# Identificação de Anomalias e Previsão em Transações Financeiras com Modelos Híbridos

Integrantes:

Davi Rodrigues de Oliveira (10403139 - 10403139@mackenzista.com.br)

Heitor Maciel de Vasconcellos Leite (10402559 - 10402559@mackenzista.com.br)

Matheus Gonçalves Mendes (10402963 - 10402963@mackenzista.com.br)

Roberto Rinco Silveira (10403243 - 10403243@mackenzista.com.br)

**Resumo:**

Este projeto tem como objetivo identificar anomalias em transações financeiras, utilizando técnicas de inteligência artificial e modelos estatísticos. A proposta combina métodos tradicionais, como ARIMA, com redes neurais profundas, como LSTM e GRU, para detectar padrões incomuns que possam indicar fraudes e projetar valores futuros dessas transações. Utilizando uma base de dados real, o estudo pretende analisar a eficácia desses modelos na detecção de comportamentos atípicos e na previsão de séries temporais, contribuindo para a mitigação de riscos e o fortalecimento da segurança no setor financeiro.

## Introdução:

## Contextualização:

Com o aumento expressivo das transações financeiras digitais, identificar padrões suspeitos se tornou uma tarefa crítica para instituições financeiras. A crescente complexidade e o volume de dados dificultam a detecção eficiente de fraudes utilizando apenas métodos tradicionais. Nesse cenário, técnicas de Inteligência Artificial (IA) surgem como aliadas poderosas para identificar anomalias de forma automática e precisa, além de prever o comportamento futuro de transações com base em dados históricos.

## Justificativa:

A detecção de fraudes é um desafio constante e de grande impacto econômico. Métodos estatísticos convencionais, embora úteis, enfrentam limitações ao lidar com séries temporais financeiras altamente voláteis e não lineares. Por isso, propõe-se neste projeto uma abordagem híbrida que combina modelos estatísticos com algoritmos de aprendizado profundo para obter maior precisão na identificação de transações atípicas e na previsão de padrões financeiros, contribuindo para decisões mais assertivas e seguras.

## Objetivo:

O principal objetivo deste projeto é desenvolver um sistema de detecção de anomalias em transações financeiras e previsão de valores futuros utilizando uma abordagem híbrida baseada em técnicas de Machine Learning e modelos estatísticos. A intenção é avaliar a

eficácia de modelos como ARIMA, LSTM e GRU aplicados a dados reais, buscando oferecer uma solução robusta para prevenção de fraudes e suporte à tomada de decisão em ambientes financeiros.

## Opção do projeto:

Este projeto segue a **opção Framework**, utilizando ferramentas e bibliotecas de Machine Learning em Python para resolver um problema de previsão e classificação em dados financeiros. O foco está na aplicação prática de modelos supervisionados e não supervisionados para detecção de anomalias e projeção de séries temporais.

## Descrição do Problema

O aumento significativo no volume de dados gerados por sistemas financeiros digitais representa um grande desafio para a detecção de atividades fraudulentas. Os métodos tradicionais, baseados em regras fixas ou análises estatísticas simples, muitas vezes não conseguem capturar padrões não lineares e comportamentos atípicos mais sutis. Isso resulta em falhas na identificação de fraudes e em falsos positivos, gerando perdas financeiras e operacionais.

Além disso, a previsão de valores futuros em séries temporais financeiras é uma tarefa complexa, pois envolve dados altamente voláteis e influenciados por múltiplos fatores. Modelos estatísticos, como o ARIMA, oferecem bom desempenho em padrões lineares e estacionários, mas apresentam limitações ao lidar com dinâmicas mais complexas. Por outro lado, modelos de aprendizado profundo, como as redes neurais LSTM e GRU, têm se mostrado eficazes em capturar essas complexidades, embora exijam maior poder computacional e volume de dados para treinamento adequado.

Neste contexto, o problema central deste projeto é desenvolver uma solução que seja capaz de identificar, com precisão e agilidade, anomalias em transações financeiras, ao mesmo tempo em que projeta o comportamento dessas transações ao longo do tempo. A abordagem híbrida proposta busca combinar a força de modelos estatísticos e de deep learning para oferecer uma alternativa mais eficiente e confiável na detecção de fraudes e na previsão de séries temporais financeiras.

## Aspectos Éticos do Uso da IA e Responsabilidade no Desenvolvimento da Solução

O uso da Inteligência Artificial na detecção de anomalias em transações financeiras levanta importantes questões éticas que devem ser consideradas durante todas as etapas do desenvolvimento da solução. Um dos principais pontos diz respeito à **privacidade dos dados**, especialmente quando se trabalha com informações sensíveis, como valores, datas e dados pessoais vinculados a transações bancárias. Para mitigar esses riscos, os dados utilizados são anonimizados e tratados conforme as diretrizes estabelecidas pela **Lei Geral de Proteção de Dados (LGPD)**.

Outro aspecto ético fundamental está relacionado à **transparência e interpretabilidade dos modelos**. Sistemas de IA, principalmente os baseados em aprendizado profundo, como LSTM e GRU, podem ser percebidos como "caixas-pretas". Por isso, é importante aplicar

métricas claras e realizar análises interpretáveis dos resultados, de forma que os responsáveis pelas decisões possam compreender e confiar nas previsões realizadas pelo sistema.

Além disso, deve-se garantir que os modelos não reproduzam **viéses históricos** existentes nos dados, o que poderia levar a decisões discriminatórias ou imprecisas. O processo de análise exploratória e pré-processamento dos dados é essencial para identificar e corrigir esses possíveis desvios.

Por fim, é responsabilidade dos desenvolvedores garantir que o sistema proposto seja utilizado de forma **ética e legal**, sem comprometer a equidade, a transparência e a segurança dos usuários. A IA deve ser uma ferramenta de apoio à tomada de decisão, e não um substituto absoluto do julgamento humano, especialmente em contextos críticos como o financeiro.

## Dataset

## Descrição dos Datasets

Neste projeto, foram utilizados três conjuntos de dados públicos amplamente reconhecidos na literatura para estudos sobre detecção de fraudes em transações financeiras. Cada dataset possui características distintas que permitem explorar diferentes estratégias de modelagem e análise.

1. **Credit Card Fraud Detection Dataset 2023**
   Este conjunto de dados contém mais de 550.000 transações de cartão de crédito, com atributos numéricos anonimizados e um rótulo binário que indica se a transação foi fraudulenta. É uma versão atualizada e expandida de datasets clássicos sobre o tema, proporcionando maior diversidade de padrões.

2. **Credit Card Fraud Detection (ULB)**
   Conjunto de dados amplamente utilizado, com 284.807 registros de transações de cartão de crédito realizadas em um período de dois dias. Contém 30 atributos transformados por PCA (preservando o sigilo dos dados), e apenas 492 transações são classificadas como fraudes, o que evidencia um forte desbalanceamento de classes.

3. **PaySim -- Synthetic Financial Dataset**
   Simulação realista de transações financeiras móveis, baseada em comportamentos de sistemas de pagamento reais. Contém diferentes tipos de transações (como transferências e pagamentos), além de dados como valor da transação, conta de origem e destino, e rótulo de fraude. Apesar de ser sintético, é extremamente útil para testes em cenários de produção.

## Análise Exploratória dos Dados

Durante a análise exploratória (EDA), foram observadas as seguintes características:

1. **Distribuição das Classes**: Os três datasets apresentam classes desbalanceadas, com fraudes representando menos de 1% dos dados. Isso exige o uso de técnicas específicas para balanceamento.

2. **Distribuição de Valores**: Os valores das transações apresentam grande variabilidade, o que pode influenciar na sensibilidade dos modelos.
3. **Atributos Temporais**: Tanto o dataset do ULB quanto o PaySim possuem variáveis de tempo, possibilitando a modelagem com séries temporais e o uso de modelos como ARIMA e redes neurais recorrentes (LSTM e GRU).
4. **Outliers**: Foram identificados valores extremos em variáveis como o valor das transações, que serão tratados para evitar distorções durante o treinamento dos modelos.

## Preparação dos Dados em Python

A preparação dos dados será realizada utilizando bibliotecas como pandas, numpy e scikit-learn, e envolverá as seguintes etapas:

1. **Limpeza dos dados**: Verificação de valores nulos, remoção de registros inconsistentes e padronização dos formatos.
2. **Normalização**: As variáveis serão escaladas entre 0 e 1 ou padronizadas com média zero e desvio padrão unitário, visando otimizar o desempenho de redes neurais.
3. **Conversão de atributos categóricos**: No dataset PaySim, colunas como o tipo de transação serão convertidas para formato numérico (via one-hot encoding).
4. **Balanceamento de classes**: Serão aplicadas técnicas como SMOTE ou undersampling para lidar com o desbalanceamento entre fraudes e transações legítimas.
5. **Criação de sequências temporais**: Para os modelos baseados em LSTM e GRU, os dados serão transformados em janelas de tempo que permitam capturar padrões sequenciais nas transações.
6. **Divisão do dataset**: Os dados serão divididos em conjuntos de treino, validação e teste, respeitando a ordem temporal dos eventos para preservar a coerência das séries temporais.

## Metodologia e Resultados Esperados

## Metodologia

Para a resolução do problema proposto, será utilizada uma **abordagem híbrida**, combinando **modelos estatísticos tradicionais** e **técnicas modernas de aprendizado profundo**, a fim de identificar anomalias em transações financeiras e realizar a previsão de valores futuros em séries temporais.

O processo será conduzido em etapas, conforme descrito a seguir:

1. **Análise Exploratória de Dados (EDA)**
   A análise inicial dos dados será utilizada para compreender a estrutura das variáveis, identificar padrões sazonais, avaliar o desbalanceamento de classes e observar a

distribuição estatística das transações. Essa etapa fornecerá os primeiros insights para guiar a modelagem.

2. **Pré-processamento dos Dados**
   Inclui a limpeza de dados, normalização de valores, balanceamento das classes (por meio de técnicas como SMOTE) e transformação dos dados em janelas temporais para uso em modelos sequenciais.

3. **Modelagem Estatística -- ARIMA**
   O modelo ARIMA será utilizado para analisar séries temporais agregadas, como a soma ou volume diário de transações, permitindo gerar previsões de curto prazo. Essa abordagem capta os padrões lineares e estacionários dos dados.

4. **Modelagem com Deep Learning -- LSTM e GRU**
   Redes neurais recorrentes do tipo LSTM e GRU serão aplicadas para capturar padrões complexos e não lineares ao longo do tempo. Esses modelos serão treinados com sequências de transações, visando detectar comportamentos fora do padrão que possam indicar fraudes.

5. **Validação e Avaliação dos Modelos**
   Os modelos serão avaliados com base em métricas como **Acurácia**, **Precisão**, **Recall**, **F1-Score** e **AUC-ROC** para a tarefa de classificação, e **MSE (Erro Quadrático Médio)** para tarefas de previsão. A validação cruzada e o uso de conjuntos separados de teste garantirão a confiabilidade dos resultados.

6. **Comparação de Abordagens**
   Será realizada uma comparação entre os modelos estatísticos e os de aprendizado profundo, individualmente e em combinação, para determinar qual abordagem apresenta melhor desempenho na detecção de anomalias e na previsão de transações.

## Resultados Esperados

Espera-se alcançar acurácia acima de 95% na detecção de fraudes com redes LSTM, e erro quadrático médio (MSE) inferior a 0.05 nas previsões com ARIMA.

Além disso, o projeto visa demonstrar que:

1. Técnicas de aprendizado profundo são mais eficazes para capturar padrões não lineares em transações financeiras.

2. A modelagem sequencial melhora a detecção de anomalias ao considerar o histórico de comportamento das contas envolvidas.

3. A previsão de transações futuras pode ser realizada com boa acurácia, oferecendo uma ferramenta adicional para planejamento financeiro e mitigação de riscos.

## Referências

As referências a seguir foram citadas ao longo do projeto e baseiam-se em fontes acadêmicas e técnicas relevantes para a fundamentação teórica e prática da proposta.

1. ALJOHANI, A. *Predictive analytics and machine learning for real-time supply chain risk mitigation and agility*. Sustainability, v. 15, n. 20, p. 15088, 2023.
2. BENTO, P. et al. *Stacking ensemble methodology using deep learning and ARIMA models for short-term load forecasting*. Energies, v. 14, n. 21, p. 7378--7378, 2021.
3. HUA, Y. *Bitcoin price prediction using ARIMA and LSTM*. E3S Web of Conferences, v. 218, p. 01050, 2020.
4. MOHAMUDALLY, N.; PEERMAMODE-MOHABOOB, M. *Building an anomaly detection engine (ADE) for IoT smart applications*. Procedia Computer Science, v. 134, p. 10--17, 2018.
5. SIRISHA, U. M.; BELAVAGI, M. C.; ATTIGERI, G. *Profit prediction using ARIMA, SARIMA and LSTM models in time series forecasting: A comparison*. IEEE Access, v. 10, p. 124715--124727, 2022.
6. ZHU, Q.; SUN, L. *Big data driven anomaly detection for cellular networks*. IEEE Access, v. 8, p. 31398--31408, 2020.

## Bibliografia

A bibliografia contempla materiais que serviram de base teórica e técnica para o desenvolvimento do projeto e aprofundamento nos temas abordados.

1. ANDRADE, M. M. de. *Como preparar trabalhos para cursos de pós-graduação*. 3. ed. São Paulo: Atlas, 1999.
2. GIL, A. C. *Como elaborar projetos de pesquisa*. 3. ed. São Paulo: Atlas, 1991.
3. GHOSH, K. *A comparison of standard statistical, machine learning and deep learning methods in forecasting the time series*. SSRN Electronic Journal, 2024.
4. KANDPAL, P. K. et al. *Time series forecasting of NSE stocks using machine learning models (ARIMA, Facebook Prophet, and Stacked LSTM)*. Lecture Notes in Networks and Systems, 2023.
5. LI, Z.; HAN, J.; SONG, Y. *On the forecasting of high-frequency financial time series based on ARIMA model improved by deep learning*. Journal of Forecasting, v. 39, n. 7, p. 1081--1097, 2020.
6. HUSSEIN, T.; BOURAS, A. *Anomaly detection: A survey*. Lecture Notes in Networks and Systems, p. 391--401, 2021.
7. Kaggle Datasets:
    - NELGIRIYEWITHANA, P. *Credit Card Fraud Detection Dataset 2023*. Disponível em: https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023

- ULB. *Credit Card Fraud Detection*. Disponível em: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
- EALAXI. *PaySim: Synthetic Financial Datasets For Fraud Detection*. Disponível em: https://www.kaggle.com/datasets/ealaxi/paysim1