31-08-2025

Agenda:

→ ML Algo highlevel overview

→ FE - III

ML Algo highlevel Overview:—

System that can learn & improve from data without explicit programming

F_1	F_2	T
weight (kg)	Height (metres)	Value - x
70	1.75	22.86
85	1.7	29.41
50	1.8	15.43

(w/h^2) →

↓

2 variable

A	B	C	D	E	F	value - x
-	-	-	-	-	-	100
-	-	-	-	-	-	102.3

ML →

A D
B E
C F
} using ML → value_x

desired that formula → pattern / relation

which will provide us

value_x given [A .... F]

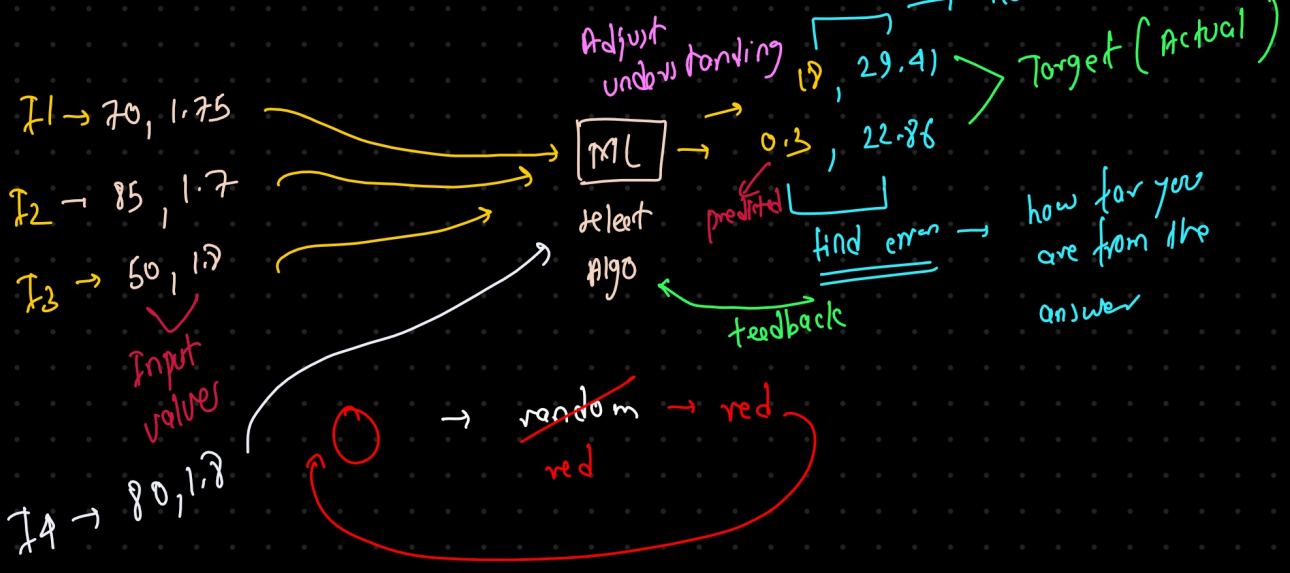
chatgpt ← ML
↓
DL
↓
Transformer based Models

Data
wiki, stackflow, x, 7, 2

Bill gates → number → chatgpt → predict

learned all the pattern
↓
very very complex formula

ML



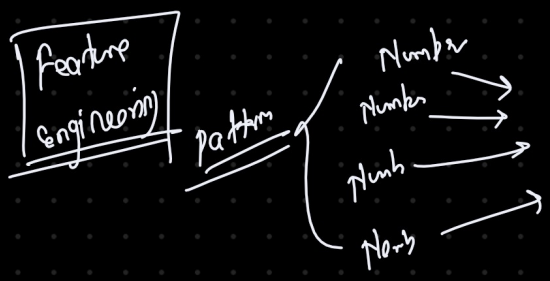
"Dog" * "Cat" → "Animal"

Human * AI * Engineer → AI Engineer

def (v1 , v2 , v3) :

=

calc & patter understand → number



Rice(kg) weight(kg) BMI

Null → | ML

Feature scaling :—

KNN → distance-based algo

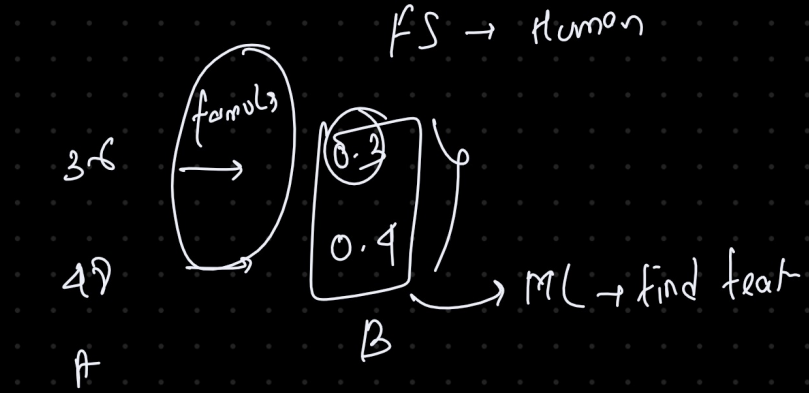
Height (metres)	weights (kg)
0.0 → 7.0	0 → 170

formula : a, b

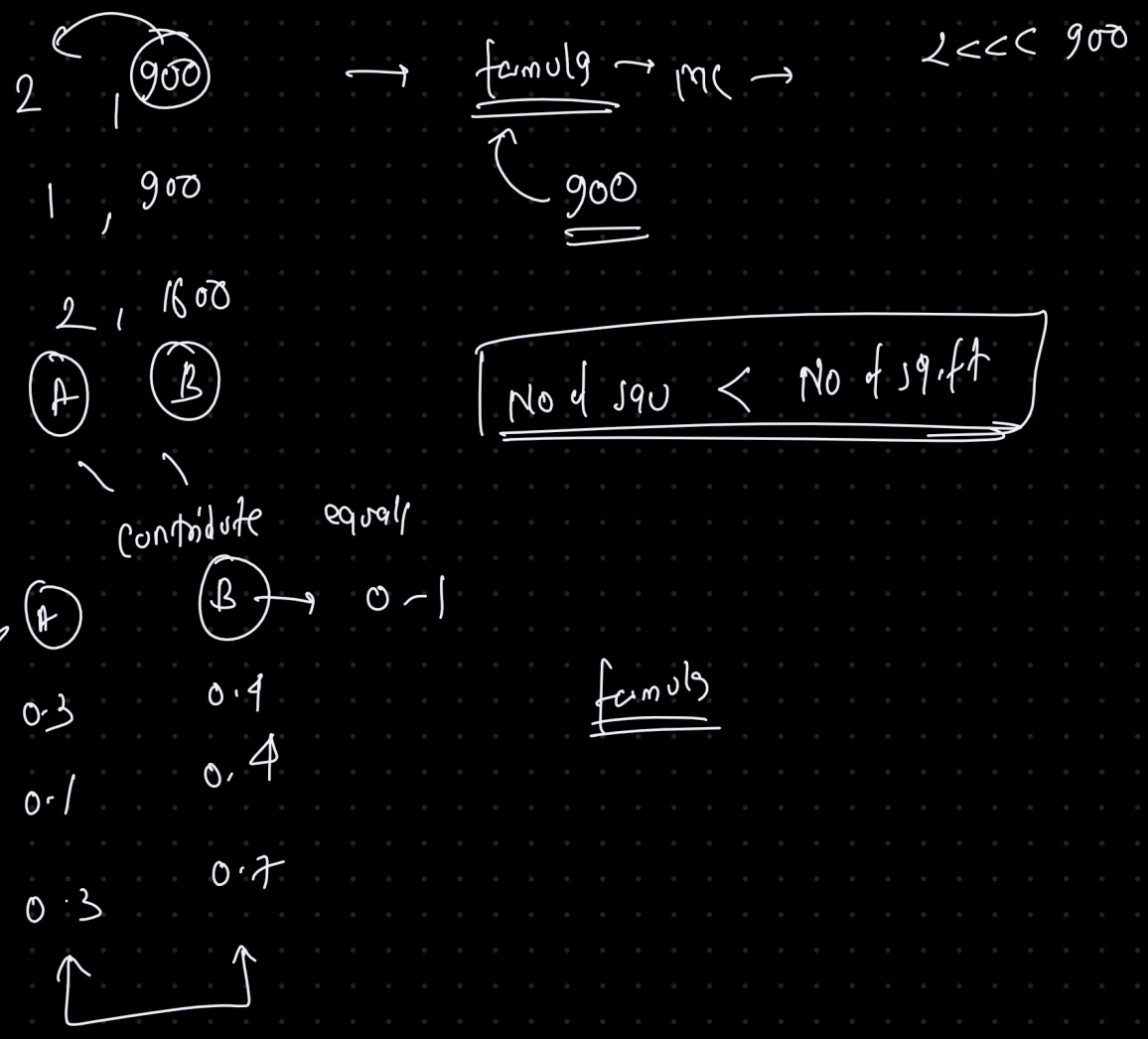
$$\begin{bmatrix} a & 7.0 \\ b & 70 \end{bmatrix}$$ (ML)

7.0 → 70
6.5 → 65
5.5 → 55
↓
10*

a → 0.7
b → 0.3



House (sq.ft)	No. of Bedrooms	Price
900	2	300k $
900	1	250k $
1600	2	500k $



	Age	Income
A	20	100,000
B	25	100,000
C	20	150,000

$$\text{distance} = \sqrt{(age_1 - age_2)^2 + (income_1 - income_2)^2}$$

distance (A, B) = 5 → close

distance (A, C) = 50,000 → far far

5 ≪ 50,000

	Age	Income
A	0.039	0.040
		0.040
B	0.135	
C	0.039	0.066

Distance (A, B) ≈ 0.097
Distance (A, C) ≈ 0.026
} Min-max scaling

Table:

	Maths (out of 100)	Science (out of 200)
(80)	0.75	(150) 0.71
60		120
90		180
0-1		0-1

Scaling # 1 Min-Max Scaling (Normalization)

→ Scale values into [0, 1]

→ Sensitive outliers

A

70 → 0

80 → 0.2

90

120 → 1

formula =

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}}$$

$$x_{new} = \frac{70 - 70}{120 - 70} = \frac{0}{\text{something}}$$

$$1 = \frac{120-70}{120-70} = 1$$

$$= \frac{80-70}{120-70} = \frac{10}{50} = 0.2$$

(2) Standardization (Z-score scaling)

* Handles outliers better than Min-Max

$$X_{new} = \frac{X_{old} - \mu}{\sigma}$$

Mean = 0, Std Dev = 1

(3) Robust scaler (Median & IQR based)

$$X_{new} = \frac{X_{old} - Median(X)}{IQR(X)} \qquad IQR = Q3 - Q1$$

(4) log transformation

$$X_{new} = \log(1+X)$$

* right skewness

(5) Others

* square root → $X_{new} = \sqrt{X_{old}}$
* cube-root → $X_{new} = x^{1/3}$
* Yeo-Johnson

* Box-Cox (λ)

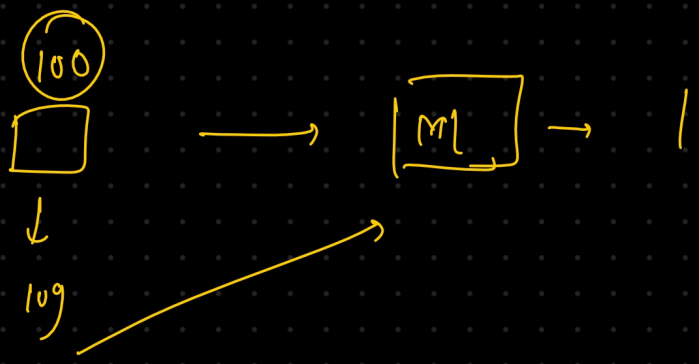
$$X_{new} = (X_{old}^{\lambda} - 1)/\lambda, \text{ if } \lambda \neq 0$$

$$\log(x) \quad \text{if } \lambda = 0$$

	No. of apples	
A	100	2.5
B	20	1.9
C	40	1.D
D	600	5f
E	589	4.2

ML

NN
CV
NLP



Raw rice count per plate chapathi

Concluding :—

- doesn't change the underlying representation
- Algo can learn patterns effectively
- Make feature more consistent across data.

Data Encoding :—

→ "Blue" , "green" , "red"

numerical data → ML

#1 label Encoding (Nominal)

"Blue" , "green" , "red"
↓ ↓ ↓
0 1 2

- Data-order dependant

#2 label Encoding (ordinal) / ordinal Enoder

"Red" , "Green" , "Yellow"
↓ ↓ ↓
0 1 2

Danger sort then in Desc :— "Red" , "Yellow" , "Green"
↓ ↓ ↓
0 1 2

Student :

Marks	Height	weigh	fav-colu	behavior
			blue	good
			pink	bad
			white	average

4 > 3 > 2 > 1 > 0

Red > Green

4 > 3

3 OHE (one-hot encoding)

A	B	C	D
0	1	2	3

A	0	0	0	1
B	0	0	1	0
C	0	1	0	0
D	1	0	0	0

category into binary dummy variable

df → color → Blue, Green, Yellow

↘ id → 1, 4

id	color	color_red	color_green	color_yellow
1	Red	0	0	1
2	Green	0	1	0
3	Yellow	1	0	0

- disadvantage : curse of dimensionality

dimension of the data

column → 1000 unique string