Encoding categorical Data

categorical data

Ordinal

There is a relationship between the categories present in your data.

For Ex UG, PG, HS

PS > UG > HS

Excellent > Good > Bad

We have an order.

Nominal

categories doesn't have any relation between them

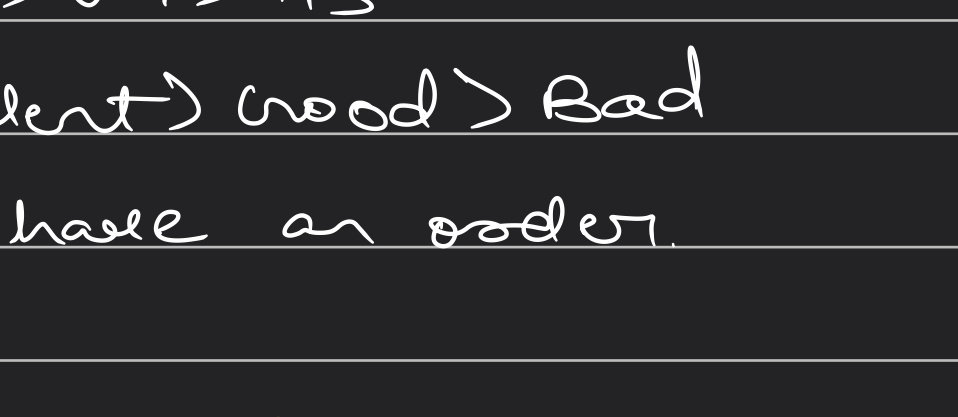
For Ex - State, Engineering Branches.

we use One Hot Encoding Here

Why Encoding

Now our categorical data is present in the form of strings & ML Algo require numerical values.

Ordinal Encoding



Whenever we have input columns, collectively called X, if there is some ordinal data present there, we use ordinal encoder there.

But if our Y (output Variable) is categorical in nature. Then we don't use ordinal encoder, we use something called "Label Encoder"

So Input - X ⇒ ordinal Encoder

Output - Y ⇒ Label Encoder

Ex Education		But we know
HighSchool	0	PG > UG > HS
PG	2	
UG	1	PG → 2
UG	1	UG → 1
PG	2	HS → 0
HighSchool	0	

+ ordinal

Done with Encoding

Nominal Encoding

is a categorical encoding where don't have order between our categorical data.

Ex Colors, Brands of Car they all have as equal importance

→ Nominal Encoding

Color	Target	Color_Y	Color_B	Color_R	Target
Yellow	0	1	0	0	0
Blue	1	0	1	0	1
Red	1	0	0	1	1

Yellow	0	1	0	0
Blue	1	0	1	0
Red	0	0	0	1

Now our string of color in categorical data got converted into vectors.

Yellow = [1, 0, 0]

Blue = [0, 1, 0]

Red = [0, 0, 1]

- Dummy Variable Trap

Here we remove one column or variable from our Input variables.

Say we N columns as Input variable but we shall only use N-1 columns.

But why do we remove this?

This is due to Multicolinearity

Input columns may have some mathematical relationship between.

If this exist then those input variable are not suitable for ML.

Because in ML we require Independent Variable as input & only our Target Variable should be dependant on "Input".

So to break this multicolinearity we

remove one column.

For Ex we remove Color_Y

?? what would be color of yellow row

→ previously.

Yellow = [1, 0, 0]

Blue = [0, 1, 0]

Red = [0, 0, 1]

Now after removing Color_Y

Yellow = [0, 0]

Blue = [1, 0]

Red = [0, 1]

This removing of column helped us reduce dimension.

- One Hot Encoding using most frequent variables

Let say we have list of Brands

say count = 40.

We need to encode them there are few brands that comes most frequent.

So what we can do is put the least frequent brands in their own catorgy called "others"

This way we reduce the dimensionality