



Big Data Real-Time Analytics Com Python e Spark 3.0

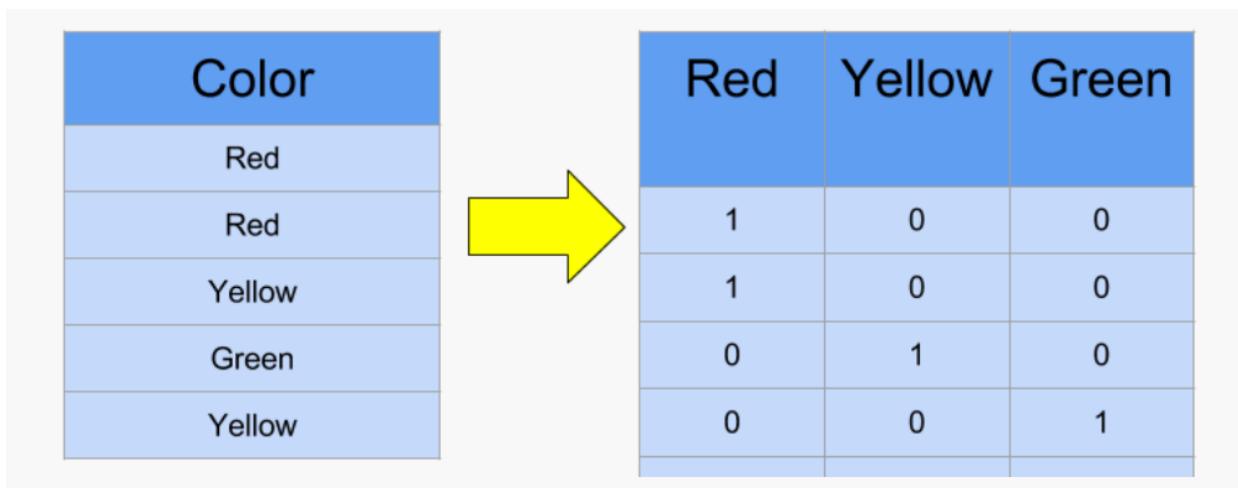
Big Data Real-Time Analytics Com Python e Spark Versão 3.0

O Que é One-Hot Encoding?

One-Hot Encoding é outra técnica popular para tratar variáveis categóricas. Ele simplesmente cria recursos adicionais com base no número de valores exclusivos no recurso categórico. Cada valor exclusivo na categoria será adicionado como um recurso (uma nova variável).

Nessa abordagem, para cada categoria de um recurso, criamos uma nova coluna (às vezes chamada de variável fictícia) com codificação binária (0 ou 1) para indicar se uma determinada linha pertence a essa categoria.

Vamos considerar a imagem abaixo. Observe que a variável Color possui 3 categorias (Red, Yellow e Green). Aplicando One-Hot Encoding 3 novas variáveis são criadas, sendo o valor 1 quando a ocorrência daquela cor e 0 quando não há ocorrência.



The diagram illustrates the One-Hot Encoding process. On the left, a table shows a single column 'Color' with five rows: Red, Red, Yellow, Green, and Yellow. A large yellow arrow points from this table to the right, indicating the transformation. On the right, a second table shows three columns: 'Red', 'Yellow', and 'Green'. The rows correspond to the same five entries as the first table. The values in the 'Red' column are 1, 1, 0, 0, and 0 respectively. The values in the 'Yellow' column are 0, 0, 1, 0, and 0 respectively. The values in the 'Green' column are 0, 0, 0, 1, and 1 respectively.

Color
Red
Red
Yellow
Green
Yellow

	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow	0	0	1

Uma potencial desvantagem desse método é um aumento significativo na dimensionalidade do conjunto de dados (que é chamado de Curse of Dimensionality).

Ou seja, a codificação one-hot é o fato de estarmos criando colunas adicionais, uma para cada valor exclusivo no conjunto do atributo categórico que gostaríamos de codificar. Portanto, se tivermos um atributo categórico que contenha, digamos, 1.000 valores exclusivos, essa codificação one-hot gerará 1.000 novos atributos adicionais e isso não é desejável.

Para simplificar, a codificação one-hot é uma ferramenta bastante poderosa, mas só é aplicável para dados categóricos que possuem um número baixo de valores exclusivos.

One-Hot Encoding é o processo de criação de variáveis fictícias (dummy variables).