Centro Universitário de Excelência
Sistemas de Informação

Pré-Processamento de dados

Autores: Lucas Ferreira Lan
Herick Marcio M. Brito
Juary dos Santos L. Júnior

Feira de Santana, 2025

Agenda

O objetivo desta apresentação é a contextualização da importância da etapa de pré-processamento, destacando como lidar com dados duplicados, ausentes, escalas diferentes e variáveis categóricas.

1. Introdução: Contextualização da importância da etapa de pré-processamento.

2. Tipos de tratamento dos dados: Apresentação das principais técnicas para o tratamento de dados.

3. Tratamento de valores ausentes: Apresentação das técnicas para o tratamento de valores ausentes.

4. Tratamento de valores com diferenças de escala: Apresentação de algumas técnicas para o tratamento de valores com diferença de escala

5. Tratamento de valores categóricos: Apresentação das principais técnicas para o tratamento de dados categóricos.

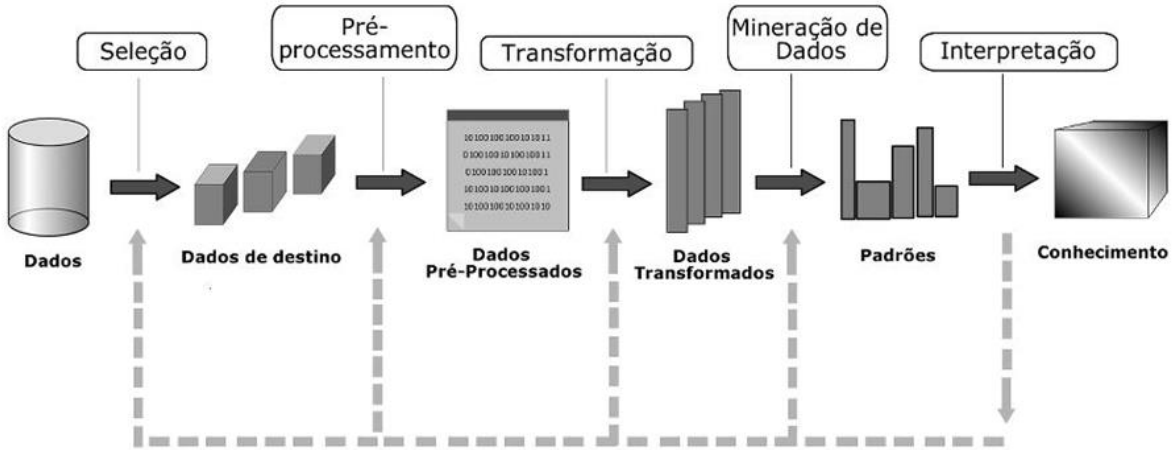
6. Resultados: Apresentação dos resultados obtidos.

Agenda

7. Considerações finais:
Considerações finais
sobre o trabalho.

Introdução

- Dados brutos provenientes das bases de dados.
- O pré-processamento é responsável pela limpeza e organização.



Pré-processamento de dados

Introdução

Benefícios do pré-processamento:

- Melhora da precisão;
- Evita resultados incorretos;
- Garantia de compatibilidade com algoritmo;
- Remoção de ruídos.

Tipos de tratamento dos dados

Valores Ausentes	Valores Duplicados	Diferença de escala	Valores categóricos
Remoção	Remoção	Padronização z-score	Codificação de Rótulo
Imputação	Agregação	Normalização Min-Max	Codificação One Hot

Valores Ausentes

Imputação

- Preenchimento dos valores nulos;
- Média, moda, mediana ou valor padrão;
- Possível distorção do resultado devido a adição de valores artificiais

x

Remoção

- Remoção das linhas com valores nulos;
- Redução do tamanho da base;
- Perda significativa de informações.

Valores Ausentes

Dataset antes do tratamento

ID_Pedido	Valor_Pedido	Avaliacao_Cliente	Tempo_Entrega_Min
201	75,00	5	40
202	25,00	3	Nulo
203	30,00	Nulo	35
204	95,00	4	55

Valores Ausentes

Pós-Imputação

ID_Pedido	Valor_Pedido	Avaliacao_Cliente	Tempo_Entrega_Min
201	75,00	5	40
202	25,00	3	**43,33**
203	30,00	**4**	35
204	95,00	4	55

Valores Ausentes

Pós-Remoção

ID_Pedido	Valor_Pedido	Avaliacao_Cliente	Tempo_Entrega_Min
201	75,00	5	40
204	95,00	4	55

Valores Ausentes

```python
colunas_verificar = self._get_target_columns(columns)

metodos_statistica = Statistics(self.dataset)

for coluna in colunas_verificar:
    valor_preenchimento = None

    if method == "mean":
        valor_preenchimento = metodos_statistica.mean(coluna)
    elif method == "median":
        valor_preenchimento = metodos_statistica.median(coluna)
    elif method == "mode":
        modas = metodos_statistica.mode(coluna)
        if modas:
            valor_preenchimento = modas[0]
    elif method == "default_value":
        valor_preenchimento = default_value
    else:
        raise ValueError(f"Método '{method}' não suportado.")

    if valor_preenchimento is not None:
        self.dataset[coluna] = [valor_preenchimento if valor_coluna is None else valor_coluna for valor_coluna in self.dataset[coluna]]
```

Imputação

Valores com diferença de escala

Z-score

- Variáveis com Média = 0 e desvio padrão = 1;
- Ideal quando os dados tem uma distribuição gaussiana;

$$z = \frac{x - \mu}{\sigma}$$

X

Min-Max

- Variáveis entre 0 e 1;
- Valor mínimo = 0, Valor máximo = 1;
- Ideal quando os dados não possuem uma distribuição gaussiana;

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

Valores com diferença de escala

Dataset antes do tratamento

ID_Pedido	Valor_Pedido	Avaliacao_Cliente
101	25.00	3
102	150.00	5
103	75.00	4
104	20.00	1

Valores com diferença de escala

Z-score	x	Min-Max
$z = (X - \mu) / \sigma$		$X' = (X - Xmin) / (Xmax - Xmin)$
$z = (25 - 67,5) / 52,26$		$X' = (25 - 20) / (150 - 20)$
$z = -0,813$		$X' = 0,038$

Valores com diferença de escala

Z-score x Min-Max

ID_Pedido	Valor_Pedido	Avaliacao_Cliente
101	-813	-169
102	1.579	1.183
103	143	507
104	-909	-1.521

ID_Pedido	Valor_Pedido	Avaliacao_Cliente
101	38	0.50
102	1.000	1
103	423	0.75
104	0	0

Valores com diferença de escala

```python
colunas_verificar = self._get_target_columns(columns)

processor = MissingValueProcessor(self.dataset)

valores_filtrados = processor.notna()

metodos_statistica = Statistics(valores_filtrados)

for coluna in colunas_verificar:
    valores = self.dataset[coluna]

    valores_numericos = [valor_numerico for valor_numerico in valores if isinstance(valor_numerico, (int, float)) and valor_numerico is not None]

    if not valores_numericos:
        continue

    media_aritmetica = metodos_statistica.mean(coluna)
    desvio_padrao = metodos_statistica.stdev(coluna)

    valores_escalonados = []
    if desvio_padrao == 0:
        valores_escalonados = [0.0 if isinstance(valor, (int, float)) else valor for valor in valores]
    else:
        for valor in valores:
            if isinstance(valor, (int, float)):
                valor_escalonado = (valor - media_aritmetica) / desvio_padrao
                valores_escalonados.append(valor_escalonado)
            else:
                valores_escalonados.append(valor)

    self.dataset[coluna] = valores_escalonados
```

Normalização Min-Max

$$z = \frac{x - \mu}{\sigma}$$

Valores categóricos

Rótulo

- Valores categóricos ordinais;
- Atribuição de um número inteiro único para cada categoria única.

x

One Hot

- Valores categóricos nominais;
- Criação de colunas binárias para cada categoria única;
- Aumento significativo do dataset.

Valores categóricos

Dataset antes da codificação

ID_Pedido	Tipo_Cozinha	Periodo_Pedido
301	Italiana	Tarde
302	Japonesa	Noite
303	Brasileira	Tarde
304	Italiana	Manhã

Valores categóricos

Dataset pós-codificação de rótulo

ID_Pedido	Tipo_Cozinha	Periodo_Pedido
301	1	1
302	2	2
303	0	1
304	1	0

Valores categóricos

Dataset pós-codificação one hot

ID_Pedido	Cozinha_Brasileira	Cozinha_Italiana	Cozinha_Japonesa	Periodo_Manhã	Periodo_Noite	Periodo_Tarde
301	0	1	0	0	0	1
302	0	0	1	0	1	0
303	1	0	0	0	0	1
304	0	1	0	1	0	0

Valores categóricos

Codificação One Hot

```
for coluna in columns:
    valores = self.dataset[coluna]
    categorias = sorted({val for val in valores if val is not None})

    for categoria in categorias:
        nova_coluna = f"{coluna}_{categoria}"

        valores_nova_coluna = [1 if valor == categoria else 0 for valor in valores]

        self.dataset[nova_coluna] = valores_nova_coluna

    del self.dataset[coluna]
```

Resultados

- Todos resultados obtidos condizentes com o esperado;
- Êxito nos testes unitários

```
............
----------------------------------------------------------------------
Ran 12 tests in 0.004s

OK
```

Resultados

```
Dataset antes do ISNA
{'idade': [20, 30, None, 50], 'salario': [500, None, 800, 1200], 'cidade': ['A', 'B', 'C', None]}


Dataset retornado pelo ISNA
{'idade': [30, None, 50], 'salario': [None, 800, 1200], 'cidade': ['B', 'C', None]}
```

```
Dataset antes do One Hot
{'idade': [20, 30, None, 50], 'salario': [500, None, 800, 1200], 'cidade': ['A', 'B', 'C', None]}


Dataset depois One Hot
{'idade': [20, 30, None, 50], 'salario': [500, None, 800, 1200], 'cidade_A': [1, 0, 0, 0], 'cidade_B': [0, 1, 0, 0], 'cidade_C': [0, 0, 1, 0]}
```

```
Dataset antes do Min-Max
{'idade': [20, 30, None, 50], 'salario': [500, None, 800, 1200], 'cidade': ['A', 'B', 'C', None]}


Dataset depois Min-Max
{'idade': [0.0, 0.3333333333333333, None, 1.0], 'salario': [0.0, None, 0.42857142857142855, 1.0], 'cidade': ['A', 'B', 'C', None]}
```

Considerações finais

- Aprendizado eficiente sobre a importância da etapa de pré-processamento;
- Consolidação de conceitos na prática com a implementação na linguagem Python;
- A implementação começou desafiadora, mas foi se tornando mais intuitiva à medida que avançávamos no projeto.

Referências

FERRARI, Daniel Gomes; CASTRO, Leandro Nunes de. *Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações.* Rio de Janeiro: LTC, 2018. E-book. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/978-85-472-0100-5. Acesso em: 8 set. 2025.

ESTATÍSTICA FÁCIL. *O que é normalização Min-Max.* Disponível em: https://estatisticafacil.org/glossario/o-que-e-normalizacao-min-max/. Acesso em: 8 set. 2025.

DATACAMP. *Normalization in machine learning.* Disponível em: https://www.datacamp.com/pt/tutorial/normalization-in-machine-learning. Acesso em: 8 set. 2025.

ESTATÍSTICA FÁCIL. *O que é Z-Score Normalization.* Disponível em: https://estatisticafacil.org/glossario/o-que-e-z-score-normalization/. Acesso em: 10 set. 2025.

DATACAMP. *One-hot encoding in Python: tutorial.* Disponível em: https://www.datacamp.com/pt/tutorial/one-hot-encoding-python-tutorial. Acesso em: 20 set. 2025.

Referências

PEDRORP. *Guia de codificadores de atributos categóricos em Machine Learning.* Medium, 2020. Disponível em: https://medium.com/@pedrorp/guia-de-codificadores-de-atributos-categ%C3%B3ricos-em-machine-learning-60a9f22c9a3b. Acesso em: 20 set. 2025.

ESTATÍSTICA FÁCIL. *O que é normalização Min-Max.* Disponível em: https://estatisticafacil.org/glossario/o-que-e-normalizacao-min-max/. Acesso em: 20 set. 2025.

DATACAMP. *Normalization vs Standardization.* Disponível em: https://www.datacamp.com/pt/tutorial/normalization-vs-standardization. Acesso em: 20 set. 2025.