* Understanding of Problem statement
* Data Ingestion.
* Data Preparation | Preprocessing & Exploratory data Analysis
  - Missing Value treatment
  - Class imbalance
  - Outlier treatment
  - Data Encoding
  - Feature Engineering / Feature extraction
    - Creating new features
    - Modify the existing
    - Selecting the right feature

* Feature extraction — process of selecting and extracting the relevant features from the raw data.

1000 features ⟶ Most important features ↓ Used to train the model.

* Curse of dimensionality
  - ↳ with increase in no. of feature :-
    1. Model training becomes computationally expensive
    2. Model interpretation becomes complex.

Area of house	Price of a house
1800	60
1900	70
—	—
—	—

f_1 f_2 f_3 f_4 f_5 - - - f_{1000}

① Creating new feature

Ex-1

Distance	time	Speed = Distance/time
50	2	50/2
100	3.5	100/3.5
150	5.8	—
110	2.1	—
—	—	—

Ex-2

temp	3 month moving avg temp
21°	Nan
22°	Nan
23°	$\frac{21+22+23}{3} = 22°$
18°	$\frac{22+23+18}{3}$
20	$\frac{23+18+20}{3}$
21	
22	
18	

Ex: Ratio

$$\text{No of room} = \frac{\text{effarea}}{\text{Area of 1 room}}$$

new feature ↓

No of Room	Area of 1 room	effective area
2	100	200
3	200	600

② Modifying the existing feature

ⓐ Changing the datatype

Ex-1

Age	Modified Age
15 yrs	15
20 yrs	20
25 yrs	25
—	—
—	—

Ex-2

object

install	install_modified
15+	15
20+	20
25+	25

eg :

Date	Day	Month	Year
15-2-2022	15	2	2022
16-2-2022	—	—	
+ -	—	—	-
- -			
=	—	.	—
=			
	—		

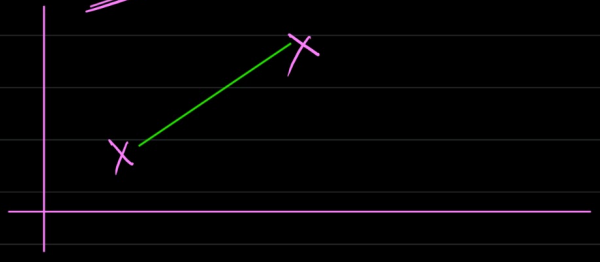
⑥ feature scaling (optional)

→ feature scaling tries to bring all the features on the same scale

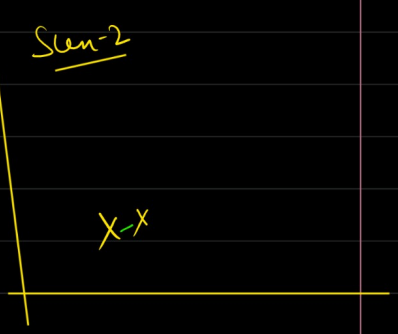
Area of house	# of rooms	Parking area	Y (price of the house)
1800	2	100	60
2500	3	50	65
3000	4	25	100
6000	1	30	-
-	5	30.5	-
=	=	=	=
=	=	=	=

Why?

Scen-1



Scen-2



* Many of Algo's are distance based, that's why if high magnitude it becomes computationally expensive.
* optimisation becomes faster.
* Interpretation becomes easier.

Types of feature scaling

① Standardisation (ML algo)

$$Z\text{ score} = \frac{x_i - \bar{x}}{\sigma}$$

SND $\Rightarrow \mu = 0, \sigma = 1$

Age	Age_Scaled
25	$\frac{25-24.6}{1.03}$
26	$\frac{26-24.6}{1.03}$
23	—
24	—
25	—

$\bar{x} = 24.6$

$\sigma = 1.03$

② Normalization [min max Scaler] $\Rightarrow [0,1]$ → DL Algorithms → PCA

Age	Age Scaled
25	$\frac{25-23}{26-23} = \frac{2}{3}$
26	$\frac{26-23}{26-23} = 1$
23	$\frac{23-23}{26-23} = 0$
24	
25	

$$x_{scaled} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

$[0,1]$

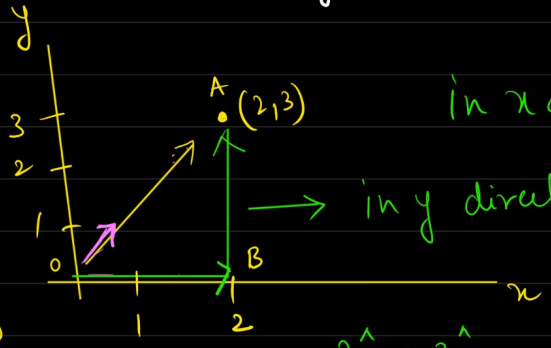
min = 23

max = 26

③ Unit vector

(2,3)

x y



in x direction = 2 units

in y direction = 3 unit

$= 2\hat{x} + 3\hat{y}$

Pythagoras theorem $\Rightarrow$

$OA^2 = OB^2 + AB^2$

$OA = \sqrt{2^2 + 3^2}$

$\|\vec{OA}\| = \sqrt{4+9} = \sqrt{13}$

$\hat{u} = \left(\frac{2}{\|\vec{OA}\|}, \frac{3}{\|\vec{OA}\|}\right)$

$\hat{u} = \left(\frac{2}{\sqrt{13}}, \frac{3}{\sqrt{13}}\right) \Rightarrow \sqrt{\left(\frac{2}{\sqrt{13}}\right)^2 + \left(\frac{3}{\sqrt{13}}\right)^2}$

$= \sqrt{\frac{4}{13} + \frac{9}{13}} = \sqrt{1} = 1$

③ Selecting the right feature (feature selection)

1000 features → Top 10 features

↓

training of model

① Filter method :

↓

Calculate Correlation with target variable.

Area of house (↓)	Area of parking (↓)	Price of house (y)

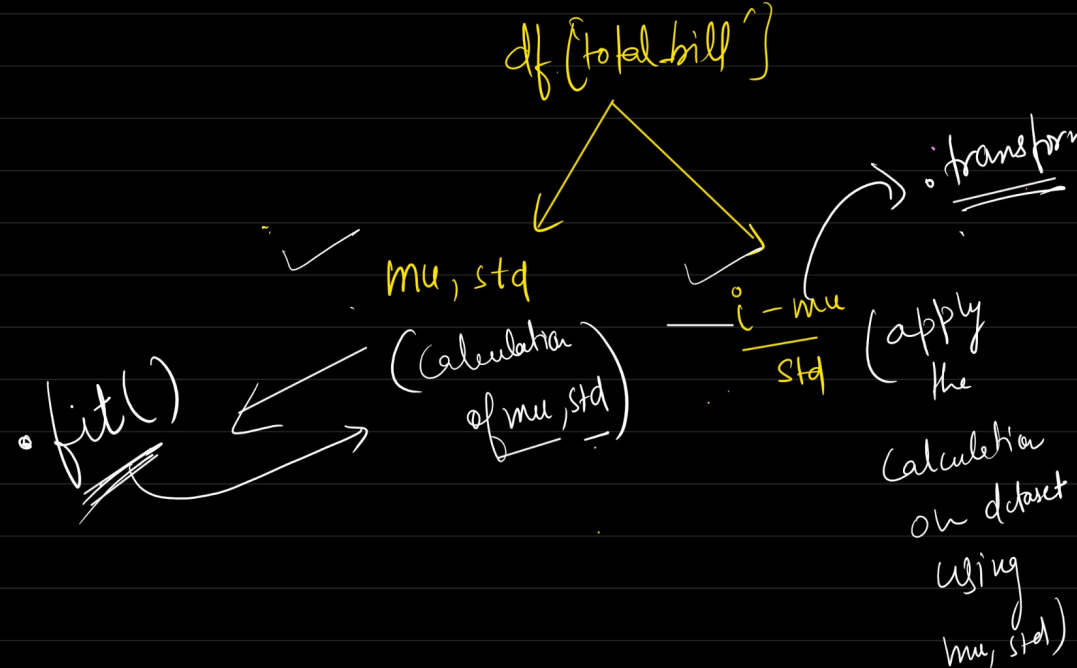
If this is more ← Corr (Area of house, Price of house)

Corr (Area of Parking, Price of house)

② Embedded method

③ Wrapper method.

* PCA



Scaling → fit (mu, sigma) data
→ transform (apply)

train → [fit + transform (mu, sigma) ↓ apply]

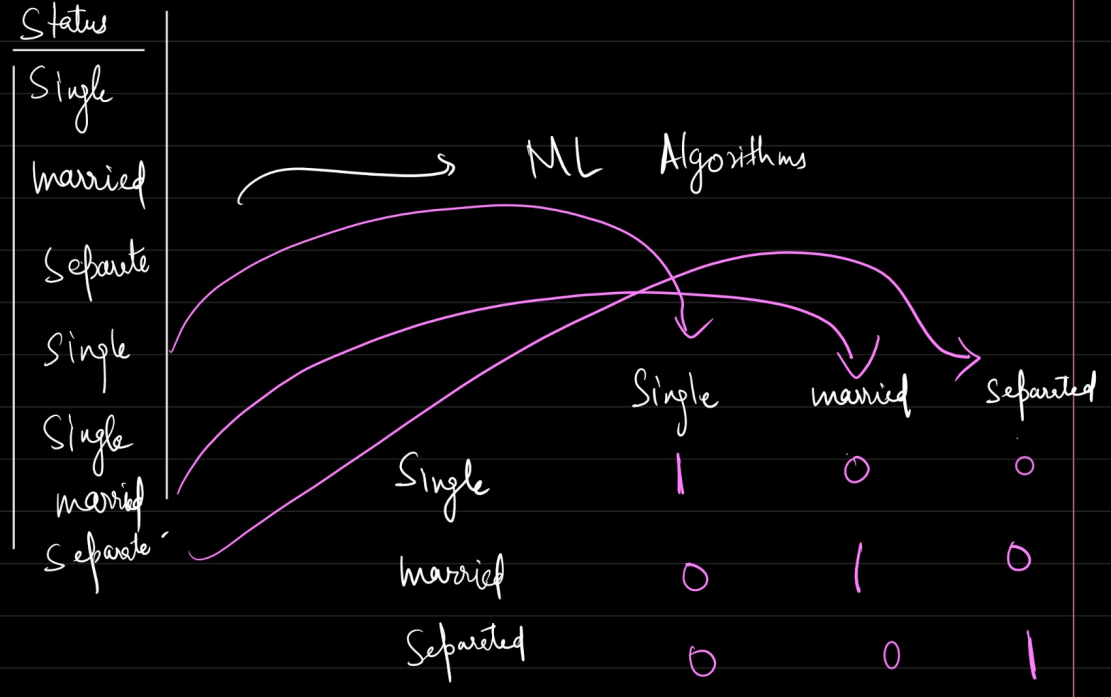
You don't know this data → representation ← test of unseen data → [transform]

Data Encoding

1. Nominal | OHE
2. Label and Ordinal
3. Target Guided ordinal encoding.

ML Algorithms ↓ numerical data

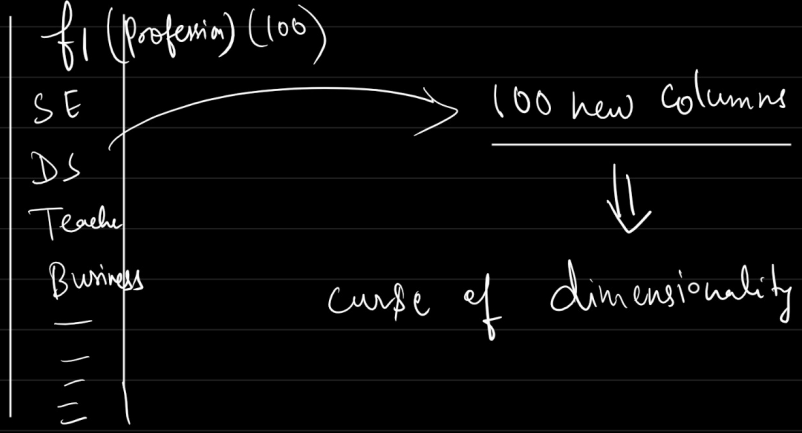
1. Nominal | OHE
   → Categorical data to Numerical data.
   → No order in the data



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

* disadvantage

→ A column has many categories



② Label and Ordinal encoding techniques

* Label encoding → assign numerical label to each category.

Red — 1
Green — 2
Yellow - 3

DS — 1
DE — 2
DA — 3
BA — 4

1
2
3
4
⋮
⋮
2

* No problem of curse of dimensionality

disadvantage - It will learn the pattern

* Ordinal encoding

High School — 1
College - 2
PostGraduate - 3

PHd - 4

③ target guided ordinal Encoding

- based on their relationship with the target variable
- Useful when we have large no of Unique Categories in Categorical value.
- Categorical group with mean/median of corresponding target variable.

time	total bill
lunch	150
dinner	180
breakfast	120.5
lunch	150
dinner	180

groupby('time').
mean