

Lista de Exercícios

Curso de Extensão: Introdução à Ciência de Dados

Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) – Campus Fortaleza

Nome do Aluno: _____

Data: _____

Professor: Valberto Feitosa

Exercícios – Feature Engineering : Atributos Categóricos e Codificação

Parte 1: Conceitual (teoria)

1. (V/F): O `OneHotEncoder` é mais indicado que o `OrdinalEncoder` para variáveis categóricas nominais, pois não impõe uma ordem entre as categorias.
2. Explique com suas palavras o motivo pelo qual o Scikit-Learn não aceita atributos do tipo `string` diretamente como entrada nos modelos.
3. Qual é o risco de aplicar codificação ordinal (`OrdinalEncoder`) em uma variável sem ordem natural entre as categorias? Dê um exemplo.
4. Cite uma vantagem e uma desvantagem da codificação one-hot.
5. (Múltipla escolha): A codificação one-hot:
 - a) Cria colunas binárias para cada categoria.
 - b) Implica ordem entre os valores.
 - c) Aumenta a dimensionalidade do conjunto de dados.
 - d) É ideal para variáveis numéricas contínuas.

Parte 2: Prática com Pandas e Scikit-Learn

6. Dado o seguinte `Series` de dados categóricos:

python

 Copiar  Editar

```
import pandas as pd
dados = pd.Series(['gato', 'cachorro', 'pássaro', 'gato', 'cachorro'])
```

- a) Aplique `factorize()` e mostre os valores codificados e as categorias.
- b) Explique se essa codificação seria apropriada para uso direto em um modelo de regressão linear.

7. Usando o mesmo conjunto acima, aplique `OrdinalEncoder` e `OneHotEncoder` e mostre a saída de cada um. Comente as diferenças entre os dois resultados.

8. Crie um `DataFrame` com uma coluna "bairro" contendo os seguintes valores:

python

Copiar Editar

```
['Centro', 'Meireles', 'Aldeota', 'Centro', 'Benfica', 'Meireles']
```

- a) Use `OrdinalEncoder` e exiba o resultado.
- b) Aplique `OneHotEncoder` e exiba o resultado usando `.toarray()` com `pandas.DataFrame`.
- c) Explique por que o uso de `OneHotEncoder` seria mais adequado nesse caso.

Parte 3- Análise e Compreensão do Dataset

Objetivo:

Compreender o conjunto de dados antes de aplicar técnicas de pré-processamento e codificação, com foco em análise de variáveis categóricas e quantitativas.

Dataset:

Housing Prices Dataset – Kaggle

1- Carregamento do Dataset:

Acesse o link acima, baixe ou carregue o dataset diretamente em um ambiente de análise como Jupyter Notebook, Google Colab ou Kaggle Notebooks.

2- identificação inicial:

- a) Qual é a **variável alvo** do dataset?
- b) Liste todas as **features (atributos explicativos)** presentes no conjunto de dados.

3- Caracterização das variáveis:

Para cada feature, responda:

- a) Qual é o **significado da variável**?
- b) Ela é **numérica ou categórica**?
- c) Se **categórica**, quais são as **categorias possíveis**?
- d) Qual é a **escala ou unidade** (quando aplicável)? Ex: metros, dólares, quantidade etc.

4- Classificação das variáveis:

Classifique as features nas seguintes categorias:

Variáveis **nominais**

Variáveis **ordinais**

Variáveis **contínuas**

Variáveis **discretas**

5- Codificação de variáveis categóricas:

Selecione pelo menos duas variáveis categóricas e, para cada uma:

- a) Indique qual técnica de codificação você aplicaria: `OrdinalEncoder` ou `OneHotEncoder`.

b) Justifique sua escolha com base na natureza da variável e no tipo de modelo de machine learning que pretende utilizar.

6- tratamento de variáveis quantitativas:

- a) Selecione **duas variáveis quantitativas** e aplique a técnica de **escalonamento adequada** (ex: normalização, padronização).
- b) Selecione **duas ou mais variáveis com valores ausentes** e **trate os dados faltantes** de forma apropriada (ex: média, mediana, interpolação, exclusão etc.).

7- Preparação da base de dados:

Com base nos passos anteriores, **prepare toda a base de dados** para uma **análise descritiva completa**.

Isso inclui:

Codificação das variáveis categóricas

Tratamento dos dados faltantes

Escalonamento das variáveis numéricas

Organização dos dados em um único DataFrame pronto para modelagem ou visualização.