INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS TERESINA CENTRAL
ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

ANA BEATRIZ BRITO DE FARIAS

APLICAÇÃO DE ALGORITMOS E MÉTODOS ESTATÍSTICOS PARA A IDENTIFICAÇÃO DE ANOMALIAS EM TRANSAÇÕES FINANCEIRAS

TERESINA, PIAUÍ
2026

ANA BEATRIZ BRITO DE FARIAS

APLICAÇÃO DE ALGORITMOS E MÉTODOS ESTATÍSTICOS PARA A IDENTIFICAÇÃO DE ANOMALIAS EM TRANSAÇÕES FINANCEIRAS

Trabalho de Conclusão de Curso (Artigo Científico) apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Teresina.

Orientador: Prof. Me. Fernando Castelo Branco Gonçalves Santana

TERESINA, PIAUÍ
2026

APLICAÇÃO DE ALGORITMOS E MÉTODOS ESTATÍSTICOS PARA A IDENTIFICAÇÃO DE ANOMALIAS EM TRANSAÇÕES FINANCEIRAS

Ana Beatriz Brito de Farias[1]
Fernando Castelo Branco Gonçalves Santana[2]

RESUMO

Com o crescimento das transações digitais e o surgimento de fraudes sofisticadas, análises manuais tornaram-se inviáveis, exigindo sistemas automatizados em tempo real. Este trabalho objetiva aplicar inteligência de dados, combinando estatística e aprendizado de máquina, na identificação de anomalias financeiras para suporte à tomada de decisão. A metodologia divide-se em duas etapas principais: primeiro, empregam-se métodos como o *Z-Score* e o Intervalo Interquartil para a análise inicial e detecção de valores extremos. Complementarmente, utiliza-se o aprendizado não supervisionado com os algoritmos *Density-Based Spatial Clustering of Applications with Noise*, *Local Outlier Factor* e *Isolation Forest*, permitindo mapear comportamentos suspeitos em grandes volumes de dados sem a necessidade de históricos prévios de fraude. O desempenho das técnicas foi avaliado em métricas e representações gráficas integradas. Conclui-se que a integração entre modelagem estatística e mineração de dados fortalece o monitoramento e a segurança dos processos, provendo maior proteção institucional contra riscos financeiros. Os resultados indicam que *Isolation Forest* obteve um melhor desempenho, dentre os algoritmos.

Palavras-chave: detecção de fraudes; aprendizado de máquina; detecção de anomalias; aprendizado não supervisionado; segurança financeira.

ABSTRACT

With the growth of digital transactions and the emergence of sophisticated fraud, manual analysis has become unfeasible, requiring automated real-time systems. This study aims to apply data intelligence, combining statistics and machine learning, to identify financial anomalies to support decision-making. The methodology is divided into two main stages: first, methods such as *Z*-Score and Interquartile Range are employed for initial analysis and detection of extreme values. Complementarily, unsupervised learning is utilized with Density-Based Spatial Clustering of Applications with Noise, Local Outlier Factor, and Isolation Forest algorithms, allowing the mapping of suspicious behaviors in large volumes of data without the need for prior fraud history. The performance of the techniques was evaluated using integrated metrics and graphical representations. It is concluded that the integration between statistical modeling and data mining strengthens monitoring and process security, providing greater institutional protection against financial risks. The results indicate that the Isolation Forest achieved the best performance among the algorithms.

Keywords: fraud detection; machine learning; anomaly detection; unsupervised learning; financial security.

---

[1] Graduanda em Análise e Desenvolvimento de Sistemas pelo Instituto Federal do Piauí (IFPI) . Email: anafariasjust@gmail.com

[2] Mestre em Ciência da Computação (UFMA, 2016). Especialista em Desenvolvimento Web (UFPI, 2004) e Graduado em Processamento de Dados (AESPI, 2002). Professor no Instituto Federal do Piauí. Email: fernandosantana@ifpi.edu.br

Data de Aprovação: 15/06/2026

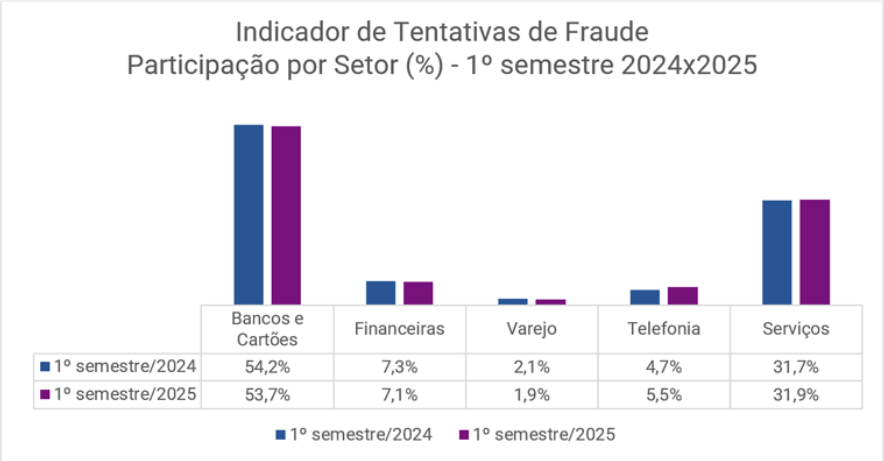
1 INTRODUÇÃO

A transformação digital tornou-se um pilar indispensável para o desenvolvimento socioeconômico global, moldando a forma como a sociedade interage com os novos canais de atendimento. Sob essa perspectiva, o avanço tecnológico observado nas últimas décadas transformou profundamente a dinâmica das transações financeiras, promovendo uma mudança estrutural que vai além da automação de processos ao inaugurar uma era de conveniência, eficiência e redução de custos operacionais (GOV.BR, 2023).

No entanto, esse cenário de expansão acelerada e o uso de inteligência artificial aplicada ao crime ampliaram significativamente as vulnerabilidades do ecossistema cibernético, “resultando em fraudes sofisticadas e prejuízos vultosos que elevaram a vigilância estratégica das instituições e autoridades reguladoras”(MONTINI, 2024).

Refletindo a gravidade desse panorama no cenário nacional, a figura 1 mostra dados do primeiro semestre de 2025 apontam que 53,7% das tentativas de fraude no país foram direcionadas a bancos e emissores de cartões, registrando uma ocorrência a cada 4,2 segundos devido ao alto potencial de retorno financeiro visado pelos criminosos. (SERASA EXPERIAN, 2025).

Figura 1 – Tentativas de fraudes financeiras no Brasil (2024-2025)



Fonte: Serasa Experian (2025).

A mineração de dados, também denominada descoberta de conhecimento em bancos de dados, constitui um conjunto essencial de técnicas fundamentadas em aprendizado de máquina e análise estatística para a identificação de padrões, correlações e anomalias em grandes volumes de dados (IBM, 2024). Impulsionada pela evolução do *big data* e do *data warehousing*, essa abordagem operacional atua de duas formas principais: descrevendo o conjunto de dados de destino ou prevendo resultados por meio de algoritmos automatizados. No contexto da segurança, a integração dessas tecnologias acelera significativamente a extração de insights relevantes e viabiliza a automação dos processos de análise, permitindo organizar e filtrar fluxos massivos de informações para trazer à tona desde comportamentos de usuários e gargalos operacionais até a ocorrência de fraudes e violações de segurança(IBM, 2024).

Para responder a esse cenário de vulnerabilidades, este trabalho tem como objetivo aplicar técnicas estatísticas e algoritmos de aprendizado de máquina não supervisionado para a identificação automatizada de fraudes financeiras em transações de cartões de crédito. A metodologia proposta compreende o emprego dos métodos *Z-Score* e Intervalo Interquartil (IQR) para o tratamento preliminar de *outliers*, combinados à execução dos algoritmos *Density-Based Spatial Clustering of Applications with Noise (DBSCAN*), *Local Outlier Factor (LOF)* e *Isolation Forest*.

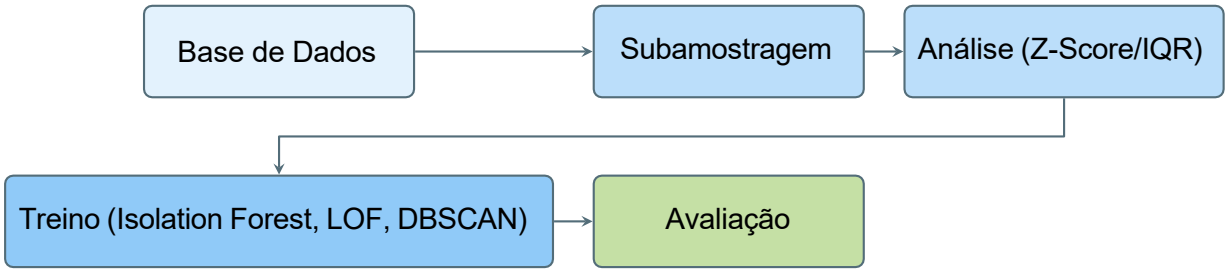
2 METODOLOGIA

A detecção de anomalias em sistemas financeiros fundamenta-se na identificação de padrões que divergem significativamente da massa crítica de dados transacionais. No contexto de fraudes com cartões de crédito, essas irregularidades manifestam-se na base de dados como *outliers*, ou seja, observações atípicas que se distanciam drasticamente do comportamento padrão da maioria das amostras. Tais pontos fora da curva podem representar tanto erros operacionais e ruídos de processamento quanto atividades criminosas estrategicamente planejadas para mimetizar o comportamento do usuário comum. Contudo, conforme apontam (HOPPEN; PRATES, 2017) , “embora o *outlier* seja tradicionalmente tratado como um ruído prejudicial à calibração de modelos estatísticos, em sistemas voltados à segurança bancária ele deixa de ser um problema de amostragem e passa a ser

exatamente o foco central da investigação".

O fluxograma ilustrado na Figura 2 mostra as etapas seguidas durante o trabalho, organizadas em cinco macroetapas sequenciais para garantir a robustez da análise. Inicialmente, a Base de Dados bruta é carregada para o ambiente de execução. Na sequência, realiza-se a Subamostragem para equalizar a proporção entre as classes, seguida pela fase de Análise (*Z-Score*/IQR), responsável pelo pré-processamento, tratamento de valores extremos e padronização das escalas numéricas. Com os dados devidamente tratados, o bloco de Treino executa simultaneamente os três algoritmos não supervisionados selecionados (*Isolation Forest*, *Local Outlier Factor* e DBSCAN). Por fim, a etapa de Avaliação mensura e compara o desempenho de cada modelo por meio de matrizes de confusão e métricas de classificação.

Figura 2 – Fluxo do Processamento de Dados



Fonte: Produzido pela autora (2026).

A base de dados utilizada apresentou um desafio inicial de desbalanceamento severo, contendo apenas 492 transações fraudulentas (0,172%) em um universo de 284.807 registros. Para viabilizar a análise técnica dos algoritmos não supervisionados e reduzir o ruído estatístico provocado pela desproporção extrema, aplicou-se a técnica de subamostragem aleatória (*Random Undersampling*). "Essa abordagem integra os métodos em nível de dados desenvolvidos para lidar com a classificação desbalanceada, atuando diretamente na transformação do conjunto de treinamento para mitigar o viés dos algoritmos em direção à classe majoritária." (SALMI, 2024). Neste estudo, a opção foi reduzir estrategicamente os registros legítimos em vez de duplicar a classe minoritária, o que permitiu suavizar o desbalanceamento original sem introduzir dados sintéticos e, simultaneamente, otimizar a eficiência computacional do pipeline de processamento. Esse procedimento

preservou a totalidade das 492 fraudes e selecionou uma amostra de 9.500 registros legítimos, totalizando um conjunto de dados de 9.992 transações. Essa configuração permitiu suavizar o desbalanceamento original para uma taxa de aproximadamente 4,92% de anomalias, mantendo a característica de raridade da fraude necessária para o teste dos modelos, ao mesmo tempo em que tornou o processamento computacional viável.

Optou-se por não utilizar técnicas de sobreamostragem sintética, como o *SMOTE*, pois a geração de dados artificiais poderia descaracterizar a raridade das anomalias, prejudicando o desempenho de algoritmos baseados em isolamento estrutural. Complementarmente à amostragem, utilizou-se o método *Z-Score* e o Intervalo Interquartil (IQR) para uma validação estatística rigorosa. Essa etapa permitiu constatar que valores extremos de transação nem sempre correspondem a atividades ilícitas, eviden-ciando a necessidade de modelos mais sofisticados que o simples filtro de valores.

A partir desse cenário controlado, a pesquisa procedeu com a aplicação comparativa de três abordagens distintas: o *Isolation Forest*, focado no isolamento de atributos; o LOF (*Local Outlier Factor*), baseado em densidade local; e o DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), que utiliza agrupamento espacial.

Essa diversidade metodológica permitiu avaliar qual técnica melhor identifica as características intrínsecas das anomalias em relação ao comportamento padrão da base, garantindo maior clareza na distinção entre classes.

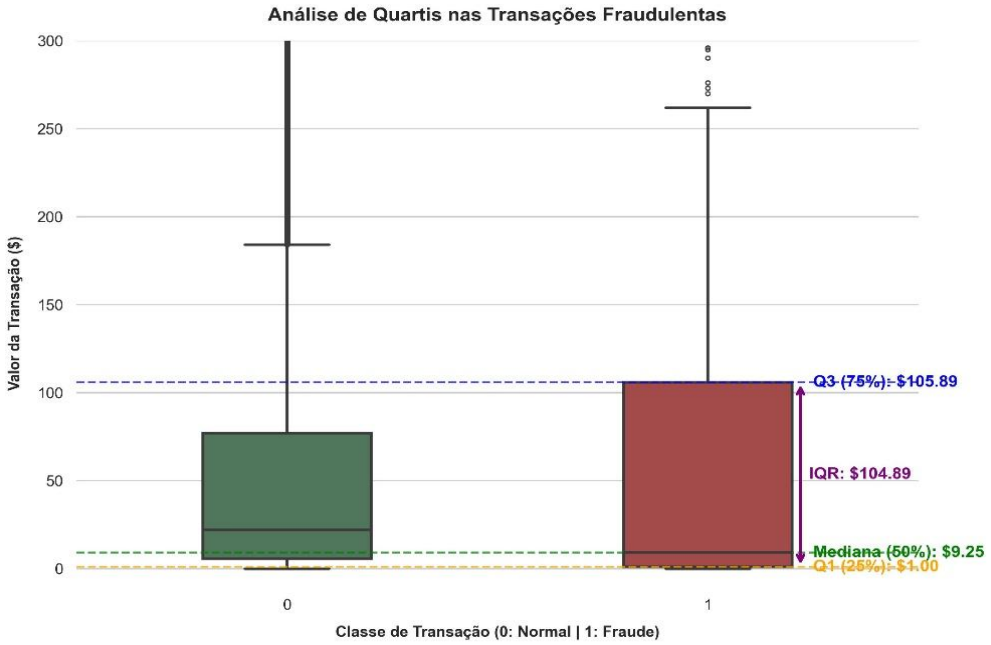
2.1 ANÁLISE DE DISPERSÃO E *OUTLIERS* - INTERVALO INTERQUARTIL E *Z-SCORE*

Para compreender a distribuição volumétrica das transações, aplicou-se uma análise estatística focada na variável Amount (valor). O objetivo foi identificar se valores extremos de consumo estariam diretamente correlacionados à ocorrência de fraudes.

Primeiramente, utilizou-se o Intervalo Interquartil (IQR),Essa técnica consiste em uma medida de variação que divide os dados ordenados em quatro partes iguais (Q1, Q2 e Q3), definindo o IQR como Q3 − Q1. Sob essa regra, qualquer valor que ultrapasse o limite de Q3 + 1, 5 × IQR ou fique abaixo de Q1 − 1, 5 × IQR é classificado

como um *outlier* (ORACLE, 2024). Conforme ilustrado na Figura 3, os dados revelaram uma mediana de $9,25 e um terceiro quartil (Q3) de $105,89, demonstrando que a grande massa das fraudes se concentra em valores baixos. O limite superior para detecção de *outliers* comuns foi estabelecido em $263,22 (Q3 +1, 5×IQR), evidenciando que os valores que ultrapassam essa barreira fogem do padrão habitual do próprio comportamento fraudulento neste conjunto de dados.

Figura 3 – Intervalo Interquartil



Fonte: Produzido pela autora(2026).

O Z-Score foi integrado logo após a análise do IQR para padronizar a escala de dispersão e quantificar a distância de cada transação em relação à média geral do dataset. Enquanto o IQR permitiu delimitar a zona central de atuação do comportamento fraudulento sem sofrer distorções causadas por valores extremos, o Z-Score foi integrado para transformar os dados em uma escala comparável com média zero e desvio padrão um. A escolha dessa medida estatística justifica-se por ela ser amplamente utilizada para avaliar a posição de um ponto de dados em relação à média de um conjunto de dados e sua dispersão, atuando como uma ferramenta fundamental para identificar valores extremos ou anômalos dentro da distribuição transacional (Comunikar Editorial, 2026).

Aplicou-se o Z-Score, estabelecendo um limiar rigoroso de 10 desvios padrão para a classificação de anomalias extremas. Para facilitar a compreensão da

distribuição estatística gerada por essa técnica, o mapeamento dos componentes analíticos e sua respectiva representação gráfica estão detalhados na Tabela 1.

Tabela 1 – Mapeamento dos Elementos Visuais do Gráfico de *Z-Score* - Figura 4

Elemento Visual	**Componente**	**Descrição e Significado Estatístico**
Eixo X	Índice de Transações	Sequência individualizada de cada operação usada como identificador único para evitar a agregação de dados e expor a dispersão real.
Eixo Y	*Z-Score*	Magnitude do desvio estatístico; valores mais elevados indicam transações que se afastam drasticamente da média populacional.
Linha Horizontal	Limiar (*Threshold*)	Linha preta tracejada no nível Y = 10 que delimita a fronteira de decisão e divide a variabilidade aceitável da anomalia severa.
Marcadores Verdes/Azuis	Zona de Normalidade	Representam a massa crítica de dados transacionais (inliers) que operam dentro da zona de confiança e da tendência central.
Mercadores Vermelhos	Anomalias Severas	Identificam os *outliers* extremos que ultrapassaram a linha de corte, possuindo probabilidade de ocorrência extremamente baixa

Fonte: Produzido pela autora (2026).

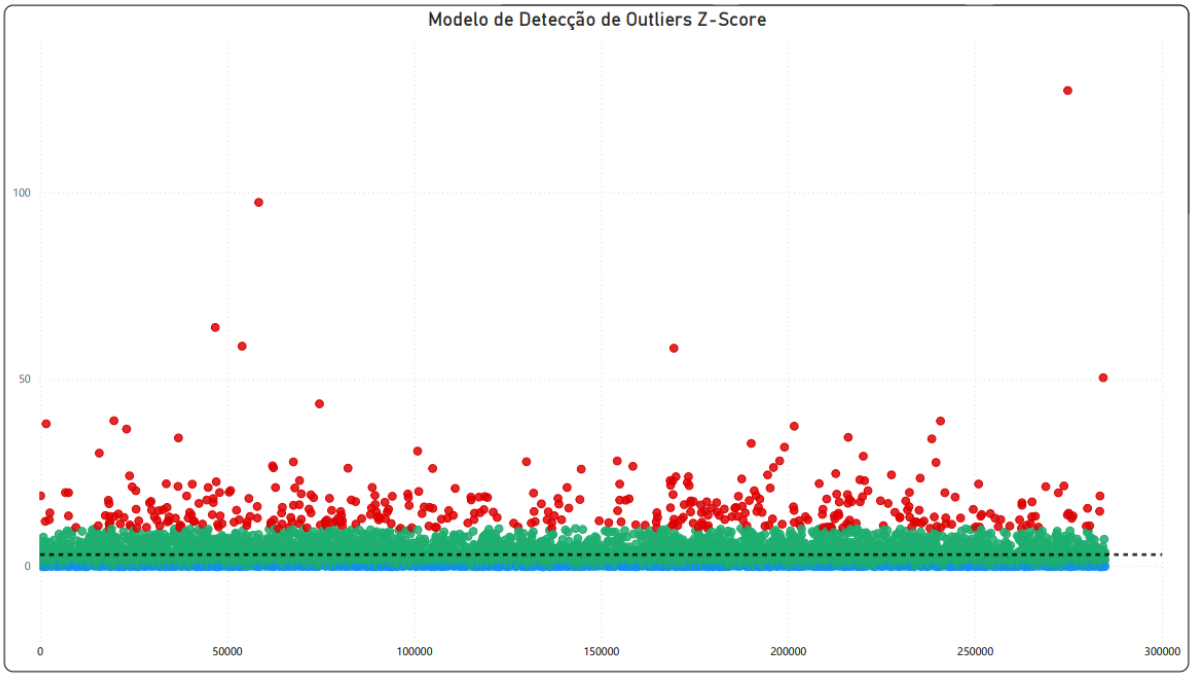
O mapeamento desses elementos sobre o conjunto de dados gerou a distribuição espacial do Z-Score, detalhada a seguir na Tabela 1. A Figura 4 apresenta visualmente essa dispersão global, permitindo a identificação imediata dos pontos que violaram o limiar estipulado.

Divergindo desse cenário de isolamento nos extremos estatísticos, ao analisar o comportamento das fraudes confirmadas sobre o eixo temporal, nota-se uma

dinâmica oposta. As fraudes reais não se concentram no topo do gráfico associadas a valores altos de desvio, mas aparecem dispersas e ocultas em meio à massa densa de transações legítimas.

Esse fenômeno revela que a fraude real é estatisticamente "discreta". O fraudador não busca o maior valor possível em uma única compra, mas sim quantias que passem despercebidas pelos filtros tradicionais de limite de crédito. A ação fraudulenta evita deliberadamente caracterizar-se como um *outlier* de valor absoluto para garantir o sucesso da transação de forma dissimulada.

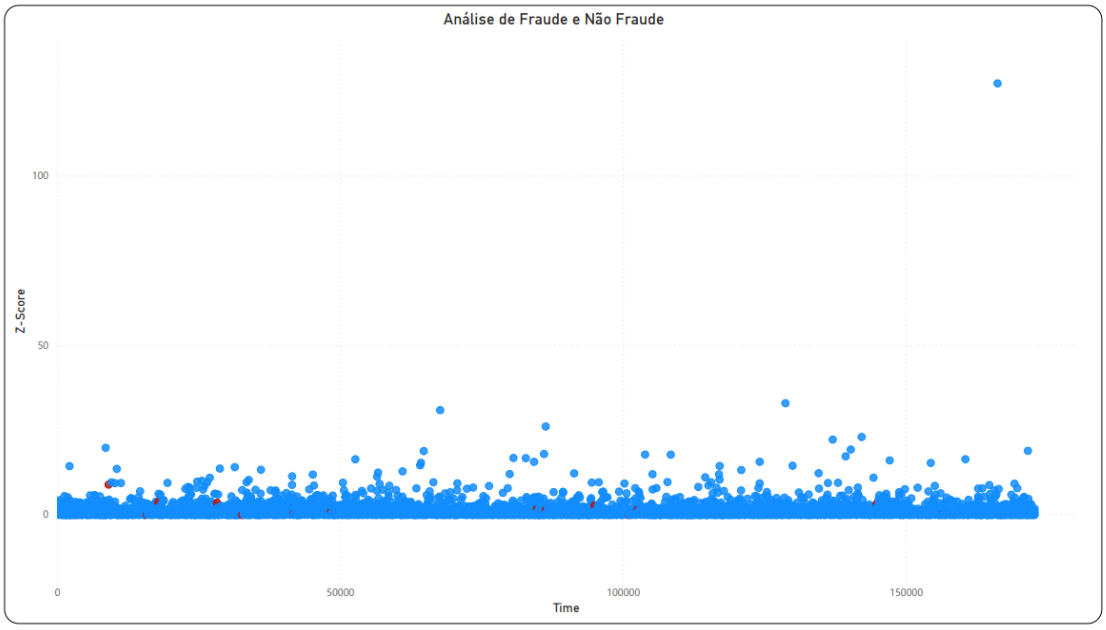
Figura 4 – Detecção de *Outliers* por meio do Escore Z



Fonte: Produzido pela autora(2026).

No gráfico da figura 5 mostra que ao plotar as fraudes confirmadas no eixo temporal observa-se um comportamento oposto. As fraudes reais não estão no topo do gráfico (valores altos), mas sim misturadas à massa densa de transações comuns. O que revela que a fraude real é estatisticamente "discreta". O fraudador não busca o maior valor possível em uma única compra, mas sim valores que passam despercebidos pelos filtros de limite de crédito. A fraude real evita ser um *outlier* de valor para garantir o sucesso da transação.

Figura 5 – Fraude e Não Fraude Real



Fonte: Produzido pela autora(2026).

A convergência dos métodos demonstrou que transações com altíssimo índice de dispersão (Z-Score > 50 ou valores muito acima do limite IQR) são, em sua maioria, falsos positivos, representando apenas gastos legítimos de alto valor. Observou-se que a fraude real é estatisticamente discreta, situando-se majoritariamente dentro ou próximo aos limites normais de consumo para evitar disparos de bloqueios automáticos. Essa evidência justifica a necessidade de algoritmos de aprendizado de máquina, visto que uma simples filtragem estatística por valor seria insuficiente para detectar o comportamento criminoso.

2.2 ALGORITMOS NÃO SUPERVISIONADOS DE DETECÇÃO DE ANOMALIAS

Como a análise univariada baseada em limites de valor não acompanha o comportamento velado das fraudes, o estudo adotou três algoritmos de aprendizado de máquina não supervisionado: O *Isolation Forest*, *Local Outlier Factor* e o DBSCAN. A escolha dessas técnicas justifica-se pela capacidade individual de cada uma de mapear o espaço amostral de forma multidimensional, buscando assinaturas de fraude que vão além do montante financeiro da transação.

O *Isolation Forest* atua isolando as anomalias por meio de árvores de decisão aleatórias, baseando-se no princípio de que as fraudes são raras e possuem atributos

discrepantes, o que permite detectá-las com alta eficiência computacional em bases transacionais massivas (WASPADA et al., 2020). O LOF, por sua vez, foca na densidade local, comparando a vizinhança de uma transação com o comportamento histórico do perfil do usuário para capturar fraudes camufladas que passariam despercebidas por análises globais (BREUNIG et al., 2000). Por fim, o DBSCAN complementa a abordagem ao agrupar transações legítimas em formatos geométricos complexos e classificar automaticamente como ruído ou anomalia qualquer registro disperso em regiões de baixa densidade espacial, sem a necessidade de estipular uma taxa de contaminação prévia (GAVA et al., 2018).

A definição e a escolha das variáveis de controle de cada modelo foram estruturadas de forma a garantir que os algoritmos operassem com base nas propriedades estatísticas da amostra obtida após o processo de subamostragem. O detalhamento das configurações específicas, os valores atribuídos a cada critério de ajuste e as respectivas premissas teóricas que nortearam essas escolhas estão apresentados na Tabela 2.

Tabela 2 – Hiperparâmetros utilizados no treinamento dos modelos

Algoritmo	**Hiperparâmetros Adotados**	**Hiperparâmetros Adotados**
Isolation Forest	*contamination* = 0.047 *random_state* = 42	Ajuste baseado na proporção da amostra tratada.
LOF	*n_neighbors* = 20 *contamination* = 0.04	Configuração padrão de vizinhança para densidade local.
DBSCAN	eps = 4.0 *min_samples* = 10	Definição do raio espacial para isolar ruídos eficientemente.

Fonte: Produzido pela autora (2026).

O algoritmo *Isolation Forest* adota uma premissa inversa à dos métodos tradicionais de modelagem. Em vez de aprender exaustivamente o comportamento das transações legítimas para depois buscar desvios, o algoritmo foca diretamente no isolamento das anomalias. Como os registros fraudulentos são raros e possuem combinações atípicas de atributos, eles são facilmente isolados nas primeiras divisões de árvores de decisão binárias geradas aleatoriamente. O modelo calcula um escore de anomalia baseado no comprimento do caminho percorrido da raiz até a folha

terminal da árvore; caminhos mais curtos indicam alta probabilidade de fraude. Na prática, este algoritmo apresentou o melhor equilíbrio operacional no projeto com um F1-Score elevado, demonstrando eficácia em capturar crimes ocultos em faixas de consumo comuns.

Para complementar o mapeamento espacial, aplicou-se o *Local outlier factor* (LOF), um método fundamentado no cálculo da densidade relativa local. O algoritmo avalia o grau de anomalia de cada transação comparando a densidade do ponto em questão com a densidade de seus vizinhos mais próximos. Enquanto as transações legítimas costumam formar agrupamentos altamente densos, as fraudes situam-se em regiões consideravelmente mais esparsas do espaço de atributos. O LOF atribui uma pontuação contextual que permite identificar anomalias de baixo valor que destoam do padrão da vizinhança ao redor. Contudo, devido à sua alta sensibilidade a pequenas variações de densidade na própria massa de transações estáveis, o modelo tendeu a registrar uma taxa de alarmes falsos superior à dos demais testados.

Por fim, utilizou-se o DBSCAN, um algoritmo de agrupamento espacial baseado em densidade projetado para descobrir estruturas de formatos livres e descartar pontos isolados diretamente como ruído. A parametrização do raio de vizinhança e do número mínimo de pontos foi calibrada para forçar o agrupamento compacto das 9.500 transações normais, segregando as fraudes como pontos não alcançáveis pelas fronteiras desse grupo principal. O DBSCAN atuou no pipeline como um mecanismo rigoroso de exclusão, alcançando um índice de Precisão excelente ao confirmar as anomalias, embora sua rigidez na formação de clusters tenha resultado em um Recall limitado, deixando de localizar uma parcela significativa das transações criminosas na base de dados, além de demandar um processo exaustivo de ajuste de hiperparâmetros para controlar o volume de falsos positivos na operação.

3 RESULTADOS

Após a normalização dos dados e o ajuste de hiperparâmetros, os algoritmos foram submetidos a testes comparativos. A Tabela 3 apresenta o desempenho bruto de cada modelo em relação à capacidade de detecção volumétrica sobre o total de fraudes reais da amostra.

Tabela 3 – Comparativo de Desempenho dos Algoritmos Não Supervisionados

Algoritmo	Premissa de Detecção	Fraudes Detectadas de 492
Isolation Forest	Isolamento Global via particionamento de atributos.	313
Local Outlier Factor	Cálculo de densidade relativa local comparada aos vizinhos.	41
DBSCAN	Agrupamento espacial baseado em densidade e identificação de ruído.	104

Fonte: Produzido pela autora (2026).

Observa-se, pelos dados expostos, uma disparidade significativa na sensibilidade dos algoritmos. O DBSCAN demonstrou uma alta rigidez na filtragem espacial, isolando um grande volume de registros como ruído genérico, porém apresentou baixa eficiência específica na identificação do comportamento fraudulento, capturando apenas 104 das 492 fraudes reais. Por outro lado, o LOF manifestou limitações ainda mais severas para este domínio, demonstrando extrema dificuldade em distinguir os padrões fraudulentos ocultos dentro da densidade local da base. Para uma compreensão técnica e aprofundada desses comportamentos, as métricas de precisão, recall e F1-Score são detalhadas a seguir na Tabela 4.

Tabela 4 – Métricas de Desempenho dos Algoritmos Testados

Algoritmo	Recall	Precisão	F1-Score
Isolation Forest	0,64	0,67	0,65
Local Outlier Factor	0,08	0,10	0,09
DBSCAN	0,21	0,73	0,33

Fonte: Produzido pela autora (2026).

Para validar a eficácia dos algoritmos não supervisionados, optou-se pela utilização das métricas de Precisão, *Recall* e *F1-Score*, visto que oferecem uma visão mais completa e equilibrada do desempenho do algoritmo em relação às classes minoritárias e majoritárias (SBC, 2024), em detrimento da acurácia tradicional, que se mostra ineficaz devido ao desbalanceamento severo dos dados transacionais.

A Precisão foi adotada para mensurar a assertividade do modelo, garantindo que transações legítimas não fossem bloqueadas erroneamente como fraudes (falsos positivos). Já o Recall foi priorizado por refletir a capacidade do sistema em detectar o maior número possível de fraudes reais (minimizando os falsos negativos). Por fim, o *F1-Score* aplicou-se como uma média harmônica entre as duas métricas anteriores, fornecendo um indicador único e equilibrado do desempenho geral dos modelos.

A análise conjunta das métricas revela que o modelo Isolation Forest consolidouse como a solução mais equilibrada, apresentando um *F1-Score* de 0,65, além de demonstrar a maior capacidade de captura real ao identificar 313 fraudes. Embora o algoritmo DBSCAN tenha alcançado a maior assertividade por alerta emitido, refletida em sua alta Precisão de 0,73, seu baixo *Recall* (0,21) resultaria em um volume elevado de fraudes perdidas (388 transações não detectadas), o que compromete a segurança institucional em um cenário bancário real. Por outro lado, o *Local Outlier Factor*(LOF) demonstrou limitações severas para este domínio, estagnando em um *F1-Score* de 0,09, o que confirma que a fraude financeira exige modelos baseados em isolamento estrutural e não apenas em densidades locais.

Em última análise, a seleção entre os modelos depende da prioridade estratégica da instituição. Contudo, para um cenário de monitoramento automatizado onde o equilíbrio entre detecção e experiência do usuário é fundamental, o *Isolation Forest* demonstrou ser a alternativa mais sustentável e robusta, oferecendo uma performance técnica superior dentro das restrições operacionais do setor bancário por apresentar métricas consistentes de sensibilidade e precisão combinadas.

4 CONSIDERAÇÕES FINAIS

O presente estudo cumpriu o objetivo de avaliar o desempenho de algoritmos de aprendizado de máquina não supervisionados na detecção de fraudes em cartões de crédito. A análise estatística por meio do Z-Score e IQR confirmou a hipótese de que a fraude real é estrategicamente discreta, mimetizando comportamentos de consumo habituais para evadir sistemas baseados estritamente em regras de montante.

Dentre os modelos testados, o Isolation Forest destacou-se pela viabilidade operacional, apresentando um F1-Score de 0,65, o que representa o melhor equilíbrio

entre a captura ativa de anomalias (Recall de 0,64) e a mitigação de falsos positivos (Precisão de 0,67), preservando a experiência do usuário. Validou-se, ainda, que a estratégia de subamostragem aleatória (Random Undersampling) é eficiente para este domínio, uma vez que atenua o desbalanceamento severo sem descaracterizar a raridade e o isolamento estrutural necessários para a classificação de fraudes genuínas.

Embora o algoritmo DBSCAN tenha demonstrado uma excelente assertividade por alerta gerado, refletida em sua Precisão (0,73), o seu baixo índice de Recall (0,21) reforça as limitações de modelos baseados estritamente em densidade espacial para este cenário, visto que a sua rigidez resultou na perda de detecção de uma parcela significativa das transações fraudulentas.

Como sugestão para trabalhos futuros, propõe-se a disponibilização do modelo em ambiente de produção. Sugere-se estruturar a solução como um serviço independente, que possa ser consultado por outros sistemas via API e protegido em contêineres. Essa abordagem garante que a ferramenta seja fácil de integrar aos sistemas bancários na nuvem, permitindo que a detecção de fraudes funcione de maneira rápida e automática para cada transação realizada.

REFERÊNCIAS

BREUNIG, M. M. et al. **LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data**. [S.l.: s.n.], 2000. p. 93–104.

Comunikar Editorial. **O que é Z-Score?** 2026. Acessado em: 1 de abr. de 2026. Disponível em: <https://comunikar.com.br/glossario/o-que-e-z-score/>.

GAVA, É. M. et al. O algoritmo Density-Based Spatial Clustering of Applications With Noise (DBSCAN) na clusterização dos indicadores de dados ambientais. In: **Anais do XI Workshop de Inteligência Artificial - UNESC**. Criciúma, SC: UNESC, 2018. Disponível em: <https://www.researchgate.net/publication/331135069>.

GOV.BR. **Tecnologia e finanças digitais: uma revolução em movimento**. 2023. Acesso em: 27 fev. 2026. Disponível em: <https://www.gov.br/investidor/pt-br/penso-logo-invisto/tecnologia-e-financas-digitais-uma-revolucao-em-movimento>.

HOPPEN, J.; PRATES, W. **Outliers, o que são e como tratálos em uma análise de dados?** 2017. <https://aquare.la/o-que-sao-outliers-e-como-trata-los-em-uma-analise-de-dados/>. Acesso em: 17 abr. 2026.

IBM. **O que é mineração de dados?** 2024. Acesso em: 27 mai. 2026. Disponível em:<https://www.ibm.com/br-pt/think/topics/data-mining>.

MONTINI, A. Bancos centrais intensificam vigilância diante de ameaças digitais com ia. **Febraban Tech**, 2024. Acesso em: 15 fev. 2026. Disponível em: <https://febrabantech.febraban.org.br/especialista/alessandra-montini/bancos-centrais-intensificam-vigilancia-diante-de-ameacas-digitais-com-ia>.

ORACLE. **Métricas do Insights: Intervalo entre Quartis (IQR)**. 2024. <https://docs.oracle.com/cloud/help/pt_BR/pbcs_common/PFUSU/insights_metrics_IQR.htm>. Acesso em: 19 abr. 2026.

SALMI, M. **Sampling for imbalance data in Python**. 2024. <https://medium.com/@salmimabrouka7/sampling-for-imbalanced-data-in-python-20fc995361db>. Acesso em: 17 abr. 2026.

SBC. Estratégias para lidar com desbalanceamento de dados em aprendizado de máquina. In: **Livro de Minicursos da Sociedade Brasileira de Computação**. SBC, 2024. Acesso em: 01 mai. 2026. Disponível em: <https://books-sol.sbc.org.br/index.php/sbc/catalog/download/152/654/1174?inline=1>.

SERASA EXPERIAN. **Recorde: quase 7 milhões de tentativas de fraude foram registradas no 1º semestre de 2025; setor bancário é principal alvo**. 2025. Disponível em: Serasa-Experian. Acesso em: 04 de abr. 2026.

WASPADA, I. et al. Performance analysis of Isolation Forest algorithm in fraud detection of credit card transactions. In: **Proceedings of the International Conference on Computer Science and Information Technology**. [S.l.: s.n.], 2020. p. 1–6.