Data Encoding

└→ Convert the string into numerical data.

① Nominal | OHE
② Label and Ordinal
③ Target guided ordinal Encoding.

Machine learning
⇓
learning patterns from the data
⇓
mathematical function.
⇓ =
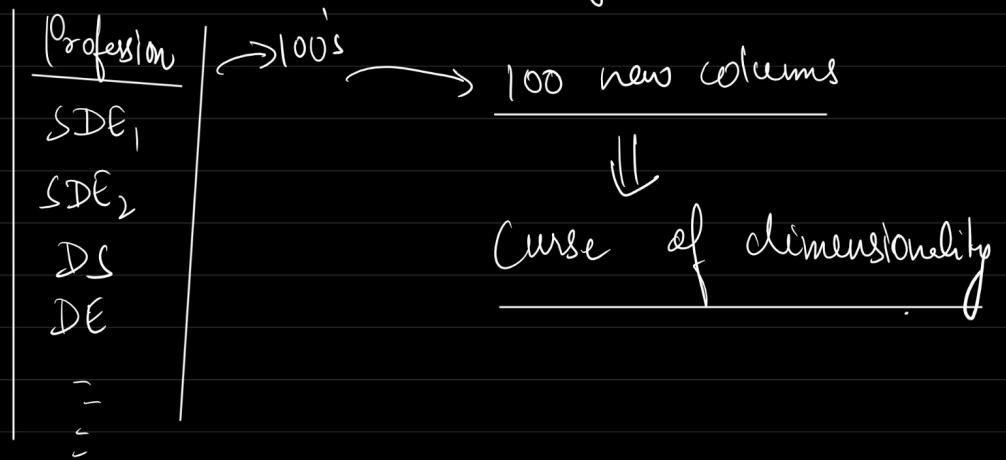numerical data

* Nominal / OHE

→ Categorical data to Numerical data
→ No order in the data

Status
Single
married
Separated
Single
married
—
—

ML Algorithms.

Dummy Variables

	Single	married	Separated
Single	1	0	0
married	0	1	0
Separated	0	0	1

Colour
red
green
yellow
—
—
—

	red	green	yellow
red	1	0	0
green	0	1	0
Yellow	0	0	1

* disadvantage → A column has many categories

Profession	→100's → 100 new columns
SDE$_1$	
SDE$_2$	
DS	
DE	

$\Downarrow$

Curse of dimensionality

* n distinct values in a column ⟶ n no of dummy variables will be created.

*

	S	M	Sep
S	1	0	0
M	0	1	0
Sep	0	0	1

n ⟶ $\underline{n-1}$

$n-1$ dummy column will explain n – original column.

* Label and ordinal encoding

→ Label encoding → assigns numerical data to each category

Red – 1
Green – 2

Yellow – 3

DS – 1
DE – 2
DA – 3

* No curse of dimensionality

* __disadvantage__ → Good for ordinal data. as for nominal data it will learn the pattern

→ <u>Ordinal Encoding</u>

HighScool — 1
Colled — 2
PG — 3
PHD — 4

* <u>Target Guided Ordinal Encoding</u>

↳ based on their relationship with the target variable

→ Useful when we have large number of Unique Categories in Categorical value.

↳ Categorical groups with mean/ median of corresponding target variable.

→150

time	total bill
lunch	8 → 180
dinner	12 → 120
breakfast	3
breakf	—
lunch	— →120
	— → 150

groupby('time')['total bill']

avg/median,

lunch - 150
dinner - 180
breakfast - 120