

Relatório Analítico: Classificação e Agrupamento em Dados de Fraude

Lucas Lopes Freitas Moura

23 de novembro de 2025

Parte 1 — Preparação da Base e Técnicas Supervisionadas

1 Visão Geral

Este documento apresenta um estudo sobre a base pública *Credit Card Fraud Detection*, onde o propósito foi preparar e modelar os dados com foco na identificação de operações fraudulentas.

A principal dificuldade encontrada nessa base é o forte desbalanceamento entre as classes: transações fraudulentas representam apenas 0,17% do total.

2 Pré-processamento Utilizado

2.1 Remoção de Redundância

Embora nenhuma coluna apresentasse valores faltantes, verificou-se a existência de registros duplicados.

- **Procedimento adotado:** Eliminaram-se **1.081 entradas repetidas**, evitando que essas repetições impactassem negativamente o treinamento.

2.2 Distribuição das Classes

A variável alvo (`Class`) confirma a assimetria entre rótulos.

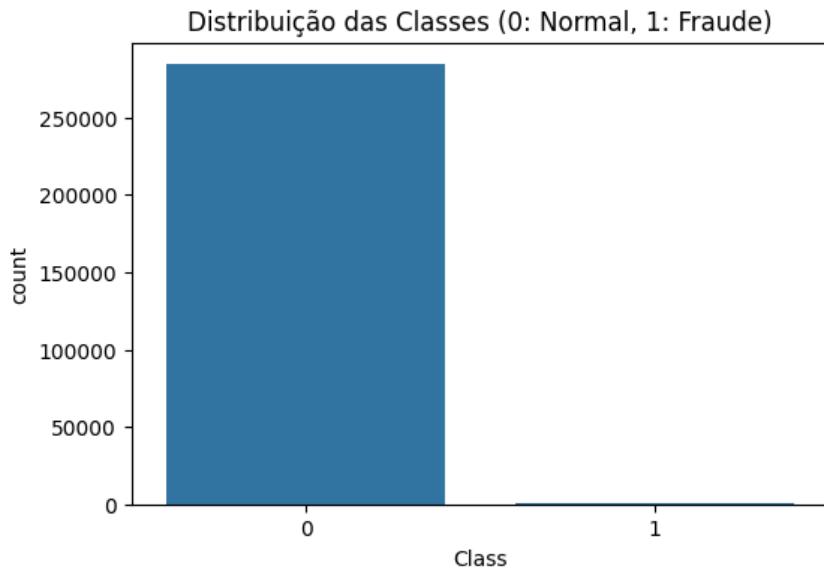


Figura 1: Distribuição das classes. A Classe 1 (fraude) aparece com frequência ínfima.

2.3 Escalonamento

Como as variáveis produzidas via PCA (V1–V28) já se encontram padronizadas, o tratamento concentrou-se em *Time* e *Amount*. Optou-se pelo uso do **RobustScaler**, considerando a presença de valores extremos.

2.4 Correlação entre Atributos

A análise da matriz de correlação evidencia a independência entre as variáveis principais, consequência direta da aplicação de PCA.

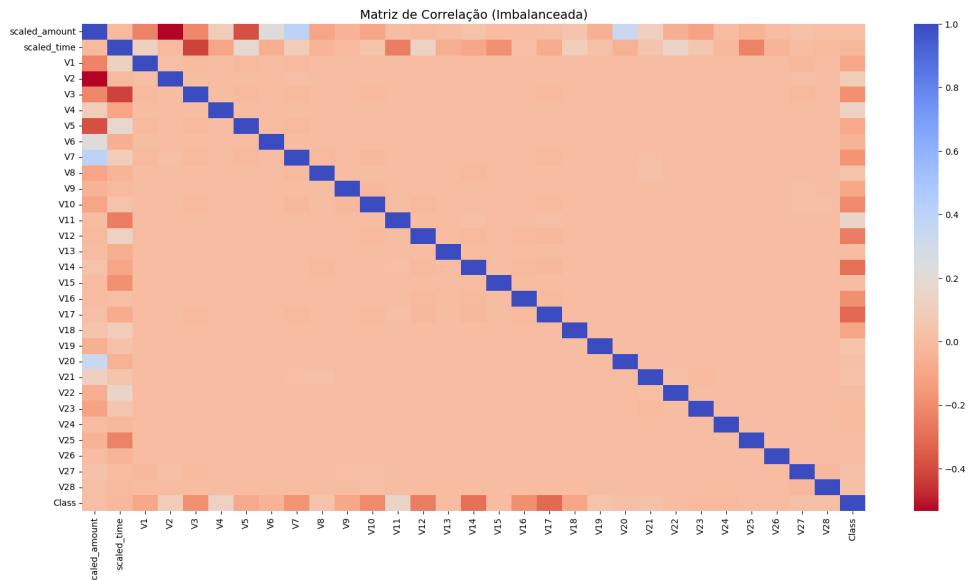


Figura 2: Mapa de correlação entre os atributos.

3 Impacto do Balanceamento

Para mitigar o desequilíbrio entre classes, foi empregado o algoritmo **SMOTE** exclusivamente no conjunto de treinamento.

- **Antes do SMOTE:** Recall de 0.58.
- **Após SMOTE:** Recall aumentou para **0.87**.

Parte 2 — Avaliação com Algoritmos de Agrupamento

1 Procedimentos Adotados

Nesta etapa investigou-se se algoritmos não supervisionados seriam capazes de segmentar naturalmente as transações em dois grupos distintos.

- A coluna *Class* foi retirada para garantir imparcialidade.
- Utilizou-se uma amostra de **10.000 registros** para viabilizar o cálculo eficiente da métrica *Silhouette Score*.

2 Desempenho dos Algoritmos

A seguir apresenta-se um resumo comparativo dos métodos analisados.

Método	Silhueta	Distribuição	Comentário
K-Means (k=2)	0.6180	C0: 9712 — C1: 288	Pouco confiável
DBSCAN	-0.0527	C0: 7263 — Ruído: 1779	Coerente
SOM (2x1)	0.4940	C0: 9353 — C1: 647	Separação artificial

Tabela 1: Resumo dos resultados de agrupamento.

3 Discussão dos Resultados

3.1 Considerações por Algoritmo

1. **K-Means:** Apesar da silhueta elevada (0.61), a separação produzida não reflete o padrão real das fraudes, agrupando majoritariamente outliers.
2. **DBSCAN:** Apresentou o resultado mais fiel. A elevada quantidade de ruído e a silhueta negativa sugerem inexistência de clusters bem definidos.
3. **SOM:** A estrutura imposta (2 neurônios) produz uma divisão artificial, confirmada pela silhueta intermediária (0.49).

3.2 Conclusão Geral

Os testes indicam que a base **não apresenta dois grupos distintos** que indiquem uma separação natural entre fraudes e transações legítimas. Isso reforça a importância de métodos supervisionados que utilizem rótulos e estratégias de balanceamento, como discutido na Parte 1.

Código Utilizado

Os scripts referentes à limpeza, preparação, visualizações e algoritmos encontram-se no endereço:

<https://github.com/lucaslopes0603/IA>