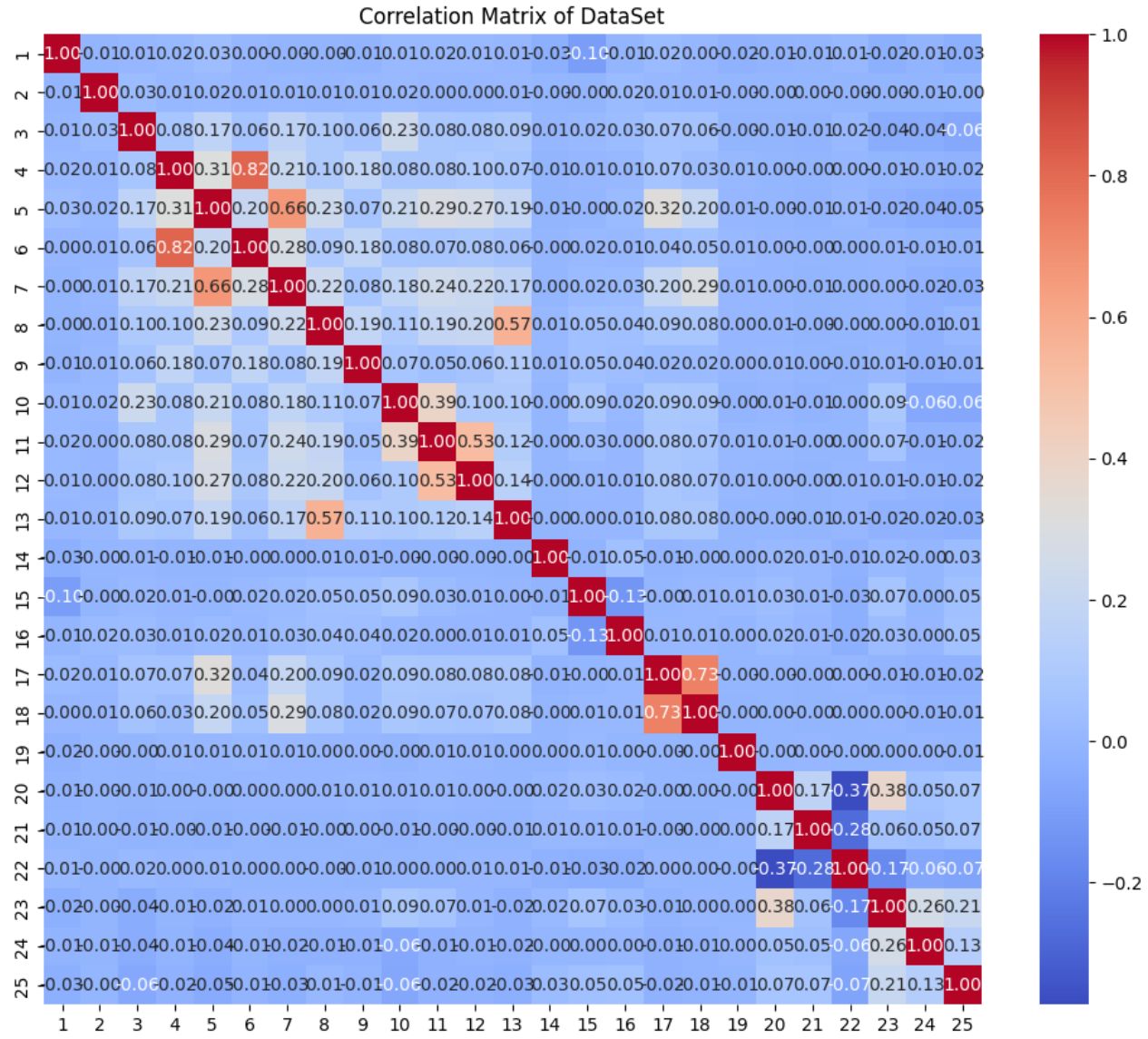
Fonte: Elaboração própria.

Figura 5.2 – Estatísticas descritivas – Box Plot – Base de treino dos modelos



Fonte: Elaboração própria.

Figura 5.3 – Matriz de correlação entre as variáveis



Fonte: Elaborado pelo autor

A matriz de correlações visualizada na Figura 5.3 indica que, de modo geral, as relações lineares entre a maioria das variáveis são fracas, com coeficientes predominantemente próximos de zero. No entanto, identificam-se focos de correlação positiva moderada a forte, como entre as variáveis 4 e 6 (0.82), 5 e 7 (0.66), e 17 e 18 (0.73), além de associações relevantes entre 8 e 13 (0.57) e 11 e 12 (0.53). Correlações negativas significativas são virtualmente inexistentes. Essa configuração aponta para a existência de multicollinearidade pontual entre alguns pares de variáveis, um fator que merece atenção na etapa de seleção de atributos e construção do modelo para assegurar sua estabilidade e interpretabilidade.

5.4. **Resultados**

Esta sessão tem o objetivo de apresentar os resultados dos modelos desenvolvidos para cálculo de escores de anomalia em contratos formalizados de operações de produto de seguridade. Será apresentado os *dataset* utilizado, incluindo suas variáveis, e os resultados obtidos sob as três perspectivas de avaliação previamente apresentadas na metodologia. As métricas de performance dos modelos, como de precisão, recall e AUC dos modelos são apresentados e discutidos de forma complementar na sessão de consistência com práticas de monitoramento de movimentações não convencionais.

5.4.1. Desenvolvimento dos modelos e discussão dos resultados

Para cada segmento, foi desenvolvido um modelo ensemble. Os três modelos escolhidos para compor o *Ensemble*, foram: *Isolation Forest*, *Copula-Based Outlier Detection* (COPOD) e *Fully connected AutoEncoder* conforme técnicas e algoritmos apresentados na seção 3 desta tese.

Iniciou-se o processo de modelagem com a seleção de atributos relevantes para a identificação de operações não convencionais. Para a execução desta etapa, escolheu-se o modelo de *Isolation Forest* por ser uma metodologia implementada com a ferramenta interpretativa SHAP, conforme Lundberg et al (2017). A análise dos *summary plots* permite observar o comportamento das variáveis criadas em relação a detecção de anomalias, passando uma maior segurança na escolha de informações relevantes dentro do contexto do projeto.

Após a seleção das informações relevantes, estas foram utilizadas para ajustar um novo *Isolation Forest*. Em seguida, esse mesmo conjunto de variáveis relevantes foi empregado no ajuste dos modelos COPOD e AutoEncoder, completando o ensemble.

Os escores foram construídos com base em variáveis relacionadas à promoção da sustentabilidade e alinhamento com o objetivo da instituição financeira. Os dados utilizados compreendem informações sobre produtos, clientes, funcionários, dependências, administradores de dependência e relacionamento com os clientes.

Desenvolvidos os três modelos individualizados, foram seguidos os três passos apresentados na metodologia para avaliação da qualidade do modelo para o uso proposto: i) interpretabilidade *Shap*; ii) consistência negocial com *Shap whaterfall*; e iii) consistência com as práticas de monitoramento de movimentações não convencionais.

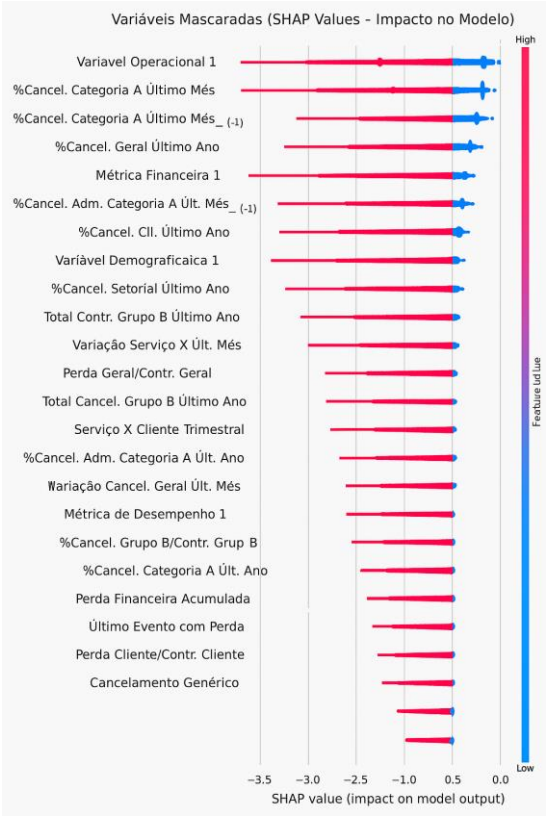
5.4.1.1.Interpretabilidade via SHAP

Observando o eixo vertical, notamos a listagem das variáveis utilizadas no modelo, enquanto o eixo horizontal indica o impacto que cada variável exerce na saída do modelo, em termos de valor de SHAP. As cores das marcas de dispersão (vermelho e azul) representam, de modo geral, valores altos e baixos das variáveis, respectivamente. Quanto maior a amplitude dos valores de SHAP para uma variável, maior a sua relevância para o comportamento do modelo na identificação de pontos atípicos.

A Figura 5.4 apresenta os resultados da análise de SHAP (*Shapley Additive Explanations*), que revela a contribuição individual de cada variável para a detecção de anomalias.

Observando o eixo vertical, notamos a listagem das variáveis utilizadas no modelo, enquanto o eixo horizontal indica o impacto que cada variável exerce na saída do modelo, em termos de valor de SHAP. As cores indicam o valor da variável: vermelho para valores altos e azul para valores baixos. Quanto maior a amplitude dos valores de SHAP para uma variável, maior a sua relevância para o comportamento do modelo na identificação de pontos atípicos.

Figura 5.4 – SHAP Value Modelo



Fonte: Elaboração própria.

A análise dos valores SHAP indica que as variáveis com maior impacto negativo no modelo (próximas de -3,5) estão fortemente associadas a fatores operacionais e comportamentais. As principais influências negativas incluem indicadores relacionados a processos internos, como atrasos, cancelamentos em categorias específicas e métricas financeiras críticas. Esses atributos sugerem que o modelo é altamente sensível a falhas em fluxos operacionais e à recorrência de perdas ou desistências, o que reflete diretamente na sua capacidade de identificar padrões anômalos.

Além disso, variáveis associadas a cancelamentos administrativos e contratuais apresentam contribuições significativas, apontando que descontinuidades abruptas, independentemente do tipo de cliente ou serviço, têm papel relevante na geração de alertas no modelo.

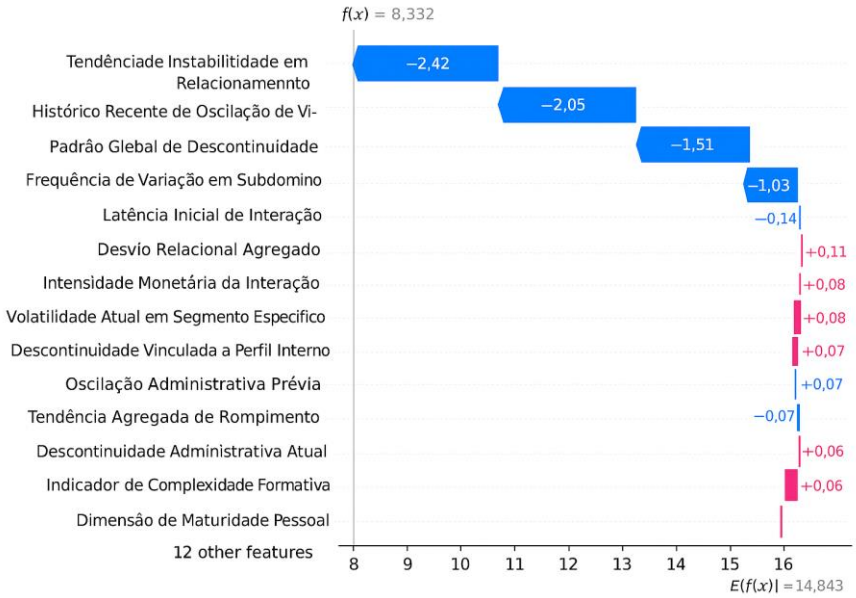
Por outro lado, variáveis com valores SHAP mais próximos de zero demonstram menor impacto na predição. Entre elas, estão características demográficas, indicadores de desempenho positivo ou eventos históricos pontuais de perda, cuja influência sobre o desfecho do modelo é limitada. Essas observações reforçam que o foco da inteligência preditiva está concentrado em evidências atuais e recorrentes de risco.

Também é importante notar a presença de variáveis que capturam oscilações comportamentais recentes, como variações mensais em serviços ou volumes de cancelamento por grupos específicos. Tais variações podem sinalizar tendências emergentes e alterações no perfil de risco ao longo do tempo.

5.4.1.2.Consistência Negocial das Variáveis com *Shap whaterfall*

Para validar a coerência das anomalias detectadas com as regras de negócio, utilizou-se a visualização *Shap Waterfall* aplicada ao modelo *Isolation Forest*. A Figura 5.5 e Figura 5.6 registram os resultados dessa abordagem, detalhando os fatores de maior escore de atipicidade calculada pelo modelo *AutoEncoder* e a maior atipicidade condicionada à classificação de anomalia do modelo COPOD.

Figura 5.5 - Exemplo SHAP Waterfall – Maior Anomalia -Autoencoder



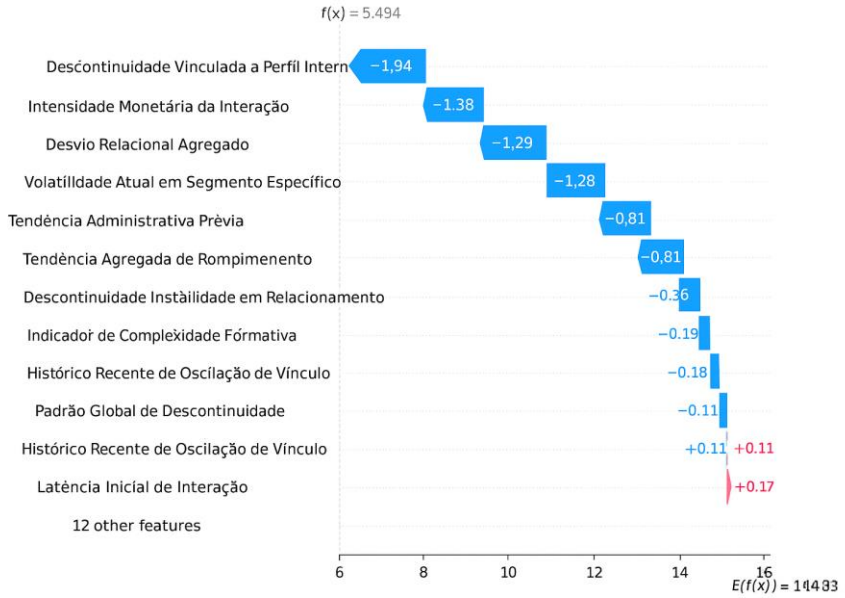
Fonte: Elaboração própria.

A análise do gráfico SHAP Waterfall apresentado na Figura 5.5 - Exemplo SHAP Waterfall – Maior Anomalia -Autoencoder, referente à maior anomalia detectada por um AutoEncoder, exemplifica os fatores críticos que contribuíram para a classificação atípica da instância. O valor esperado do modelo, E(f(x)) = 14.843, contrasta significativamente com a saída f(x) = 8.332, indicando um desvio acentuado característico de anomalia. Os valores SHAP negativos destacam as variáveis que mais reduziram a previsão do modelo, sendo “Tendência de Instabilidade em Relacionamento" (SHAP = -2.42) e "Histórico Recente de Oscilação de Vida" (SHAP = -2.05) as principais responsáveis por essa redução. Essas variáveis sugerem problemas operacionais, como atrasos em processos críticos e taxas elevadas de cancelamento, que se alinharam a padrões anômalos. Além disso, variáveis como "Descontinuidade Administrativa Atual" e "Dimensão de Maturidade Pessoal" apresentaram impacto neutro ou positivo limitado (SHAP próximo de zero ou positivo), insuficientes para compensar os efeitos negativos dominantes.

A técnica SHAP Waterfall, neste contexto, descreve visualmente como cada variável contribui para a diferença entre a saída esperada e o resultado observado, oferecendo transparência à decisão do AutoEncoder. No cenário de detecção de anomalias, essa interpretabilidade é crucial para validar se o modelo identifica padrões coerentes com riscos operacionais ou financeiros. O AutoEncoder, ao reconstruir os dados com alto erro nesta instância, sinaliza uma discrepância estrutural, enquanto o SHAP quantifica o papel específico de cada variável nessa discrepância. Essa combinação permite não apenas a identificação da anomalia, mas também a priorização de ações

corretivas com base nas variáveis mais impactantes, reforçando a eficácia do modelo em cenários práticos de gestão de riscos.

Figura 5.6 - Exemplo SHAP Waterfall – Maior Anomalia – COPOD



Fonte: Elaboração própria.

A análise da **Erro! Fonte de referência não encontrada.**Figura 5.6, referente ao exemplo *SHAP Waterfall* aplicado à maior anomalia detectada pelo modelo COPOD, evidencia os principais fatores que contribuíram para a classificação atípica da instância. O valor esperado E(f(x)) = 14.843 contrasta drasticamente com f(x) = 5.484, indicando um desvio significativo. As variáveis com maior impacto negativo são " Descontinuidade Vinculada a Perfil Interno" (SHAP = -1.94) e "Intensidade Monetária da Interação" (SHAP = -1.38).

Esses valores destacam que contratos com prejuízos acumulados e intensidade monetária altos são críticos para a detecção da anomalia. Adicionalmente, métricas percentuais como "indicador de Complexidade" (0.104) e "Desvio Relacional Agregado" (1.29) reforçam padrões de risco operacional e financeiro. Variáveis como "Latência Inicial de Interação" (+.11) e Historico Recente de Oscilação de Vínculo (-0,11)" têm contribuições neutras ou pouco expressivas, confirmando que o modelo prioriza desvios associados a custos diretos e inadimplência.

A partir da análise individualizada entende-se que os resultados dos testes do modelo COPOD retornou valores mais consistentes com a expectativa dos analistas sobre a possibilidade de anomalias. Enquanto o caso da Figura 5.5**Erro! Fonte de referência não encontrada.** foi considerado anômalo por conta de percentuais de cancelamentos elevados em uma baixa quantidade de contratações, o da Figura 5.6 considerou um comportamento atípico o conjunto entre as diversas

variáveis.

5.4.1.3. Consistência com Práticas de Monitoramento de Movimentações não Convencionais

Por se tratar de um modelo não supervisionado, a ausência de rótulos (*labels*) para as anomalias é uma das maiores dificuldades em sistemas reais de detecção de fraudes ou anomalias. Não ter um "*ground truth*" dificulta o treinamento do modelo (ou a escolha de hiperparâmetros) e principalmente a avaliação quantitativa do desempenho.

Uma das estratégias mais viáveis é a amostragem dirigida com especialistas. Nela, subconjuntos de dados (estratificados ou não) são selecionados aleatoriamente para análise humana. Por exemplo, em cenários antifraude, analistas especializados podem classificar manualmente registros como "normais" ou "anômalos", criando um '*testbed'* limitado. Embora o custo operacional seja elevado — especialmente em grandes volumes —, mesmo amostras modestas permitem estimar métricas parciais e ajustar o modelo. Contudo, é crucial reconhecer que não há soluções universais: a avaliação permanece incompleta sem anomalias previamente validadas.

Nesse contexto, alocar recursos para rotulação pontual, priorizando casos críticos ou suspeitos, emerge como uma prática equilibrada. Ao direcionar esforços para rotular um subconjunto estratégico, obtém-se uma base para calcular indicadores de desempenho e refinar o algoritmo iterativamente, mitigando riscos de supervisão enviesada ou incompleta.

Considerando a verificação manual realizada pelos especialistas, para geração de uma base-teste, identificou-se entre as operações confirmadas como anômalas aquelas que efetivamente resultaram em resgates de perdas operacionais para os clientes no resgate. Para efeito de verificação manual, forma analisadas operações que foram indicadas como anômalas pelas três modelos. Além disso, como critério negocial de marcação, foram consideradas casos frequentemente seguidos de recontratações do mesmo produto, evidenciando um descasamento de interesses e caracterizando práticas não sustentáveis.

Esse padrão reforça a importância da análise de monitoramento sob a perspectiva da perda operacional, permitindo a identificação de movimentações não convencionais que podem comprometer a integridade do sistema e a segurança financeira dos clientes.

Deste modo, dentre as operações que foram marcadas como anômalas, 39,5% realmente geraram perdas, ou seja, a taxa de acerto dos modelos de anomalias com perdas operacionais, foi de

0.394464. Em outras palavras, de todas as operações que ele apontou como casos de anomalias nas operações, aproximadamente quarenta porcento geraram perda financeira para os clientes.

Este critério negocial é caracterizado pela política de relacionamento com o cliente, que visa ofertar produtos de acordo com o perfil e necessidade do cliente. Um dos indícios de negócios não sustentáveis é quando o produto é resgatado/cancelado antes do prazo e geram perdas financeiras para o cliente. Conforme informado pelos especialistas, a taxa de aproximadamente 40% é bastante superior ao método tradicional de amostra estratificada de controles internos, com uso de regras, que possui média de sucesso de 15% a 20%.

A detecção de anomalias em cenários não supervisionados representa um desafio significativo, especialmente quando há limitações de recursos para a verificação manual das ocorrências identificadas. Nesse contexto, a estratégia adotada priorizou a precisão do modelo, buscando reduzir a quantidade de falsos positivos e, assim, minimizar a necessidade de revisões manuais excessivas.

Os modelos para anomalias adotados neste experimento buscaram melhor precisão embora o recall fique baixo. Isso significa que das operações que o modelo aponta como anômalas, a maior parte deles é realmente passível de perdas operacionais. Em contraposição, há outras perdas operacionais de valor do cliente que não são detectados pelo modelo indicadas pelo baixo recall. Por fim, para avaliação final sob a ótica da perda operacional, foi considerado o horizonte temporal dentro de um período de até 15 meses após a proposta, o que pode ter trazer desvio da marcação de baixo recall, pois muitos resgates do produto são exercidos em 12 meses, gerando perda operacional, mas não necessariamente uma anomalia para o modelo. Os resultados da matriz de confusão entre as indicações do modelo de anomalias versus as apuradas pelos especialistas constam na Tabela 5.2.

Tabela 5.2 - Performance dos modelos de anomalias após verificação dos especialistas

Métrica	**Valor**
Precisão	0.313543
Recall	0.092194
F1-Score	0.142490
Acurácia	0.612729
Especificidade	0.891790
AUC	0.491992

Fonte: elaborado pelo autor.

Os resultados apresentados na Tabela 5.2 indicam que a precisão do modelo foi de 0,3135, sugerindo que aproximadamente 31% das anomalias identificadas realmente correspondem a padrões atípicos confirmados pelos especialistas. Entretanto, o recall de 0,0922 revela que o modelo conseguiu detectar apenas 9,2% das perdas operacionais no produto existentes, demonstrando uma limitação na abrangência da detecção.

A especificidade de 0,8918 indica que a maioria dos casos normais foi corretamente classificada, enquanto a acurácia global de 61,27% reforça a importância de analisar métricas mais específicas, uma vez que a predominância de casos normais pode inflacionar esse indicador.

Dada a priorização da precisão, o modelo evita um volume excessivo de falsos positivos, o que é crucial quando há restrições para a revisão manual das anomalias detectadas. No entanto, a baixa pontuação F1-Score de 0,1425 evidencia um desequilíbrio entre precisão e recall, indicando que muitos outros casos de perdas não foram identificadas. Esse resultado sugere a necessidade de ajustes no modelo para melhorar a detecção das anomalias sem comprometer excessivamente a precisão.

O AUC (0.4919), um indicador da capacidade do modelo de distinguir padrões normais de anômalos, mostra um desempenho próximo ao aleatório, o que sugere dificuldades na separação entre esses grupos. Esse resultado pode estar relacionado à natureza dos dados e à ausência de um treinamento supervisionado, que poderia fornecer padrões mais claros para a classificação.

Em análise do sentido negocial do observado alto número de resgates gerando perda operacional para o cliente, os especialistas alegaram que há diversos motivos para que este fenômeno ocorra sem caracterizar uma operação anômala ou não sustentável.

Estratégias como o refinamento dos limiares de decisão, o uso de técnicas de aprendizado semissupervisionado ou a combinação de diferentes algoritmos de detecção podem ser consideradas para aprimorar o desempenho global do sistema. Dessa forma, a abordagem adotada deve equilibrar a detecção eficiente de anomalias e a viabilidade operacional da revisão manual, garantindo um impacto positivo na mitigação de riscos.

5.5. **Conclusão modelos de anomalias**

Este trabalho aplicou modelos de detecção de anomalias ao desenvolver modelos empíricos não supervisionados, voltados para produtos de seguridade no setor financeiro. Utilizando dados de

um banco nacional e técnicas de aprendizado de máquina, os modelos propostos buscam superar as limitações das abordagens tradicionais de monitoramento interno, oferecendo soluções inovadoras para desafios práticos identificados tanto na academia quanto na indústria. A implementação bem-sucedida desses sistemas não apenas amplia a capacidade de identificar movimentações não convencionais, como perdas financeiras e operacionais, vendas casadas ou operações canceladas, mas também fortalece a segurança contra fraudes, a confiança dos clientes e a eficiência operacional, alinhando-se às demandas por métodos mais robustos e adaptáveis.

A carência de dados rotulados ("*ground truth*") em modelos de detecção de anomalias não supervisionados representa um desafio crítico, pois inviabiliza a validação direta de desempenho e a calibração precisa dos algoritmos. Sem referências claras de normalidade ou anomalia, métricas tradicionais — como precisão e recall — tornam-se inacessíveis, comprometendo a confiança nas previsões do modelo. Essa limitação exige abordagens criativas para simular um conjunto de validação confiável, garantindo que o modelo não apenas identifique padrões atípicos, mas também se alinhe a expectativas práticas do domínio de aplicação.

A análise de interpretabilidade via *SHAP Values* demonstrou, de forma transparente, a contribuição individual de variáveis na detecção de anomalias, destacando que métricas operacionais críticas — como atrasos em etapas processuais, taxas elevadas de desistência em períodos recentes e perdas financeiras acumuladas — tiveram impacto negativo significativo (valores próximos de -3,5), sinalizando padrões associados a riscos imediatos. Por outro lado, atributos demográficos ou de desempenho positivo mostraram relevância marginal (SHAP próximo de zero), enquanto oscilações temporais em indicadores setoriais reforçaram a sensibilidade do modelo a mudanças comportamentais abruptas. A técnica SHAP evidenciou a priorização do algoritmo por fatores de risco operacional e financeiro, traduzindo complexidades técnicas em insights acionáveis para gestão proativa de anomalias, alinhada a estratégias de mitigação de perdas e otimização de processos críticos. Tais apontamentos são condizentes com o sentido negocial esperado para as variáveis.

Em sequência, a análise de consistência negocial por meio do método *SHAP Waterfall*, identificou os principais fatores associados às anomalias detectadas pelos modelos. No primeiro caso, a discrepância entre o valor esperado (E(f(x)) = 14,843) e a saída observada (f(x) = 8,332) foi influenciada por variáveis operacionais críticas, como indicadores de cancelamento em períodos recentes (SHAP = -2,42) e atrasos em etapas processuais-chave (SHAP = -1,03), refletindo possíveis falhas em processos sistêmicos. No segundo modelo, o desvio significativo foi atribuído

principalmente a métricas financeiras críticas, como perdas contratuais acumuladas (SHAP = -1,94) e indicadores de inadimplência recorrente (SHAP = -1,38). Em ambos os cenários, variáveis com valores absolutos elevados (e.g., transações de alto montante) tiveram impacto reduzido, evidenciando que os modelos priorizam padrões de risco associados a custos diretos e interrupções operacionais, em detrimento de atributos contextuais ou métricas neutras.

Neste contexto, a abordagem adotada para a detecção de anomalias priorizou a precisão do modelo, reduzindo a quantidade de falsos positivos e minimizando a necessidade de revisões manuais excessivas. Os resultados indicam que, embora a precisão tenha sido relativamente alta (0,3135), o recall foi baixo (0,0922), sugerindo que apenas uma pequena parcela das perdas operacionais foi efetivamente identificada. A especificidade de 0,8918 demonstra que a maioria dos casos normais foi corretamente classificada, enquanto a acurácia global de 61,27% reforça a importância de considerar métricas específicas para avaliar o desempenho real do modelo. No entanto, a baixa pontuação F1-Score (0,1425) revela um desequilíbrio entre precisão e recall, indicando que muitas perdas operacionais aparecem, mas o modelo não os classifica diretamente como anomalia. Tal fato pode ser explicado pelo critério da equipe de especialistas sobre a marcação da perda operacional e financeira pelo cliente que pode ser diferente do critério monetário, onde muitos resgates podem ter sido realizados de forma consciente. Junta-se ao fato do horizonte temporal do modelo (15 meses) poder apresentar desvio na marcação de resgates visto ações automáticas após 12 meses.

Apesar das limitações inerentes à ausência de dados rotulados, a integração de amostragem especializada e pseudo-supervisão oferecem um caminho viável para validar modelos não supervisionados. Essas abordagens não substituem um "*ground truth*" ideal, mas permitem criar marcos de referência dinâmicos, ajustáveis conforme novos dados ou feedback humano são incorporados. Ao combinar automação com intervenção estratégica, é possível transformar desafios de rotulação em oportunidades para desenvolver sistemas mais robustos e adaptáveis, capazes de evoluir em ambientes de incerteza e complexidade crescente.

Dessa forma, é essencial buscar um equilíbrio entre a detecção eficiente e a viabilidade operacional, garantindo um impacto positivo na mitigação de riscos. Tais modelos apresentaram ganhos consistentes pois permitiram identificar variáveis na detecção de anomalias visando a prevenção de perdas operacionais em negócios não sustentáveis. Adicionalmente, foi verificadaa consistência negocial, indicando futuras monitoramentos e retroalimentando o processo de indução da contratação deste tipo de operações, bem como trouxe aprimoramento na precisão e redução de

custos com ganhos operacionais no trabalho de amostragem.

Diante desse cenário, é necessário aprimorar o modelo para melhorar a detecção das anomalias sem comprometer excessivamente a precisão. A análise do fenômeno revelou que nem todos os resgates que geram perdas operacionais são, de fato, anomalias ou operações insustentáveis, o que reforça a complexidade do problema.

Para otimizar o desempenho do sistema, pode-se explorar estratégias como o ajuste dos limiares de decisão, a adoção de aprendizado semissupervisionado e a combinação de diferentes algoritmos. Para futuras pesquisas, sugerimos aprimorar o modelo atual com finetuning, utilizando o feedback dos especialistas para calibrar as múltiplas camadas da rede neural do modelo de anomalias, garantindo que ele reconheça os casos específicos detectados por eles e atenda melhor aos objetivos da área de controles internos.

Ao integrar técnicas modernas de ciência de dados, o estudo visa aprimorar a abrangência do monitoramento de riscos operacionais e legais, como práticas ilícitas ou negócios insustentáveis, que podem gerar perdas para clientes e instituições. A análise empírica dos resultados demonstra o potencial dos modelos para atuar preventivamente, mitigando riscos de conduta e retroalimentando processos de gestão de risco e controles internos. Assim, o trabalho não apenas confirma a eficácia das abordagens propostas, mas também estabelece um marco para a aplicação de inteligência artificial em contextos financeiros complexos, reforçando a necessidade de inovação contínua na detecção proativa de anomalias.

5. Considerações Finais

A presente tese aborda a utilização de inteligência artificial na premente necessidade do setor financeiro por métodos e modelos inovadores para a detecção de fraudes e comportamentos atípicos, com ênfase na superação do desafio do desbalanceamento de classes. A motivação central reside na constante evolução das táticas fraudulentas, exigindo soluções computacionais avançadas para proteger instituições e clientes. A pesquisa investigou e desenvolveu abordagens que mitiguem as limitações dos métodos tradicionais, explorando o potencial da inteligência artificial, especialmente no que concerne à geração de dados sintéticos para o balanceamento de classes e à aplicação de técnicas de aprendizado não supervisionado.

Para atingir seu objetivo principal esta tese estruturou-se em uma abordagem de múltiplos artigos e nos estudos de caso, utilizou-se de base de dados de uma instituição financeira nacional. Inicialmente,

Os resultados apresentados nos artigos que compõem esta tese oferecem contribuições significativas para o campo da ciência de dados aplicada à detecção de fraudes. A exploração de técnicas de IA generativa para o balanceamento de dados representa um avanço promissor para lidar com a escassez de exemplos da classe minoritária, um problema recorrente na área. Tal técnica também pode ser aplicada em diversos outros fins, como disponibilização de *dataset* de forma a obedecer a privacidade dos dados. Adicionalmente, o desenvolvimento de um modelo de detecção de anomalias não supervisionado demonstra o potencial de identificar padrões atípicos sem a necessidade de dados previamente rotulados, o que pode ser particularmente útil em cenários onde a rotulagem é dispendiosa ou inviável.

A revisão da literatura realizada, objeto do capítulo 3, visou mapear a produção científica voltada à detecção de fraudes em bancos, adotando uma abordagem que combinou revisão sistemática da literatura e análise de redes complexas. Com base em 227 estudos até dezembro de 2023, foram identificadas conexões entre os trabalhos por meio de citações, coautorias e a recorrência de palavras-chave, além de uma classificação detalhada segundo doze características. Essa metodologia permitiu não apenas a síntese do conhecimento atual, mas também a identificação de lacunas críticas, evidenciando a necessidade de novas abordagens que integrem métodos preditivos e análises descritivas para aprimorar a eficiência dos sistemas de segurança bancária.

Os resultados indicam que técnicas de aprendizado de máquina, frequentemente combinadas com métodos estatísticos ou regras de negócios em modelos híbridos, são as mais

utilizadas para a detecção de fraudes. Além disso, o estudo destaca a ausência de publicações voltadas à prevenção de fraudes decorrentes da engenharia social, bem como destacou os desafios comuns na literatura sobre o tema, em destaque a base de dados desbalanceada, a necessidade de dados rotulados para modelos de classificação e interpretabilidade. Em síntese, a revisão sistemática não só consolida os métodos comprovadamente eficazes na detecção e análise de fraudes, mas também propõe uma agenda de pesquisa que incentive o desenvolvimento de modelos generativos e abordagens inovadoras para lidar com os desafios dos dados desbalanceados e ampliar o escopo da prevenção de fraudes.

No desenvolvimento de modelos para atuar na geração de dados sintéticos apresentado no capítulo 4, a análise comparativa dos métodos tradicionais de inteligência artificial para geração de dados sintéticos evidenciou inovações significativas, como o desenvolvimento do modelo Aurora de IA Generativa.

Os modelos SMOTE, GAN e VAE demonstraram eficácia na geração de dados sintéticos, cada um com vantagens específicas. O GAN apresentou a melhor correspondência estatística com os dados originais, enquanto o SMOTE se destacou na preservação das correlações entre variáveis, conforme evidenciado pelas métricas de Jensen-Shannon e Kullback-Leibler.

A integração de modelos de linguagem de grande porte (LLMs) com abordagens de recuperação aumentada (RAG) utilizando modelos como GPT-4o e Gemini 1.5 Pro, superou limitações técnicas tradicionais, gerando dados com alta similaridade e baixa divergência estatística. O uso de frameworks como LangChain e plataformas como LMStudio contribuiu para a criação de conjuntos de dados mais balanceados, impulsionando os métodos de oversampling e aprimorando o desempenho preditivo.

O small language model (SLM) Aurora desenvolvido destacou-se pelo seu caráter inovador ao ser especificamente refinado por meio de fine-tuning, direcionado para a criação de dados sintéticos de alta fidelidade. Essa abordagem personalizada possibilitou a replicação não apenas das características estatísticas, mas também da estrutura intrínseca dos dados reais, comprovada por métricas quantitativas robustas. O modelo Aurora demonstrou alta eficácia na geração de dados sintéticos, com baixa divergência estatística (Jensen-Shannon e Kullback-Leibler) e elevada similaridade com os dados reais. Esses resultados reforçam seu potencial como uma ferramenta promissora para aplicações preditivas, contribuindo significativamente para o avanço da literatura sobre modelagem sintética.

Destaca-se também como um achado relevante deste estudo, que a aplicabilidade local de um modelo de IA Generativa SLM para criação de dados sintéticos mostrou-se eficaz, mesmo diante de limitações de hardware e arquitetura. Verificou-se que os resultados são influenciados pela limitação da quantidade tokens disponíveis por modelo, ambiente de aplicação dos modelos, memória, docker de IDE de código, GPUs e outras questões de infraestrutura.

Por fim, no capítulo 5 este trabalho desenvolveu e aplicou modelos empíricos não supervisionados de detecção de anomalias em produtos de seguridade no setor financeiro, utilizando dados de um banco nacional e técnicas avançadas de aprendizado de máquina. Os métodos propostos buscam superar as limitações das abordagens tradicionais de monitoramento interno, ampliando a capacidade de identificar movimentações atípicas que podem resultar em perdas financeiras e operacionais. A análise detalhada, utilizando técnicas como SHAP e SHAP Waterfall, evidenciou a importância de variáveis operacionais e financeiras críticas, demonstrando de forma transparente a influência de indicadores como atrasos processuais, cancelamentos recentes e perdas acumuladas, além de revelar o impacto de fatores de risco e a necessidade de calibrar os modelos para alinhar os resultados às expectativas práticas do domínio.

Além disso, a pesquisa destacou a importância de equilibrar a precisão do modelo com a viabilidade operacional, evidenciando desafios como a ausência de dados rotulados que dificulta a validação direta dos algoritmos. Estratégias inovadoras, como a integração de amostragem especializada, pseudo-supervisão e ajustes de limiares de decisão, foram propostas para aprimorar a detecção das anomalias sem comprometer a precisão, bem como a utilização de métodos semissupervisionados e fine-tuning com *feedback* de especialistas. Tais avanços não só fortalecem a mitigação de riscos operacionais e a segurança contra fraudes, mas também estabelecem um marco para a aplicação contínua de inteligência artificial em contextos financeiros complexos, promovendo melhorias significativas nos processos de controle e gestão de riscos.

Por fim, cabe ressaltar que este trabalho de pesquisa buscou atender aos requisitos de originalidade, ineditismo e relevância ao investigar e propor soluções para os desafios persistentes na detecção de fraudes bancárias. As contribuições da tese abrangem desde o levantamento do estado da arte até o desenvolvimento e aplicação empírica de modelos inovadores, com foco no tratamento do desbalanceamento de classes através da geração de

dados sintéticos com IA generativa e na exploração de modelos não supervisionados. Os resultados obtidos possuem implicações práticas relevantes para a indústria financeira, auxiliando, com uso de inteligência artificial no aprimoramento de modelos de mitigação de fraudes, além de fomentar futuras pesquisas na área.

6. Referências Bibliográficas

Adeojo, J. (2024). Can synthetic data boost machine learning performance? Towards Data Science. Disponível em https://towardsdatascience.com/can-synthetic-data-boost-machine-learning-performance-6b4041e75dda. Acessado em 07/07/2024.

Akila , S., & Reddy, EUA (2017, novembro). Risk based bagged ensemble (RBE) para detecção de fraudes em cartões de crédito. Em 2017 Conferência Internacional sobre Computação Inventiva e Informática (ICICI) (pp. 670-674). IEEE.

Ashfaq, T., Khalid, R., Yahaya, AS, Aslam, S., Azar, AT, Alsafari , S., & Hameed, IA (2022). Um mecanismo eficiente de detecção de fraude baseado em aprendizado de máquina e blockchain. Sensores, 22(19), 7162.

Babu, SK, & Vasavi, S. (2017). Análise preditiva como serviço sobre evasão fiscal usando o processo de regressão gaussiana. Helix, 7(5), 1988-1993.

Bakumenko , A., & Elragal , A. (2022). Detecção de anomalias em dados financeiros usando algoritmos de aprendizado de máquina. Sistemas, 10(5), 130.

Bank for International Settlements - BIS (2021). Revisions to the Principles for the Sound Management of Operational Risk. Basel Committee on Banking Supervision

Barbosa, FT, Lira, AB, Oliveira, OBD, Santos, LL, Santos, IO, Barbosa, LT, ... & Sousa-Rodrigues, CFD (2019). Tutorial para realização de revisão sistemática e meta-análise com estudos de anestesia intervencionista. Revista Brasileira de Anestesiologia, 69, 299-306.

Bej, S., Davtyan, N., Wolfien, M., Nassar, M., & Wolkenhauer, O. (2021). LoRAS: An oversampling approach for imbalanced datasets. Machine Learning, 110, 279-301.

Bernard, P., El Mekkaoui De Freitas, N., & Maillet, BB (2019). Um indicador de detecção de fraude financeira para investidores: um IDeA . Annals of Operations Research, 1-24.

Bessis, J. (2011). *Risk management in banking*. John Wiley & Sons.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer google schola, 2, 5-43.

Bonacich, P. (1987). Power and centrality: A family of measures. American journal of sociology, 92(5), 1170-1182.

Böse, B., Avasarala, B., Tirthapura, S., Chung, YY e Steiner, D. (2017). Detectando ameaças internas usando rabanete: um sistema para detecção de anomalias em tempo real em fluxos de dados heterogêneos. IEEE Systems Journal, 11(2), 471-482.

Boyle, DM, DeZoort , FT e Hermanson, DR (2015). O efeito do uso de modelos alternativos de

fraude nos julgamentos de risco de fraude dos auditores. Journal of Accounting and Public Policy, 34(6), 578-596.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees: Taylor & Francis.

Brookshear, J. G. (2013). Ciência da Computação-: Uma Visão Abrangente. Bookman Editora.

Brown, T., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical statistics for data scientists: 50+ essential concepts using R and Python. O'Reilly Media.

CAF. (2023) Mapa da Identidade e da Fraude - 3º trimestre de 2023. Disponível em: https://www.caf.io/recursos-pt/mapa-da-identidade-e-da-fraude-3o-trimestre-de-2023. Acesso em: 1 jul. 2024.

Cai, X., Xiao, M., Ning, Z., & Zhou, Y. (2023, December). Resolving the imbalance issue in hierarchical disciplinary topic inference via llm-based data augmentation. In 2023 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 1424-1429). IEEE.

Can, B., Yavuz, AG, Karsligil , EM, & Guvensan , MA (2020). Um olhar mais atento sobre as características das transações fraudulentas com cartões. Acesso IEEE, 8, 166095-166109.

Carminati, M., Caron, R., Maggi, F., Epifani , I., & Zanero , S. (2015). BankSealer : Um sistema de apoio à decisão para análise e investigação de fraudes bancárias online. computadores e segurança, 53, 175-186.

Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407.

Chang, C. Y., Hsu, M. T., Esposito, E. X., & Tseng, Y. J. (2013). Oversampling to overcome overfitting: exploring the relationship between data set composition, molecular descriptors, and predictive modeling methods. Journal of chemical information and modeling, 53(4), 958-971.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Cheah, P. C. Y., Yang, Y., & Lee, B. G. (2023). Enhancing financial fraud detection through addressing class imbalance using hybrid SMOTE-GAN techniques. International Journal of Financial Studies, 11(3), 110. https://doi.org/10.3390/ijfs11030110

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

785-794).

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. P. (2024). A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, *57*(6), 137. https://doi.org/10.1007/s10462-024-10759-6

Cui, J., Yan, C., & Wang, C. (2021). LEMBRE-SE : Criação de perfil de comportamento multicontextual baseado em incorporação de métrica de classificação para detecção de fraudes bancárias on-line. IEEE Transactions on Computational Social Systems, 8(3), 643-654.

Damenu , TK, & Beaumont, C. (2017). Analisando a segurança da informação em um banco usando a metodologia de sistemas flexíveis. Informação e Segurança de Computadores, 25(3), 240-258.

Damodaran, A. (2012). Investment valuation: Tools and techniques for determining the value of any asset. John Wiley & Sons.

Darwish, SM (2020). Uma abordagem inteligente de detecção de fraudes em cartões de crédito baseada na fusão semântica de dois classificadores. Soft Computing, 24(2), 1243-1253.

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). Mathematics for machine learning. Cambridge University Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Ding, Y., Ni, J., Wang, Y., Liu, G., & Cheng, Y. (2023). Credit Card Fraud Detection Based on Improved Variational Autoencoder Generative Adversarial Network. Journal of Intelligent & Fuzzy Systems, 44(3), 4587-4596.

Eom, G., & Byeon, H. (2023). Searching for optimal oversampling to process imbalanced data: Generative adversarial networks and synthetic minority over-sampling technique. Mathematics, 11, 3605. https://doi.org/10.3390/math11163605

Esen , MF, Bilgic , E., & Basdas , U. (2019). Como detectar informações privilegiadas corporativas ilegais? Uma abordagem de mineração de dados para detectar transações internas suspeitas. Sistemas Inteligentes em Contabilidade, Finanças e Gestão, 26(2), 60-70.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).

Fahim, M., & Sillitti , A. (2019). Técnicas de detecção, análise e previsão de anomalias em

ambiente iot : uma revisão sistemática da literatura. Acesso IEEE, 7, 81664-81681.

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3), 238-247.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 36(4), 193-202.

Furnell, S., & Clarke, N. (2004). Towards a framework for understanding electronic deception and scams. In Proceedings of the 6th Australian Information Warfare and Security Conference.

Galvão, MCB, & Ricarte, ILM (2019). Revisão sistemática da literatura: conceituação, produção e publicação. Logeion: Filosofia da informação, 6(1), 57-73.

Gartner. (2023). O que é inteligência artificial? Disponível em< https://www.gartner.com.br/pt-br/temas/inteligencia-artificial>

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc."

Goldberg, Y. (2016). A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, 57, 345-420.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

Hadnagy, C. (2011). Social Engineering: The Art of Human Hacking. Wiley.

Han, J., Kamber, M., & Mining, D. (2006). Concepts and techniques. Morgan kaufmann, 340, 94104-3205.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Hewapathirana , IU (2019). Detecção de mudanças em redes dinâmicas atribuídas. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1286.

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial Fraud:: A Review of Anomaly Detection Techniques and Recent Advances.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics

(Volume 1: Long Papers). *arXiv preprint arXiv:1801.06146*.

Hsin , YY, Dai, TS, Ti , YW, Huang, MC, Chiang, TH e Liu, LC (2022). Engenharia de recursos e estratégias de reamostragem para fraude de transferência de fundos com dados de transação limitados e um modi operandi não homogêneo no tempo . Acesso IEEE, 10, 86101-86116.

Hull, J. (2012). Gestão de riscos e instituições financeiras,+ Web Site (Vol. 733). John Wiley & Filhos.

Jorion , P. (2000). Lições de gestão de risco da gestão de capital de longo prazo. Gestão financeira europeia, 6(3), 277-300.

Kandpal, N., et al. (2023, July). Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning* (pp. 15696-15707). PMLR.

Khodabandehlou , S., & Golpayegani , SAH (2022). Detecção de manipulação de mercado: Uma revisão sistemática da literatura. Sistemas Especialistas com Aplicações, 118330.

Kim, AC, Kim, S., Park, WH, & Lee, DH (2014). Modelo de detecção de fraudes e crimes financeiros usando análise forense de malware. Ferramentas e aplicativos multimídia, 68, 479-496.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kumar, G., Muckley , CB, Pham, L., & Ryan, D. (2019). Modelos de alerta para fraude podem proteger os clientes idosos de uma instituição financeira?. O Jornal Europeu de Finanças, 25(17), 1683-1707.

Labanca , D., Primerano , L., Markland-Montgomery, M., Polino , M., Carminati, M., & Zanero , S. (2022). Amaretto: uma estrutura de aprendizado ativo para detecção de lavagem de dinheiro. Acesso IEEE, 10, 41720-41739.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Legrand A., Niepceron B., Cournier A. and Trannois H., "Study of Autoencoder Neural Networks for Anomaly Detection in Connected Buildings," 2018 IEEE Global Conference on Internet of Things (GCIoT), 2018, pp. 1-5.

Lewis, P. et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS*, 33, 9459–9474. *Advances in neural information processing systems*, *33*, 9459-9474.

Li, T., Kou, G., Peng, Y., & Philip, SY (2021). Uma abordagem integrada de detecção, otimização e interpretação de cluster para dados financeiros. Transações IEEE sobre cibernética, 52(12), 13848-13861.

Li, Zheng & Zhao, Yue & Botta, Nicola & Ionescu, Cezar & Hu, Xiyang. (2020). COPOD: Copula-Based Outlier Detection

Liang, P. et al. (2022). Holistic Evaluation of Language Models. *arXiv preprint arXiv:2211.09110*.

Liberati , A., Altman, DG, Tetzlaff , J., Mulrow, C., Gøtzsche , PC, Ioannidis, JP, ... & Moher, D. (2009). A declaração PRISMA para relatar revisões sistemáticas e meta-análises de estudos que avaliam intervenções de saúde: explicação e elaboração. Anais de medicina interna, 151(4), W-65.

Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation forest.” Data Mining, 2008. ICDM'08. Eighth IEEE International Conference

Liu, Y., Li, Z., Huang, J., & Wu, S. (2012). Analytical models of machine learning are automated systems that can accurately and efficiently identify anomalies in large volumes of data. Journal of Data Science, 10(3), 521-536.

Locher, C. (2005). Methodologies for evaluating information security investments-What Basel II can change in the financial industry. ECIS 2005 Proceedings. 122. https://aisel.aisnet.org/ecis2005/122

Lokanan , M., Tran, V., & Vuong, NH (2019). Detecção de anomalias em demonstrações financeiras usando algoritmo de aprendizado de máquina: o caso de empresas listadas vietnamitas. Jornal Asiático de Pesquisa Contábil, 4(2), 181-201.

Lundberg, Scott & Lee, Su-In. (2017). A Unified Approach to Interpreting Model Predictions.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Marco, R., Ahmad, S. S. S., & Ahmad, S. (2023). Improving Conditional Variational Autoencoder with Resampling Strategies for Regression Synthetic Project Generation. International Journal of Intelligent Engineering and Systems, 16(4). DOI: 10.22266/ijies2023.0831.30.

Martin, NC, Santos, LRD, & Dias Filho, JM (2004). Governança empresarial, riscos e controles internos: a emergência de um novo modelo de controladoria. Revista Contabilidade & Finanças, 15, 22/07.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5, 115-133.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mitchell, T. M. (1997). Does machine learning really work?. AI magazine, 18(3), 11-11.

Mitnick, K., & Simon, W. (2002). The Art of Deception: Controlling the Human Element of Security. Wiley.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020, April). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In 2020 11th international conference on information and communication systems (ICICS) (pp. 243-248). IEEE.

Mosavi, A., Faghan, Y., Ghamisi, P., Duan, P., Ardabili, SF, Salwana, E., & Band, SS (2020). Revisão abrangente de métodos de aprendizado por reforço profundo e aplicações em economia. Matemática, 8(10), 1640.

Mu, E., & Carroll, J. (2016). Desenvolvimento de um modelo de decisão de risco de fraude para priorizar casos de risco de fraude em empresas de manufatura. Jornal Internacional de Economia da Produção, 173, 30-42.

Mujahid, M., Kına, E. R. O. L., Rustam, F., Villar, M. G., Alvarado, E. S., De La Torre Diez, I., & Ashraf, I. (2024). Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering. Journal of Big Data, 11(1), 87.

Munir, U., & Manarvi , I. (2010). Avaliação de risco de segurança da informação para o setor bancário - Um estudo de caso de bancos paquistaneses. Jornal Global de Ciência da Computação e Tecnologia, 10(10), 44-55.

Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

Nami, S., & Shajari, M. (2018). Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors. Expert Systems with Applications, 110, 381-392.

Nanduri , J., Liu, YW, Yang, K., & Jia, Y. (2020). Detecção de fraude de comércio eletrônico por meio de ilhas de fraude e modelo de aprendizado de máquina multicamada . Em Avanços em Informação e Comunicação: Anais da Conferência sobre o Futuro da Informação e Comunicação de 2020 (FICC), Volume 2 (pp. 556-570). Springer International Publishing.

Nesvijevskaia , A., Ouillade , S., Guilmin , P., & Zucker, JD (2021). O trade-off precisão versus interpretabilidade no modelo de detecção de fraudes. Dados e Política, 3, e12.

Neunzig, C., Möllensiep, D., Hartmann, M., Kuhlenkötter, B., Möller, M., & Schulz, J. (2023). Enhanced classification of hydraulic testing of directional control valves with synthetic data generation. Production Engineering, 17(5), 669-678.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining

techniques in financial fraud detection: A classification framework and an academic review of literature. Decision support systems, 50(3), 559-569.

Nicholls, J., Kuppa, A., & Le-Khac, NA (2021). Cibercrime financeiro: uma pesquisa abrangente de abordagens de aprendizado profundo para lidar com o cenário de crimes financeiros em evolução. Ieee Access, 9, 163965-163986.

Nonnenmacher, J., & Gómez, JM (2021). Detecção não supervisionada de anomalias para auditoria interna: revisão da literatura e agenda de pesquisa. Jornal Internacional de Pesquisa em Contabilidade Digital, 21.

Oliveira, AL, & Meira, SR (2006). Detecção de novidades em séries temporais através de previsão de redes neurais com intervalos de confiança robustos. Neurocomputing, 70(1-3), 79-92.

Omidi , M., Min, Q., Moradinaftchali , V., & Piri, M. (2019). A eficácia de métodos preditivos em fraudes em demonstrações financeiras. Dinâmica discreta na natureza e na sociedade, 2019.

Pandey, M. (2010). Um modelo para gerenciamento de risco de fraude online usando validação de transação. The Journal of Operational Risk, 5(1), 49.

Passos de Oliveira, M., Gonzalez, P. L., Tessmann, M. S., & de Abreu Pereira Uhr, D. (2022). The greatest co-authorships of finance theory literature (1896–2006): scientometrics based on complex networks. Scientometrics, 127(10), 5841-5862.

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, *2*(11), 559-572.

Pourhabibi , T., Ong, KL, Kam, BH e Boo, YL (2020). Detecção de fraude: Uma revisão sistemática da literatura sobre abordagens de detecção de anomalias baseadas em gráficos. Sistemas de Apoio à Decisão, 133, 113303.

PRABHA, N., & MANIMEKALAI, S. (2021). UMA REVISÃO ABRANGENTE DAS TÉCNICAS DE RASTREAMENTO E DETECÇÃO DE ATIVIDADES FRAUDULENTAS.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal

of Machine Learning Research, 21(1), 5485-5551.

Rahman, M., Ming, TH, Baigh , TA, & Sarker , M. (2021). Adoção de inteligência artificial em serviços bancários: uma análise empírica. International Journal of Emerging Markets, (ahead-of-print).

Ravi, H. (2021). Inovação em bancos: fusão de inteligência artificial e blockchain. Asia Pacific Journal of Innovation and Entrepreneurship, 15(1), 51-61.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP*, 3982–3992.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.

Russell, S. J., & Norvig, P. (2010). Artificial intelligence a modern approach. London.

Saha , P., Bose, I., & Mahanti , A. (2016). Um esquema baseado em conhecimento para avaliação de risco no processamento de empréstimos por bancos. Sistemas de Apoio à Decisão, 84, 78-88.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

Schmidhuber, M., & Kruschwitz, U. (2024). Llm-based synthetic datasets: Applications and limitations in toxicity detection. LREC-COLING 2024, 37.

Seagriff , T., & Lord, S. (2011). Técnicas de pesquisa operacional suave: usos atuais e futuros. Na conferência YoungOR 17.

Serasa Experian. (2021). Pesquisa Global de Identidade e Fraude 2021 Proteger e promover o engajamento dos clientes na nova era digital. Disponível em: https://www.serasaexperian.com.br/images-cms/wp-content/uploads/2021/06/Pesquisa-Global-de-Identidade-e-Fraude-2021.pdf. Acesso em: 1 jul. 2024.

Shen, T., et al. (2024, August). Retrieval-augmented retrieval: Large language models are strong zero-shot retriever. In Findings of the Association for Computational Linguistics ACL 2024 (pp. 15933-15946).

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., & Downs, J. (2009). Who falls for phish?:

A demographic analysis of phishing susceptibility and effectiveness of interventions. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 373-382).

Siddaway , AP, Wood, AM e Hedges, LV (2019). Como fazer uma revisão sistemática: um guia de boas práticas para conduzir e relatar revisões narrativas, meta-análises e meta-sínteses. Revisão anual de psicologia, 70, 747-770.

Silva, W., Kimura, H., & Sobreiro , VA (2017). Uma análise da literatura sobre risco financeiro sistêmico: uma pesquisa. Journal of Financial Stability, 28, 91-114.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. nature, 529(7587), 484-489.

Smith, J., Johnson, R., & Brown, L. (2024). Synthetic Oversampling Theory and Practical Using LLMs. Journal of Artificial Intelligence Research, 58(3), 123-145. https://doi.org/10.1016/j.jair.2024.03.015.

Stajano, F., & Wilson, P. (2010). Understanding scam victims: Seven principles for systems security. Communications of the ACM, 53(3), 134-140.

Stojanović , B., Božić , J., Hofer-Schmitz, K., Nahrgang , K., Weber, A., Badii , A., ... & Runevic , J. (2021). Siga a trilha: Aprendizado de máquina para detecção de fraude em aplicações Fintech. Sensores, 21(5), 1594.

Sudjianto , A., Nair, S., Yuan, M., Zhang, A., Kern, D., & Cela -Díaz, F. (2010). Métodos estatísticos para combater crimes financeiros. Technometrics , 52(1), 5-19.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Tanaka, F. H. K. dos S., & Aranha, C. (2019). Data Augmentation Using GANs. 2019 16th International Conference on Machine Learning and Applications (ICMLA), 10-15. IEEE. https://doi.org/10.1109/ICMLA.2019.00010

Teng, HW, & Lee, M. (2019). Procedimentos de estimativa do uso de cinco métodos alternativos de aprendizado de máquina para prever a inadimplência do cartão de crédito. Revisão de Políticas e Mercados Financeiros da Bacia do Pacífico, 22(03), 1950021.

Ti , YW, Hsin , YY, Dai, TS, Huang, MC e Liu, LC (2022). Geração de recursos e comparação de contribuições para detecção de fraudes eletrônicas. Scientific Reports, 12(1), 18042.

TUKEY, John W. **Exploratory data analysis**. 1977.

Turing, A. M. (2009). Computing machinery and intelligence (pp. 23-65). Springer Netherlands.

van Wageningen- Kessels , F., Van Lint, H., Vuik , K., & Hoogendoorn , S. (2015). Genealogia de

modelos de fluxo de tráfego. EURO Journal on Transportation and Logistics, 4(4), 445-473.

Van den Goorbergh, R., van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, *29*(9), 1525-1534. DOI: 10.1093/jamia/ocac093

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Veigas, K. C., Regulagadda, D. S., & Kokatnoor, S. A. (2021). Optimized Stacking Ensemble (OSE) for Credit Card Fraud Detection using Synthetic Minority Oversampling Model. Indian Journal of Science and Technology, 14(32), 2607-2615. https://doi.org/10.17485/IJST/v14i32.807

Wei, W., Li, J., Cao, L., Ou, Y., & Chen, J. (2013). Detecção eficaz de fraudes bancárias on-line sofisticadas em dados extremamente desequilibrados. World Wide Web, 16, 449-475.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: a comprehensive review. Computers & security, 57, 47-66.

WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. London, UK: Springer-Verlag, 2000. p. 29-39.

Wu, DD, Olson, DL e Luo, C. (2014). Uma abordagem de suporte à decisão para gerenciamento de risco de contas a receber. Transações IEEE em Sistemas, Homem e Cibernética: Sistemas, 44(12), 1624-1632.

Yang, Y., Khorshidi, H. A., & Aickelin, U. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: Insights for medical problems. Frontiers in digital health, 6, 1430245.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

Zafari , B., Ekin , T., & Ruggeri, F. (2022). Fronteiras de decisão multicritério para detecção de anomalias de prescrição ao longo do tempo. Journal of Applied Statistics, 49(14), 3638-3658.

Zamini, M., & Hasheminejad, S. M. H. (2019). A comprehensive survey of anomaly detection in banking, wireless sensor networks, social networks, and healthcare. Intelligent Decision Technologies, 13(2), 229-270.

Zhang, L., Zhang, W., McNeil, MJ, Chengwang , N., Matteson, DS e Bogdanov, P. (2021).

AURORA: uma estrutura unificada for Detecção de anomalias em séries temporais multivariadas. Mineração de dados e descoberta de conhecimento, 35(5), 1882-1905.

Zhang, Y. F., Lu, H. L., Lin, H. F., Qiao, X. C., & Zheng, H. (2022). The Optimized Anomaly Detection Models Based on an Approach of Dealing with Imbalanced Dataset for Credit Card Fraud Detection. Mobile Information Systems, 2022.

Zhou, R., Zhang, Q., Zhang, P., Niu, L., & Lin, X. (2021). Anomaly detection in dynamic attributed networks. Neural Computing and Applications, 33, 2125-2136.

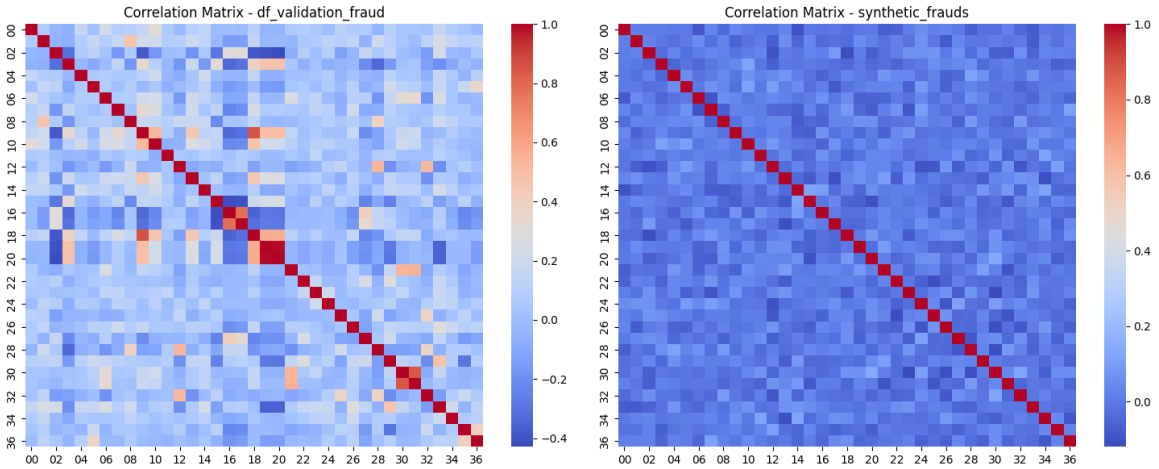
Zhou, Y., Guo, C., Wang, X., Chang, Y., & Wu, Y. (2024). A survey on data augmentation in large model era. Journal of LaTeX Class Files, 14(8), 1-33.

Zhu, Y., & Yang, K. (2019). Aprendizagem ativa tripartida para descoberta interativa de anomalias. Acesso IEEE, 7, 63195-63203.

Zioviris, G., Kolomvatsos, K., & Stamoulis, G. (2022). Credit card fraud detection using a deep learning multistage model. The Journal of Supercomputing, 78(12), 14571-14596.

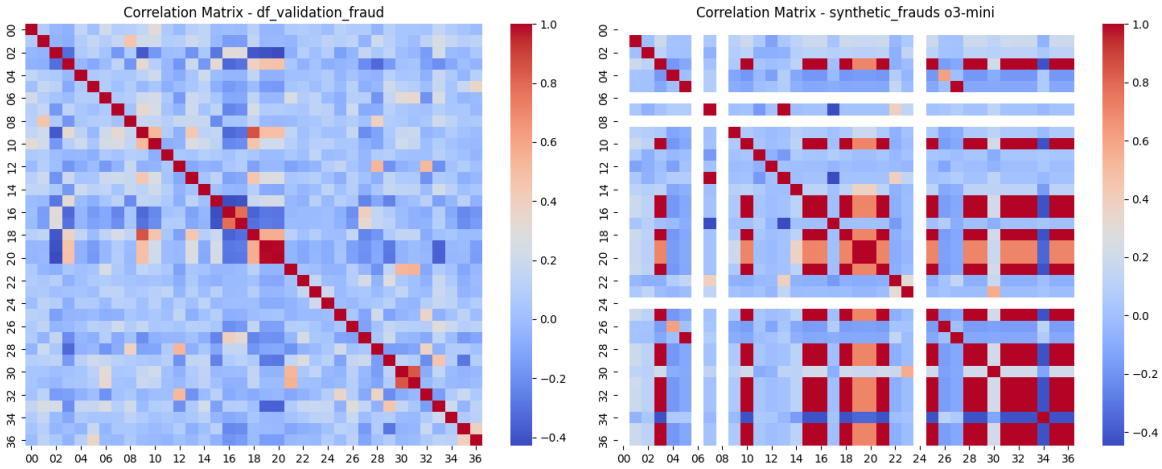
Anexo 1 – Aplicação de LLM via Prompt

Figura 0.1 – Correlação entre as variáveis reais e geradas – LLM GPT4 ADA



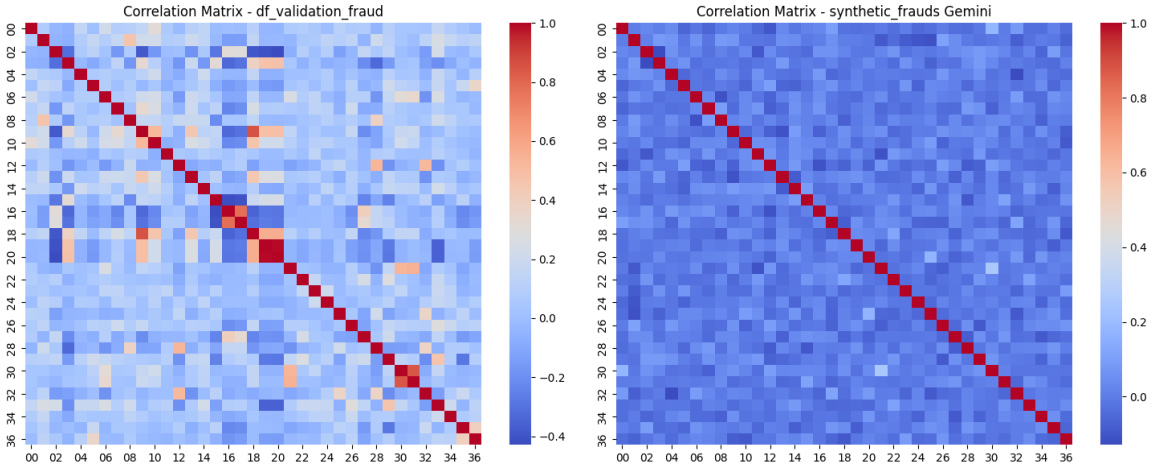
Fonte: elaboração própria

Figura 0.2 – Correlação entre as variáveis reais e geradas – o3-mini-high



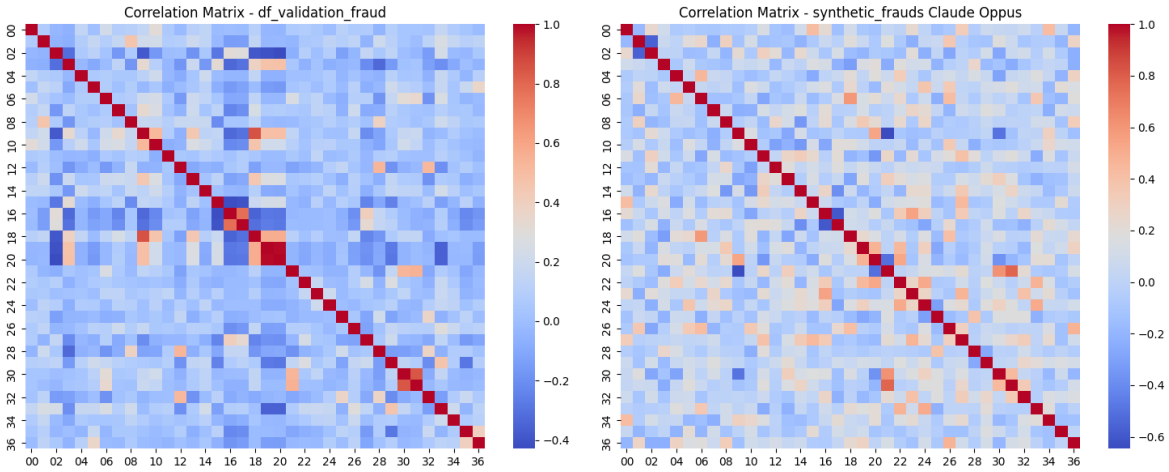
Fonte: elaboração própria

Figura 0.3 – Correlação entre as variáveis reais e geradas – LLM Gemini



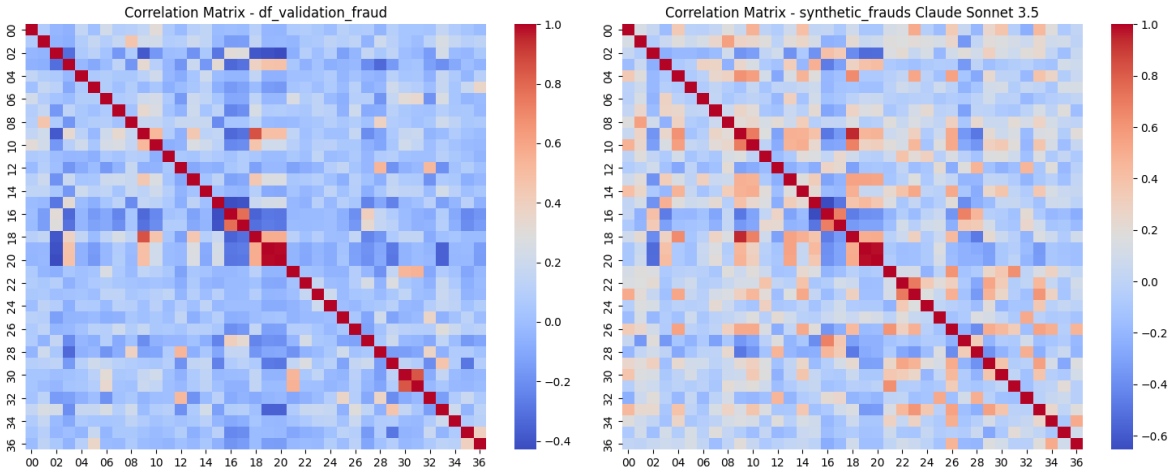
Fonte: elaboração própria

Figura 0.4 – Correlação entre as variáveis reais e geradas – LLM Claude 3 Opus



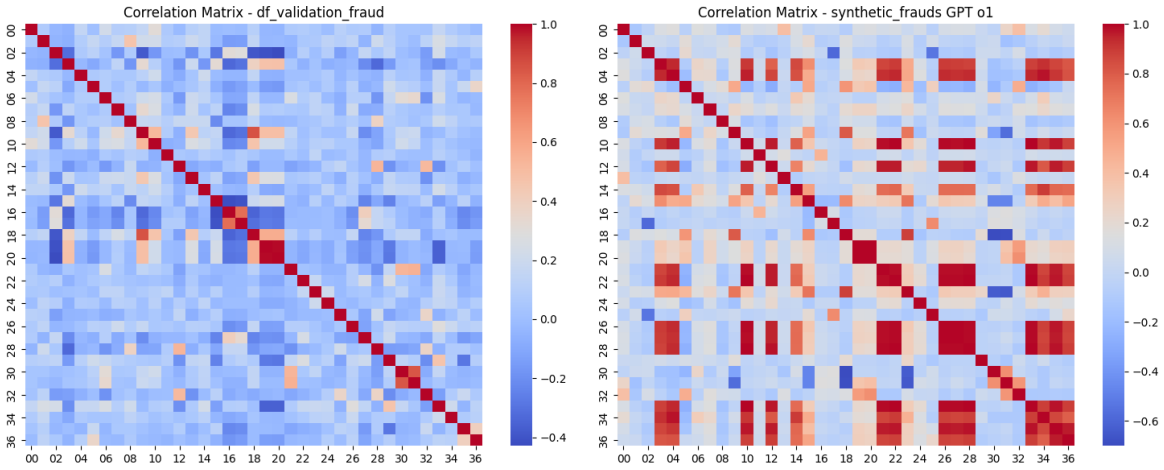
Fonte: elaboração própria

Figura 0.5 – Correlação entre as variáveis reais e geradas – LLM Claude 3.5 Sonnet



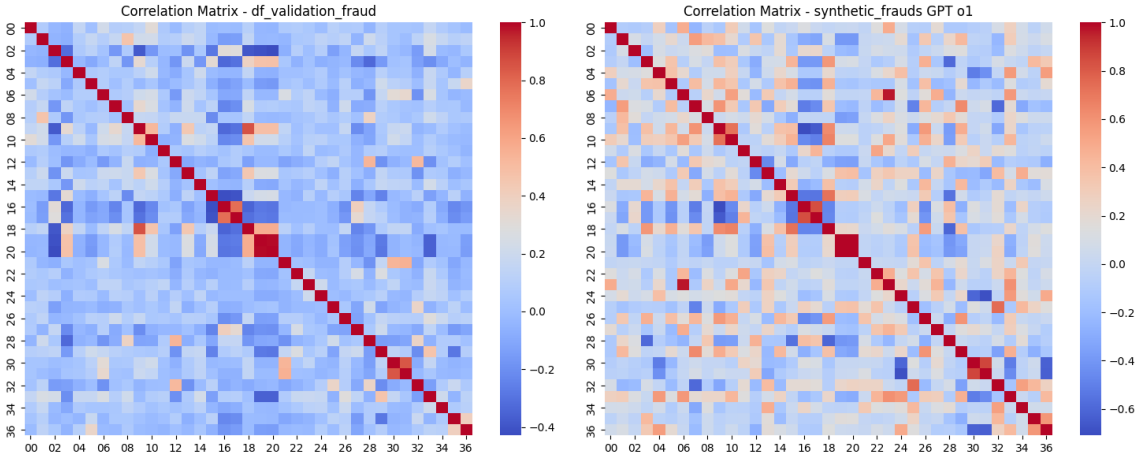
Fonte: elaboração própria

Figura 0.6 – Correlação entre as variáveis reais e geradas – LLM GPT o1



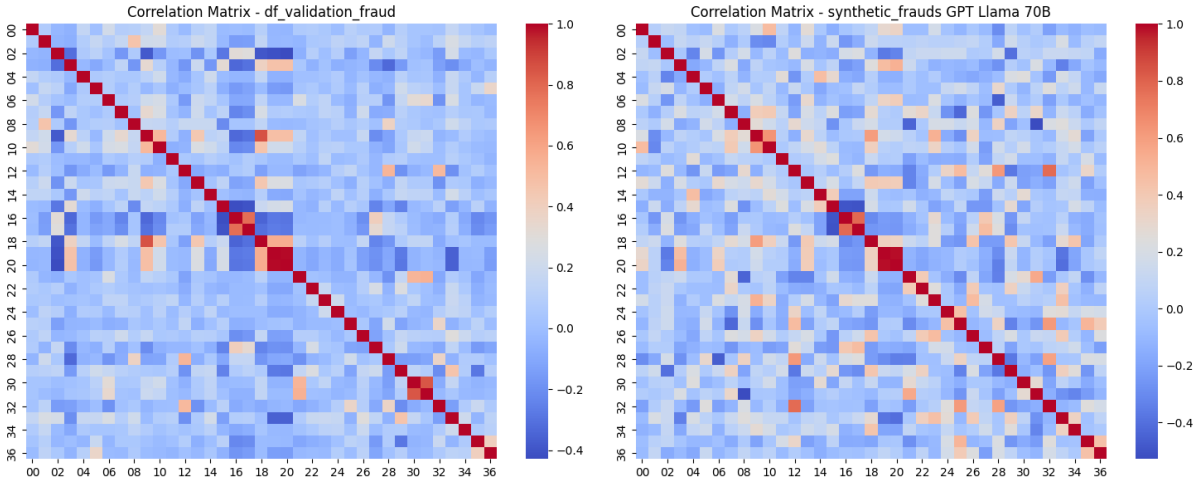
Fonte: elaboração própria

Figura 0.7 – Correlação entre as variáveis reais e geradas – LLM Gemini 2.0 advanced



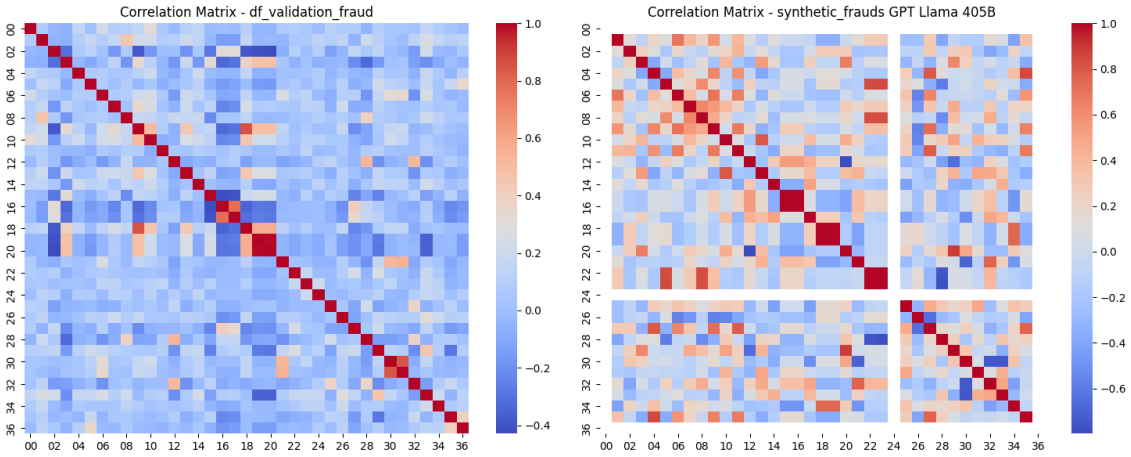
Fonte: elaboração própria

Figura 0.8 – Correlação entre as variáveis reais e geradas – LLM Llama 3.1 70B



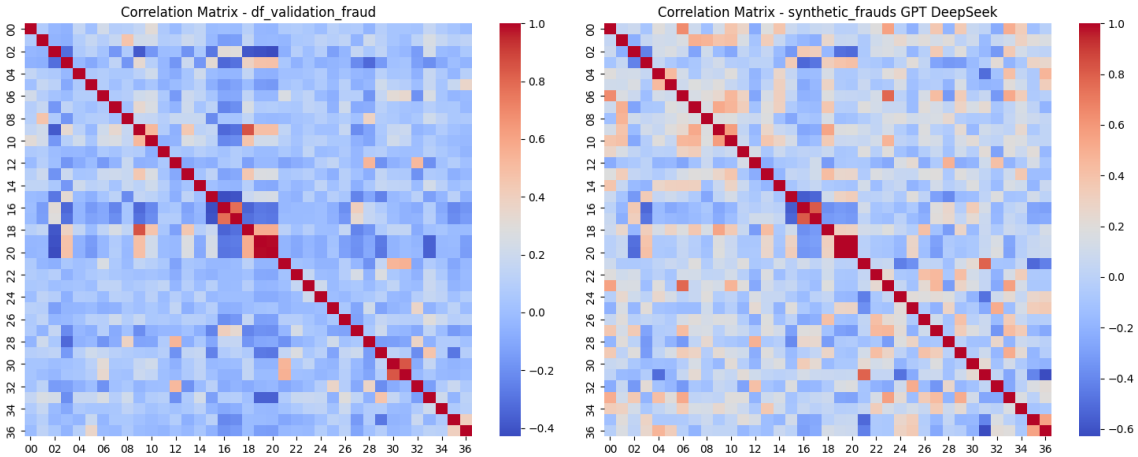
Fonte: elaboração própria

Figura 0.9 – Correlação entre as variáveis reais e geradas – LLM Llama 3.1 405B



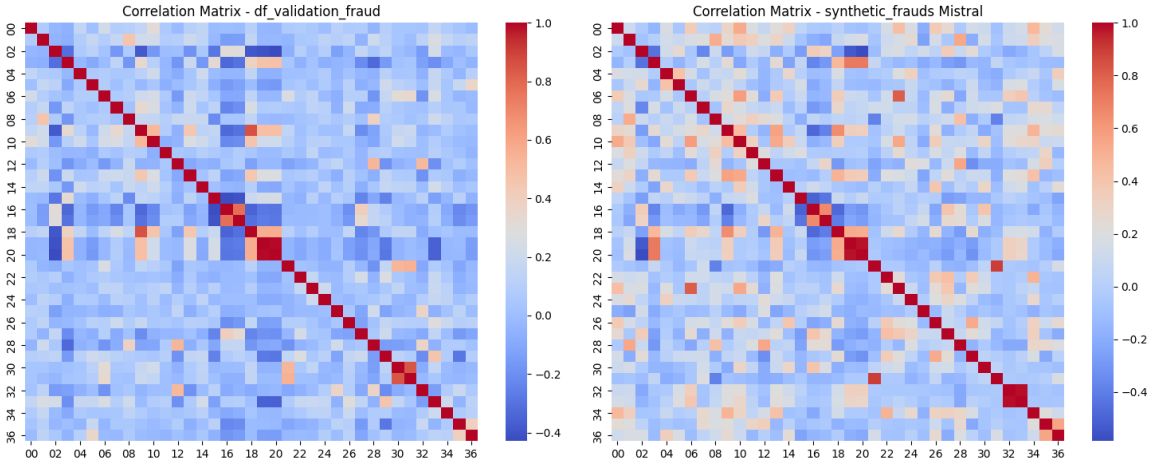
Fonte: elaboração própria

Figura 0.10 – Correlação entre as variáveis reais e geradas – LLM DeepSeek



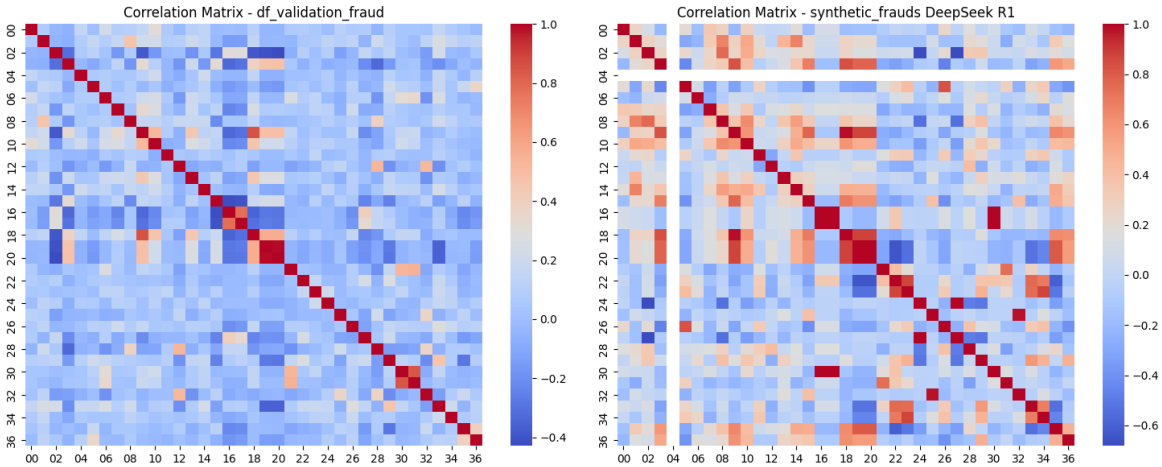
Fonte: elaboração própria

Figura 0.11 – Correlação entre as variáveis reais e geradas – LLM Mistral



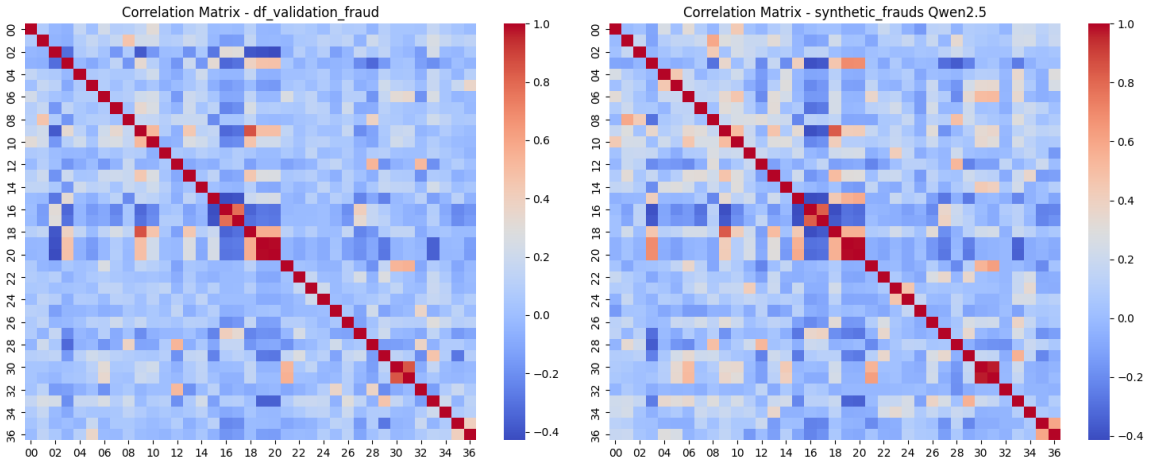
Fonte: elaboração própria

Figura 0.12 – Correlação entre as variáveis reais e geradas – DeepSeek R1



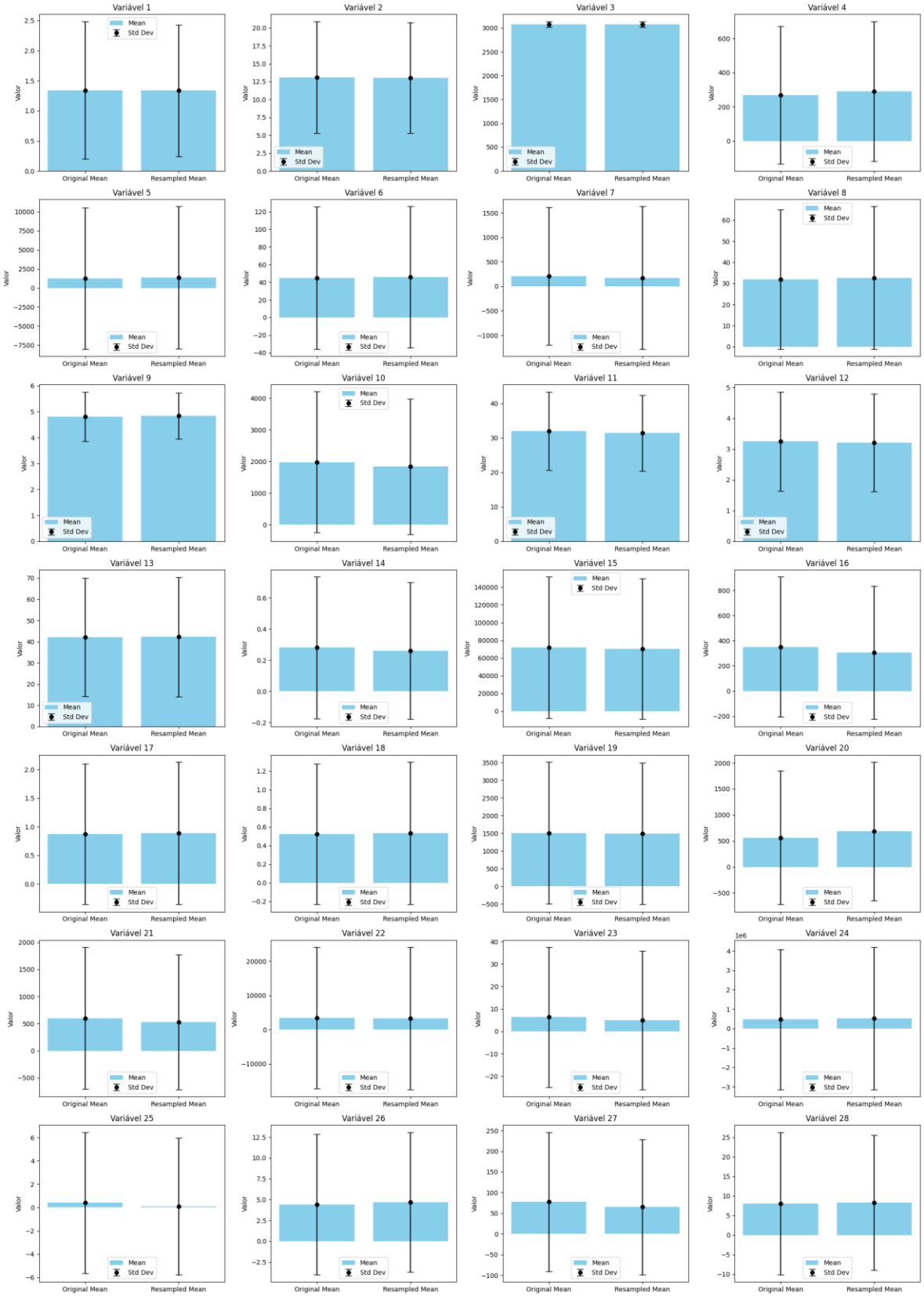
Fonte: elaboração própria

Figura 0.13 – Correlação entre as variáveis reais e geradas – Qwen2.5



Fonte: elaboração própria

Figura 0.14 – Estatísticas descritivas – Box Plot – Modelo GPT ADA

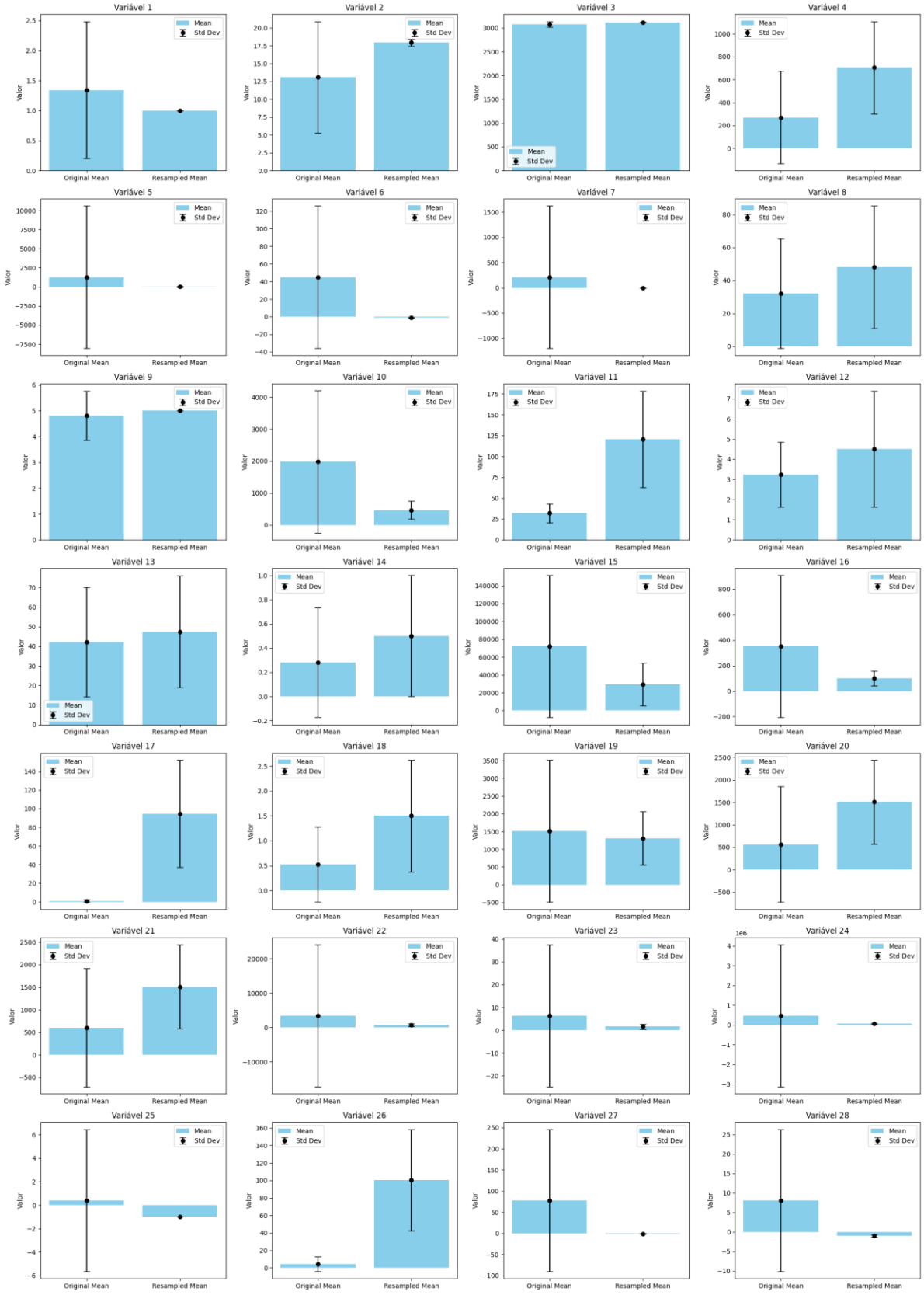


Fonte: elaboração própria

Figura 0.15 – Distribuição das variáveis - Modelo GPT ADA

Fonte: elaboração própria

Figura 0.16 – Estatísticas descritivas – Box Plot – Modelo GPT o3-mini-high



Fonte: elaboração própria

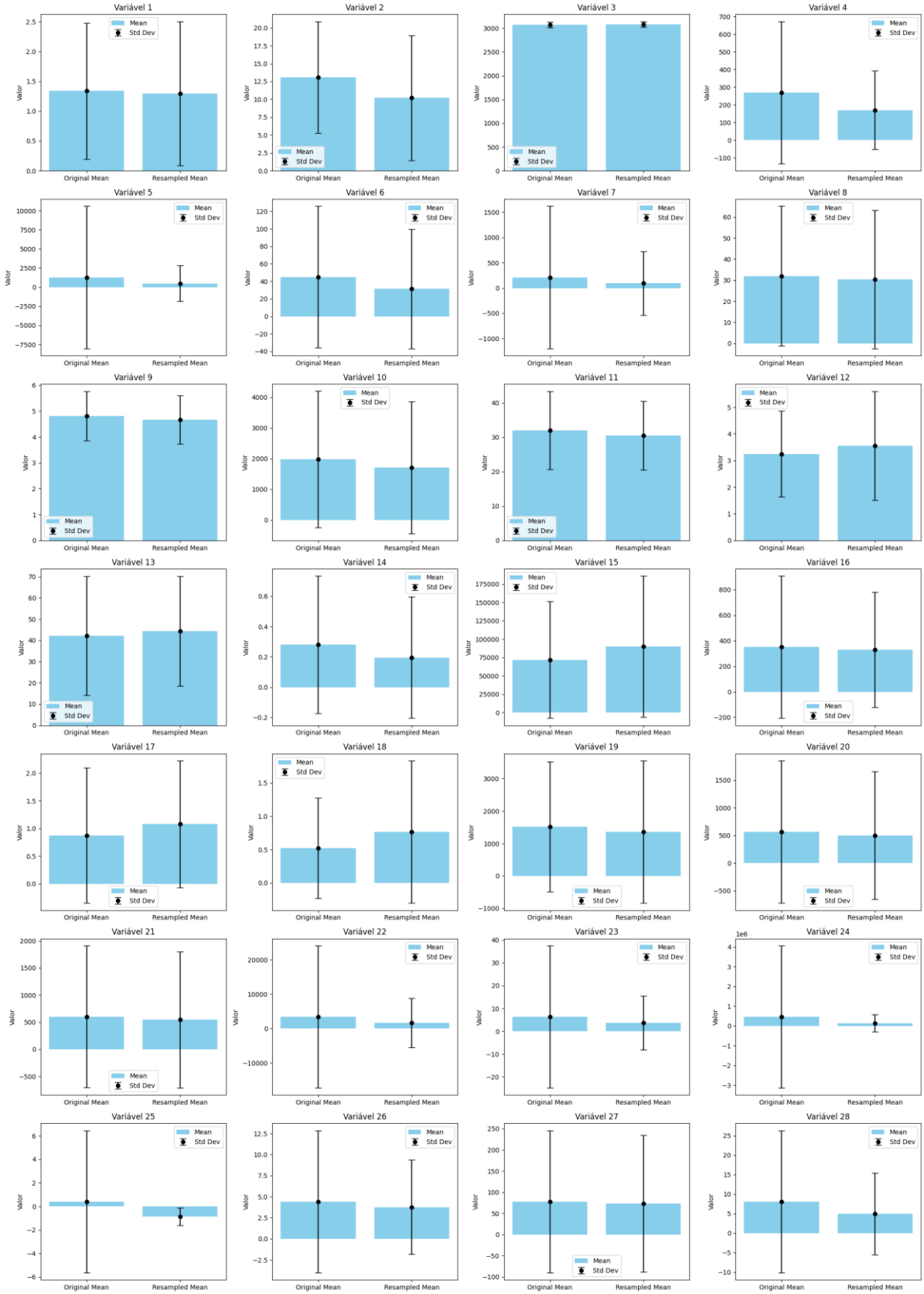
Figura 0.17 – Distribuição das variáveis – Modelo GPT o3-mini-high

Fonte: elaboração própria

Figura 0.18 – Distribuição das variáveis – Modelo GPT Gemini

Fonte: elaboração própria

Figura 0.19 – Estatísticas descritivas – Box Plot – Modelo Claude Sonnet 3.5



Fonte: elaboração própria

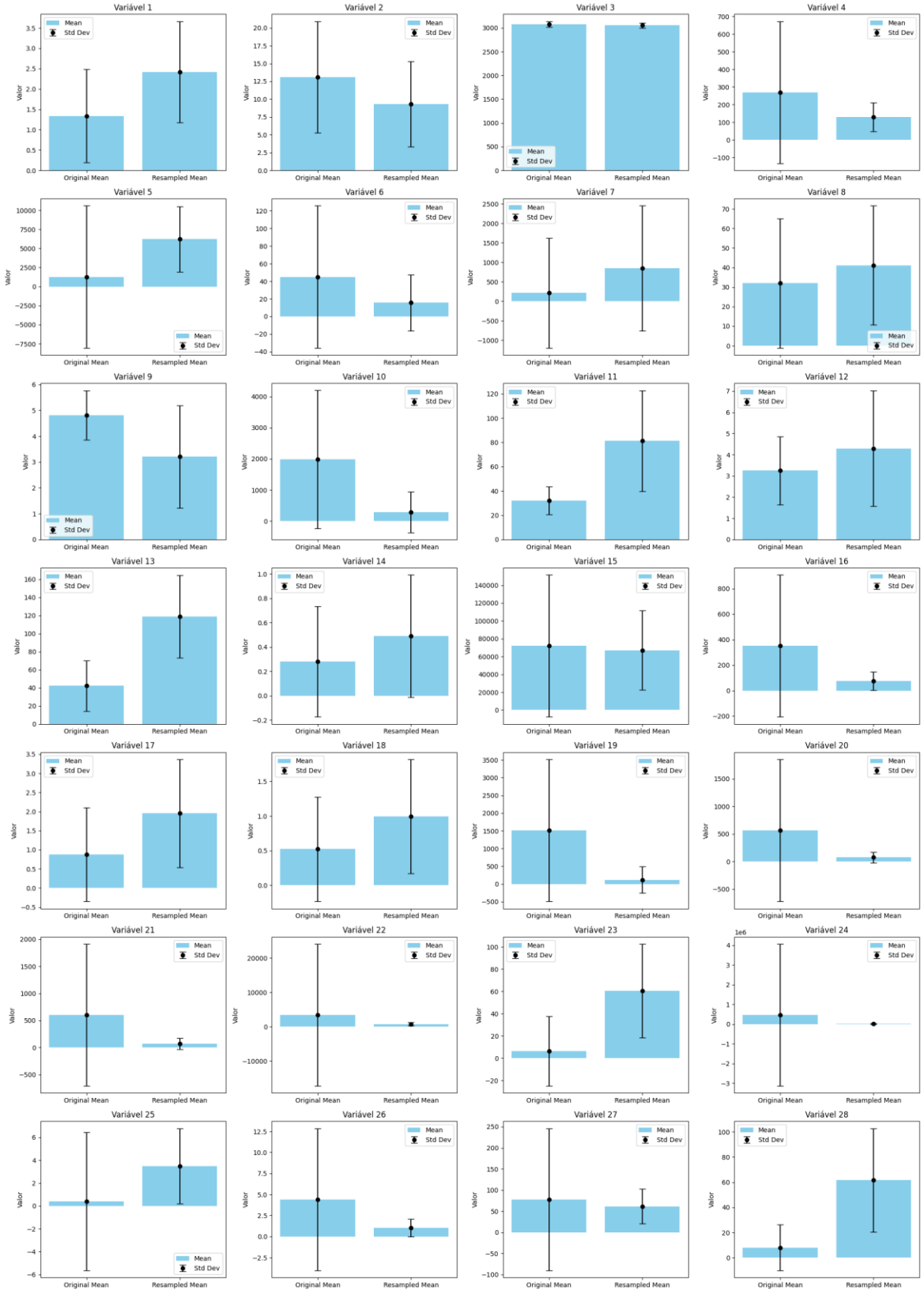
Figura 0.20 – Distribuição das variáveis - Modelo Claude Sonnet 3.5

Fonte: elaboração própria

Figura 0.21 – Distribuição das variáveis modelo GPT Claude 3 Opus

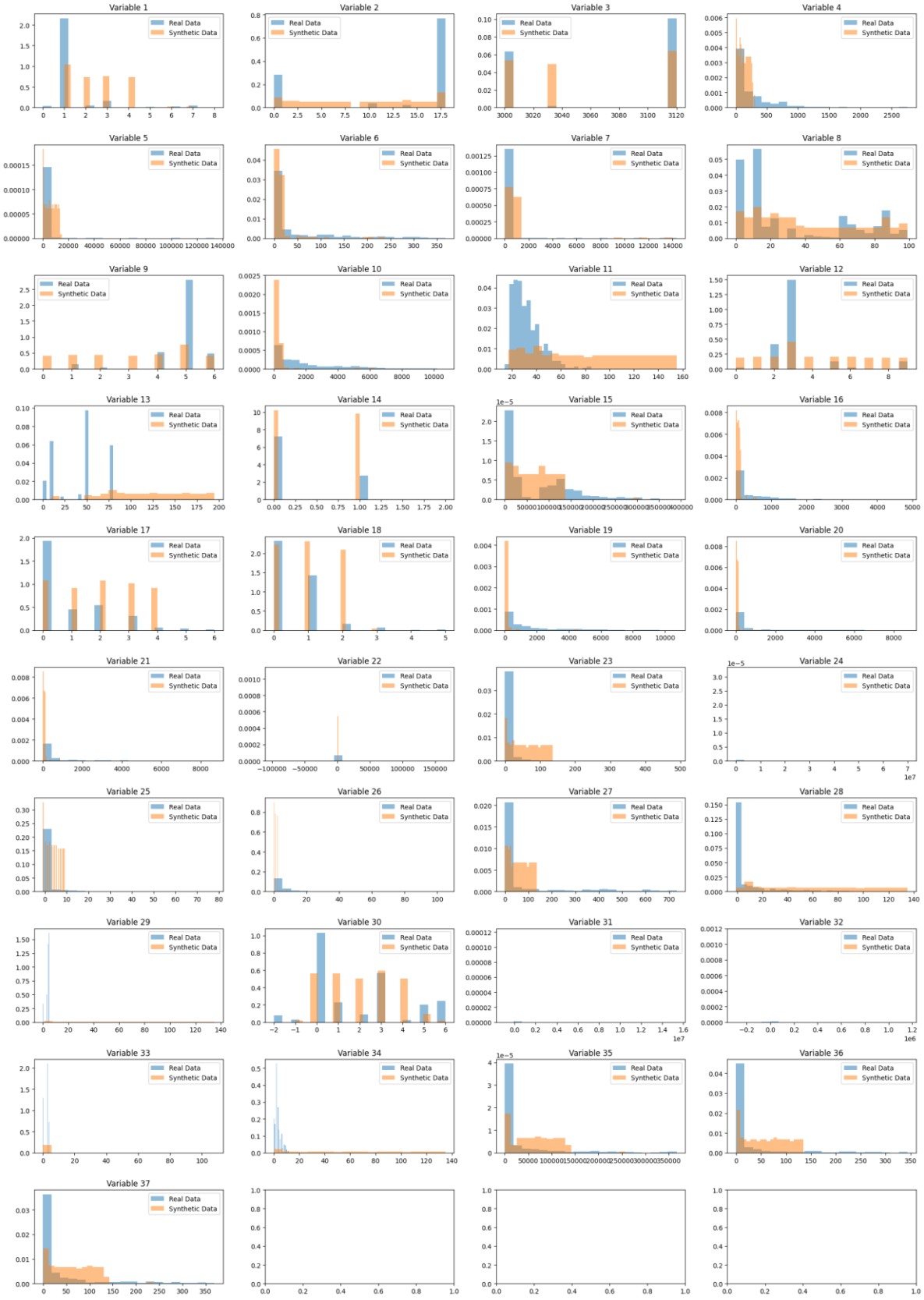
Fonte: elaboração própria

Figura 0.22 – Estatísticas descritivas – Box Plot – Modelo GPT o1



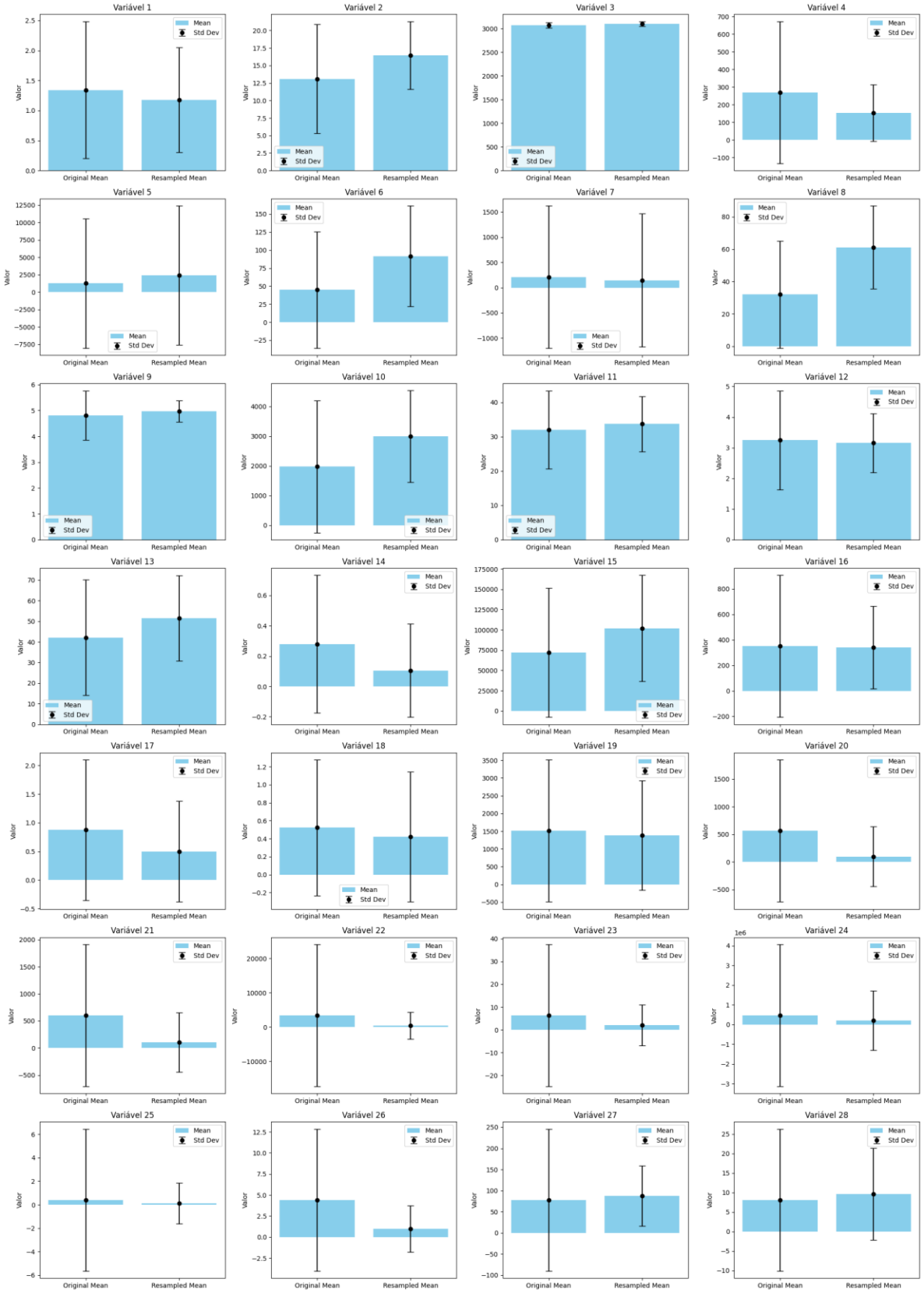
Fonte: elaboração própria

Figura 0.23 – Distribuição das variáveis – Modelo GPT o1



Fonte: elaboração própria

Figura 0.24 – Estatísticas descritivas – Box Plot – Modelo Gemini 2.0

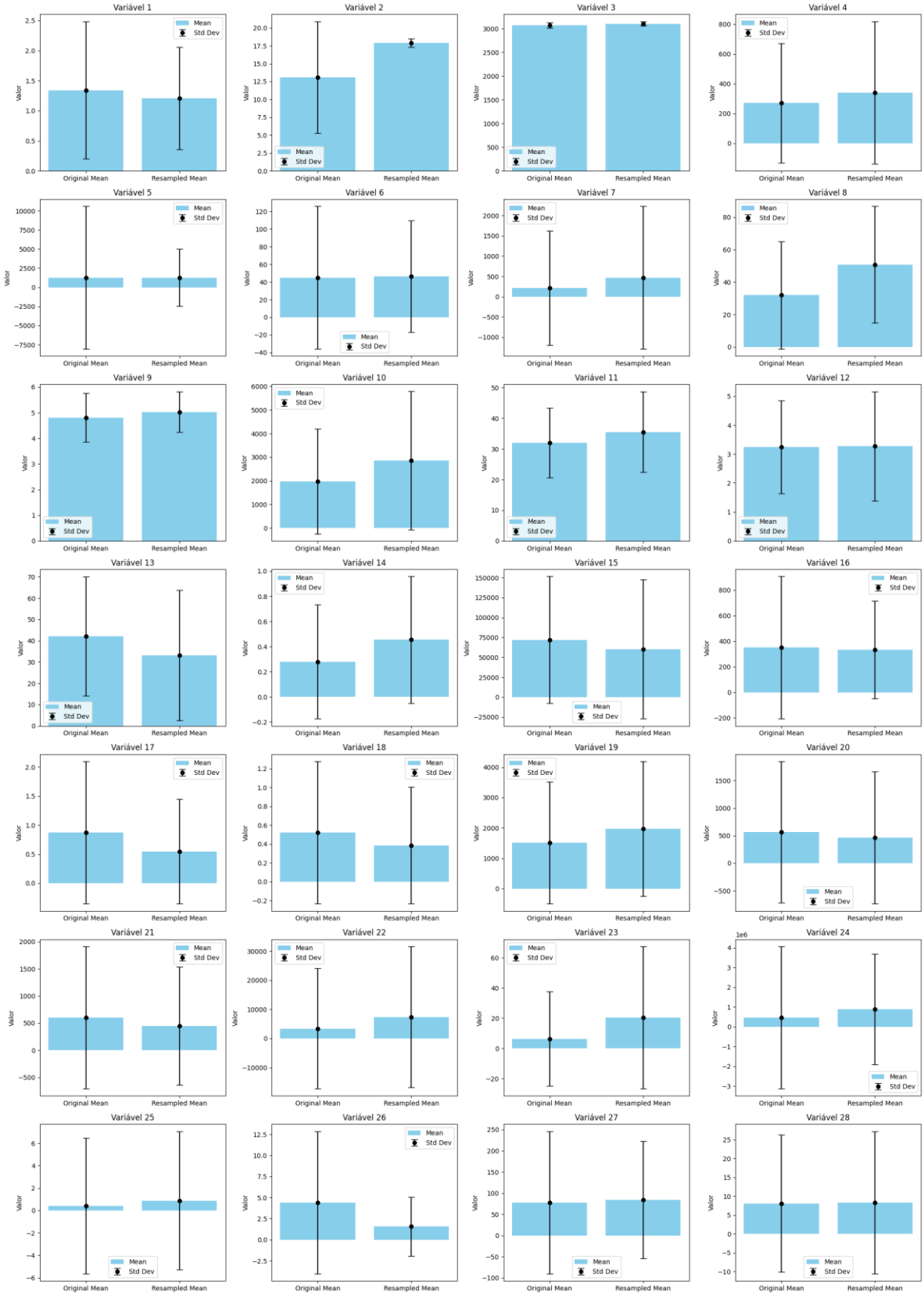


Fonte: elaboração própria

Figura 0.25 – Distribuição das variáveis – Modelo Gemini 2.0

Fonte: elaboração própria

Figura 0.26 – Estatísticas descritivas – Box Plot – Modelo Llama 3.1 70B

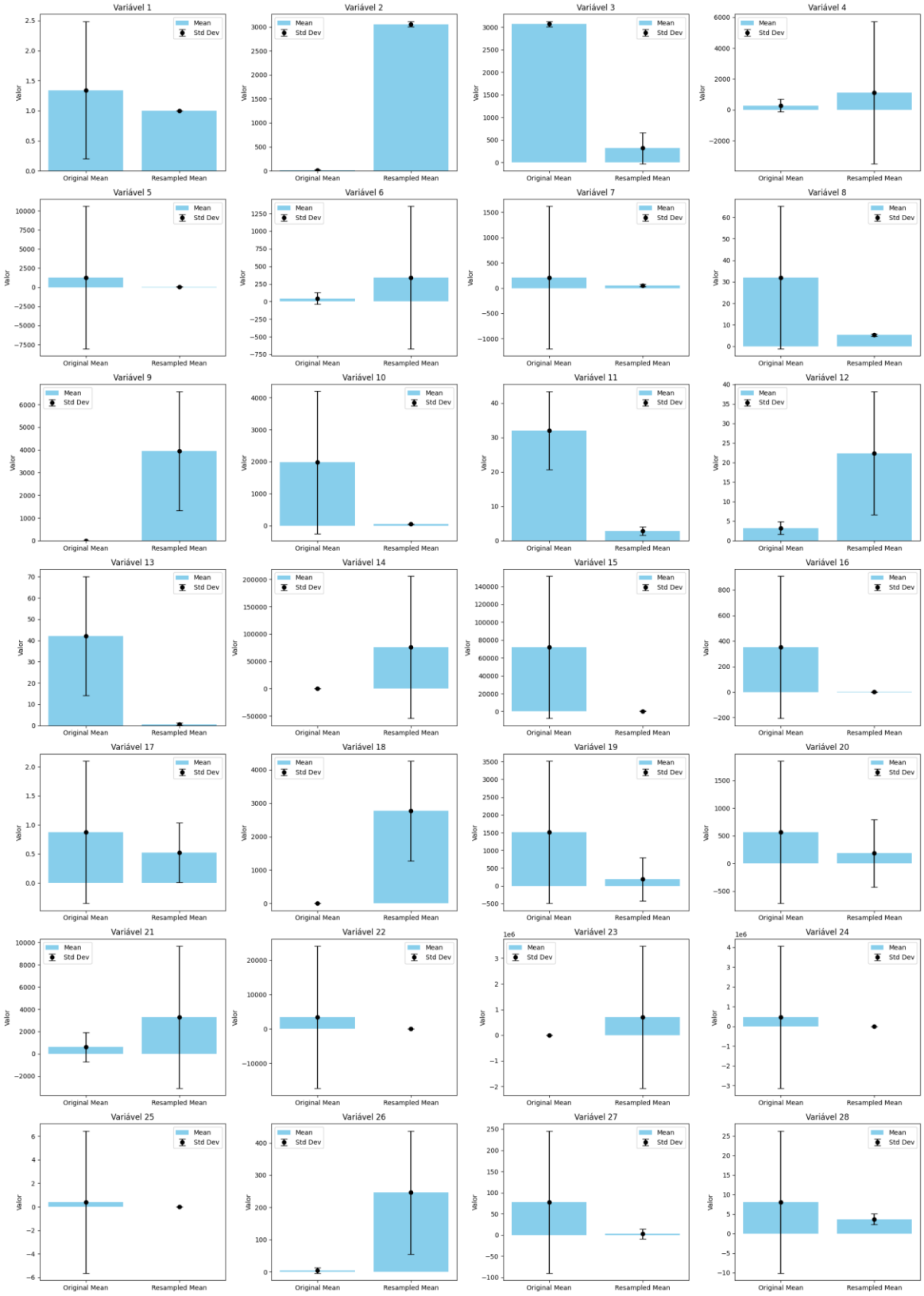


Fonte: elaboração própria

Figura 0.27 – Distribuição das variáveis – Modelo Lhama 3.1 70B

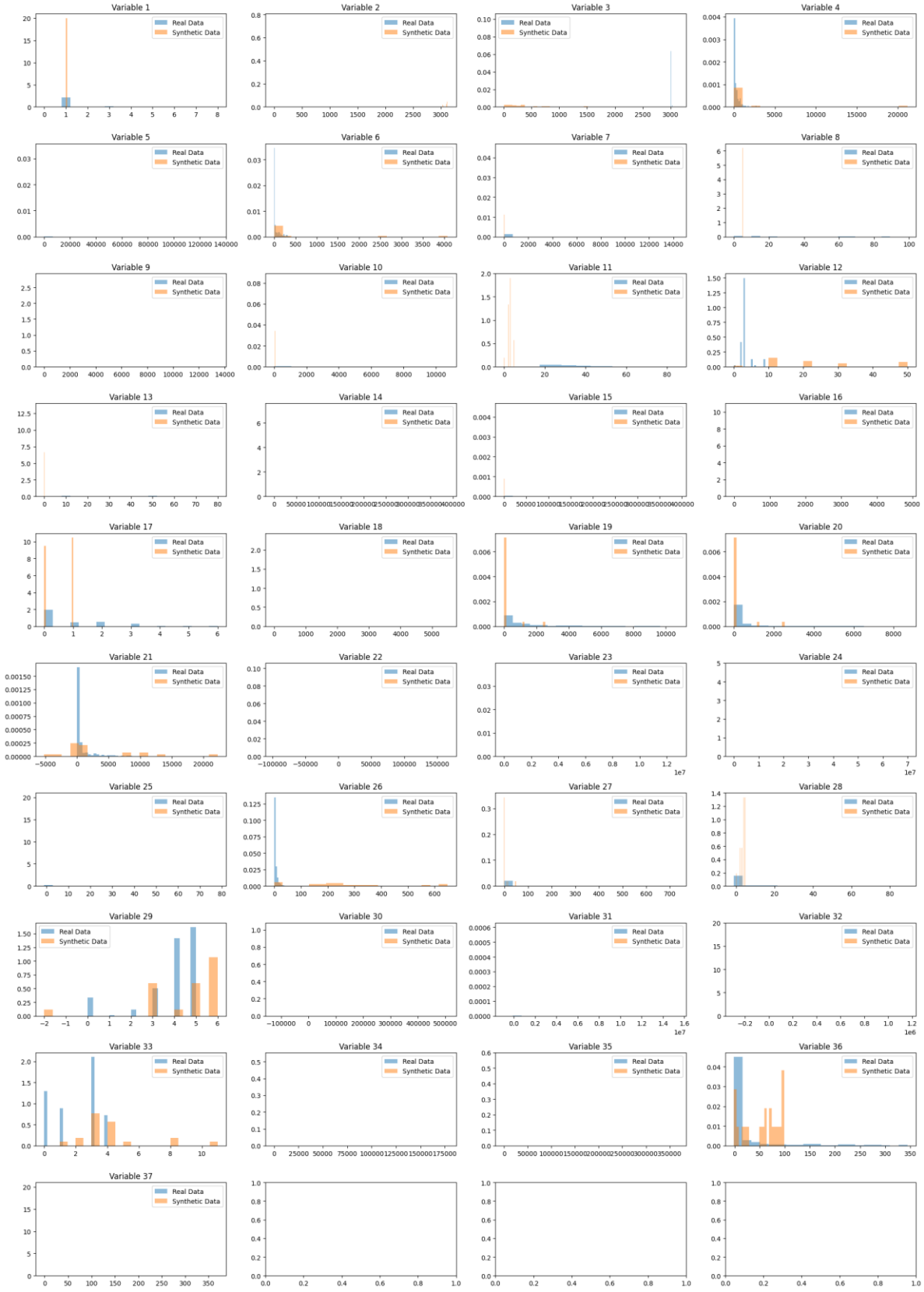
Fonte: elaboração própria

Figura 0.28 – Estatísticas descritivas – Box Plot – Modelo Llama 3.1 405B



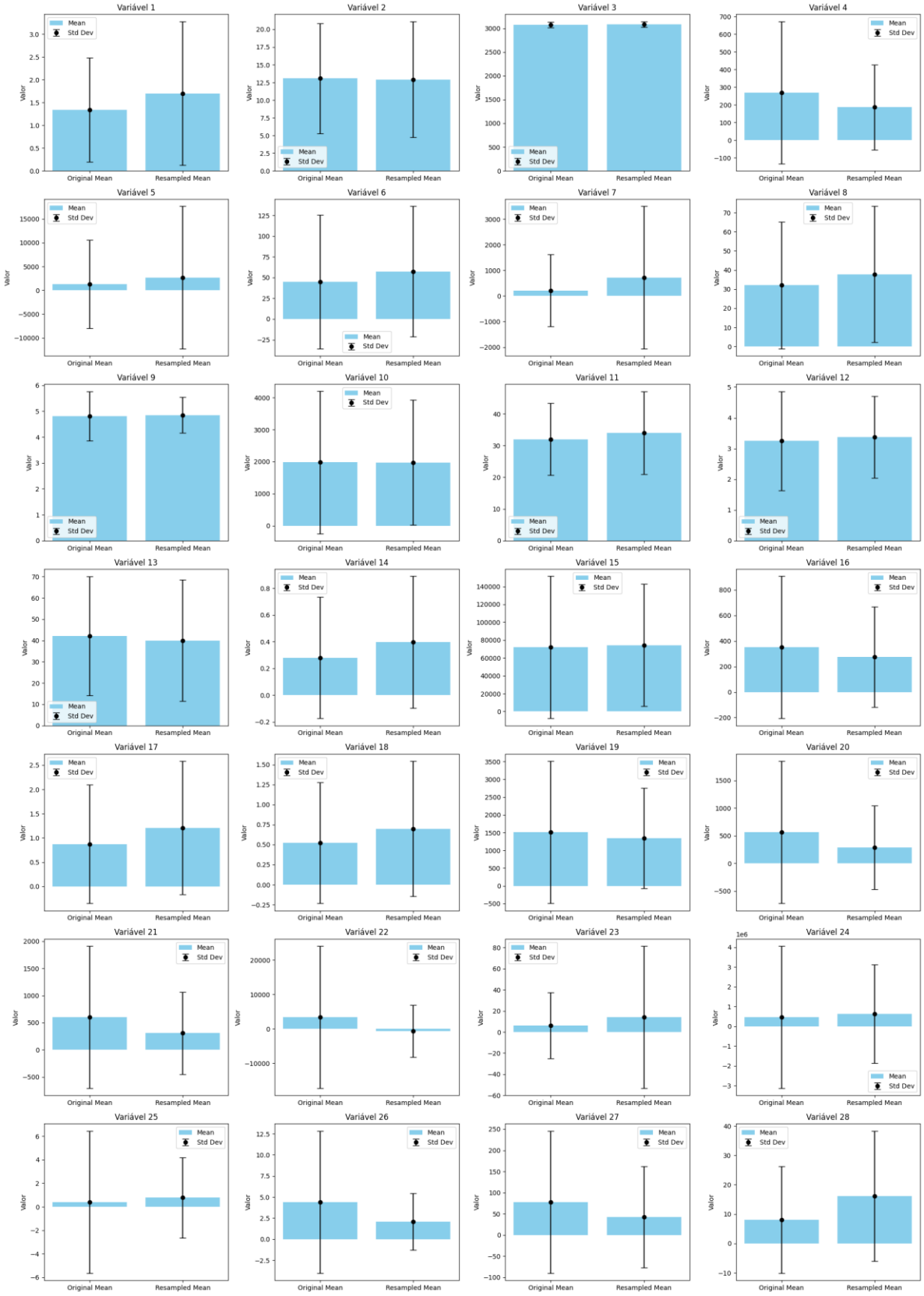
Fonte: elaboração própria

Figura 0.29 – Distribuição das variáveis - Llama 3.1 405B



Fonte: elaboração própria

Figura 0.30 – Estatísticas descritivas – Box Plot – Modelo Deepseek V3

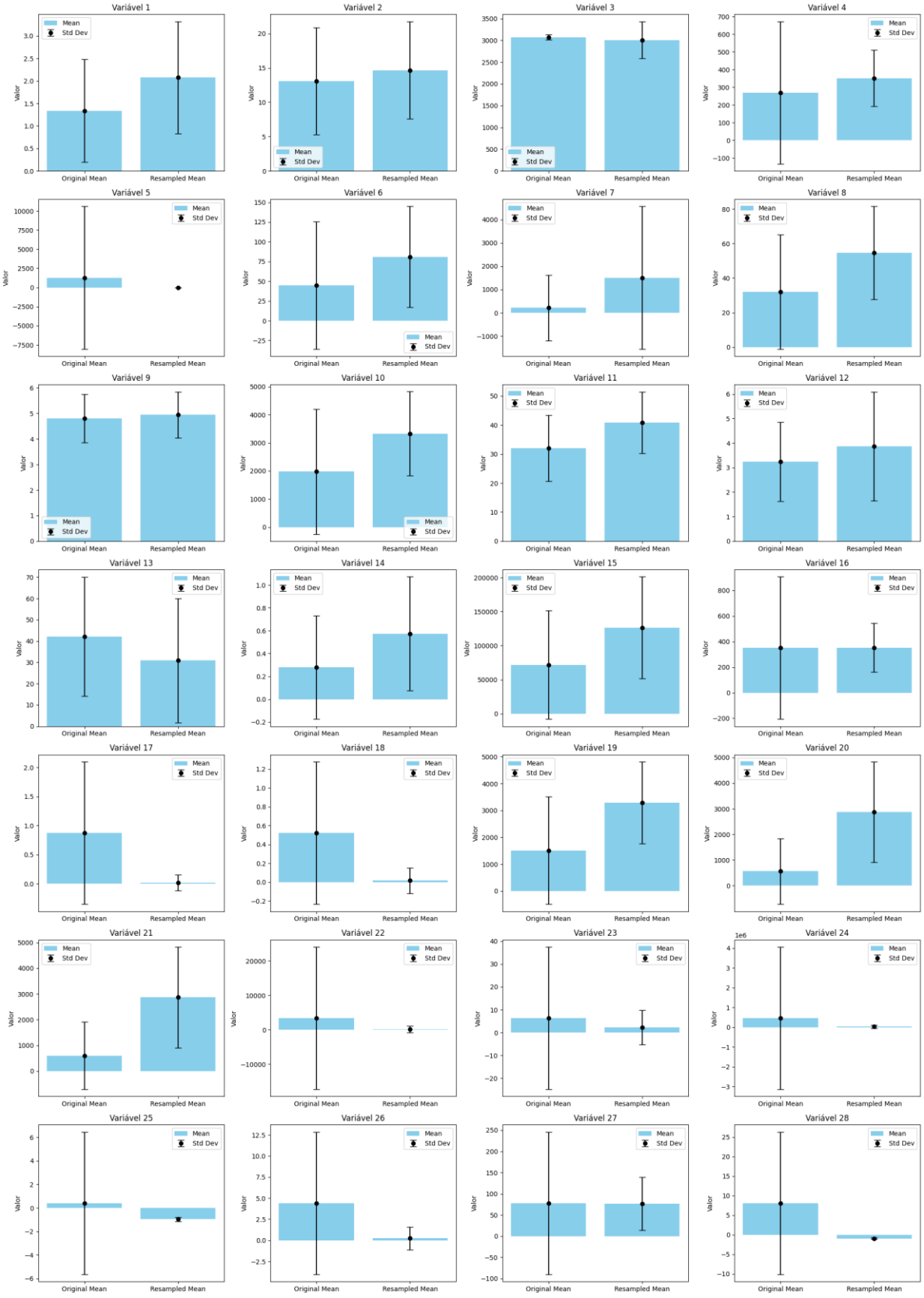


Fonte: elaboração própria

Figura 0.31 – Distribuição das variáveis – Modelo DeepSeek V3

Fonte: elaboração própria

Figura 0.32 – Estatísticas descritivas – Box Plot – Modelo DeepSeek R1

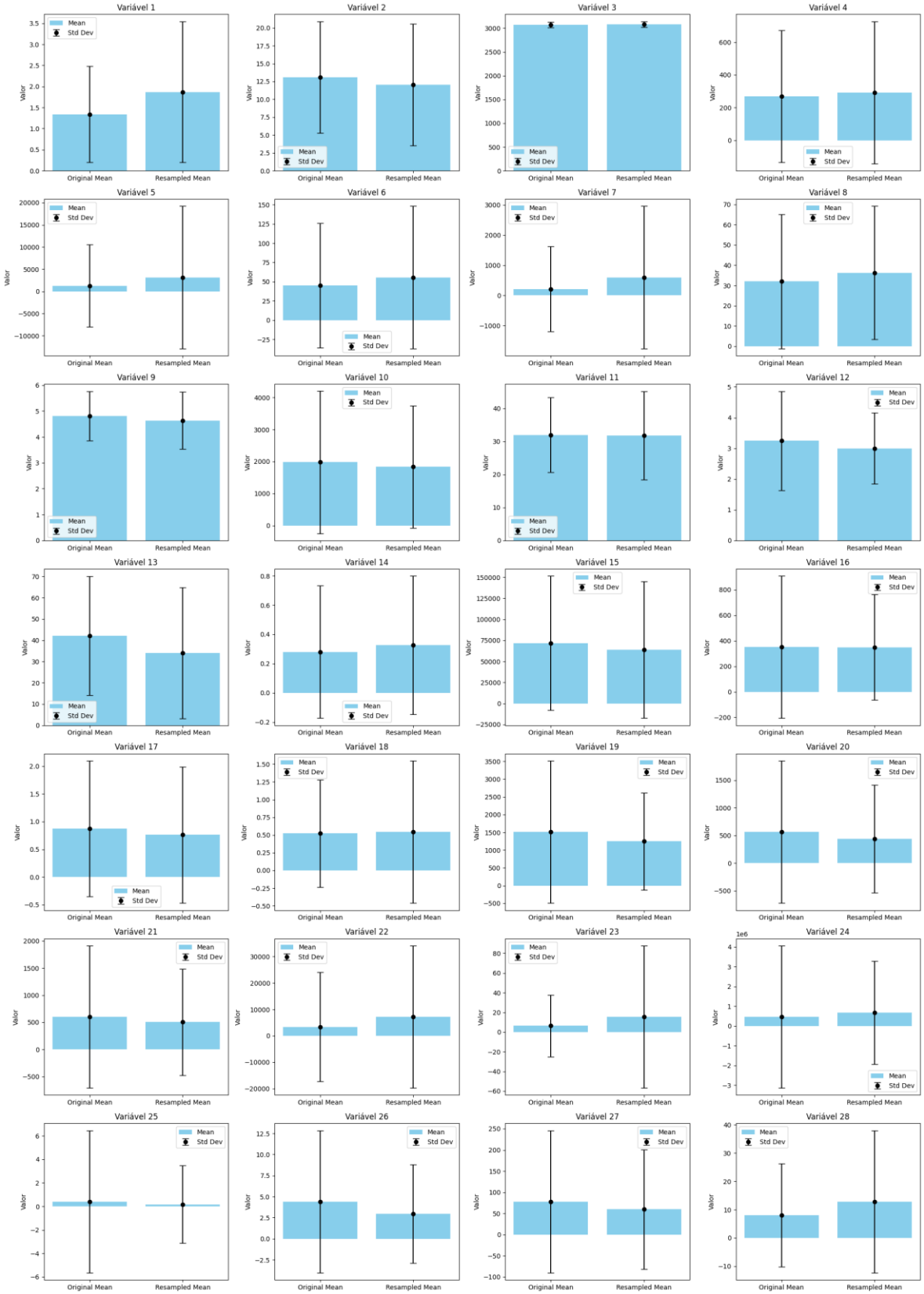


Fonte: elaboração própria

Figura 0.33 – Distribuição das variáveis – Modelo DeepSeek R1

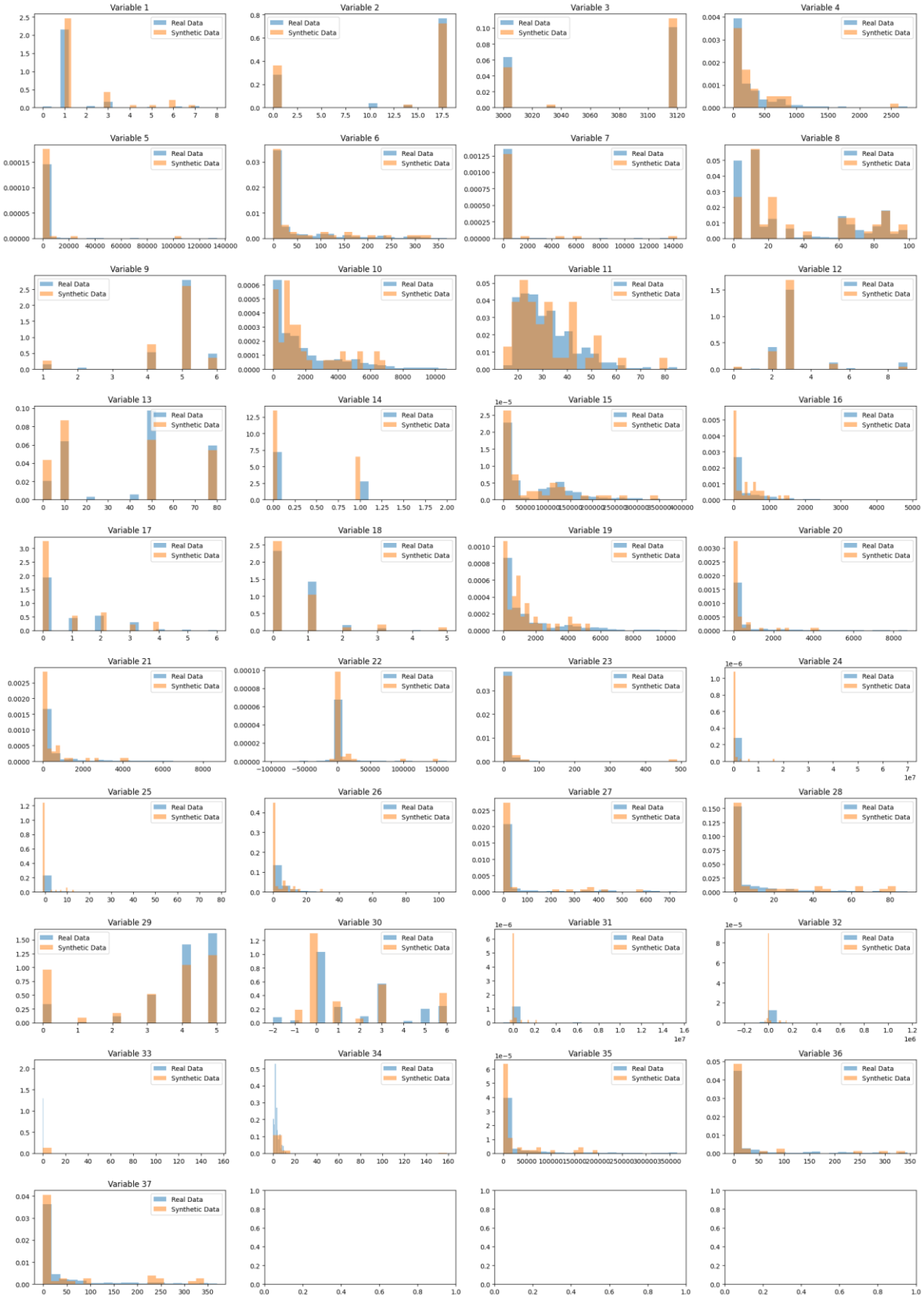
Fonte: elaboração própria

Figura 0.34 – Estatísticas descritivas – Box Plot – Modelo Mistral



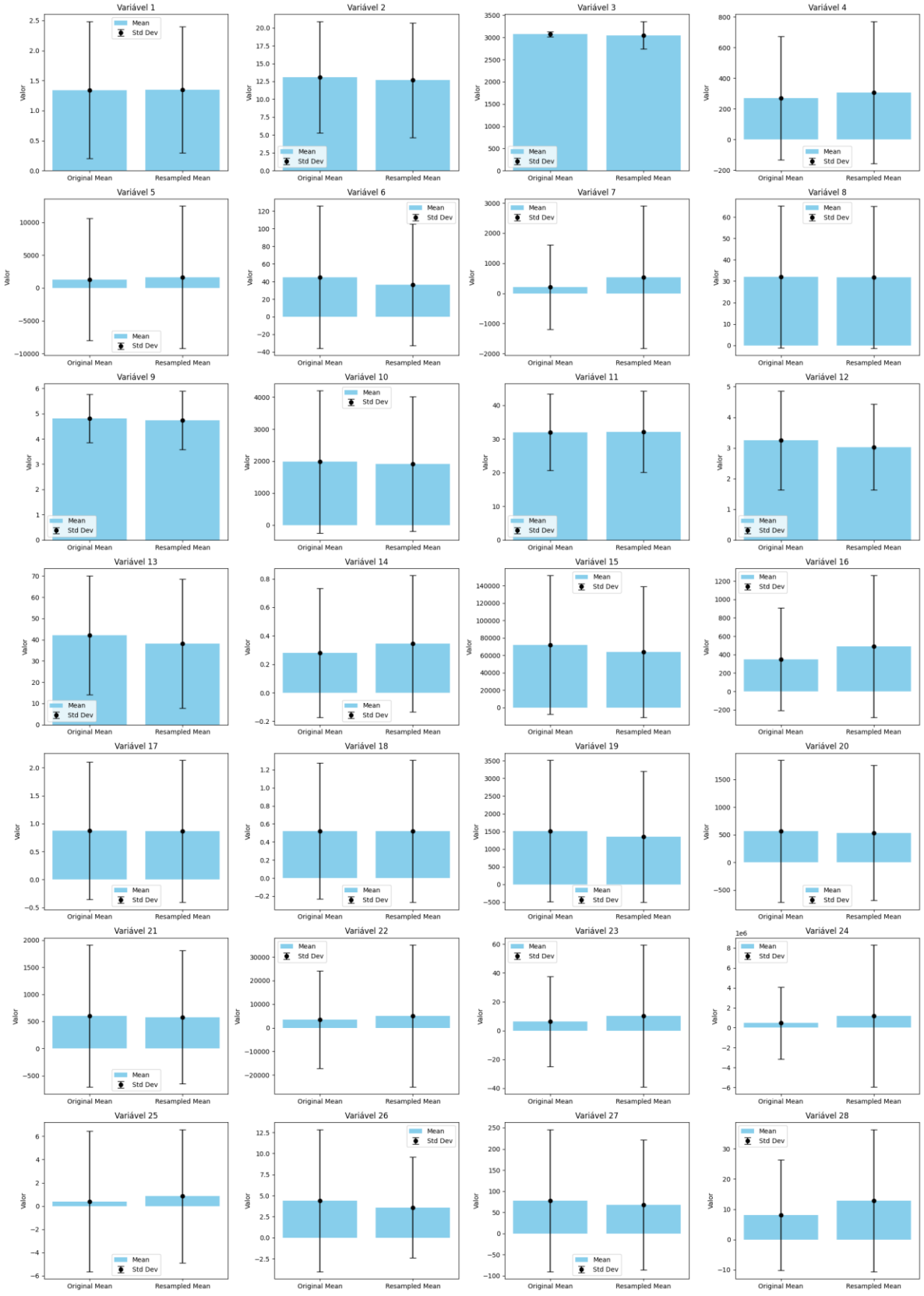
Fonte: elaboração própria

Figura 0.35 – Distribuição das variáveis – Modelo Mistral



Fonte: elaboração própria

Figura 0.36 – Estatísticas descritivas – Box Plot – Modelo Qwen2.5



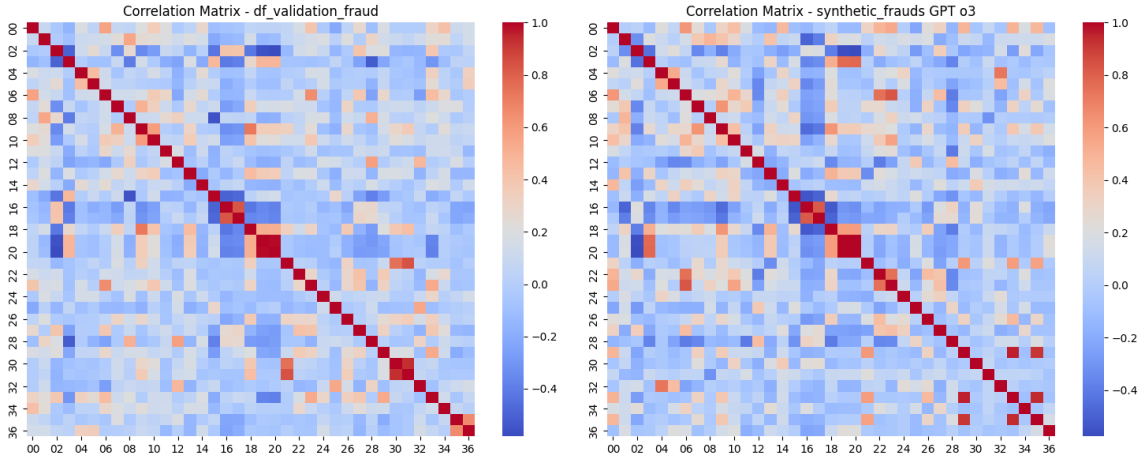
Fonte: elaboração própria

Figura 0.37 – Distribuição das variáveis – Modelo Qwen 2.5

Fonte: elaboração própria

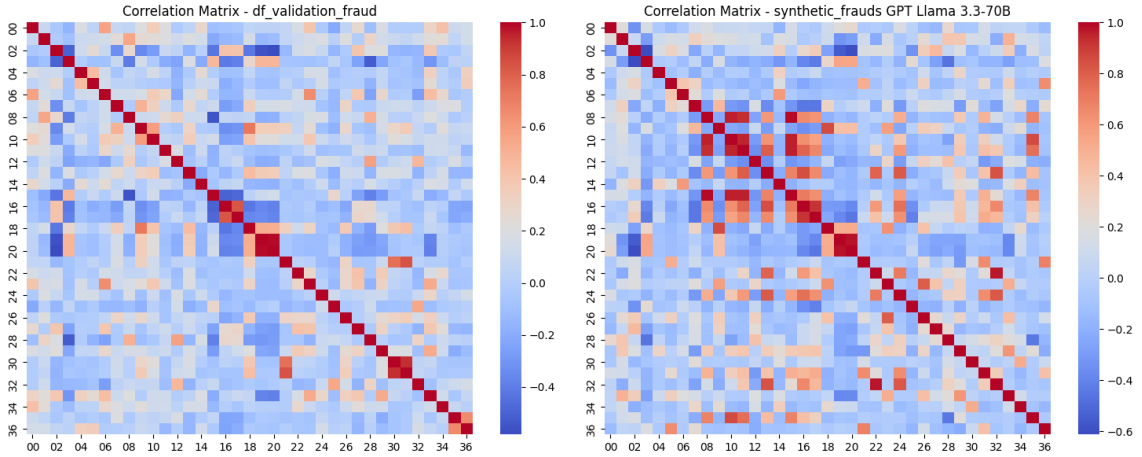
Anexo 2 – Aplicação de LLM com RAG

Figura 0.1 – Correlação entre as variáveis reais e geradas – LLM GPTo3



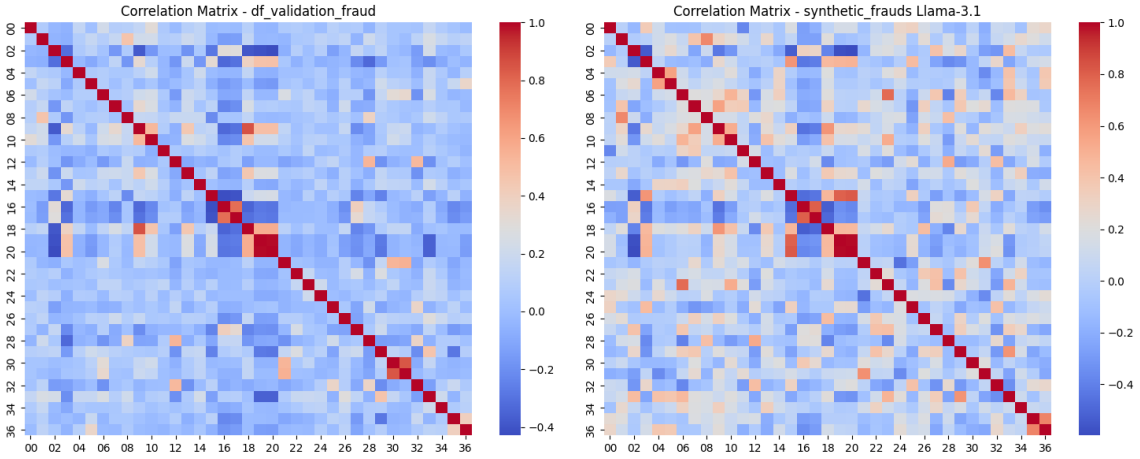
Fonte: elaboração própria

Figura 0.2 – Correlação entre as variáveis reais e geradas – LLM Llama 3.3 70B



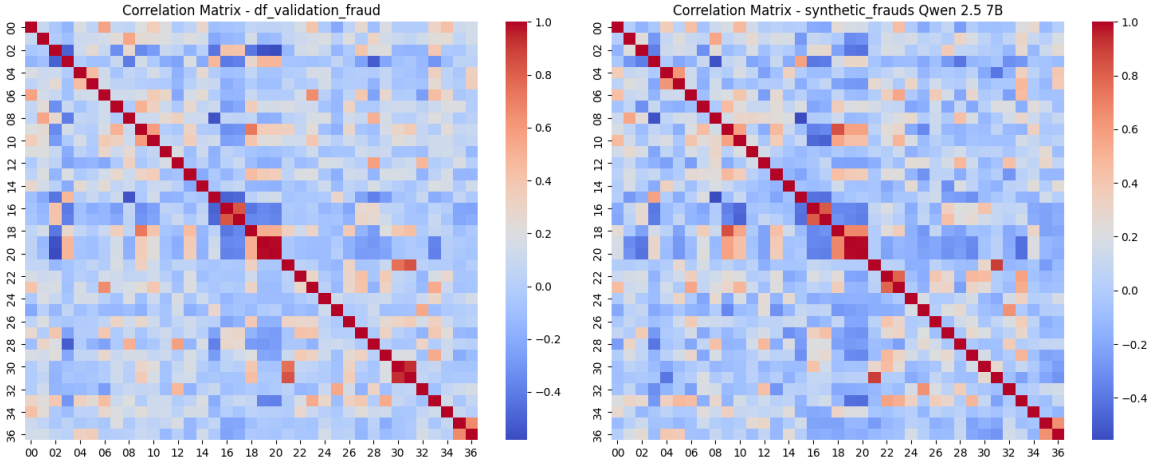
Fonte: elaboração própria

Figura 0.3 – Correlação entre as variáveis reais e geradas – LLM Llama 3.1



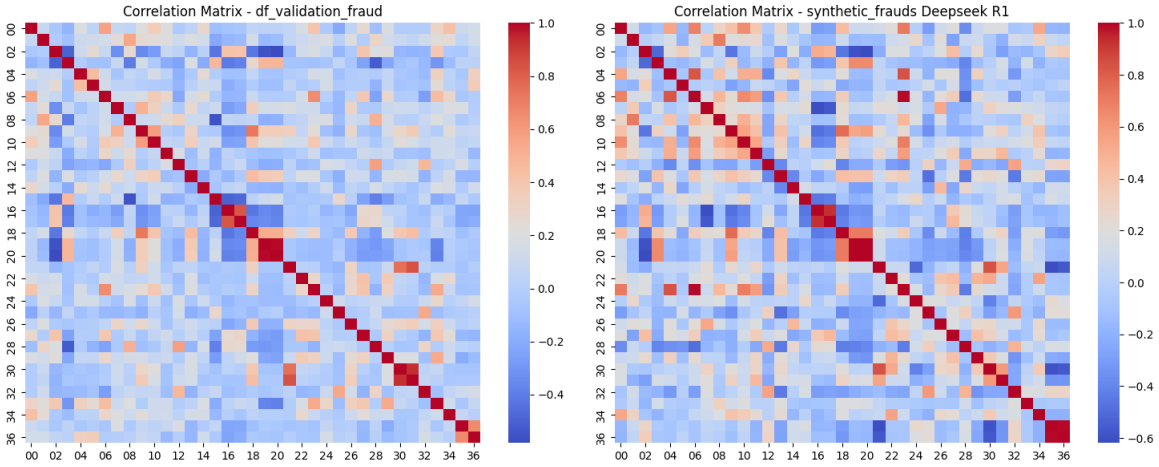
Fonte: elaboração própria

Figura 0.4 – Correlação entre as variáveis reais e geradas – LLM Qwen 2.5 7B



Fonte: elaboração própria

Figura 0.5 – Correlação entre as variáveis reais e geradas – LLM DeepSeek R1



Fonte: elaboração própria