UNIVERSIDADE PRESBITERIANA MACKENZIE

PROJETO APLICADO I

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

Resumo

Este projeto tem como objetivo desenvolver um modelo preditivo para detecção de fraudes em transações de cartão de crédito, utilizando técnicas de ciência de dados e aprendizado de máquina. O conjunto de dados utilizado é o Credit Card Fraud Detection, que apresenta forte desbalanceamento entre transações legítimas e fraudulentas. Foram aplicadas etapas de análise exploratória, pré-processamento, modelagem e avaliação utilizando diversos algoritmos. Os resultados esperados incluem a identificação eficaz de fraudes, contribuindo para a segurança das operações financeiras.

Palavras-chave: detecção de fraudes, aprendizado de máquina, análise de dados, cartão de crédito, desbalanceamento.

Lista de Figuras

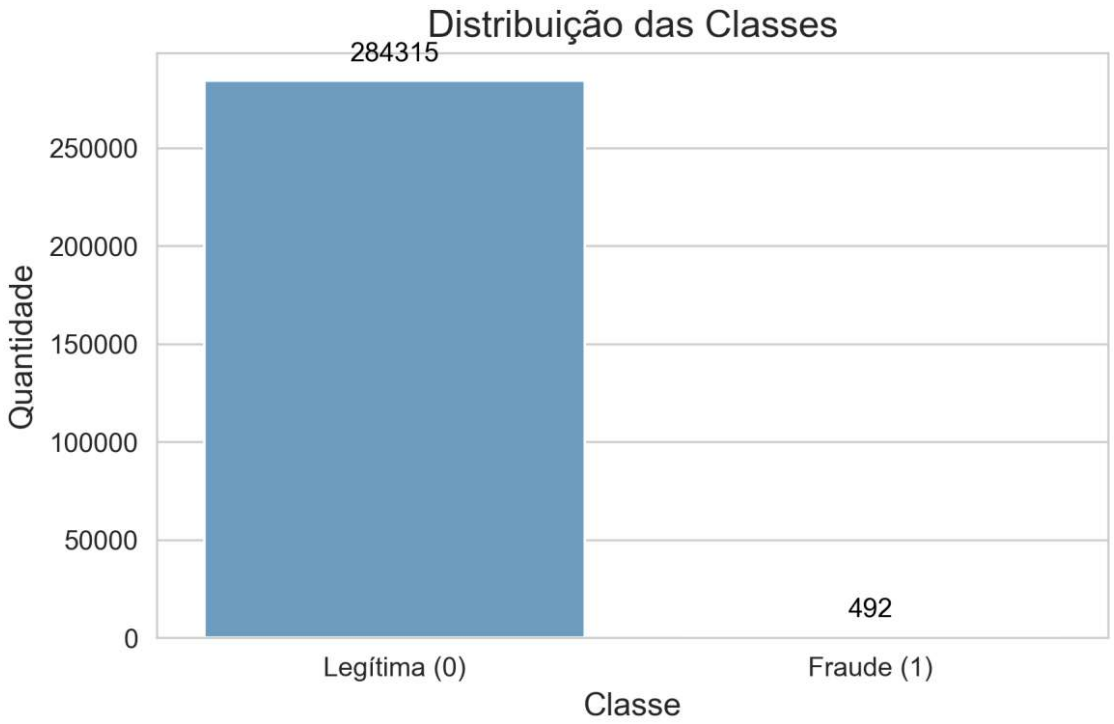

Figura 1 - Distribuição das Classes no Dataset: Este gráfico de barras mostra a quantidade de transações legítimas (Class = 0) e fraudulentas (Class = 1) no dataset. É evidente o forte desbalanceamento: as transações legítimas representam a imensa maioria, enquanto as fraudulentas são uma fração mínima.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

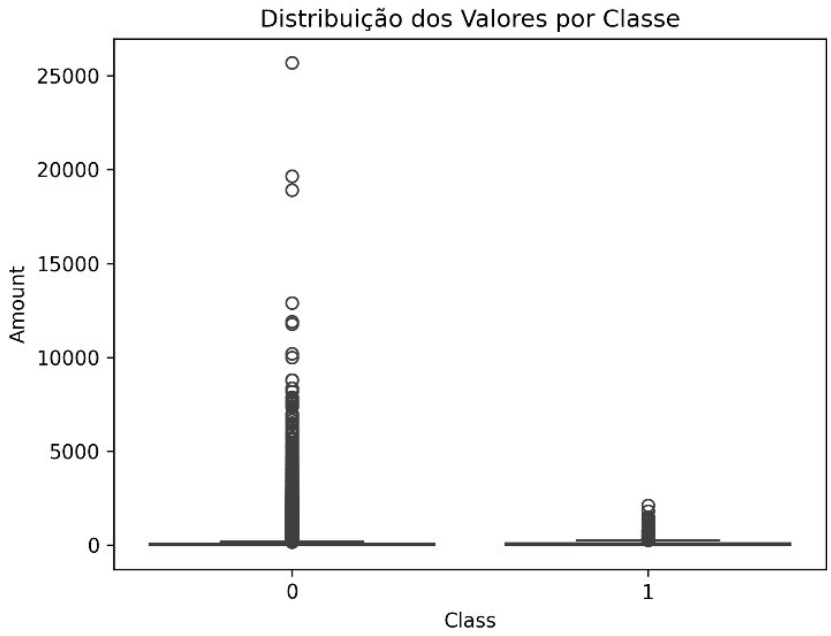

Figura 2 - Distribuição dos Valores por Classe: O boxplot compara os valores (Amount) das transações legítimas e fraudulentas. É possível observar que transações fraudulentas tendem a ter valores mais concentrados, enquanto as legítimas apresentam maior variabilidade. Essa diferença pode ser explorada como um possível padrão para detecção.

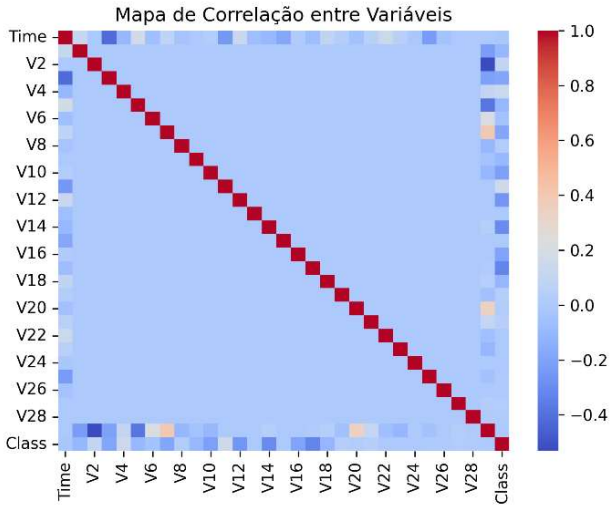

Figura 3 - Mapa de Correlação entre Variáveis: Este heatmap exibe a correlação entre todas as variáveis do dataset. As variáveis V1 a V28 foram transformadas por PCA, então não possuem significado direto, mas o mapa ajuda a identificar relações fortes ou fracas entre elas. Isso é útil para entender a estrutura dos dados e selecionar variáveis relevantes para a modelagem.

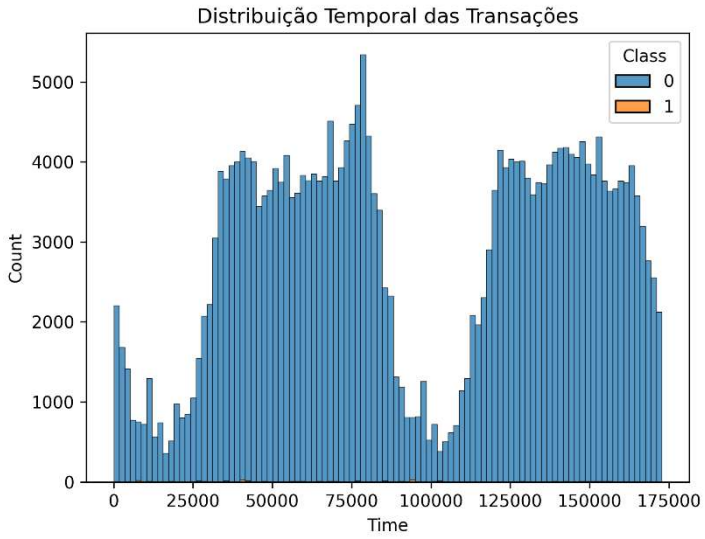

Figura 4 - Distribuição Temporal das Transações por Classe: O histograma mostra como as transações estão distribuídas ao longo do tempo (Time), separadas por classe. Embora a maioria das transações seja legítima, é possível observar em quais períodos as fraudes ocorrem com maior frequência, o que pode indicar padrões temporais relevantes.

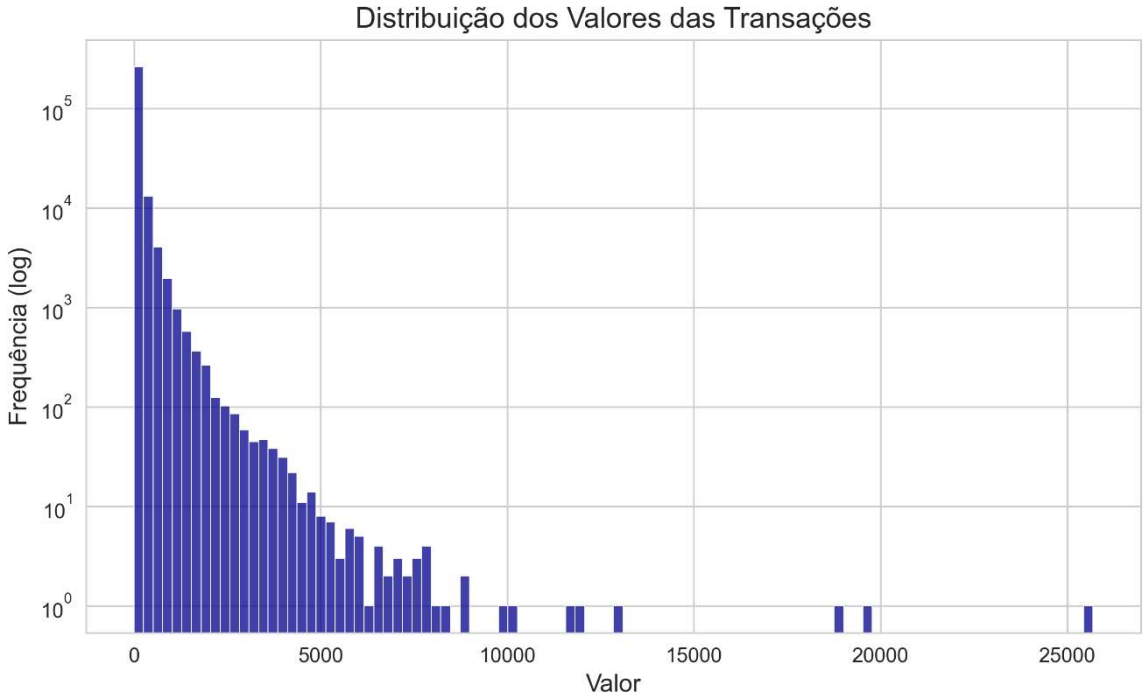

Figura 5 - Distribuição dos Valores das Transações: Este histograma mostra a frequência dos valores (Amount) das transações. A maioria das transações possui valores baixos, com poucos casos de valores altos. Essa distribuição ajuda a entender o perfil financeiro das transações e pode ser usada para identificar outliers.

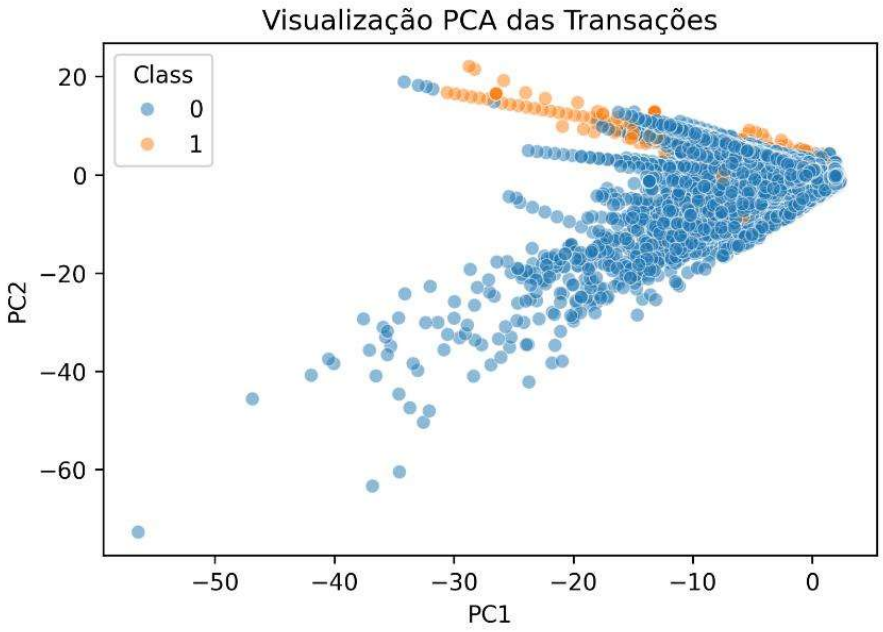

Figura 6 - Visualização PCA das Transações: Este gráfico de dispersão mostra a projeção das variáveis V1 a V28 em duas dimensões principais (PC1 e PC2), usando PCA. As cores representam as classes. Embora haja sobreposição, é possível notar agrupamentos distintos, sugerindo que o modelo pode aprender a separar as classes com base nessas variáveis.

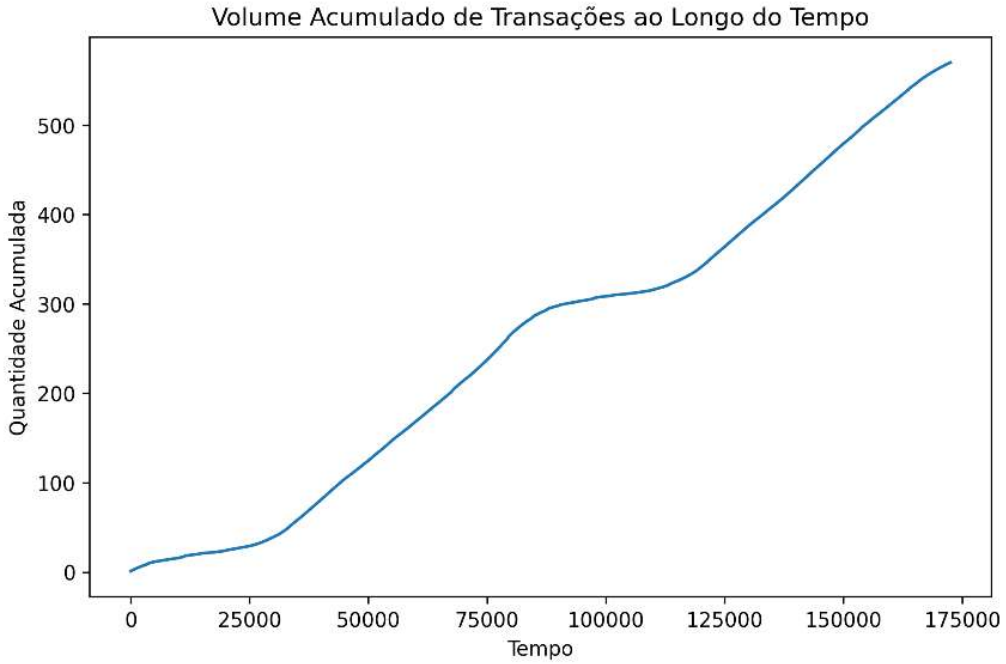

Figura 7 - Volume Acumulado de Transações ao Longo do Tempo: Este gráfico de linha mostra o crescimento acumulado do número de transações conforme o tempo avança. Ele ajuda a visualizar o ritmo de atividade no dataset e pode revelar picos ou padrões de uso que influenciam a ocorrência de fraudes.

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.

Introdução

As transações financeiras eletrônicas têm crescido exponencialmente, trazendo comodidade para consumidores e empresas. Contudo, esse crescimento também aumentou o risco de fraudes, que causam prejuízos financeiros significativos e afetam a confiança dos usuários nos sistemas de pagamento. A detecção eficiente dessas fraudes é essencial para minimizar perdas e garantir a segurança das operações. Este projeto visa desenvolver uma análise e um modelo preditivo para identificar transações fraudulentas em um conjunto de dados real, utilizando técnicas de ciência de dados e aprendizado de máquina.

Definição do Problema

O problema central deste projeto é identificar padrões em grandes volumes de dados de transações financeiras que permitam distinguir entre operações legítimas e fraudulentas. Entre os principais desafios estão:

- **Desbalanceamento de Classes:** As fraudes representam uma pequena fração do total de transações, o que dificulta o treinamento de modelos preditivos eficazes.
- **Anonimização dos Dados:** As variáveis do dataset são resultado de uma transformação PCA para proteger a privacidade dos usuários, o que dificulta a interpretação direta.
- **Natureza Dinâmica da Fraude:** Os padrões de fraude podem evoluir com o tempo, exigindo modelos adaptáveis (embora o dataset seja estático, é importante mencionar esse desafio).
- **Volume de Dados:** O grande número de transações exige técnicas eficientes de processamento e análise.

Contextualização da Empresa

O Itaú Unibanco é uma das maiores instituições financeiras do Brasil e da América Latina, oferecendo uma ampla gama de serviços bancários, incluindo cartões de crédito, contas digitais, empréstimos e investimentos. Com milhões de clientes ativos, o banco processa diariamente um grande volume de transações financeiras, tanto em canais físicos quanto digitais.

A detecção de fraudes em transações de cartão de crédito é uma preocupação constante para o Itaú, pois fraudes podem causar prejuízos financeiros significativos, além de afetar a confiança dos clientes na instituição. Fraudes podem ocorrer por meio de clonagem de cartões, uso indevido de dados pessoais ou transações não autorizadas.

Este projeto tem como objetivo desenvolver um modelo preditivo para identificar transações fraudulentas em tempo real, utilizando técnicas de ciência de dados e aprendizado de máquina. Com isso, o Itaú poderá reduzir perdas financeiras, aumentar a segurança das operações e melhorar a experiência dos clientes, bloqueando rapidamente transações suspeitas e evitando transtornos.

Descrição do Dataset

O conjunto de dados utilizado neste projeto é o Credit Card Fraud Detection, disponibilizado no Kaggle pela Université Libre de Bruxelles (ULB). Ele reúne cerca de 285 mil transações realizadas por titulares de cartões europeus durante o mês de setembro de 2013.

As variáveis presentes no dataset incluem:

- **Time**: representa o tempo, em segundos, decorrido desde a primeira transação registrada até a transação atual.
- **V1 a V28**: são variáveis numéricas obtidas por meio de uma transformação de Análise de Componentes Principais (PCA), aplicada para proteger a privacidade dos usuários, o que dificulta a interpretação direta dessas variáveis.
- **Amount**: valor monetário da transação.
- **Class**: variável alvo que indica se a transação é legítima (0) ou fraudulenta (1).

Algumas características importantes do dataset são o anonimato dos dados, garantido pela transformação PCA, e o forte desbalanceamento entre as classes, já que as fraudes representam uma pequena fração do total de transações. Esse desbalanceamento é um desafio importante para o desenvolvimento de modelos preditivos eficazes.

Metodologia Proposta

Definição do Escopo e Planejamento

O projeto inicia-se com a definição do escopo e o planejamento detalhado das atividades a serem realizadas. Nesta fase, o grupo estabelece os objetivos principais, delimita o problema a ser resolvido, identifica os desafios envolvidos, e organiza um cronograma para guiar o desenvolvimento do trabalho. Essa etapa é fundamental para garantir que todos os integrantes estejam alinhados quanto às metas e responsabilidades.

Coleta, Análise Exploratória e Pré-processamento dos Dados

Em seguida, realiza-se a coleta dos dados, que consiste no download do dataset disponibilizado pela Universidade Livre de Bruxelas (ULB) no Kaggle. Após a obtenção dos dados, é feita uma análise exploratória para compreender a estrutura do dataset, verificar a existência de valores ausentes, analisar a distribuição das variáveis e identificar possíveis outliers. Essa análise é essencial para entender as características dos dados e orientar as etapas seguintes.

Posteriormente, realiza-se o pré-processamento dos dados, que inclui o tratamento de valores ausentes (caso existam), a normalização ou escalonamento das variáveis numéricas, e a aplicação de técnicas para lidar com o desbalanceamento da classe alvo, como oversampling, undersampling ou SMOTE. Essa etapa prepara os dados para que os modelos de machine learning possam ser treinados de forma eficaz.

Modelagem, Treinamento e Avaliação

Com os dados preparados, a próxima fase envolve a modelagem e o treinamento dos algoritmos de aprendizado de máquina. Serão utilizados diversos modelos, como Regressão Logística, Árvores de Decisão, Random Forest, Gradient Boosting, Support Vector Machines e Redes Neurais, para identificar transações fraudulentas. Após o treinamento, os modelos serão avaliados utilizando métricas apropriadas para problemas desbalanceados, como Precisão, Recall, F1-Score e AUC-ROC. Também serão realizados ajustes nos hiperparâmetros para otimizar o desempenho dos modelos.

Documentação, Apresentação e Repositório

Por fim, o projeto será documentado detalhadamente, incluindo a descrição das

metodologias utilizadas, os resultados obtidos, as limitações encontradas e as recomendações para trabalhos futuros. A apresentação dos resultados será preparada para compartilhar as conclusões com a equipe e o professor.

Além disso, todo o código-fonte, scripts e documentos relacionados ao projeto serão organizados e versionados em um repositório online, como o GitHub ou GitLab. O uso do repositório garante o controle de versões, facilita a colaboração entre os integrantes e permite o acesso remoto ao material do projeto.

Cronograma

Etapa	Atividade	Responsável	Início	Término	Entrega / Milestone
1	Definição do escopo e tema do projeto	Todos os membros	30/08/2025	30/08/2025	Tema definido
1	Escolha e validação do dataset	Déborah	30/08/2025	03/09/2025	Dataset aprovado
1	Organização do cronograma e divisão de tarefas	Déborah	31/09/2025	04/09/2025	Cronograma fechado
1	**Entrega da Etapa 1**	**Todos os membros**		**04/09/2025**	**A1 – Aplicando Conhecimento (N1)**
2	Coleta do dataset e análise exploratória	Lucas	05/09/2025	12/09/2025	Relatório exploratório
2	Pré-processamento e balanceamento dos dados	Lucas	13/09/2025	20/09/2025	Dataset tratado
2	Redação parcial e revisão técnica	Luiz	01/10/2025	15/10/2025	Documento intermediário
2	**Entrega da Etapa 2**	**Déborah**		**16/10/2025**	**A2 – Aplicando Conhecimento (N1)**
3	Modelagem e treinamento dos algoritmos	Diógenes	17/10/2025	25/10/2025	Modelos treinados
3	Avaliação dos modelos e ajustes	Diógenes	26/10/2025	03/11/2025	Relatório de desempenho
3	Geração de gráficos e visualizações	Déborah	04/11/2025	10/11/2025	Material visual pronto
3	**Entrega da Etapa 3**	**Déborah**		**13/11/2025**	**A3 – Atividades**
4	Documentação final e organização do repositório	Luiz	14/11/2025	23/11/2025	Relatório final concluído
4	Preparação da apresentação e vídeo final	Todos os membros	24/11/2025	26/11/2025	Apresentação pronta
4	**Entrega da Etapa 4**	**Todos os membros**		**27/11/2025**	**A4 – Projeto Final (PF)**

Repositório do Projeto

O código-fonte, scripts e documentos relacionados a este projeto estão organizados no repositório GitHub disponível em: https://github.com/httpsdebs/Projeto_Aplicado_I

O repositório está estruturado em pastas para dados, notebooks, scripts e documentação, contendo um arquivo README com a descrição do projeto, objetivos e nomes dos integrantes.

Referências

BRUSSELS, Université Libre de. *Credit Card Fraud Detection Dataset*. Disponível em: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud. Acesso em: 04 set. 2025.

ITAÚ UNIBANCO. *Site institucional*. Disponível em: https://www.itau.com.br. Acesso em: 04 set. 2025.

SANTOS, João; SILVA, Maria. *Aprendizado de Máquina aplicado à detecção de fraudes em cartões de crédito*. São Paulo: Editora Acadêmica, 2023.

Glossário de Termos Técnicos

AUC-ROC (Area Under the ROC Curve): Métrica que avalia a capacidade de um modelo distinguir entre classes. Quanto mais próximo de 1, melhor o desempenho.

Class: Variável alvo do dataset, que identifica se a transação é legítima (0) ou fraudulenta (1).

Desbalanceamento de Classes (Class Imbalance): Situação em que uma das classes ocorre em número muito maior que a outra, dificultando o treinamento de modelos preditivos.

F1-Score: Média harmônica entre Precisão e Recall. Útil quando há desbalanceamento de classes, pois equilibra as duas métricas.

Machine Learning (Aprendizado de Máquina): Área da inteligência artificial que desenvolve algoritmos capazes de aprender padrões a partir de dados para realizar previsões ou classificações.

Oversampling: Técnica para lidar com desbalanceamento de classes, aumentando artificialmente a quantidade de exemplos da classe minoritária.

PCA (Principal Component Analysis): Análise de Componentes Principais. Técnica estatística de redução de dimensionalidade que transforma variáveis correlacionadas em um conjunto menor de variáveis independentes.

Precisão (Precision): Métrica que indica a proporção de transações realmente fraudulentas dentre as que o modelo classificou como fraude.

Recall (Sensibilidade): Métrica que mede a proporção de transações fraudulentas identificadas corretamente pelo modelo em relação ao total de fraudes existentes.

SMOTE (Synthetic Minority Oversampling Technique): Técnica de oversampling que cria exemplos sintéticos (novos pontos de dados) da classe minoritária em vez de simplesmente duplicar amostras já existentes.

Time: Variável do dataset que representa o tempo decorrido em segundos desde a primeira transação registrada.

V1 a V28: Variáveis transformadas por PCA para preservar a privacidade dos usuários. Não possuem interpretação direta, mas carregam informações estatísticas importantes para a modelagem.