Técnicas de Pré-Processamento de Dados para Datasets de Detecção de Fraudes e Expectativa de Vida
Um estudo prático e comparativo

Davi S. da Cruz[1], **Sávio H. B. Oliveira**[1], **Yure G. V. Pires** [1]

[1]Departamento de Informática – Instituto Federal do Maranhão (IFMA)
Coelho Neto - MA

{davi.cruz,saviobarradas,yurepires}@acad.ifma.edu.br

Abstract. *This study aims to present the results obtained after applying data processing techniques to two distinct datasets: one focused on fraud detection and the other related to life expectancy analysis. The data preparation steps included the removal and treatment of missing values, normalization of numerical attributes, and the detection and treatment of outliers. These processes to which the variables were subjected enabled the construction of more consistent, cohesive, suitable datasets for further analysis and predictive model development. The results demonstrate a significant improvement in data quality and reveal relevant patters that may contribute to the understanding of the analyzed phenomena, highlighting the importance of preprocessing as fundamental step in projects that deal with large volumes and high diversity of data.*

Resumo. *Este trabalho tem como objetivo apresentar os resultados obtidos após a aplicação de técnicas de pré-processamento em duas bases de dados distintas: uma voltada à detecção de fraudes e outra relacionada à análise da espectativa de vida. As etapas de preparação dos dados incluíram a remoção e tratamento de valores ausentes, a normalização e/ou padronização de atributos numéricos e detecção e tratamento de outliers. Esses processos a qual as variáveis foram submetidas permitiu a obtenção de conjuntos de dados mais consistentes, coesos e adequados para análisesposteriores e construção de modelos preditivos. Os resultados demonstram uma melhora significativa na qualidade dos dados e revelam padrões relevantes que podem contribuir para a compreensão dos fenômenos analisados, destacando a importância do pré-porcessamento como etapa fundamental em projetos que trabalham com grande diversidde e quantidade de dados.*

1. Introdução

A qualidade dos dados desempenha um papel crucial no desempenho de modelos análiticos e preditivos em projetos de aprendizado de maquina. Dados brutos freequentemente apresentam imperfeições, como vários valores ausentes, inconsistências, variáveis categóricas não codificadas e a presença de ouliers, que podem comprometer significativamente os resultados das análises. Neste contexto, o pré´processamento de dados surge como um etapa essencial para assegurar a integridade e a confiabilidade das informações utlizadas, a fim de garantir melhores resultados em um futuro uso para o treinamento de modelos de aprendizado de máquina. Este artigo tem como propósito descrever e avaliar

os impactos do pré-processamento em duas bases de dados com finalidades distintas: uma voltada a detecção de fraudes e outra relacionada à expectativa de vida. O trabalho contempla as etapas de limpeza de dados, normalização, tratamento de outliers, remoção e tratamento de valores nulos. Em virtude da aplicação dessas técnicas, foi possível transformar os dados brutos em conjuntos mais apropriados para o aprendizado de máquina. A análise dos resultados evidencia como a preparação adequada dos dados pode influenciar positivamente a extração de conhecimento e a geração de insitghts relevantes para os contextos estudados.

2. Datasets

Neste trabalho foram escolhidos dois datasets que a equipe julgou apresentarem caracteríticas favoráveis para a exemplificar os processos realizados no pré-processamento de dados. Desse modo, os bancos de dados usados neste trabalho possuem algumas colunas com dados nulos, algum que não possuem normalização e a presença de outliers. Tais aspectos nos datasets permitem a aplicação de técnicas de pré-processamento, visando tornar os dados presentes nesses bancos de dados o mais apropriados para o treinamento de modelos de aprendizado de máquina.

2.1. Life Expectancy

Este conjunto de dados foi extraído do Observatório Global da Saúde (GHO) da Organização Mundial da Saúde (OMS), com informações complementares coletadas do site das Nações Unidas. Nele estão reunidos dados reerentes à expectativa de vida e diversos fatores de saúde e econômicos de 193 países, ao longo do período de 2000 a 2015. Os dados estão mesclados em um único arquivo que contem 22 colunas e 2.938 registros. Tais colunas representam 20 variáveis preditoras que abrangem aspectos diversos como imunização, mortalidade, economia e fatores sociais. Todas esas variáveis foram selecionadas com base na sua relevância crítica para compreensão dos determinantes da expectativa de vida.

Uma análise incial, constata-se a presença de valores ausentes, principalmente nas variáveis de população, cobertura vacinal contra hepatite B e Produto Interno Bruto (PIB). Tal conjunto de dados fornece uma base sólida para aplicação de técnicas de aprendizado de máquina, sendo fundamental a realização de um pré-processamento adequado para garantir a qualidade, a coerência e a integridade dos dados que serão utilizados.

2.2. Fraud Detection

O conjunto de dados foi elaborado com base em padrões do mundo real, totalizando 51.000 registros de transações financeiras. Cada transação é rotulada como fraudulenta ou legítima, o que viabiliza a aplicação de algoritmos de aprendizado de máquina voltado para tarefas de classificação binária. O dataset conta com informações detalhadas sobre o comportamento do usuário, métodos de pagamento, uso de dispositivos e o contexto transacional.

O conjunto de dados é estruturado em 12 colunas, incluindo identificadores únicos, variáveis categóricas e numéricas. Entre os atributos mais relevantes, detacam-se: *Transaction_Amount* (valor de transação), *Transaction_Type* (tipo de operação), *Time_of_Transaction* (horário da transação), *Device_Used* (dispositivo utilizado), *Previous_Fraudulent_Transactions* (histórico de fraudes associadas ao usuário) e *Payment_Method* (forma de pagamento). A variável Fraudulent assume valores binários,

indicando se a transação foi indicada como fraudulenta (1) ou legítima (0).
A inspeção inicial dos dados aponta valores ausentes em aproximadamente 5% das entradas nas colunas *Transaction_Amount*, *Time_of_Transaction*, *Device_Used*, *Location* e *Payment_Method*. As demais colunas apresentam dados completos. A presença de variáveis categóricas indica a necessidade de codificação para o uso em modelos preditivos, e o tratamento dos dados ausentes se mosta essencial para garantir a consistência da base.

3. Metodologia

O pré-processamento dos dados é uma etapa crítica na análise de dados, principalmente ao trabalhar com dados do mundo real, onde a qualidade dos dados frequentemente esta comprometida [Han et al. 2012] . O pipeline de pré-processamento implementado neste trabalho segue as etapas conceituais apresentadas em Ferreira et al. (2020) [Lenz et al. 2020], no capítulo "Fundamentos de Aprendizagem de Máquina", onde são detalhadas as fases de: Remoção de valores ausentes; Normalização e padronização de atributos; Detecção e tratamento de outliers; Codificação de variáveis categóricas.

Neste trabalho, dois conjuntos de dados distintos foram submetidos às etapas de pré-processamento, sendo grande parte delas comuns em ambos tratamentos dos dados. Entretanto, em razão da natureza distinta destes datasets diferentes pipelines foram seguidas para alcançar os resultados desejados nos conjuntos de dados. Utilizamos as bibliotecas Pandas [pandas development team 2020], Numpy [Harris et al. 2020] e MatplotLib[Hunter 2007] para executar o pre-processamento dos dados em Python.

3.1. Visualização de Dados

Trata-se do primeiro contato com o database, crucial para obter o panorama geral do conjunto de dados, que indicará quais etapas de pré-processamento poderão ser aplicadas no mesmo. Segundo [Ferreira et al. 2021], uma análise exploratória bem conduzida reduz o risco de decisões equivocadas na fase de limpeza e tratamento.

- **Visualização univariada e bivariada**
  - Histogramas, boxplots e distribuições de densidade (com **matplotlib**) para inspecionar a forma, dispersão e presença de assimetria em cada variável, ;
  - Gráficos de barras e heatmaps para frequências de categorias e correlações iniciais.
- **Inspeção de padrões de falta de dados**
  - Mapas de missing (e.g., **missingno.matrix**) e tabelas de contagem (**df.isnull().sum()**) para identificar variáveis e subgrupos (como "Developed" vs. "Developing") com alto índice de ausência.

3.2. Análise Multivariada

- **Avaliação de colinearidade**
  - Verificação de variáveis altamente correlacionadas para possível redução de dimensionalidade ou remoção de atributos redundantes.

3.3. Tratamento de Valores Faltantes

- **Diagnóstico**
  - Identificação de padrões de missing "completos por grupo" (e.g., todas as observações de um país ou uma categoria sem dados) vs. missing esporádico.
- **Estratégias de imputação com pandas e numpy**
  - **Imputação por estatística descritiva**
    - ***df[col].fillna(df[col].median())*** ou ***df[col].fillna(df[col].mean())*** para preencher valores numéricos, optando pela mediana em presença de outliers.
  - **Imputação condicional**
    - Agrupamento (df.groupby([...])['col'].transform('median')) para calcular medianas específicas por subgrupo (e.g., país-ano ou nível de desenvolvimento).
  - **Placeholder para variáveis categóricas**
    - df['cat'].fillna('Unknown') para preservar o registro e sinalizar ausência como categoria distinta.

3.4. Normalização de Dados

- **Escalonamento com pandas, numpy ou Scikit-learn**
  - Min–Max Scaling: $((x–min)/(max–min))$ via sklearn.preprocessing.MinMaxScaler ou operações vetorizadas em numpy para ajustar atributos contínuos ao intervalo $[0, 1]$;

3.5. Discretização de Dados

- **Binning de variáveis contínuas**
  - Utilização de pd.cut() ou pd.qcut() para criar faixas (bins) de valores uniformes ou quantis, transformando atributos contínuos em ordinais.
  - Exploração de discretização com base em cortes estatísticos (quartis, decis) para facilitar modelos que se beneficiam de categorias ou para reduzir ruído.

3.6. Detecção e Tratamento de Outliers

- **Métodos estatísticos com pandas e numpy**
  - **IQR (Interquartile Range):**
    1. Cálculo de Q_1 e Q_3 com df[col].quantile([0.25,0.75]);
    2. Definição de limites: $[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$;
    3. Identificação de outliers via máscaras booleanas e remoção ou winsorização.
  - **Z-score:**
    1. Cálculo de $z = (x–\mu)/\sigma$ com numpy; filtragem de registros com $|z| > limiar$ (e.g., 3).
- **Abordagens de tratamento**
  - Remoção pura de pontos extremos quando justificado pela análise de domínio;
  - Verificação utilizando percentis extremos para mitigar influência sem remover observações.

4. Resultados e Discussões

4.1. Limpeza e Completude de Dados

- **Valores ausentes**
  - *Life Expectation Database:* 100% das 2.938 observações completadas via imputação condicional (mediana/média por país–ano e nível de desenvolvimento);
  - *Fraudes Detection Database:* $\sim$ 5% de missing nas colunas transacionais preenchidos com medianas (numéricas) e placeholders "Unknown" (categóricas).

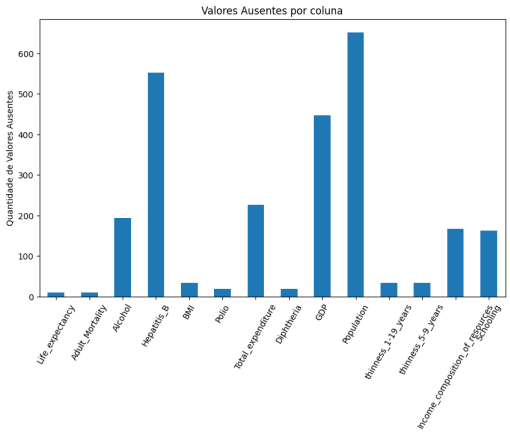

Figure 1. Life Expectancy

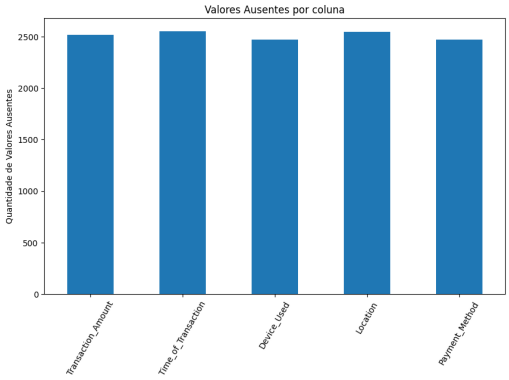

Figure 2. Fraude Detection

Figure 3. Colunas com valores ausentes por dataset

- **Duplicatas**
  - *Fraudes Detection Database:* remoção de transações idênticas legítimas, preservando todas as fraudulentas; base final ≈ 50.600 registros.

A estrategia de imputação garantiu 0% de lacunas em ambas as bases, mas com lógicas adptada ao conjunto de dados: no dataset de expectativa de vida, reforçando tendências históricas e heterogeneidade; e no dataset de detecção de fraudes, sinalizando ausências sem perder instâncias.

4.2. Homogeinização e Escalonamento

- **Min-Max Scaling** aplicado em todas as variáveis contínuas em ambas as bases, levando atributos com PIB e valor da transação ao intevalo $[0, 1]$.

A uniformização de escalas é vital para algoritmos sensíveis a magnitudes (KNN, SVM, redes neurais). As transformações extra no estudo de expectativa de vida preparam a base para escolha de modelo conforme distribuição dos dados.

4.3. Codificação e Discretização

- *Fraudes Detection Database:* one-hot encoding de *Transaction_Type*, *Payment_Method* e *Device_Used*, resultando em 24 atributos finais.

A codificação expandiu o espaço de características no dataset Fraudes Detection, capturando detalhes de cada modalidade de pagamento e dispositivo. Na base de saúde, a ausência de categorias textuais simplificou o pipeline.

4.4. Detecção de Outliers e Padrões de Anomalia

- Life Expectation Database: outliers não tratados explicitamente, pois o foco estava em manter tendências e variabilidade inerente aos países;
- Fraudes Detection Database: IQR aplicado a *Transaction_Amount* identificou $\sim X \cdots$ outliers, com taxa de fraude significativamente maior que a média.

A concentração de fraudes em valores extremos de transação sinaliza potencial de enriquecer o conjunto de features ou de criar filtros seletivos para modelos de detecção.

5. Conclusão

O desenvolvimento de um pipeline de pré-processamento robusto, estruturado em etapas bem definidas de inspeção, tratamento de valores ausentes, codificação, escalonamento e gerenciamento de outliers e duplicatas, mostrou-se essencial para garantir a qualidade e a confiabilidade dos dois conjuntos de dados estudados. No dataset de expectativa de vida, a imputação condicional por grupos preservou as tendências temporais e as diferenças entre países desenvolvidos e em desenvolvimento; já no dataset de detecção de fraudes, a combinação de placeholders, mediana e dummy encoding assegurou a integridade das observações, sobretudo da classe minoritária.

A padronização das variáveis numéricas via Min–Max Scaling (com experimentos adicionais de Standard e Power Transforms na base de saúde) uniformizou as amplitudes, preparando os dados para algoritmos sensíveis a escalas diferentes. Paralelamente, a análise exploratória e multivariada inicial permitiu fundamentar cada decisão de tratamento de outliers e escolha de técnicas de discretização, evitando a introdução de vieses e a perda de informações relevantes.

Em síntese, essas etapas de pré-processamento não apenas aumentam o poder preditivo dos modelos de aprendizado de máquina a seguir aplicados, mas também viabilizam interpretações mais confiáveis e reproduzíveis dos resultados. Como desdobramento natural, a próxima fase envolverá a seleção e validação de algoritmos de classificação e regressão, empregando validação cruzada e métricas apropriadas, de modo a explorar plenamente o ganho obtido pela preparação cuidadosa dos dados.

References

Ferreira, R. G. C., Marque, L. T., de Miranda, L. B. A., Pinto, R. A., Pessutto, L. R. C., Pereira, M. A., and de Andrade, A. L. C. (2021). *Preparação e Análise Exploratória de Dados*. Sagah, Porto Alegre.

Han, J., Kamber, M., and Pei, J. (2012). 3 - data preprocessing. In Han, J., Kamber, M., and Pei, J., editors, *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 83–124. Morgan Kaufmann, Boston, third edition edition.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R'ıo, J. F., Wiebe, M., Peterson, P., G'erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Lenz, M. L., Neumann, F. B., Santarelli, R., and Salvador, D. (2020). *Fundamentos de Aprendizagem de Máquina*. Sagah, Porto Alegre.

pandas development team, T. (2020). pandas-dev/pandas: Pandas.