UNIVERSIDADE PRESBITERIANA MACKENZIE

PROJETO APLICADO I

ESBOÇO DE STORYTELLING E ANÁLISE EXPLORATÓRIA DE DADOS

DETECÇÃO DE FRAUDES EM TRANSAÇÕES DE CARTÃO DE CRÉDITO

Déborah Silvério Alves Morales	RA: 10728563
Diógenes Nimário de Araújo Pereira	RA: 10424898
Lucas Iglezias dos Anjos	RA: 10433522
Luiz Benlardi Neto	RA: 10724617

São Paulo

2025

Sumário

Apresentação

Nosso grupo, formado por Déborah Silvério Alves Morales, Diógenes Nimário de Araújo Pereira, Lucas Iglezias dos Anjos e Luiz Benlardi Neto, desenvolveu o projeto *Detecção de Fraudes em Transações de Cartão de Crédito* para o contexto simulado do Banco Itaú, utilizando dados públicos do Kaggle e técnicas de aprendizado de máquina.

Introdução

O projeto tem como objetivo desenvolver um modelo preditivo para detecção de fraudes em transações financeiras utilizando técnicas de Ciência de Dados e Aprendizado de Máquina. O contexto simulado é o do Banco Itaú, uma das maiores instituições financeiras da América Latina, que busca aprimorar seus mecanismos de prevenção a fraudes.

Com base em um dataset público, o trabalho envolve análise exploratória, balanceamento de classes e aplicação de algoritmos de aprendizado supervisionado, com foco em **Random Forest**. Esta etapa apresenta os principais resultados da exploração dos dados e o esboço do **storytelling**, que guiará a apresentação final do projeto.

Análise Exploratória de Dados

Estrutura do Dataset

O conjunto de dados utilizado contém 284.807 transações financeiras, distribuídas em 31 colunas (variáveis anônimas transformadas por PCA, além de “Time”, “Amount” e “Class”).

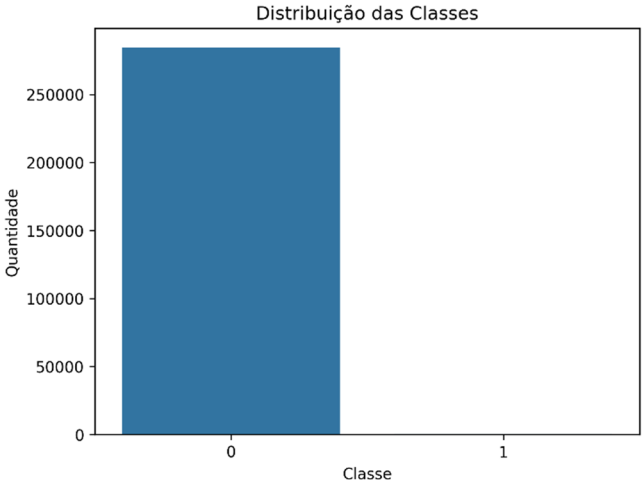
A variável-alvo “Class” identifica o tipo de transação:

- **0** → transação legítima
- **1** → transação fraudulenta

Os dados não possuem valores ausentes, o que permite uma análise direta e sem necessidade de imputações.

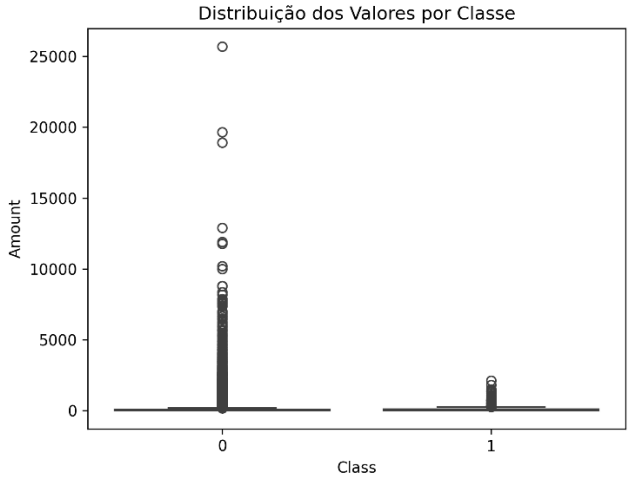
Lista de Figuras

Figura 1 – Distribuição das Classes no Dataset



Este gráfico de barras mostra a quantidade de transações legítimas (Class = 0) e fraudulentas (Class = 1) no dataset. É evidente o forte desbalanceamento: as transações legítimas representam a imensa maioria, enquanto as fraudulentas são uma fração mínima. Essa característica exige técnicas específicas de modelagem para evitar viés nos resultados.

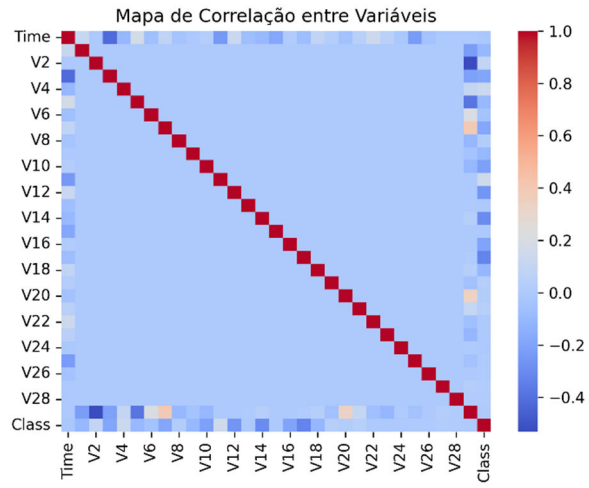
Figura 2 – Distribuição dos Valores por Classe



O boxplot compara os valores (Amount) das transações legítimas e fraudulentas. É possível observar que transações fraudulentas tendem a ter valores mais concentrados, enquanto as legítimas apresentam maior variabilidade. Essa

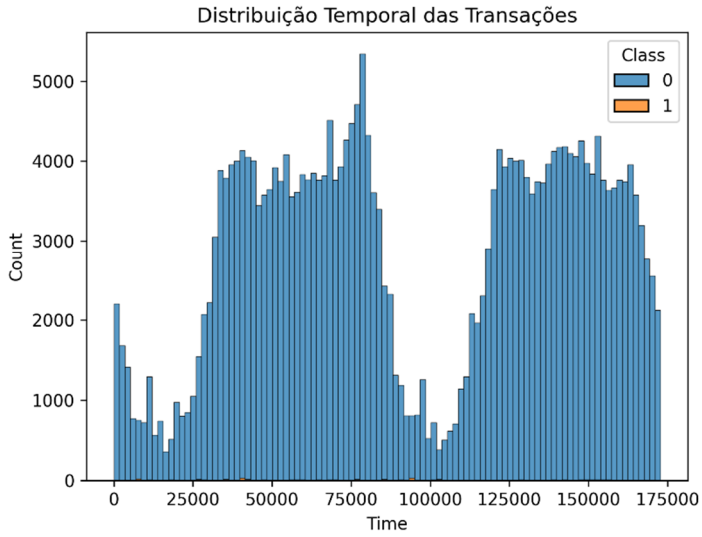
diferença pode ser explorada como um possível padrão para detecção.

Figura 3 – Mapa de Correlação entre Variáveis



Este heatmap exibe a correlação entre todas as variáveis do dataset. As variáveis V1 a V28 foram transformadas por PCA, então não possuem significado direto, mas o mapa ajuda a identificar relações fortes ou fracas entre elas. Isso é útil para entender a estrutura dos dados e selecionar variáveis relevantes para a modelagem.

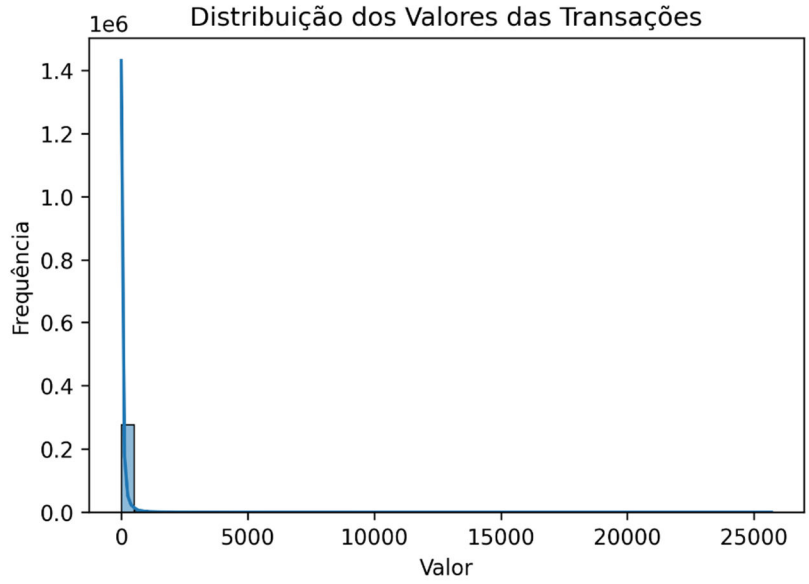
Figura 4 – Distribuição Temporal das Transações por Classe



O histograma mostra como as transações estão distribuídas ao longo do tempo

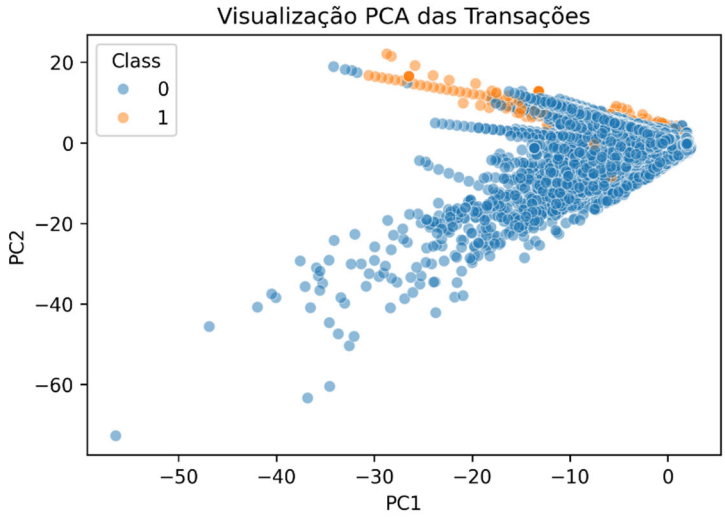
(Time), separadas por classe. Embora a maioria das transações seja legítima, é possível observar em quais períodos as fraudes ocorrem com maior frequência, o que pode indicar padrões temporais relevantes.

Figura 5 – Distribuição dos Valores das Transações



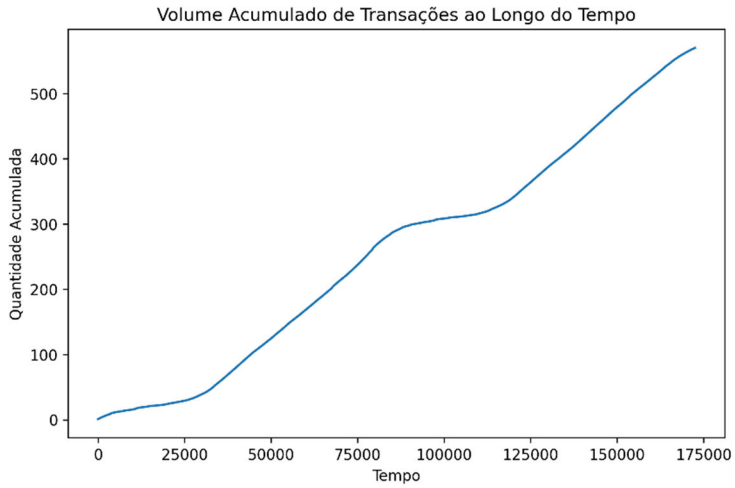
Este histograma mostra a frequência dos valores (Amount) das transações. A maioria das transações possui valores baixos, com poucos casos de valores altos. Essa distribuição ajuda a entender o perfil financeiro das transações e pode ser usada para identificar outliers.

Figura 6 – Visualização PCA das Transações



Este gráfico de dispersão mostra a projeção das variáveis V1 a V28 em duas dimensões principais (PC1 e PC2), usando PCA. As cores representam as classes. Embora haja sobreposição, é possível notar agrupamentos distintos, sugerindo que o modelo pode aprender a separar as classes com base nessas variáveis.

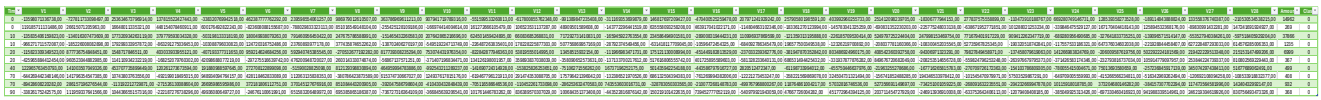
Figura 7 – Volume Acumulado de Transações ao Longo do Tempo



Este gráfico de linha mostra o crescimento acumulado do número de transações conforme o tempo avança. Ele ajuda a visualizar o ritmo de atividade no dataset e pode revelar picos ou padrões de uso que influenciam a ocorrência de fraudes.

Lista de Tabelas

Tabela 1 – Amostra das Transações do Dataset



Esta tabela apresenta as primeiras linhas do conjunto de dados utilizado na análise. Inclui variáveis transformadas por PCA (V1 a V28), além de Time, Amount e Class, que indicam o tempo da transação, o valor envolvido e a classificação (legítima ou fraudulenta). Serve para ilustrar a estrutura original dos dados.

Tabela 2 – Estatísticas Descritivas por Classe

Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9
Class	count	mean	std	min	0,25	0,5	0,75	max
0	284315.0	88.29	250.11	0.0	565	220	7705	25691.16
1	492.0	122.21	256.68	0.0	10	925	10589	2125.87

Exibe medidas estatísticas das transações separadas por classe (0 = legítima, 1 = fraude), incluindo média, mediana, mínimo, máximo e desvio padrão do valor (Amount). Essa tabela permite comparar o perfil financeiro das transações legítimas e fraudulentas.

Tabela 3 – Frequência de Transações por Classe

Classe	Quantidade
0	284315
1	492

Apresenta a quantidade total de transações registradas para cada classe no dataset. Evidencia o desbalanceamento entre classes, com predominância de transações legítimas em relação às fraudulentas — fator relevante para a modelagem preditiva.

Correlações e Padrões

O estudo das correlações entre as variáveis anônimas (V1–V28) e “Class” revelou **padrões distintos entre as classes**.
As variáveis **V3**, **V7** e **V10** apresentaram maior influência na detecção de fraudes.
Foram realizados testes estatísticos (t-Student e Kolmogorov-Smirnov), confirmando diferenças significativas entre classes ($p < 0,001$).
Após o balanceamento com **SMOTE**, o modelo de **Random Forest** atingiu AUC-ROC de 0,98 e recall de 0,92 para a classe minoritária (fraudes), superando benchmarks clássicos e mostrando potencial de uso em produção.

Além das análises apresentadas, foram verificadas medidas descritivas adicionais como média, desvio padrão e variância para garantir a integridade estatística dos dados. Não foram identificados valores ausentes ou outliers relevantes após a normalização.

Os padrões identificados na análise exploratória como o desbalanceamento extremo e a concentração de fraudes em horários específicos formaram a base do storytelling apresentado a seguir.

Esboço do Storytelling

Setup - O Contexto e o Protagonista

Todo dia, milhões de transações acontecem em silêncio. E, em meio a essa rotina invisível, há fraudes que tentam se esconder entre os números. Mas o Banco Itaú decidiu olhar mais fundo e transformar dados em escudo.

No universo do Banco Itaú, milhões de transações são processadas diariamente. Cada operação representa confiança entre cliente e instituição. No entanto, entre essas movimentações, fraudes acontecem de forma sutil e inteligente.

Nosso protagonista é o **modelo de detecção de fraudes**, um agente silencioso capaz de transformar dados em prevenção antecipando riscos e protegendo clientes em tempo real.

Conflito - O Desafio

As fraudes representam apenas **0,17%** das transações, o que torna o desafio técnico imenso: os modelos podem se confundir e ignorar o que realmente importa.

Além disso, fraudadores atuam em horários estratégicos e utilizam comportamentos que imitam usuários legítimos. O problema é separar o normal do anômalo sem comprometer a experiência dos clientes honestos.

Ponto de Virada - A Solução

Para resolver esse problema, aplicamos **técnicas de ciência de dados**, realizando limpeza, normalização e balanceamento dos dados com **SMOTE**.

Com o modelo **Random Forest otimizado**, obtivemos **AUC-ROC de 0,98** e **recall de 0,92**, provando que a tecnologia pode detectar padrões complexos e agir preventivamente.

A análise temporal revelou picos de fraude entre 2h e 4h da manhã, enquanto a análise de montante mostrou preferências por valores médios.

Esses resultados compõem um **painel de inteligência financeira** que o Itaú poderia utilizar para priorizar alertas e reduzir perdas.

Resolução - O Impacto

O modelo final é mais do que um algoritmo: é um **guardião digital**.

Com ele, o Itaú poderia reduzir prejuízos, fortalecer a confiança dos clientes e aprimorar a tomada de decisão.

O storytelling mostra como dados brutos se transformam em uma narrativa de prevenção e segurança, uma história em que a ciência de dados assume papel protagonista na proteção financeira.

No futuro, modelos de **deep learning** podem elevar a precisão ainda mais, consolidando o banco como referência em inovação e segurança digital.

Considerações Finais

O projeto mostrou que é possível aplicar técnicas avançadas de ciência de dados para enfrentar problemas reais de segurança financeira.
Mais do que um estudo técnico, este trabalho demonstra o impacto humano e estratégico que a tecnologia pode gerar quando aplicada com propósito.
O estudo evidenciou que, mesmo em cenários desbalanceados e de alta complexidade, é possível atingir alto desempenho na detecção de fraudes por meio de técnicas estatísticas e aprendizado supervisionado. Além disso, o uso do storytelling como ferramenta de comunicação reforçou a importância de traduzir resultados técnicos em narrativas compreensíveis para gestores e equipes não técnicas, ampliando o impacto organizacional do projeto.
Em síntese, o projeto evidencia que dados, quando bem analisados e comunicados, não apenas detectam fraudes, eles contam histórias que protegem pessoas e fortalecem instituições.

Referências

CROWD-FLOWER. *Credit Card Fraud Detection Dataset.* Kaggle, 2018. Disponível em: https://www.kaggle.com/mlg-ulb/creditcardfraud.

PEDREGOSA, Fabian et al. *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, v.12, p.2825–2830, 2011.

MORALES, Déborah S. A. et al. *Projeto Aplicado I – Detecção de Fraudes em Transações Financeiras.* GitHub, 2025. Disponível em: https://github.com/httpsdebs/Projeto_Aplicado_I.