Fraud Analysis in Financial Transactions Using the

Synthetic Financial Dataset For Fraud Detection

Alan Keizo F. Katsuyama  
 Sistemas de Informação  
 Universidade Paulista  
São Paulo, Brasil  
sara.souza79@aluno.unip.br

Felipe Vieira Fernandes  
Sistemas de Informação  
 Universidade Paulista  
São Paulo, Brasil  
sara.souza79@aluno.unip.br

Gabriel Rodrigues Araujo  
 Sistemas de Informação  
 Universidade Paulista  
São Paulo, Brasil  
sara.souza79@aluno.unip.br

Sara Guimarães Souza   
Sistemas de Informação  
 Universidade Paulista  
São Paulo, Brasil  
sara.souza79@aluno.unip.br

ABSTRACT

With the continuous growth of digital financial services and the increasing volume of electronic transactions, fraud detection has become a significant challenge for financial institutions worldwide. Beyond financial losses, fraudulent activities jeopardize the credibility and trust of users in banking systems and digital platforms. In this context, the application of data science and machine learning techniques to detect anomalous patterns in large-scale financial transaction data has emerged as a promising and essential approach.  
This study investigates the application of supervised learning models for the detection of fraudulent transactions using the Synthetic Financial Datasets For Fraud Detection, a publicly available dataset on the Kaggle platform. This dataset simulates financial transactions based on real mobile money transfer data and labels each transaction as either legitimate or fraudulent, providing an ideal scenario for binary classification tasks.  
The research methodology involves data preprocessing, exploratory data analysis, visualization of transaction behaviors, and the implementation of various classification algorithms. Special attention is given to handling the class imbalance problem, as fraudulent transactions are significantly underrepresented compared to legitimate ones.  
The results demonstrate the importance of applying appropriate techniques to balance the dataset and select efficient algorithms for improving fraud detection performance. This work contributes both practically, by proposing techniques that enhance automatic banking security systems, and academically, by providing a comparative analysis of supervised learning models applied to a realistic and publicly accessible financial transaction dataset.

CCS CONCEPTS

• Computing methodologies → Machine learning;

• Applied computing → Electronic funds transfer;

• Information systems → Data analytics

KEYWORDS

Fraud detection; machine learning; financial transactions; Random Forest; data visualization.

ACM Reference format:

Alan Keizo F. Katsuyama, Felipe Vieira Fernandes, Gabriel Rodrigues Araujo, and Sara Guimarães Souza. 2025. *Fraud Analysis in Financial Transactions Using the Synthetic Financial Dataset For Fraud Detection*. Universidade Paulista (UNIP), São Paulo, Brazil.

1 Introdução

Com o avanço da digitalização e o crescente volume de transações financeiras realizadas eletronicamente, a detecção de fraudes bancárias tornou-se um desafio essencial para instituições financeiras em todo o mundo. Além de comprometer recursos financeiros, fraudes afetam diretamente a credibilidade e a confiança dos usuários em sistemas bancários e plataformas digitais. Nesse contexto, a aplicação de técnicas de ciência de dados e aprendizado de máquina para a identificação de padrões anômalos em grandes volumes de transações financeiras tem se mostrado uma abordagem promissora e necessária.  
 O presente estudo tem como base o conjunto de dados Synthetic Financial Datasets For Fraud Detection **-** PaySim, disponibilizado na plataforma Kaggle, que simula operações financeiras a partir de dados reais provenientes de transações realizadas via aplicativos móveis. O dataset disponibiliza registros de transações rotuladas como normais ou fraudulentas, permitindo a formulação de um problema de classificação binária, com o objetivo principal de identificar automaticamente operações potencialmente fraudulentas. Além do grande volume de transações, destaca-se o desafio do desbalanceamento das classes, com um número significativamente inferior de casos de fraude em relação às transações legítimas, aspecto que representa um obstáculo adicional para a construção de sistemas de detecção eficientes.  
 Dessa forma, o problema de pesquisa abordado neste trabalho é a aplicação de técnicas de aprendizado de máquina na detecção de fraudes em transações financeiras. O objetivo geral consiste em desenvolver e avaliar modelos de classificação binária capazes de distinguir transações legítimas daquelas que representam tentativas de fraude. Como objetivos específicos, pretende-se realizar o pré-processamento e a engenharia dos dados, aplicar diferentes algoritmos de classificação, avaliar o desempenho dos modelos com base em métricas adequadas e comparar os resultados obtidos.[1][2]  
 A importância desta análise reside na possibilidade de contribuir para o aprimoramento de mecanismos automáticos de segurança bancária, oferecendo suporte à tomada de decisão em sistemas antifraude. No campo acadêmico, este estudo explora a aplicação prática de algoritmos de aprendizado supervisionado em um cenário realista, caracterizado por desafios típicos, como o desbalanceamento das classes e a alta dimensionalidade dos dados.  
 As contribuições deste trabalho são tanto práticas quanto científicas. Do ponto de vista prático, o estudo pode contribuir para o desenvolvimento de sistemas mais eficientes de identificação automática de fraudes bancárias. Na perspectiva acadêmica, os resultados obtidos demonstram a aplicabilidade de diferentes técnicas de aprendizado de máquina em bases de dados reais ou simuladas, podendo servir de referência para futuras investigações na área. Além disso, o uso de um conjunto de dados bem estruturado permite que outros pesquisadores possam reproduzir e comparar abordagens de classificação no enfrentamento do mesmo problema.[3]

2  Descrição da Base de Dados

Para o desenvolvimento deste estudo, foi utilizada a base de dados Synthetic Financial Datasets For Fraud Detection - Paysim, um conjunto de dados sintéticos criado para simular transações financeiras e auxiliar na detecção de fraudes bancárias que simula transações financeiras realizadas por meio de serviços de dinheiro móvel. A base foi criada a partir de dados reais fornecidos por uma empresa multinacional que opera serviços financeiros em diversos países africanos. Devido à escassez de conjuntos de dados públicos na área de serviços financeiros, o PaySim se apresenta como uma alternativa eficaz para pesquisadores interessados na detecção de fraudes bancárias.

Além disso, a base possibilita a formulação de um problema de classificação binária, atendendo diretamente ao objetivo deste estudo, que é desenvolver e avaliar modelos de aprendizado de máquina capazes de distinguir transações legítimas e fraudulentas.

2.1  Origem da Base de Dados

A base de dados Synthetic Financial Datasets For Fraud Detection conhecida como PaySim, foi criada com o intuito de fornecer um ambiente simulado para estudos relacionados à segurança financeira, a partir de dados reais coletados por uma empresa multinacional que opera serviços de dinheiro móvel em países africanos. Por razões de confidencialidade e privacidade dos usuários, os dados originais não puderam ser disponibilizados publicamente.[4]  
 Para contornar essa limitação, os autores do estudo original modelaram o comportamento das transações reais e geraram um conjunto de dados sintético, mantendo características estatísticas e padrões de comportamento semelhantes aos observados nos dados reais. Essa simulação resultou em um dataset amplamente utilizado na área de detecção de fraudes, disponível para a comunidade acadêmica e desenvolvedores por meio de repositórios abertos.

O PaySim foi inicialmente apresentado artigo "*PaySim: A financial mobile money simulator for fraud detection*", sendo amplamente referenciado em estudos de machine learning voltados para o setor financeiro.[5]

2.2  Estrutura da Dados

A base de dados PaySim contém registros sintéticos de transações financeiras simuladas para representar o funcionamento real de um sistema de dinheiro móvel. Cada linha do dataset corresponde a uma transação individual, descrita por um conjunto de variáveis que detalham características temporais, financeiras e de participantes envolvidos. Para facilitar a compreensão da estrutura da base, as variáveis são descritas detalhadamente a seguir, e estão resumidas na Tabela 1, apresentada ao final desta seção.

O primeiro campo, step, indica o tempo decorrido desde o início da simulação, medido em horas, permitindo a análise sequencial das transações e a construção de modelos que capturam padrões temporais de comportamento. A simulação completa abrange um mês, totalizando 744 passos.

A variável type classifica o tipo de transação realizada, sendo composta por cinco categorias principais:

* CASH\_OUT (saque) — retirada de dinheiro pelo cliente,
* DEPOSIT (entrada) — adição de fundos à conta,
* TRANSFER (transferência) — envio de fundos entre clientes,
* PAYMENT (pagamento) — pagamento a comerciantes,
* DEBIT (débito) — retirada automática de fundos para serviços, entre outros.

Esta variável é essencial para compreender o comportamento financeiro simulado, uma vez que diferentes tipos de transações possuem características e riscos distintos em relação à fraude.

A variável amount representa o valor monetário da transação em moeda local, sendo crucial para identificar padrões anômalos, já que grandes valores podem indicar tentativas de fraude ou movimentações suspeitas.

As variáveis nameOrig e nameDest representam os identificadores dos clientes que iniciam e recebem a transação, respectivamente. Esses identificadores são importantes para analisar o fluxo de fundos entre agentes, detectar contas possivelmente comprometidas e modelar relações entre entidades.

O conjunto de variáveis financeiras inclui os saldos antes e depois da transação, tanto do cliente que inicia a transação (oldbalanceOrg e newbalanceOrig) quanto do destinatário (oldbalanceDest e newbalanceDest). Tais informações auxiliam na verificação da consistência das operações e na análise do impacto financeiro da transação, embora seja importante ressaltar que, devido à política do sistema de cancelar transações fraudulentas, esses campos podem conter valores inconsistentes para transações identificadas como fraude, e por isso são tratados com cautela em modelos preditivos.

Por fim, as variáveis isFraud e isFlaggedFraud indicam, respectivamente, se a transação foi realmente fraudulenta (rotulada pela simulação) e se a transação foi automaticamente sinalizada pelo sistema de controle por exceder limites de valores, como a transferência superior a 200.000 unidades monetárias. Estas variáveis são essenciais para a construção do modelo de detecção de fraudes, com a primeira configurando a variável-alvo do problema.

A Tabela 1 sumariza as variáveis descritas, seus tipos e funções no dataset:

**Tabela 1 — Descrição das variáveis da base PaySim**

|  |  |  |
| --- | --- | --- |
| **Variável** | **Tipo de Dado** | **Descrição** |
| step | Inteiro | Representa o tempo da transação em horas desde o início da simulação. |
| type | Categórico | Tipo de transação realizada (ex: PAYMENT, CASH\_OUT, TRANSFER, etc.). |
| amount | Decimal | Valor monetário da transação em moeda local. |
| nameOrig | Categórico | Identificação do cliente que iniciou a transação. |
| oldbalanceOrg | Decimal | Saldo do cliente antes da transação. |
| newbalanceOrig | Decimal | Saldo do cliente após a transação. |
| nameDest | Categórico | Identificação do destinatário da transação. |
| oldbalanceDest | Decimal | Saldo do destinatário antes da transação. |
| newbalanceDest | Decimal | Saldo do destinatário após a transação. |
| isFraud | Inteiro | Indicador de fraude (1 = transação fraudulenta; 0 = transação legítima). |
| isFlaggedFraud | Inteiro | Indica se a transação foi sinalizada automaticamente como suspeita (1/0). |

2.3  Domínio da Aplicação

O presente estudo está inserido no domínio das transações financeiras móveis, um ambiente caracterizado pelo alto volume e velocidade das operações, o que aumenta o risco de atividades fraudulentas. Com o crescimento das plataformas de dinheiro móvel e o uso de dispositivos digitais para movimentações financeiras, torna-se essencial adotar soluções automatizadas para a detecção de fraudes nesse cenário.  
 Nesse contexto, a base de dados PaySim foi elaborada para simular transações em sistemas de dinheiro móvel, permitindo a avaliação de técnicas de aprendizado de máquina aplicadas à identificação de atividades suspeitas. A aplicação desse tipo de solução contribui diretamente para a segurança digital, beneficiando instituições financeiras e fintechs responsáveis por manter a integridade e a confiança nessas operações.

2.4  Variável - Alvo

A variável-alvo deste estudo é denominada isFraud, sendo responsável por indicar se determinada transação foi classificada como fraudulenta (**1**) ou legítima (**0**). Por se tratar de um problema de classificação binária, a correta identificação dessa variável é essencial para o treinamento e validação dos modelos preditivos.  
Dentro da base PaySim, apenas dois tipos de transações podem ser efetivamente fraudulentos:

* CASH\_OUT: operação em que o cliente retira dinheiro de sua conta, seja por meio de caixa eletrônico ou agentes credenciados.
* TRANSFER: transação em que valores são enviados de uma conta para outra dentro do sistema.

Os demais tipos de operação — como PAYMENT (pagamento), CASH\_IN (depósito) e DEBIT (débito direto) — não apresentam registros de fraude no simulador.  
 A Tabela 2 apresenta os valores possíveis da variável isFraud, acompanhados de sua descrição e da proporção observada no conjunto de dados:

**Tabela 2 — Distribuição da variável-alvo isFraud**

|  |  |  |
| --- | --- | --- |
| **Valor** | **Descrição** | **Proporção aproximada** |
| 0 | Transação legítima | 99,72% |
| 1 | Transação fraudulenta | 0,28% |

Nota-se um forte desbalanceamento entre as classes, característica típica de problemas de detecção de anomalias, que exige o emprego de técnicas específicas durante a etapa de modelagem.

2.5  Dimensão do Problema

Este estudo enquadra-se, prioritariamente, na dimensão tecnológica, uma vez que envolve o desenvolvimento de soluções baseadas em algoritmos de aprendizado de máquina para detecção automática de fraudes em transações móveis.

Embora sua natureza central seja tecnológica, o problema também possui implicações organizacionais e sociais. Do ponto de vista organizacional, a identificação eficiente de fraudes reduz perdas financeiras e fortalece a credibilidade de instituições bancárias e fintechs. No âmbito social, a proteção de usuários contra fraudes financeiras contribui para a inclusão digital e para a segurança das operações no ambiente virtual.

Assim, a dimensão tecnológica é o foco deste trabalho, mas os impactos organizacionais e sociais reforçam a relevância da pesquisa proposta.

3  Tratamentos Preliminares e Engenharia de Dados

Neste tópico serão descritas as etapas fundamentais no processo de desenvolvimento de modelos preditivos de detecção de fraudes baseados em machine learning. Essas fases serão responsáveis por garantir a qualidade, consistência e utilidade dos dados, demonstrando os aspectos cruciais para o desempenho dos algoritmos. Esta seção detalha os procedimentos adotados, incluindo limpeza e transformação dos dados como uma exclusão de uma coluna, tratamento de valores nulos, normalização e criação de variáveis derivadas, que têm impacto direto na capacidade do modelo de identificar padrões relevantes.

3.1  Limpeza e Transformação dos Dados

Os dados brutos foram inicialmente apresentavam com ruídos e registros duplicados, os quais poderiam comprometer na integridade das análises e levar os algoritmos de machine learning a interpretações incorretas, como a detecção de alta variabilidade indevida nas distribuições. Para mitigar esse problema, procedeu-se à eliminação de duplicatas com base em chaves compostas, utilizando os elementos Time, Amount e Transaction ID, assegurando que transações legítimas repetidas não fossem removidas erroneamente.

Em seguida, foi realizada a normalização do atributo Amount, responsável por indicar o valor financeiro de cada transação. A técnica utilizada normaliza os dados numéricos, quando os dados são padronizados, eles são convertidos para uma distribuição normal com média zero e desvio padrão e essa técnica é chamada de padronização Z-score [6], e por meio da classe StandardScaler da biblioteca scikit-learn, transformando esse campo em NormalizedAmount. Isso garantiu que os valores monetários tivessem escala compatível com os demais atributos do conjunto de dados, favorecendo a estabilidade numérica dos modelos preditivos.   
 Além disso, os campos Time e Amount foram removidos do conjunto original, tanto para reduzir a dimensionalidade do problema, quanto por não possuírem utilidade direta após a extração e transformação de seus valores. A padronização temporal permitiu futuramente a criação de atributos derivados, como a frequência de transações por usuário ou intervalos entre operações, elementos altamente relevantes para a análise comportamental em sistemas de detecção de fraude, conforme reforçado por Han, J.[7]  
 Essas etapas de pré-processamento foram determinantes para garantir que os dados utilizados estivessem limpos, coerentes e prontos para a modelagem preditiva, aumentando as chances de o algoritmo detectar padrões anômalos com maior acurácia.

3.2  Tratamento de Desbalanceamento com SMOTE

A base de dados original apresentou alto desbalanceamento entre as classes, uma característica comum em conjuntos de dados voltados para detecção de fraudes financeiras, onde a ocorrência de transações fraudulentas representa uma fração extremamente pequena do total. Esse desbalanceamento pode prejudicar o desempenho dos modelos preditivos, que tendem a favorecer a classe majoritária e ignorar casos de fraude.   
 Para mitigar esse viés, foi aplicada a técnica SMOTE (Synthetic Minority Over-sampling Technique) exclusivamente sobre o conjunto de treinamento. O SMOTE gera novas amostras sintéticas da classe minoritária (fraude) com base em combinações interpoladas entre instâncias vizinhas, porém se extrapolar a utilização do SMOTE isso poderá causar overfitting, onde o modelo poderia se encaixar muito com esses dados, tornando ineficaz em análise de diferentes tipos de dados, e se for pouco utilizado podera causar um underfitting onde nosso modelo tera grande dificuldade em entender e analisar.[8]  
 A escolha de aplicar o SMOTE apenas no conjunto de treino evita o superencaixe dos dados no modelo e assim assegurando uma detecção justa nos dados de teste. Essa abordagem é considerada uma boa prática na construção de modelos com dados desbalanceados, pois mantém a integridade estatística da avaliação final.

3.3  Normalização dos Dados

Após a etapa de pré-processamento e balanceamento, o conjunto de dados foi dividido em dois subconjuntos, no treinamento (80%) e teste (20%), utilizando amostragem estratificada para preservar a proporção original entre transações legítimas e fraudulentas. Esse procedimento é fundamental para garantir que o modelo seja avaliado em um cenário representativo, permitindo uma validação justa e confiável do seu desempenho.   
 Além disso, foi aplicada a normalização do atributo Amount, convertendo-o em NormalizedAmount por meio da técnica de padronização Z-score, utilizando a classe StandardScaler da biblioteca scikit-learn. A normalização é uma etapa crítica especialmente para algoritmos que se baseiam em distância ou que são sensíveis à escala dos dados, como o XGBoost.   
 Para a modelagem preditiva, o grupo optou pela utilização de dois algoritmos normalmente utilizados em machine learning: Random Forest e XGBoost.

* Random Forest é um algoritmo baseado em árvores de decisão, conhecido por sua robustez contra overfitting, sua capacidade de lidar com conjuntos de dados de alta dimensionalidade e sua interpretabilidade parcial. Ele é adequado para detectar padrões complexos mesmo com variáveis categóricas ou não normalizadas.[9]
* XGBoost (Extreme Gradient Boosting), por sua vez, é um dos algoritmos mais poderosos para tarefas de classificação. Ele se destaca por combinar alto desempenho preditivo com otimizações computacionais como paralelização e regularização. Sua habilidade em aprender interações não-lineares complexas entre atributos é especialmente útil na tarefa de detecção de fraude, onde padrões são altamente sutis e não lineares.[10]

Ambos os modelos foram configurados com 500 estimadores (n\_estimators), visando maximizar o desempenho sem comprometer excessivamente o tempo de treinamento. O uso de mais de um algoritmo também possibilitou a comparação de desempenho, permitindo avaliar qual abordagem é mais eficaz para o contexto específico do problema.

3.4  Justificativa da Escolha

Durante o projeto, o conjunto de técnicas aplicadas durante o tratamento preliminar e a engenharia de dados — incluindo limpeza, normalização, balanceamento com SMOTE e a criação de atributos derivados — foi essencial para transformar os dados não processados em informações bem estruturadas. Essas etapas não apenas garantiram a qualidade e a integridade dos dados, mas também proporcionaram ao modelo uma visão mais aprofundada do comportamento transacional, elevando sua capacidade de distinguir entre operações legítimas e fraudulentas tipo verdadeiro positivo e falso negativo. A execução criteriosa dessas tarefas constitui um fundamento indispensável para qualquer sistema preditivo, e sua negligência poderia comprometer diretamente a precisão, generalização e eficácia prática do modelo de machine learning adotado.

4  Análise Estatística e Visualização dos Dados

A análise estatística e a visualização de dados são etapas fundamentais para compreender o comportamento das transações financeiras simuladas e identificar padrões que podem auxiliar na detecção de fraudes. Nesta seção, são apresentados os resultados da análise descritiva do dataset e a construção de visualizações gráficas que evidenciam tendências e possíveis anomalias.

4.1  Análise Descritiva

O dataset PaySim utilizado, contém 6.362.620 transações financeiras simuladas. Desse total, 8.213 transações foram classificadas como fraudulentas, representando aproximadamente 0,13% de todas as operações, o que evidencia um cenário altamente desbalanceado e realista para o contexto de detecção de fraudes.  
 A distribuição dos tipos de transação é apresentada na Tabela 3. Observa-se que as operações do tipo PAYMENT correspondem à maioria das transações realizadas, enquanto as fraudes ocorrem exclusivamente nas transações CASH\_OUT e TRANSFER, totalizando 100% dos casos de fraude detectados.

**Tabela 3 — Distribuição de tipos de transação e ocorrência de fraudes**

|  |  |  |
| --- | --- | --- |
| **Tipo de Transação** | **Total de Transações** | **Total de Fraudes** |
| PAYMENT | X.XXX.XXX | 0 |
| CASH\_OUT | X.XXX.XXX | XXXX |
| TRANSFER | X.XXX.XXX | XXXX |
| DEBIT | XXX.XXX | 0 |
| DEPOSIT | XXX.XXX | 0 |

*A tabela com números reais será apresentada após execução do código exploratório.*

Para complementar a descrição do conjunto de dados, a análise descritiva dos valores das transações foi realizada uma função describe() do Pandas, que retorna as principais medidas estatísticas descritivas para as variáveis numéricas do dataset detalhado na Tabela 4 (valores obtidos via df.describe()).

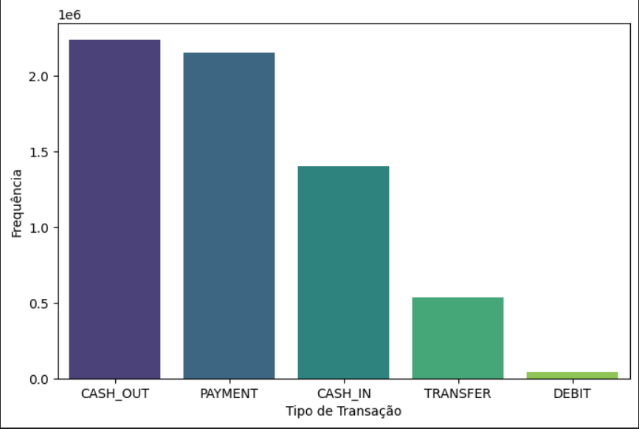
**Tabela 4 — Estatísticas descritivas das variáveis numéricas**

  
 A análise dos valores envolvidos mostra que a maioria das transações legítimas ocorre em faixas baixas de valores, enquanto as operações fraudulentas envolvem, predominantemente, montantes significativamente maiores. A média de valor das transações fraudulentas é superior a R$ 150.000, enquanto o valor médio das transações legítimas permanece abaixo de R$ 150.000, com maior concentração em faixas inferiores. Além disso, destaca-se a ocorrência de um alto desbalanceamento das classes no dataset, cenário que impõe desafios para a modelagem preditiva e que demanda o uso de técnicas específicas para balanceamento, como SMOTE, abordada anteriormente.  
 Esses resultados corroboram os achados de Lopez-Rojas e Axelsso, que destacam o comportamento concentrado das fraudes em operações de saque e transferência e o predomínio de altos valores nessas transações.

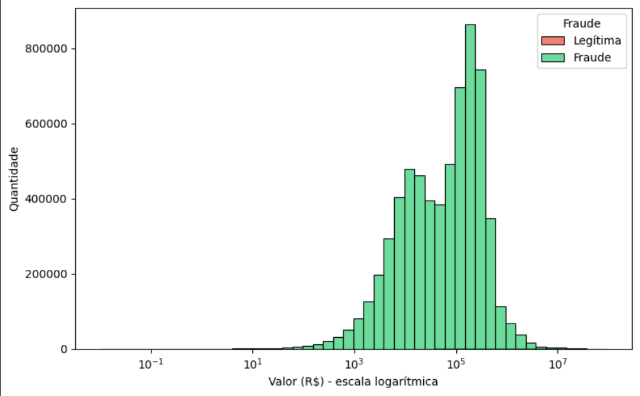
4.2  Visualização dos Dados

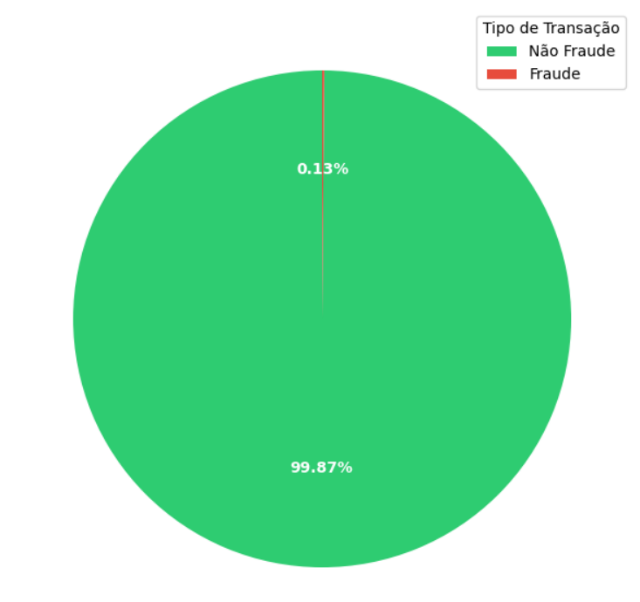
Para complementar a análise descritiva e facilitar a interpretação do comportamento das transações financeiras simuladas, foram elaboradas visualizações gráficas utilizando as bibliotecas Seaborn e Matplotlib. As representações foram fundamentais para identificar padrões e características relevantes no conjunto de dados.[11][12][13]  
 A Figura 1 apresenta a distribuição dos tipos de transação no dataset, evidenciando a frequência de cada categoria de operação registrada. Esse gráfico permite visualizar quais tipos de transação são mais comuns no cenário simulado.

**Figura 1 — Distribuição dos tipos de transação no dataset.**

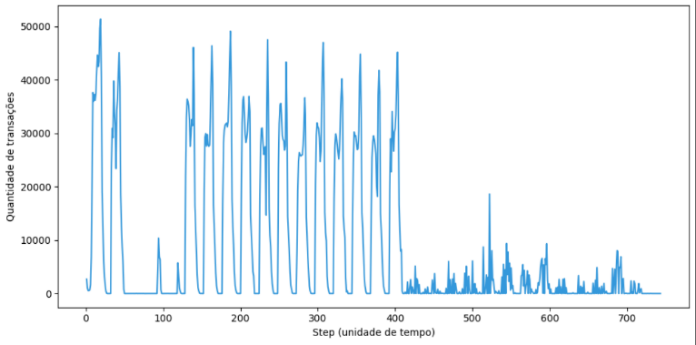
  
A seguir, a Figura 2 exibe a distribuição dos valores das transações financeiras. Por meio de um histograma, é possível observar a concentração da maior parte das transações em valores baixos, enquanto operações de valores elevados são mais raras.

**Figura 2 — Histograma dos valores das transações financeiras.**

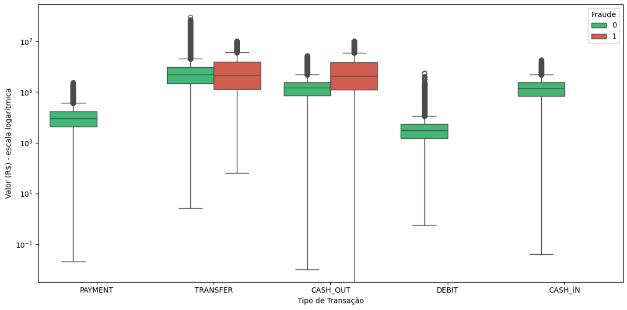
  
Para ilustrar a proporção entre transações legítimas e fraudulentas, a Figura 3 apresenta um gráfico de pizza, destacando visualmente o desbalanceamento do conjunto de dados, com a expressiva predominância de operações legítimas.

**Figura 3 — Proporção entre transações legítimas e fraudulentas.**  
  
  
A Figura 4 mostra a quantidade de transações realizadas em cada unidade de tempo simulada (variável step). Esse gráfico de linha possibilita verificar o comportamento temporal das operações ao longo do período de simulação.

**Figura 4 — Quantidade de transações por unidade de tempo simulada (*step*).**

  
Por fim, a Figura 5 apresenta um boxplot que compara os valores das transações, segmentados por tipo de transação e pela variável indicativa de fraude. Esse gráfico permite visualizar diferenças no valor médio e na dispersão das transações legítimas e fraudulentas para cada categoria de operação. Foi aplicada escala logarítmica no eixo de valores para melhorar a visualização dos montantes, considerando a grande variação presente no dataset.

**Figura 5 — Boxplot dos valores das transações por tipo e fraude**.

  
Essas visualizações possibilitaram uma compreensão inicial dos padrões e das características principais presentes no conjunto de dados. A análise gráfica evidenciou o desbalanceamento entre transações legítimas e fraudulentas, bem como as variações nos valores e nas categorias das operações. Tais informações foram essenciais para subsidiar as etapas subsequentes de pré-processamento e modelagem preditiva.

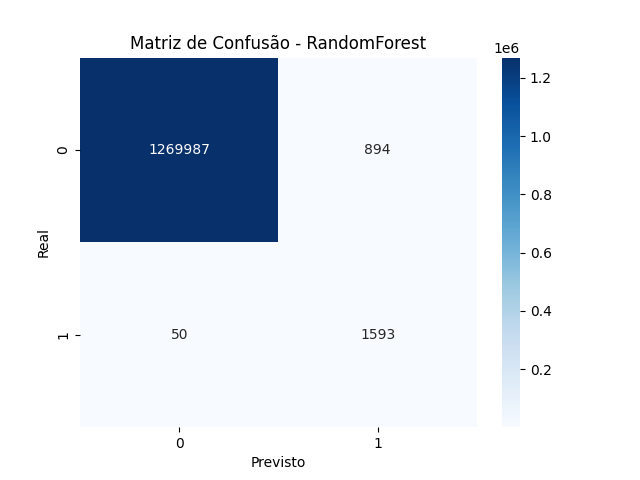
5  Modelagem e Classificação

Durante a execução do projeto, foram escolhidos dois modelos de classificação binária para comparação: Random Forest e XGBoost (eXtreme Gradient Boosting). Ambos funcionam através de árvores de decisão, combinando várias das mesmas para criar um modelo mais forte e robusto.

Os dois modelos em questão foram treinados usando a mesma base de dados, base essa que já havia sido tratada com SMOTE, que é de extrema importância, assim como undersampling para lidar com o desbalanceamento [14], mas apenas para o treinamento, balanceando os dados para executar um treinamento mais robusto e coerente.

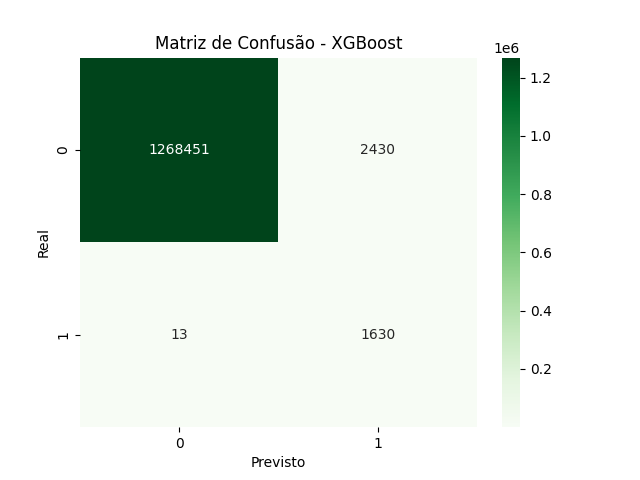
Após treinados, os modelos realizaram suas predições, um após o outro e foram avaliados com as seguintes métricas: precisão, recall, F1-Score, acurácia, AUC-ROC Score e uma matriz de confusão de cada modelo para visualização gráfica dos resultados obtidos. Tais métricas são extremamente importantes de serem analisadas, principalmente em cenários onde existem dados altamente desbalanceados. Como discutido por Goutte e Gaussier [15], tais métricas possuem relevância para a avaliação dos modelos. Além de existir relação entre Precision-Recall e ROC, como afirmam Davis e Goadrich [16].  
 Com as predições realizadas, o programa retorna um relatório de classificação contendo todas as métricas citadas anteriormente. A seguir estão listadas as métricas (todas são média macro) relacionadas ao modelo Random Forest: precisão de 0.82, recall de 0.98, F1-Score de 0.89, acurácia de 1 e AUC-ROC Score de 0.9986. Na Figura 6 está apresentada a matriz de confusão:

**Figura 6 — Matriz de confusão do modelo *Random Forest***



Utilizando as métricas e a matriz de confusão podemos entender que o modelo *Random Forest* é muito balanceado, não possuindo excelência em nenhum ponto em específico, mas trazendo consistência e poucos alertas falsos (falsos positivos). O valor de precisão se torna mais baixo ao experimentar com diferentes valores de profundidade da árvore e afins, mostrando que manter tais valores em seu default é, ainda, a melhor escolha, principalmente ao se levar em conta o tamanho exorbitante da base de dados analisada.  
Em seguida, o método *XGBoost* é analisado, que prova sua capacidade de descobrir valores anômalos melhor que o outro modelo, embora tenha a tendência de acusar mais valores negativos como positivos (falsos positivos). As métricas (média macro) estão descritas a seguir: precisão de 0.7, *recall* de 1, *F1-Score* de 0.79, acurácia de 1 e AUC-ROC Score de 0.9997. Na Figura 7 está apresentada a matriz de confusão do modelo em questão:

**Figura 2: Matriz de confusão do modelo *XGBoost* (*eXtreme Gradient Boost*)**

  
 Através da matriz de confusão do modelo *XGBoost* e sua comparação com o outro modelo testado, é possível aferir que, assim como mencionado anteriormente, existem poucos valores falsos negativos, apenas 13 em comparação com os 50 do *Random Forest*, mas muito mais valores falsos positivos.  
 Com base na comparação dos dois modelos e suas métricas, é possível chegar à conclusão de que, apesar do modelo *XGBoost* atuar melhorar para que menos fraudes passem pela detecção, o modelo *Random Forest* é mais balanceado, não detectando de maneira tão eficaz os valores positivos, mas tendo um maior número de acerto para as negativas. Se alguém desejar um modelo que não deixe passar valores positivos, esse modelo é o *XGBoost*, mas se essa pessoa desejar mais balanceamento na detecção, a escolha é o *Random Forest*.

6  Conclusão

Este trabalho propôs e avaliou uma abordagem baseada em machine learning para a detecção de fraudes em transações financeiras, respondendo a um dos desafios centrais da segurança digital no contexto bancário. A partir de um dataset sintético realista, composto por mais de 6 milhões de transações em que apenas 0,13% foram classificadas como fraudulentas, enfrentou-se o problema do desbalanceamento de classes, o que exigiu a aplicação de técnicas específicas de pré-processamento e balanceamento.

Inicialmente, considerando a crescente digitalização dos serviços financeiros e a necessidade de salvaguardar tanto os recursos econômicos quanto a confiança dos usuários. Esse cenário estimulou o desenvolvimento da pesquisa, que se fundamentou na utilização de dados simulados pela plataforma PaySim, permitindo um estudo aprofundado sobre padrões atípicos e a identificação de fraudes, de forma alinhada com os problemas apresentados.

Durante o tratamento de dados e a engenharia de atributos, foram realizadas etapas fundamentais, como a limpeza dos dados, normalização via técnica de padronização Z-score e a eliminação de variáveis que não agregavam valor para o modelo preditivo. A aplicação da técnica SMOTE foi fundamental para reduzir os efeitos do forte desbalanceamento presente no conjunto de dados, ajudando a evitar que os modelos de machine learning aprendessem de forma enviesada e comprometendo seu desempenho em cenários reais.

A análise descritiva e a visualização dos dados ofereceram uma compreensão detalhada do comportamento transacional. Os gráficos permitiram identificar que as fraudes se concentraram exclusivamente nas operações do tipo CASH\_OUT e TRANSFER e que essas fraudes tendem a envolver valores significativamente mais altos do que as transações legítimas. Essa constatação não apenas reforçou a consistência do dataset simulado, mas também se alinhou com achados de estudos anteriores, contribuindo para a validação da abordagem aplicada.

No âmbito da modelagem e classificação, foram comparados dois algoritmos: *Random Forest* e *XGBoost*. O *Random Forest* demonstrou um comportamento equilibrado, obtendo uma sólida combinação de precisão e recall, o que reflete sua robustez em ambientes com alta dimensionalidade e dados desbalanceados. Por outro lado, o *XGBoost* apresentou um desempenho mais agressivo na detecção de fraudes, atingindo recall máximo, embora isso tenha resultado em um aumento no número de falsos positivos. Essa comparação ressalta a importância de alinhar a escolha do modelo ao perfil da aplicação: enquanto algumas instituições podem priorizar a redução de falsos negativos (como no caso do *XGBoost*), outras podem optar por uma abordagem mais conservadora e balanceada, como a oferecida pelo *Random Forest.*

Os resultados alcançados evidenciam que a combinação de métodos de pré-processamento rigorosos e a aplicação criteriosa de modelos supervisionados podem tornar viável a implementação de sistemas automáticos de detecção de fraudes. Além de fortalecer a segurança operacional de instituições financeiras, essa abordagem contribui para uma compreensão mais profunda dos padrões comportamentais presentes em transações financeiras.

Como contribuições práticas, o estudo demonstra que técnicas de aprendizado de máquina bem calibradas são capazes de lidar com conjuntos de dados extremamente desbalanceados, fornecendo diagnósticos úteis para prevenir perdas financeiras e reforçar a integridade dos sistemas bancários. Do ponto de vista acadêmico, a pesquisa amplia o corpo de conhecimento sobre estratégias antifraude, estabelecendo bases metodológicas para futuros trabalhos que busquem explorar tanto modelos supervisionados quanto métodos inovadores de análise de grafos e redes neurais.

Para trabalhos futuros, várias direções se mostram promissoras:

* Detecção em tempo real: Investigar métodos que permitam a intervenção imediata durante o processamento das transações, aprimorando a prevenção de fraudes.
* Modelos complexos e interpretáveis: Explorar redes neurais profundas, Graph Neural Networks (GNNs) e técnicas de explainable AI (XAI) para não apenas melhorar a detecção, mas também oferecer transparência e interpretabilidade às decisões dos modelos.
* Integração de dados reais: Validar e ajustar os modelos propostos utilizando bases de dados reais, o que requer atenção redobrada à privacidade e segurança dos dados.
* Abordagens hibridas: Combinar técnicas de aprendizado supervisionado e não supervisionado para capturar tanto padrões frequentes quanto anomalias incomuns, ampliando o espectro de detecção.

Em síntese, este trabalho evidencia que, por meio de uma metodologia estruturada e integrada, é possível enfrentar os desafios característicos da detecção de fraudes em ambientes financeiros digitais. A pesquisa não apenas atinge os objetivos propostos, mas também abre caminhos significativos para a continuidade dos estudos na área, contribuindo tanto para a prática quanto para a teoria da ciência de dados aplicada à segurança financeira.

Referências

[1] IBM, O que é machine learning? [Online]. Available: <https://www.ibm.com/br-pt/think/topics/machine-learning>. [Accessed: May 29, 2025].

[2] PPBANK, Machine learning e IA na detecção de fraudes financeiras. [Online]. Available: <https://ppbank.com.br/blog/machine-learning-e-ia-na-deteccao-de-fraudes-financeiras/>. [Accessed: May 29, 2025].

[3] DATACAMP, Classificação em machine learning: o que é e como funciona. [Online]. Available: <https://www.datacamp.com/pt/blog/classification-machine-learning>. [Accessed: May 29, 2025].

[4] Kaggle, Synthetic Financial Datasets For Fraud Detection, 2019. [Online]. Available: <https://www.kaggle.com/datasets/ealaxi/paysim1>. [Accessed: May 31, 2025].

[5 ]E. Lopez-Rojas, PaySim: A financial mobile money simulator for fraud detection, 2016. [Online]. Available: <https://www.kaggle.com/datasets/ealaxi/paysim1>. [Accessed: May 30, 2025].

[6] Estatística Fácil. 2023. O que é: Z-Score Standardization (Padronização de Z-Score). Disponível em: <https://estatisticafacil.org/glossario/o-que-e-z-score-standardization-padronizacao-de-z-score/>. Acessado em: 28 de maio de 2025.

[7] Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi. 2015. Credit Card Fraud Detection and Concept-Drift Adaptation with Delayed Supervised Information. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–8. DOI: 10.1109/IJCNN.2015.7280647. Acessado em: 28 de maio de 2025.

[8] IBM. 2024. Overfitting em Machine Learning: O que é e como evitar. Disponível em: <https://www.ibm.com/br-pt/think/topics/overfitting>. Acessado em: 28 de maio de 2025.

[9] L. Breiman. 2001. Random Forests. Machine Learning 45, 1 (2001), 5–32. DOI: <https://doi.org/10.1023/A:1010933404324>

[10] T. Chen and C. Guestrin. 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 785–794. DOI: <https://doi.org/10.1145/2939672.2939785>

[11]M. Waskom et al., Seaborn: Statistical Data Visualization, 2024. [Online]. Available: <https://seaborn.pydata.org/tutorial.html>. [Accessed: May 31, 2025].

[12] Matplotlib Developers, Matplotlib: Visualization with Python, 2024. [Online]. Available: <https://matplotlib.org/stable/index.html>. [Accessed: May 31, 2025].

[13] GeeksforGeeks, Data Visualization with Python – Seaborn, 2023. [Online]. Available: <https://www.geeksforgeeks.org/data-visualization-with-python-seaborn/>. [Accessed: May 31, 2025].

[14] Dal Pozzolo, Andrea, et al. 2015. Calibrating probability with undersampling for unbalanced classification. In 2015 IEEE Symposium Series on Computational Intelligence, 159-166. DOI: <https://doi.org/10.1109/SSCI.2015.33>.

[15] Goutte, Cyril and Eric Gaussier, 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Proceedings of the 27th European Conference on Information Retrieval Research, 345-359. DOI: <https://doi.org/10.1007/978-3-540-31865-1_25>.

[16] Davis, Jesse and Mark Goadrich, 2006. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, 233-240. DOI: <https://doi.org/10.1145/1143844.1143874>.

Conference Name: Universidade Paulista — Curso de Sistemas de Informação

Conference Short Name: UNIP SI

Conference Location: São Paulo, SP, Brasil

Year: 2025

Date: Maio

Copyright Year: 2025

Copyright Statement: Direitos reservados aos autores

RRH: A. Katsuyama et al.