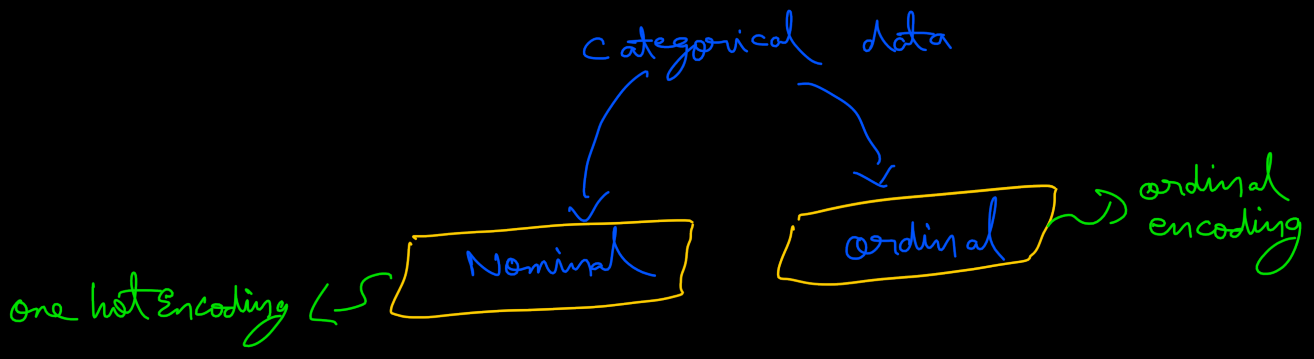
One Hot Encoding (Handling categorical data)

we always need to convert categorical data into numerical data because Most of the machine learning algorithms do no accept String. And in real word categorical data are mostly present in the forms of String. So as a ML Engineer its our duty to Convert these String into numbers.



One Hot Encoding

Color	Target
yellow	0
yellow	1
Blue	1
yellow	1
Red	1
yellow	0
Red	1
Red	0
yellow	1

one HoT Encoding

Color_Y	Color_B	Color_R	Target
1	0	0	0
1	0	0	1
0	1	0	1
1	0	0	1
0	0	1	1
1	0	0	1
0	0	1	1

what we did here?
↳ we Converted a String into a vector. like

[1, 0, 0, 0] → yellow
[0, 1, 0, 0] → Blue
[0, 0, 1, 0] → Red

Dummy variable Trap

Multicollinearity : when independent variables are dependent upon each other and has some mathematical relationship then we say it is a condition of Multicollinearity.

If we have multicollinearity then linear algorithms like linear Regression and logistic regression will not perform well.

If we focus above, after one hot encoding all three columns Color_Y, Color_B and Color_R together has a mathematical relationship that is sum of all three columns equal to 1 every time. IT means it is case of multicollinearity. for this we remove any one column. And we represent 3 columns with the help of (3-1) = 2 columns.

Color_Y, Color_B and Color_R all these columns are dummy columns and because of these problem of multicollinearity happens thats why it is called as Dummy variable Trap.

One Hot Encoding using most frequent variables

Here what we do is that when we have too many categories present in any categorical feature, on that time we select only most frequent variables on basis of domain knowledge and keep other rest categories in one category (may be other).

for example follow Jupyter Notebook