O Que é Transformar Uma Variável?

Idade (string)	Idade (numérico)
35	35
28	28
43	43

Precisamos transformar a variável para outro tipo de acordo com a operação a ser realizada no processo de análise.

O Que é Transformar Uma Variável?

Idade (numérico)	Faixa Etária
35	De 30 a 39 Anos
28	De 20 a 29 Anos
43	De 40 a 49 Anos

Pode ser necessário gerar gráficos ou análises por categoria e nesse caso faria mais sentido transformar a variável idade para o tipo qualitativo (categórico).

O Que é Transformar Uma Variável?

Cor do Automóvel	Cor do Automóvel Codificada
Azul	1
Verde	2
Prata	3

Pode ser necessário usar essa variável para construir um modelo de Machine Learning e para esse caso temos que codificar a variável para sua representação numérica correspondente.

O Que é Transformar Uma Variável?

Resultado do Exame	Resultado Positivo	Resultado Negativo	Resultado Não Conclusivo
Positivo	1	0	0
Negativo	0	1	0
Não Conclusivo	0	0	1

Pode ser necessário usar apenas uma categoria da variável (Resultado Positivo, por exemplo) e nesse caso temos que transformar a variável.

O Que é Encoding (Codificação)?

Gênero	Gênero Codificado
Masculino	0
Feminino	1
Outros	2

Codificação (Encoding) é a transformação de variáveis categóricas em contrapartes binárias ou numéricas.

Um exemplo é tratar o gênero como 0, 1 e 2.

Variáveis categóricas (qualitativas) podem ser codificadas com muitos métodos diferentes.

Normalmente não faz sentido codificar variáveis quantitativas.

Existem pelo menos 3 tipos principais de encoding de variáveis: Count/Frequency Encoding, Label Encoding e One-Hot Encoding.

Cada tipo pode ser usado dependendo do objetivo da análise e do formato da variável. Normalmente não há um tipo melhor do que o outro e o importante é justificar sua escolha ao aplicar um dos métodos.

Você pode encontrar diversos métodos com diferentes nomes, mas que são essencialmente variações dos 3 métodos que serão estudados neste capítulo.

Vamos estudar cada um dos tipos, agora na sequência.

Análise de Dados com Linguagem Python

Count/Frequency Encoding

Count/Frequency Encoding

Cor Automóvel	Cor Automóvel Codificada Pela Contagem	Cor Automóvel Codificada Pela Frequência
Verde	3	0.3
Azul	2	0.2
Prata	5	0.5
Prata	5	0.5
Verde	3	0.3
Prata	5	0.5
Azul	2	0.2
Prata	5	0.5
Verde	3	0.3
Prata	5	0.5

Count Encoding substitui as categorias pela contagem das observações dessa categoria no conjunto de dados.

Da mesma forma, podemos substituir a categoria pela frequência - ou porcentagem - de observações no conjunto de dados e nesse caso teríamos Frequency Encoding.

Ou seja, se 5 de nossas 10 observações mostram a cor prata, substituiríamos prata por 5 se estivermos fazendo a codificação de contagem ou por 0.5 se substituíssemos pela frequência.

Count/Frequency Encoding

Cor Automóvel	Cor Automóvel Codificada Pela Contagem	Cor Automóvel Codificada Pela Frequência
Verde	3	0.3
Azul	2	0.2
Prata	5	0.5
Prata	5	0.5
Verde	3	0.3
Prata	5	0.5
Azul	2	0.2
Prata	5	0.5
Verde	3	0.3
Prata	5	0.5

Usamos esse tip de encoding quando a variável possui um número muito alto de categorias.

Limitação: Se duas categorias diferentes aparecem a mesma quantidade de vezes no conjunto de dados, ou seja, aparecem no mesmo número de observações, serão substituídas pelo mesmo número, podendo, portanto, perder informações valiosas.

Análise de Dados com Linguagem Python

Label Encoding

Label Encoding

Cor do Automóvel	Cor do Automóvel Codificada
Azul	1
Verde	2
Prata	3

Label Encoding (ou Integer Encoding) é a substituição de uma categoria por sua representação numérica correspondente.

Label Encoding

Cor do Automóvel	Cor do Automóvel Codificada
Azul	1
Verde	2
Prata	3

Usamos Label Encoding quando temos um número baixo de categorias.

Limitação: Label Encoding não é adequado para modelos lineares como Regressão Logística.

ate 15 categorias

Análise de Dados com Linguagem Python

One-Hot Encoding

One-Hot Encoding

Resultado do Exame	Resultado Positivo	Resultado Negativo	Resultado Não Conclusivo
Positivo	1	0	0
Negativo	0	1	0
Não Conclusivo	0	0	1

One-Hot Encoding é amplamente usado em Processamento de Linguagem Natural e técnicas de processamento de texto em geral.

One-Hot Encoding

Resultado do Exame	Resultado Positivo	Resultado Negativo	Resultado Não Conclusivo
Positivo	1	0	0
Negativo	0	1	0
Não Conclusivo	0	0	1

Usamos One-Hot Encoding quando precisamos estabelecer uma relação binária entre as categorias.

Limitação: expande a dimensão à medida que o número de colunas aumenta, o que pode levar a diversos problemas no processo de análise e modelagem preditiva.

Técnicas Ideais de Encoding Para Variáveis Categóricas

Posso criar minha própria regra de encoding?

Sim, você pode. Mas certifique-se de não perder em explicabilidade.

Técnicas Ideais de Encoding Para Variáveis Categóricas

Qual a técnica ideal de codificação de variáveis?

Não há! Vai depender do objetivo.

Técnicas Ideais de Encoding Para Variáveis Categóricas

Sempre tenho que aplicar encoding nas variáveis categóricas?

Não. Depende do objetivo.

Técnicas Ideais de Encoding Para Variáveis Categóricas

Devo testar mais de uma técnica e verificar o resultado?

Se houver tempo ou dúvidas sobre qual técnica usar, pode ser uma opção.

Técnicas Ideais de Encoding Para Variáveis Categóricas

Toda técnica de encoding modifica a variável, como contorno isso?

O ideal é aplicar transformação somente se for necessário e tentar gerar o menor impacto possível, mantendo sempre a explicabilidade.