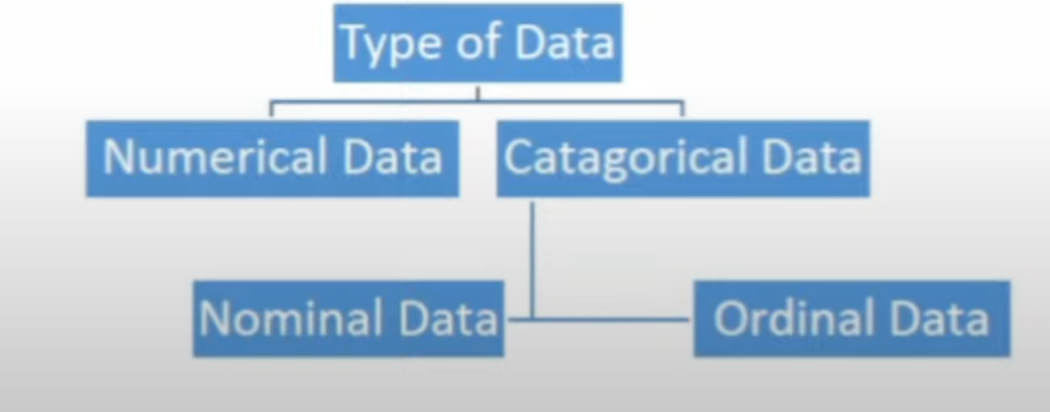
ML-9

Handling Categorical Data

Categorical data of of two types:

- Ordinal Data
- Nominal Data



▼ Nominal Data

it is the data that consists of categories which cant be ordered or ranked. also called nominal scale.

It cant be ranked or measured in any way. it can be both qualitative or quantitative .

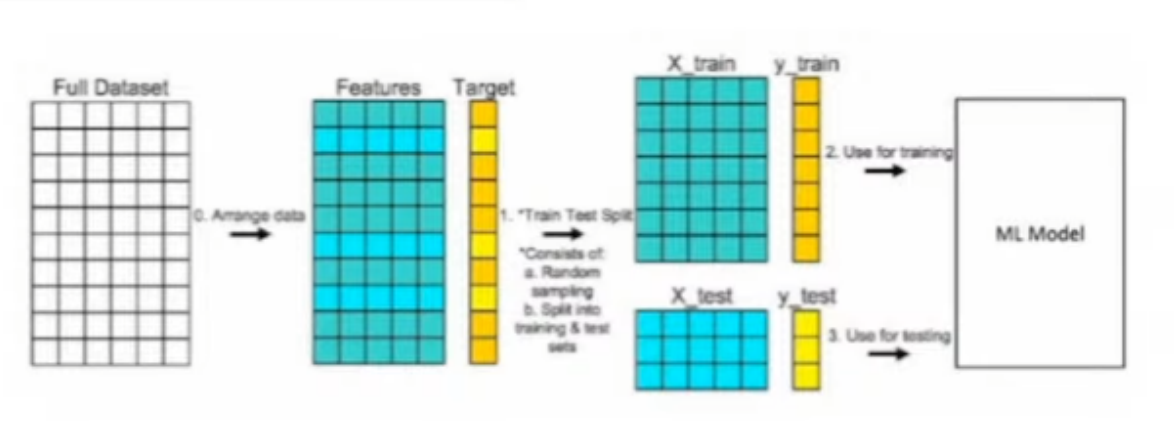
Some examples : words , letters , symbols , sex , state of our engineering branch.

▼ Ordinal Data

it has a natura order . often used in surveys , questionaries , and at the fields of finance and economics .

It stands out because it is next to impossible to differentiate between the two data values.

for example: cloth size , exam grade or division etc.



Q) what is encoding?

→ data encoding involves converting a sequence of text characters into binary code , so computers , which operate using binary numbers can process, store ,or transmit that textual information. Decoding occurs when that information is then translated from binary form into a readable version

▼ Ordinal Encoding

"Ordinal Encoding"

When we have a feature where variables have some order/rank.

Original Encoding	Ordinal Encoding
Poor	1
Good	2
Very Good	3
Excellent	4

▼ **Nominal Encoding**

Color
Red
Red
Yellow
Green
Yellow

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

So, in simpler terms:

- Use **One-Hot Encoding** when you have many options and no specific order matters.
- Use **Ordinal Encoding** when you have only a few options and there's a clear ranking or order between them.

Multicollinearity in Nominal Encoding

Color	Target
Yellow	0
Yellow	1
Blue	1
Yellow	1
Red	1
Yellow	0
Red	1
Red	0
Yellow	1
Blue	0

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	0
0	0	1	1
0	0	1	0
1	0	0	1
0	1	0	0

usually, if we have n categories in the column. we make (n-1) columns. if we will not do this there comes an error of multicollinearity .

These columns are known as **dummy variables**. So, this is called **Dummy variable trap**.

Label Encoding

used for dependent variables . Therefor e, it can be only used in classification.