UNIVERSIDADE DE BRASÍLIA
Faculdade do Gama

Aprendizado de Máquina

Mini Trabalho 4

Preparação dos dados

Grupo - 1

Ana Clara Barbosa Borges
André Emanuel Bispo da Silva
Artur Handow Krauspenhar
Gabriel Moura dos Santos
João Artur Leles Ferreira Pinheiro
João Pedro Anacleto Ferreira Machado

Brasília, DF
2025

1. Qualidade da limpeza dos dados

O Dataset obtido do kaggle apresenta boa qualidade, com features numéricas pré-processadas por PCA (V1-V28), ausência de missing values e a variável alvo (Class) que foi adequadamente codificada. Durante a análise exploratória, identificamos a presença de linhas duplicadas. No entanto, devido à ausência de contexto financeiro que permitisse avaliar sua finalidade, optamos por mantê-las. Essa decisão visa evitar a perda de informações potencialmente relevantes para a identificação de fraudes.

Outliers também foram mantidos, pois, no contexto de transações financeiras, comportamentos atípicos podem representar justamente os padrões que caracterizam ações fraudulentas. Considerando que fraudes são naturalmente raras em comparação com transações legítimas, a remoção desses pontos poderia comprometer a eficácia do modelo.

Além disso, como não há variáveis categóricas não codificadas, o pré-processamento concentrou-se em garantir a integridade e a consistência dos dados, sem a necessidade de transformações adicionais que alterassem sua estrutura original.

Por fim, vale destacar que a decisão de manter duplicatas e outliers foi feita de forma a priorizar a detecção de fraudes. Apesar de essa escolha poder impactar levemente o balanceamento do modelo, ela reforça o foco na preservação de possíveis padrões relevantes para a classificação.

2. Normalização e padronização

2.1 Normalização

A normalização, especificamente através da técnica Min-Max Scaling, ajusta os valores das features para um intervalo predefinido, tipicamente [0, 1] (ou [-1, 1] se existirem valores negativos). A fórmula para essa transformação é:

$$Xc = \frac{X - Xmin}{Xmax - Xmin}$$

Onde X_min e X_max são os valores mínimo e máximo da feature, respectivamente.

A principal motivação para a normalização é evitar que features com ordens de grandeza muito diferentes tenham uma influência desproporcional no modelo. Conforme ilustrado por IPNET Growth Partner (2021), em modelos como regressão

linear ou algoritmos baseados em distância (ex: K-Nearest Neighbors - KNN) e otimização por gradiente (ex: Redes Neurais, SVM), uma feature com valores numericamente maiores pode dominar o processo de aprendizado, não por sua relevância intrínseca, mas apenas por sua escala. A normalização mitiga isso, ajudando a garantir que todas as features contribuam de forma mais equilibrada.

Essa técnica é particularmente útil quando a distribuição dos dados é desconhecida ou não se assemelha a uma distribuição Gaussiana. Contudo, é importante notar que a normalização Min-Max é sensível a outliers, pois os valores extremos (X_min, X_max) determinam diretamente os limites da escala transformada (IPNET GROWTH PARTNER, 2021; SHAIBU, 2024b).

2.2 Padronização

A padronização, por sua vez, transforma os dados para que tenham média 0 e desvio padrão 1, subtraindo a média e dividindo pelo desvio padrão de cada feature (cálculo do Z-score). Essa técnica é frequentemente recomendada para algoritmos que assumem dados centrados em zero ou são sensíveis à variância, como na Análise de Componentes Principais (PCA), e para modelos otimizados por gradiente (Regressão Logística, SVM, Redes Neurais), pois ajuda a acelerar a convergência e equilibra a influência das features. Assim como a normalização, a padronização altera a escala e a centralização dos dados, mas não a ordem relativa dos valores dentro de uma feature, tornando-a também não essencial para árvores de decisão, que são inerentemente robustas a essas transformações por se basearem em thresholds (SHAIBU, 2024b).

Bibliografia

SHAIBU, S. **Normalização vs. Padronização: Como saber a diferença**. Disponível em: <https://www.datacamp.com/pt/tutorial/normalization-vs-standardization>.

IPNET GROWTH PARTNER. A importância da normalização e padronização dos dados em Machine Learning. **Parceiro de crescimento IPNET**, Medium, 15 fev. 2021. Disponível em: https://medium.com/ipnet-growth-partner/padronizacao-normalizacao-dados-machine-learning-f8f29246c12.