Detecção de Fraudes em Transações: Uma Abordagem de Classificação com Análise Exploratória e Pré-processamento

1st Lucas Gabriel Nunes Geremias
Pontifícia Universidade Católica do Paraná
Curitiba, Brasil
S.gabriel9@pucpr.edu.br

2nd Drayan Silva Magalhães
Pontifícia Universidade Católica do Paraná
Curitiba, Brasil
Drayan.Silva@pucpr.edu.br

3rd Joel Sepulveda Martins
Pontifícia Universidade Católica do Paraná
Curitiba, Brasil
joel.sepulveda@pucpr.edu.br

4th Lucca Lucchin de Campos Costa
Pontifícia Universidade Católica do Paraná
Curitiba, Brasil
lucca.lucchin@pucpr.edu.br

5th João Vitor Zambão
Pontifícia Universidade Católica do Paraná
Curitiba, Brasil
Joao.Zambao@pucpr.edu.br

***Abstract*—This report presents a study on fraud detection in financial transactions using Data Science techniques. The main objective was to develop a robust model capable of identifying fraudulent transactions from a real dataset. Initially, an exploratory data analysis (EDA) was performed to understand the structure, characteristics, and potential anomalies of the dataset. Subsequently, various pre-processing techniques were applied to normalize data and prepare the dataset for modeling. For the construction of the predictive model, Decision Tree was chosen, evaluated through two validation strategies: hold-out and cross-validation with StratifiedKFold, aiming to ensure the model's robustness and generalization, given the imbalanced nature of fraud data. The obtained results demonstrate the effectiveness of the proposed approach in identifying fraudulent patterns, contributing to the security and integrity of financial transactions.**

***Index Terms*—Data Science, exploratory data analysis, Machine Learning, fraud detection**

I. Introducão

Este relatório apresenta um estudo sobre a detecção de fraudes em transações financeiras, um desafio crítico no cenário da segurança de dados e finanças digitais. Desenvolvido como requisito da disciplina de Data Science, o trabalho foca na aplicação de técnicas analíticas e preditivas para identificar padrões de fraude. Para isso, foi utilizada a base de dados CreditCard, que compreende um conjunto de transações de cartão de crédito realizadas na Europa em setembro de 2013. O principal desafio deste dataset reside em sua característica desbalanceada, onde a vasta maioria das transações é legítima (indicada por 0) e uma pequena parcela é fraudulenta (indicada por 1), exigindo abordagens cuidadosas de modelagem.

O presente estudo visa, primeiramente, realizar uma análise exploratória de dados, empregando técnicas de visualização e estatísticas descritivas para compreender a estrutura do dataset, identificar anomalias e extrair insights relevantes. Em seguida, detalha-se a aplicação de pré-processamentos essenciais para preparar os dados, abordando o desbalanceamento inerente e otimizando-os para o treinamento de modelos. Por fim, o relatório descreve o desenvolvimento e a avaliação de um modelo preditivo utilizando algoritmos de Machine Learning, com o objetivo de classificar transações de forma eficaz.

Para uma apresentação clara e sistemática das metodologias e resultados, este relatório está estruturado nas seguintes seções:

II) Análise Exploratória de Dados: Detalha a investigação inicial do dataset para identificar padrões, anomalias e insights relevantes.

III) Protocolo de Validação: Descreve as estratégias utilizadas para avaliar a performance e a robustez do modelo desenvolvido.

IV) Modelo de Machine Learning Escolhido: Apresenta o algoritmo de aprendizado de máquina selecionado e sua justificativa.

V) Resultados Obtidos: Discute as descobertas e a performance do modelo desenvolvido, apresentando as métricas de avaliação.

II. Analise Exploratória de Dados

A. Analise das Classes da Base

B. Análise da Distribuição das Classes

Dada a natureza de **classificação** do problema de detecção de fraudes, a análise exploratória teve início com a avaliação

da **distribuição das classes** na base de dados. Esta investigação revelou um **severo desbalanceamento** entre o número de ocorrências das classes, um fator crítico para a modelagem preditiva.

Conforme evidenciado na Tabela I e ilustrado na Figura 1, as transações legítimas (Classe 0) constituem a esmagadora maioria do *dataset*, representando aproximadamente 99,83% do total de 284.315 instâncias. Por outro lado, as transações fraudulentas (Classe 1) somam apenas 492 ocorrências, o que corresponde a meros 0,17% do conjunto de dados. Tal desequilíbrio impõe um desafio considerável ao treinamento de modelos de Machine Learning, pois algoritmos não ajustados podem tender a classificar a maioria das transações como não fraudulentas, comprometendo a capacidade de detectar fraudes reais. Este cenário ressalta a necessidade de estratégias específicas de pré-processamento e avaliação para lidar com *bases de dados* desbalanceadas.

TABLE I
DISTRIBUIÇÃO DAS CLASSES

Classe	*Numero de Instancias*	*% total de instancias*
0 - Não Fraudulentas	284.315	99,3%
1 - Fraudulentas	492	0,17%

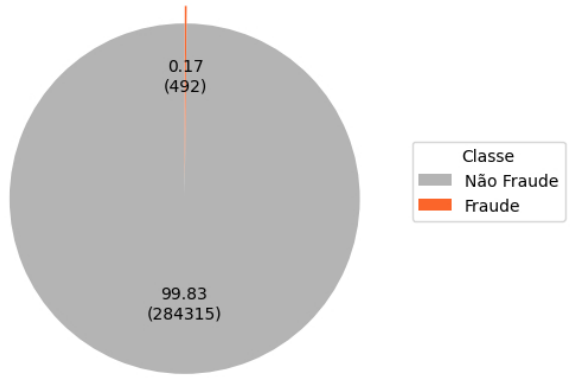

Fig. 1. Gráfico de pizza demonstrando o desbalanceamento das classes.

C. Analises Univariada

Esta seção detalha a análise univariada das variáveis presentes no *dataset*, com o objetivo de compreender suas distribuições individuais e identificar padrões relevantes para a detecção de fraudes. As variáveis explicativas originais da base de dados são:

- **Time**: Tempo decorrido em segundos desde a primeira transação registrada no *dataset*.
- **Amount**: Valor da transação em euros.
- **V1, V2, . . . , V28**: Variáveis numéricas anonimizadas, resultantes de uma transformação PCA (Principal Component Analysis). Estas variáveis foram criadas para proteger a privacidade dos usuários e a confidencialidade de informações sensíveis.

Para facilitar a interpretação e a identificação de *insights*, duas novas variáveis foram criadas:

- **Time Interval**: Intervalo de tempo, em horas, desde a primeira transação. Derivada da variável *Time*.
- **Amount Interval**: Intervalo dos valores das transações, categorizando a variável *Amount*.

É importante salientar que o *dataset* não possui valores nulos em nenhuma de suas variáveis, garantindo a completude dos dados.

A análise das variáveis *V1* a *V28* não proporcionou *insights* significativos nesta fase exploratória, devido à sua natureza anonimizada e à ausência de informações sobre as variáveis originais que as compuseram. Consequentemente, o foco da análise univariada recaiu sobre as variáveis **Amount**, **Time**, e suas versões discretizadas, **Amount Interval** e **Time Interval**, que se mostraram mais elucidativas.

1) Análise da Variável Amount e Amount Interval: As Figuras 2 e 3 ilustram a distribuição da variável *Amount* e sua versão em intervalos. Ambas as representações indicam que uma **grande concentração de transações possui valores próximos a 5 euros**. Este padrão sugere que transações com valores significativamente mais altos podem ser consideradas potenciais *outliers*, uma hipótese corroborada pelo gráfico de caixa da Figura 4, que visualiza a presença de valores extremos acima do limite superior definido pelo método de Tukey.

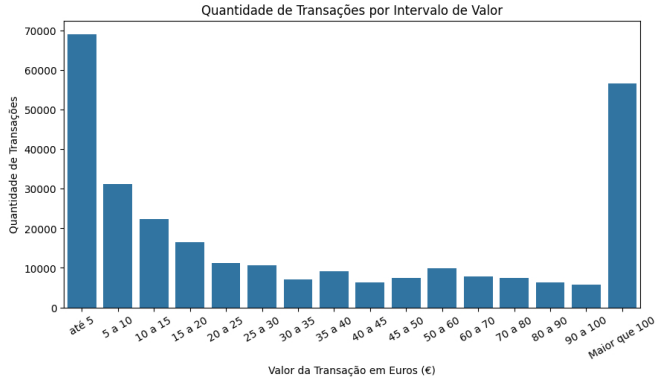

Fig. 2. Gráfico de barras demonstrando a distribuição da variavel Amount Interval

2) Análise da Variável Time e Time Interval: A análise da variável *Time* e *Time Interval* revelou padrões interessantes na distribuição das transações ao longo do tempo. Conforme a Figura 5, que apresenta a distribuição da variável *Time Interval* (já convertida para horas na sua criação), observa-se uma maior concentração de transações nos intervalos de **10 a 24 horas e de 30 a 48 horas** desde a primeira transação. A distribuição contínua da variável *Time* é visualizada no histograma da Figura 6. Diferentemente da variável *Amount*,

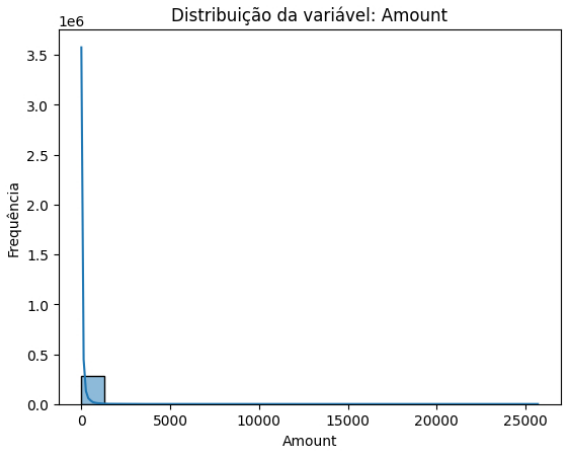

Fig. 3. Histograma com KDE demonstrando a distribuição da variavel Amount

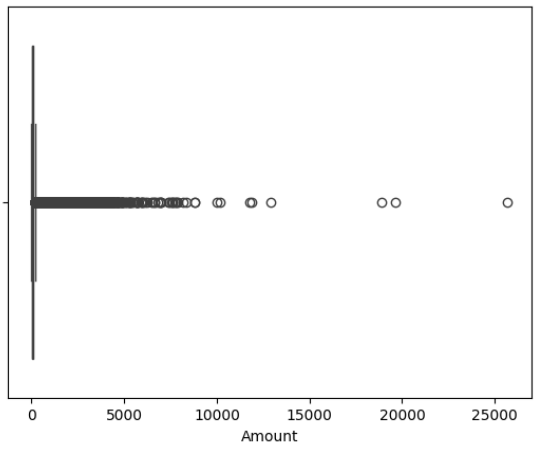

Fig. 4. Gráfico de caixa demonstrando a suspeita d eoutliers para transações com valores grandes

a aplicação do método de Tukey para detecção de *outliers* na variável *Time* não indicou a presença de valores atípicos, como pode ser verificado no gráfico de caixa da Figura 7, sugerindo uma distribuição mais homogênea em relação aos extremos.

D. Analise Multivariada

Após a análise exploratória univariada, prosseguimos para a análise multivariada, com o objetivo de investigar as relações entre as variáveis e aprofundar a compreensão do comportamento dos dados no *dataset*, especialmente em relação à detecção de fraudes. Nesta fase, formulamos as seguintes hipóteses:

1) **Relação entre Tempo e Valor da Transação:** Existe uma correlação positiva entre o valor das transações (*Amount*) e o tempo decorrido desde a primeira transação (*Time*), ou seja, espera-se que os valores das transações tendam a aumentar à medida que o tempo avança.

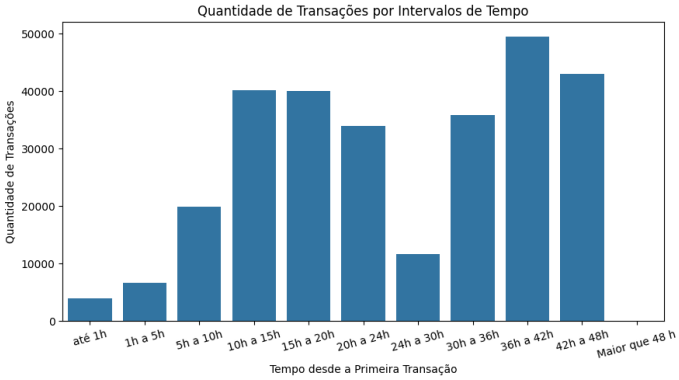

Fig. 5. Gráfico de barras demonstrando a distribuição na variavel time interval

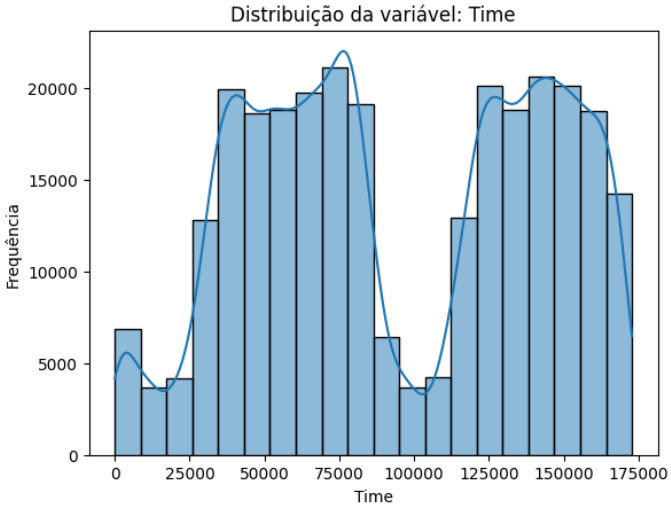

Fig. 6. Histograma demonstrando a distribuição da variavel time

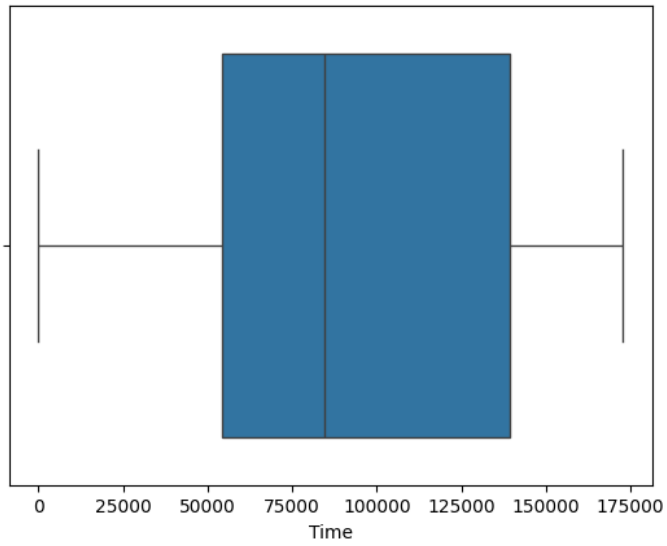

Fig. 7. Gráfico de caixa demonstrando a ausência de suspeita de outliers

2) **Tempo das Transações Fraudulentas:** As transações fraudulentas ocorreram em um período mais distante da primeira transação registrada no *dataset* em comparação com as transações legítimas.
3) **Valor das Transações Fraudulentas:** As transações fraudulentas possuem, em média, um valor inferior ao das transações não fraudulentas.
4) **Intervalo de Tempo de Transações Fraudulentas vs. Não Fraudulentas:** O intervalo de tempo mais frequente para a ocorrência de transações fraudulentas difere do intervalo mais frequente para as transações não fraudulentas.
5) **Intervalo de Valor de Transações Fraudulentas vs. Não Fraudulentas:** O intervalo de valor mais frequente para as transações fraudulentas difere do intervalo mais frequente para as transações não fraudulentas.
6) **Concentração de Transações por Intervalos de Valor e Tempo:** Existe uma concentração de transações (sejam elas fraudulentas ou não) em determinados intervalos de tempo e valor.

1) Hipótese 1: Relação entre Tempo e Valor da Transação: A primeira hipótese levantada propunha que o valor das transações (*Amount*) aumentaria com o tempo decorrido desde a primeira transação registrada no *dataset* (*Time*). A motivação para essa hipótese era a possibilidade de que transações mais recentes pudessem ter valores maiores, resultando em uma separação clara dos dados iniciais no que se refere ao valor.

Para validar esta hipótese, utilizou-se um **gráfico de dispersão**, uma ferramenta visual eficaz para identificar a existência e a natureza de correlações entre duas variáveis contínuas. Conforme ilustrado na Figura 8, o gráfico de dispersão não revela um padrão discernível ou uma tendência de aumento ou diminuição dos valores de transação em relação ao tempo. Este achado foi corroborado pelo **coeficiente de correlação**, que confirmou a ausência de uma correlação linear significativa entre a variável *Time* e a variável *Amount*. Portanto, a hipótese de que os valores das transações aumentam com o tempo não é suportada pelos dados.

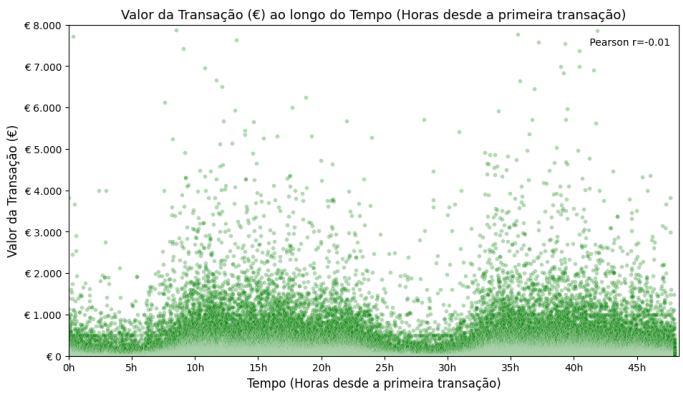

Fig. 8. Gráfico de dispersão demonstrando a não correlação entre as variáveis

2) Hipótese 2: Tempo das Transações Fraudulentas: A segunda hipótese investigou se as transações fraudulentas ocorreriam em um período mais distante da primeira transação registrada no *dataset*. A validação dessa hipótese era relevante, pois um padrão temporal distinto para fraudes poderia simplificar sua detecção por modelos de Machine Learning.

Para testar essa premissa, optou-se pela utilização de um **gráfico de caixa (boxplot)**, ferramenta visual ideal para comparar a distribuição de uma variável numérica entre diferentes categorias, permitindo a análise de quartis e a identificação de tendências centrais e de dispersão. A Figura 9 apresenta o gráfico de caixa comparando a distribuição da variável *Time* para transações fraudulentas e não fraudulentas.

Contrariando a hipótese inicial, a análise do gráfico de caixa revela que a distribuição temporal das transações fraudulentas está, na verdade, concentrada em períodos **mais próximos da primeira transação** do que as transações legítimas. Observa-se que tanto o segundo quartil (mediana) quanto o terceiro quartil para a classe de transações fraudulentas são menores em comparação com os respectivos quartis das transações não fraudulentas. Isso significa que aproximadamente 75% das transações fraudulentas ocorreram em um intervalo de tempo mais inicial do que 75% das transações legítimas. Essa distinção temporal, apesar de oposta à hipótese original, pode ser um indicador valioso para a classificação futura pelo modelo de Machine Learning.

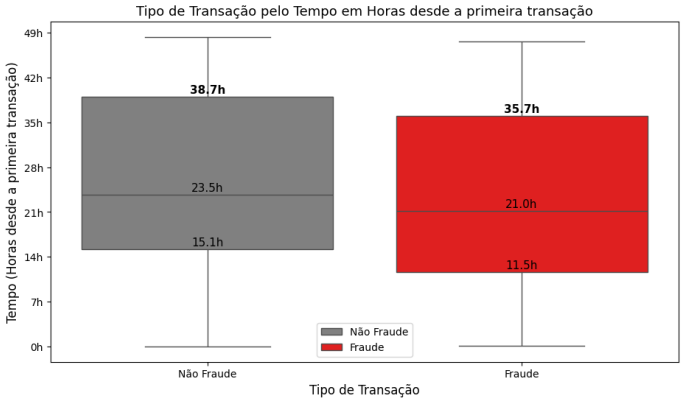

Fig. 9. Gráfico de caixa demonstrando as diferenças entre os quartis das duas classes

3) Hipótese 3: Valor das Transações Fraudulentas: Esta hipótese investigou se as transações fraudulentas tenderiam a apresentar valores monetários menores. A premissa é que fraudes de menor valor teriam maior probabilidade de passar despercebidas, enquanto transações de grande valor poderiam chamar mais atenção e serem detectadas mais facilmente.

Para testar esta hipótese, foram utilizados **gráficos de caixa (boxplot)** e **violino**, apresentados nas Figuras 10 e 11, respectivamente. Essas visualizações são eficazes para comparar a distribuição de uma variável numérica entre diferentes grupos e identificar a presença e concentração de *outliers*.

Os resultados obtidos **corroboram a hipótese** de que transações fraudulentas tendem a ter valores menores. Embora o terceiro quartil das transações fraudulentas possa ser ligeiramente maior em alguns casos, a análise geral dos

gráficos, especialmente a distribuição dos *outliers*, revela uma concentração muito maior de transações com valores extremamente altos na classe de transações **não fraudulentas** do que na classe de transações fraudulentas. Isso sugere que valores de transação muito elevados são, de fato, menos propensos a serem fraudulentos, o que pode ser um forte indicativo para a classificação do modelo.

É fundamental, contudo, considerar um **potencial viés** nesta análise devido ao severo desbalanceamento das classes. A pequena quantidade de instâncias fraudulentas pode influenciar a representação de sua distribuição de valores. Caso mais dados para a classe minoritária (fraude) estivessem disponíveis, a distribuição e a presença de *outliers* poderiam exibir um comportamento diferente, afetando as conclusões sobre a prevalência de valores altos em fraudes.

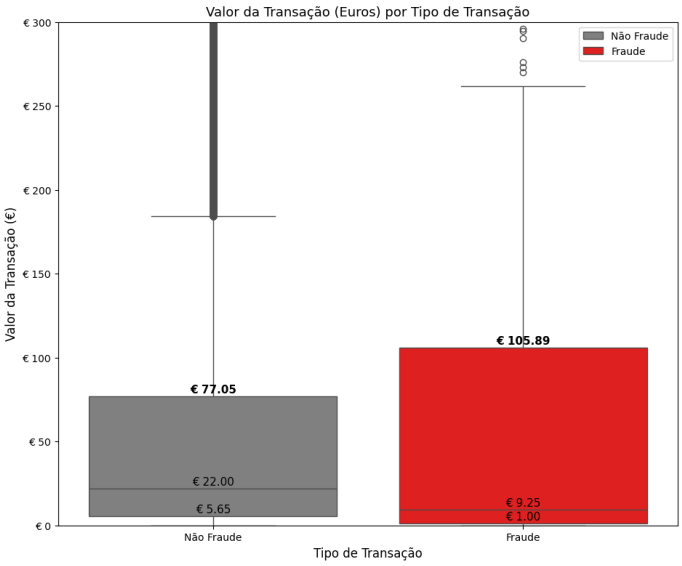

Fig. 10. Gráfico de caixa demonstrando as diferenças entre os quartis das duas classes

4) Hipótese 4: Intervalo de Tempo de Transações Fraudulentas vs. Não Fraudulentas: O objetivo desta hipótese foi investigar se a distribuição de transações fraudulentas em intervalos de tempo seria distinta da distribuição de transações não fraudulentas. A identificação de padrões temporais específicos para fraudes pode ser um fator discriminatório relevante para o modelo de Machine Learning.

Para analisar essa diferença, utilizou-se um **mapa de calor**, como apresentado na Figura 12. Este gráfico permite visualizar a concentração de transações em diferentes intervalos da variável *Time Interval* para cada classe (*fraudulentas* e *não fraudulentas*).

Os resultados confirmam a hipótese: os intervalos de tempo mais frequentes para transações fraudulentas são, de fato, diferentes dos observados para transações legítimas. As transações fraudulentas exibem maior frequência nos intervalos de 10h a 15h e de 24h a 30h. Em contraste, as transações não fraudulentas concentram-se predominantemente nos intervalos de 36h a 42h e 42h a 48h. Essa distinção clara no

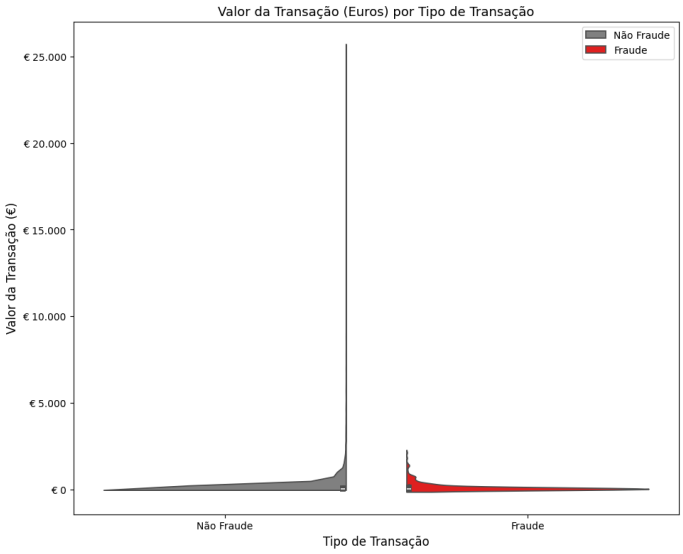

Fig. 11. Gráfico de violino mostrando a diferença na dispersão dos valores

comportamento temporal das duas classes reforça o potencial preditivo da variável *Time Interval* para a detecção de fraudes.

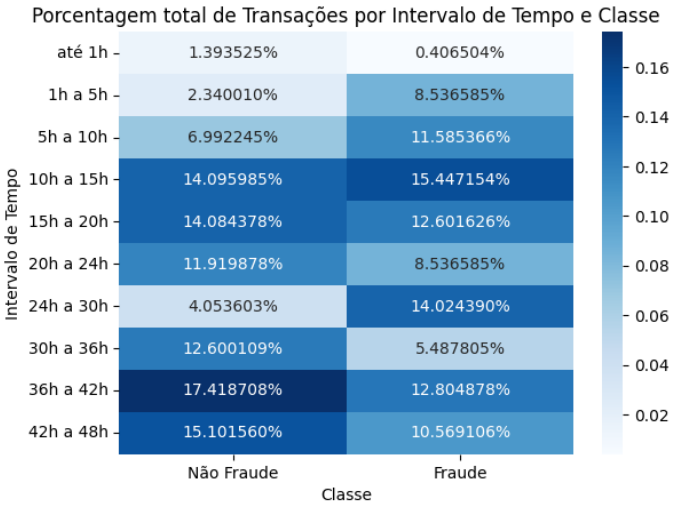

Fig. 12. Mapa de Calor evidenciando os intervalos mais concentrados para as duas classes

5) Hipótese 5: Intervalo de Valor de Transações Fraudulentas vs. Não Fraudulentas: Esta hipótese buscou verificar se a distribuição de transações fraudulentas em diferentes intervalos de valor seria distinta daquela observada para transações não fraudulentas. O objetivo era identificar um padrão de valores específico para fraudes, que pudesse atuar como um preditor relevante para o modelo de Machine Learning.

Para essa análise, um **mapa de calor** foi utilizado, conforme apresentado na Figura 13. O gráfico permite visualizar a frequência de ocorrência de transações em cada intervalo da variável *Amount Interval*, segmentado por classe (fraudulenta ou não fraudulenta).

Os resultados, contudo, não corroboram a hipótese. A análise do mapa de calor revela que os intervalos de valor mais frequentes são, em grande parte, os mesmos para ambas as classes de transações. Embora exista uma ligeira elevação na frequência para transações fraudulentas no intervalo de 90 a 100 euros, essa diferença não se mostra significativa no panorama geral da distribuição. Conclui-se, portanto, que a variável *Amount Interval*, por si só, oferece pouca distinção entre as classes de transações fraudulentas e não fraudulentas.

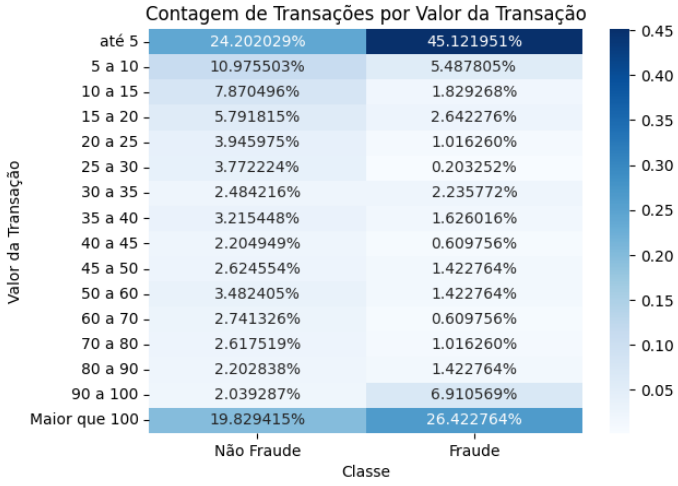

Fig. 13. Mapa de Calor evidenciando os intervalos de valor mais concentrados para as duas classes

6) Hipótese 6: Concentração de Transações por Intervalos de Valor e Tempo: O objetivo desta hipótese foi investigar se existem regiões de maior densidade de transações ao analisar conjuntamente os intervalos de valor (*Amount Interval*) e tempo (*Time Interval*). Compreender essas concentrações é crucial para desvendar o comportamento geral dos dados e identificar padrões de distribuição que podem ser relevantes para a detecção de fraudes.

Para esta análise bivariada, utilizou-se um **mapa de calor**, apresentado na Figura 14. Este tipo de gráfico é ideal para visualizar a frequência ou densidade de ocorrências em duas dimensões categóricas ou binned.

Os resultados confirmam a hipótese: existe uma concentração notável de transações em combinações específicas de intervalos de valor e tempo. Por exemplo, transações com valores de até 5 euros, combinadas com intervalos de tempo de 10h a 24h e de 30h a 48h, representam aproximadamente 20% do total de transações da base de dados. Essa descoberta reforça a existência de padrões de comportamento predominantes no *dataset*, o que pode ser um *insight* valioso para futuras etapas de modelagem.

III. Protocolo de Validação

Esta seção detalha o **protocolo de validação** empregado no estudo, abordando as etapas de pré-processamento e seleção de características (features) que foram cruciais para preparar o *dataset* e otimizar o desempenho do modelo de classificação de fraudes.

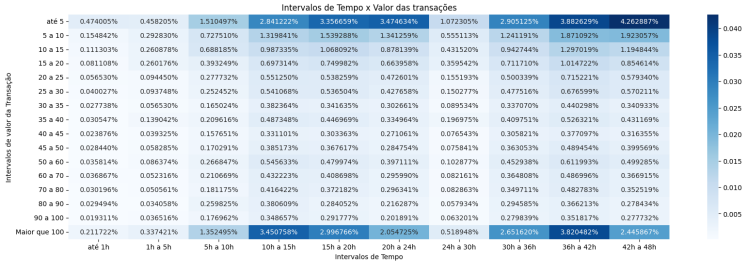

Fig. 14. Mapa de Calor evidenciando os intervalos de valor mais concentrados da base

1) Pré-processamento dos Dados: O pré-processamento dos dados foi uma etapa fundamental para adequar o *dataset* às exigências dos algoritmos de Machine Learning e para lidar com as características específicas das variáveis. As seguintes transformações foram aplicadas:

- **Codificação de Variáveis Categóricas:** As variáveis de intervalo que foram criadas (*Amount Interval* e *Time Interval*) foram transformadas utilizando *OrdinalEncoder*. Esta técnica atribui um valor numérico sequencial a cada categoria, preservando a ordem, quando aplicável.
- **Transformação de Normalização (Amount):** A variável *Amount* foi submetida à transformação *PowerTransformer* com o método *Yeo-Johnson*. O objetivo foi reduzir a assimetria da distribuição da variável, tornando-a mais próxima de uma distribuição normal e, assim, melhorando o desempenho de algoritmos sensíveis à distribuição dos dados.
- **Escalonamento Min-Max:** Por fim, todas as variáveis numéricas do *dataset* foram escalonadas para o intervalo de 0 a 1 utilizando *MinMaxScaler*. Esta etapa garante que todas as características contribuam de forma equitativa para o modelo, evitando que variáveis com maiores escalas dominem o processo de aprendizado.

2) Seleção de Características (Feature Selection): Para a **seleção de características**, adotou-se uma abordagem baseada em filtro, empregando o método *SelectKBest*. Esta técnica ranqueia as características de acordo com uma métrica estatística, permitindo a retenção das mais relevantes para o problema de classificação.

Com base nos resultados do filtro, conforme ilustrado na Figura 15, as seguintes colunas foram identificadas como menos relevantes e, consequentemente, removidas do *dataset* para o treinamento do modelo: *Amount*, *V22*, *V23*, *V24*, *V25*, *V26*, *V28*, *Time*, *V15*, *V13*, *V27*, *V8*, *V20*, *Amount Interval* e *Time Interval*. A exclusão dessas características visa reduzir a dimensionalidade dos dados, mitigar ruídos e melhorar a eficiência e o desempenho preditivo do modelo.

3) Estratégias de Validação: Para a avaliação do modelo de classificação, foram empregadas duas estratégias complementares, visando garantir a robustez e a confiabilidade dos resultados:

- **Métrica de Avaliação: F1-Score.** Dada a natureza altamente desbalanceada do *dataset* (transações fraudu-

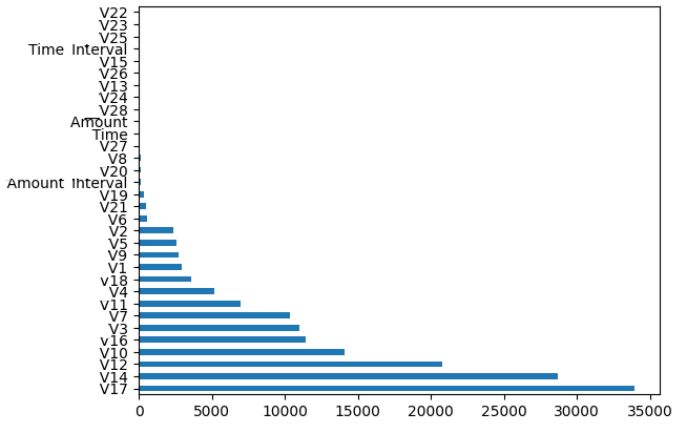

Fig. 15. Gráfico de barras com resultado do Filtro

lentas versus não fraudulentas), o **F1-Score** foi selecionado como a métrica principal de avaliação. Esta escolha é justificada por sua capacidade de ponderar a precisão (proporção de verdadeiros positivos entre todas as classificações positivas) e o recall (proporção de verdadeiros positivos entre todos os positivos reais). Diferentemente da acurácia simples, o F1-Score é menos sensível ao domínio da classe majoritária. Além disso, a dificuldade em determinar se falsos positivos (transações legítimas classificadas como fraude) ou falsos negativos (fraudes não detectadas) possuem um impacto mais severo na aplicação real, levou à escolha do F1-Score, que busca um equilíbrio entre essas duas preocupações.

- **Métodos de Validação:**
  - **Hold-out:** Uma porção do *dataset* foi separada para validação, garantindo que o modelo fosse avaliado em dados não vistos durante o treinamento.
  - **Validação Cruzada Estratificada com 10 Dobras:** Para obter uma estimativa mais robusta do desempenho do modelo e mitigar a variância associada à divisão única do hold-out, foi aplicada a validação cruzada com 10 dobras (k-fold cross-validation). A natureza **estratificada** dessa validação é crucial em datasets desbalanceados, pois assegura que a proporção de classes (fraudulentas e não fraudulentas) seja mantida em cada dobra, evitando cenários onde dobras possam conter poucas ou nenhuma instância da classe minoritária.

IV. Modelo de Machine Learning Escolhido

Para a tarefa de classificação de fraudes, o modelo de **Árvore de Decisão** foi selecionado. Esta escolha baseia-se na sua simplicidade e robustez, características que o tornam particularmente adequado para lidar com grandes volumes de dados. Modelos baseados em árvores oferecem uma vantagem significativa em termos de velocidade de treinamento e inferência quando comparados a algoritmos como K-Nearest Neighbors (KNN), cuja complexidade computacional aumenta consideravelmente com o tamanho do *dataset*, ou classificadores mais simples como Naive Bayes, que podem não capturar a complexidade das relações nos dados de forma tão eficaz.

Os parâmetros definidos para o modelo de Árvore de Decisão foram:

- **Critério de Divisão (*criterion*):** Entropia. Este critério mede a impureza da informação de um nó, buscando criar divisões que resultem em nós filhos com maior homogeneidade em relação às classes.
- **Profundidade Máxima (*max_depth*):** 5. A profundidade máxima da árvore foi limitada a 5 para controlar a complexidade do modelo, prevenir o sobreajuste (*overfitting*) e manter a interpretabilidade.
- **Estado Aleatório (*random_state*):** 42. A definição de um estado aleatório fixo (42) garante a reprodutibilidade dos resultados, assegurando que, a cada execução, as divisões da árvore sejam construídas de maneira consistente.

V. Resultados Obtidos

Esta seção apresenta os resultados do desempenho do modelo de Árvore de Decisão na tarefa de detecção de fraudes, utilizando as estratégias de validação definidas anteriormente. O foco da análise recai sobre as métricas de Precision, Recall e F1-Score, além da interpretação das matrizes de confusão.

A. Avaliação com Hold-out

A avaliação inicial do modelo utilizando a estratégia **Hold-out** demonstrou os seguintes resultados para a classe minoritária (fraude):

- **Precision:** 81,63%
- **Recall:** 81,08%
- **F1-Score:** 81,35%

A **matriz de confusão** correspondente a essa avaliação é apresentada na Figura 16. A análise da matriz permite visualizar o número de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos, fornecendo um detalhamento da capacidade do modelo em classificar corretamente as transações fraudulentas e legítimas.

B. Avaliação com Validação Cruzada Estratificada (10 Dobras)

Para uma avaliação mais robusta e para mitigar a variância da divisão única do Hold-out, o modelo foi submetido à **Validação Cruzada Estratificada com 10 Dobras**. Os resultados médios obtidos para a classe minoritária (fraude) foram:

- **Precision Médio:** 89,56%
- **Recall Médio:** 77,84%
- **F1-Score Médio:** 83,21%

A **matriz de confusão média** resultante das 10 dobras é ilustrada na Figura 17. É possível observar que a média dos resultados da validação cruzada indica um Precision superior ao do Hold-out, embora o Recall médio seja ligeiramente menor. O F1-Score médio, que busca um equilíbrio entre essas métricas, demonstra um desempenho consistente e promissor do modelo na detecção de fraudes.

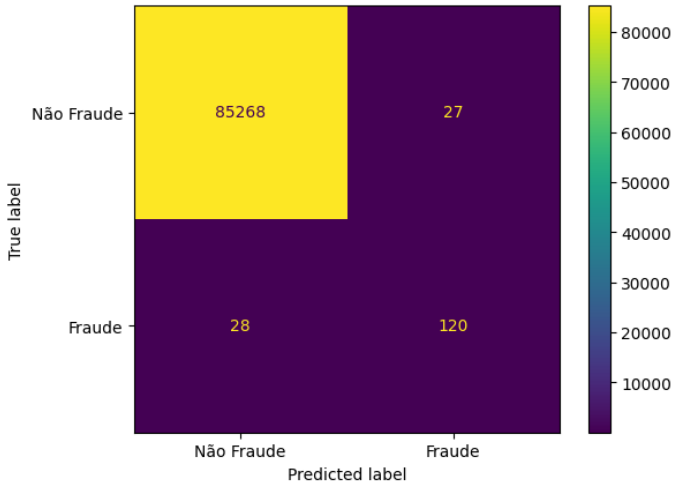

Fig. 16. Matriz de confusão do Hold-out

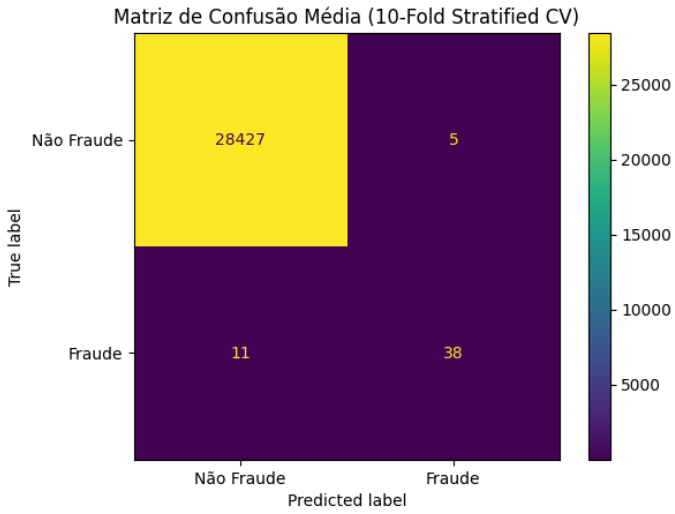

Fig. 17. Matriz de confusão média das 10 dobras

VI. Conclusão

A elaboração deste relatório proporcionou uma valiosa oportunidade para consolidar e aplicar os conhecimentos adquiridos na disciplina de Data Science a um problema complexo e de grande relevância no mundo real: a detecção de fraudes financeiras. A experiência demonstrou a importância crítica de cada etapa do pipeline de análise de dados, desde a análise exploratória até as fases de pré-processamento.

Neste estudo, mesmo um modelo de classificação considerado simples, como a **Árvore de Decisão**, foi capaz de alcançar um desempenho notável na base de dados desbalanceada, com um F1-Score médio de 83,21% na validação cruzada. Este resultado ressalta a importância fundamental de todas as etapas preparatórias. Ele ilustra que, em muitos cenários, a **qualidade e o pré-processamento adequado dos dados** podem ser mais determinantes para o sucesso de um modelo preditivo do que a complexidade do algoritmo em si. Frequentemente, um modelo mais simples, alimentado por dados bem tratados, pode superar um "hipermodelo" complexo que opera com dados de baixa qualidade ou inadequadamente preparados.

Em síntese, o trabalho reforça a premissa de que a excelência em Data Science não reside apenas na escolha de algoritmos avançados, mas principalmente na meticulosa atenção às fases de exploração e transformação dos dados, que pavimentam o caminho para soluções robustas e eficazes.

References

[1] Pandas Development Team, "pandas: powerful Python data analysis and manipulation," 2024. [Online]. Available: https://pandas.pydata.org

[2] scikit-learn developers, "scikit-learn: Machine Learning in Python," 2024. [Online]. Available: https://scikit-learn.org/stable/

[3] NumPy Development Team, "NumPy: the fundamental package for scientific computing with Python," 2024. [Online]. Available: https://numpy.org

[4] ULB MLG, "Credit Card Fraud Detection Dataset," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud