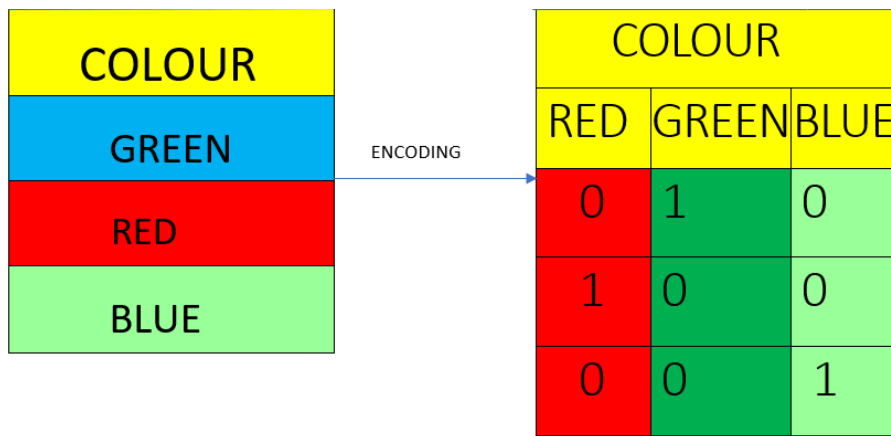


Encoding: One Hot, Label Encoder,
Getdummies

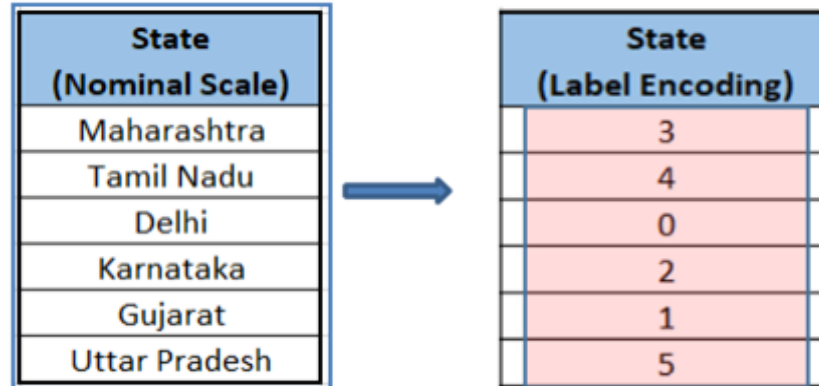
One Hot Encoding

Toman variables numéricas para medir la distancia. Sin embargo, existe la manera de trabajar con variables categóricas haciéndolas variables dummy, a través del uso de la técnica de transformación de datos **One Hot Encoding (OHE)**. No permite convertir strings directamente.



Label Encoder

Sin embargo One Hot Encoding es la única técnica para transformar variables categóricas, existe una alternativa para reducir el problema de multidimensionalidad cuando tenemos muchas categorías en una variable a través del uso de la técnica de transformación de datos **Label Encoder (LE)**.



The diagram illustrates the transformation of categorical data from a nominal scale to label encoding. It consists of two tables connected by a blue arrow pointing from left to right.

State (Nominal Scale)
Maharashtra
Tamil Nadu
Delhi
Karnataka
Gujarat
Uttar Pradesh

State (Label Encoding)
3
4
0
2
1
5

Getdummies()

Es un método similar a One Hot Encoding el cual transforma variables categórica en dummies. La diferencia es que One Hot Encoding almacena la transformación en un objeto. Una vez que se tiene la instancia `OneHotEncoder()`, se puede guardar para ser usada más tarde en las siguientes fases de manipulación de datos

Pandas Get Dummies

Turn your Categorical Column (Ex: "Name")...

Index	Name	8/6/2020
0	Liho Liho	\$234.54
1	Chambers	\$45.74
2	The Square	\$56.22
3	Liho Liho	\$32.31

...Into Dummy Indicator Columns

Index	Liho Liho	Chambers	The Square	8/6/2020
0	1	0	0	\$234.54
1	0	1	0	\$45.74
2	0	0	1	\$56.22
3	1	0	0	\$32.31

Ventajas y Desventajas Getdummies()

Ventajas	Desventajas
Es un método sencillo para convertir categorías a números	Agrega muchas columnas binarias si tienes muchas categorías en una variable
No requiere que las categorías sean mayores o iguales a cero como One Hot Encoding	Necesita aplicarse individualmente a cada columna o variable categórica
Fue la primera versión de estandarización de variables categóricas	Requiere de mayor tiempo de ejecución