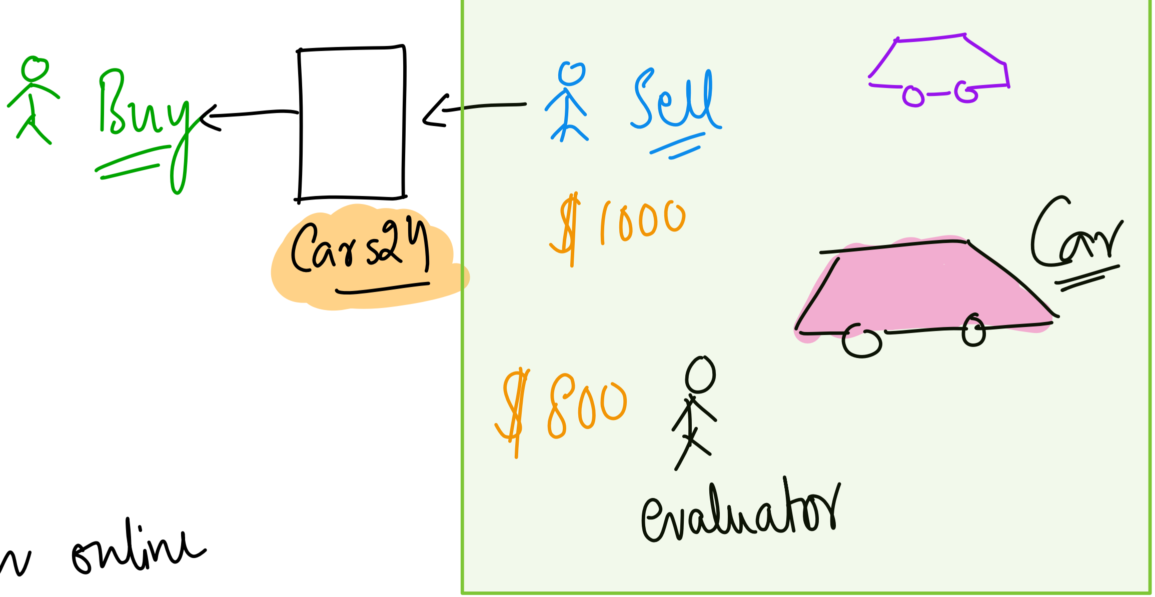
LINEAR REGRESSION - 1

Cars24

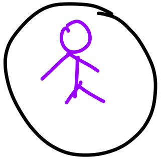
Optimising Cost



① Develop system online

Analysis ↳ data about the car

1000 requests →



Computer Vision ↳ Photographs

↳ Insurance data }

{↳ Police Record

↳ pollution record. }

{↳ Service Record.

Output → Continuous → Selling Price } Regression.

Do we know the target variable → YES } Supervised

Cat

Elephant

dog

● 0
● 1
● 0
● 2
● 0
● 1
● 0
● 2
● 0
● 1
● 2
● 0
● 2

Order

1000

0

500

1000

Label Encoding

2 Categories

Target Encoding

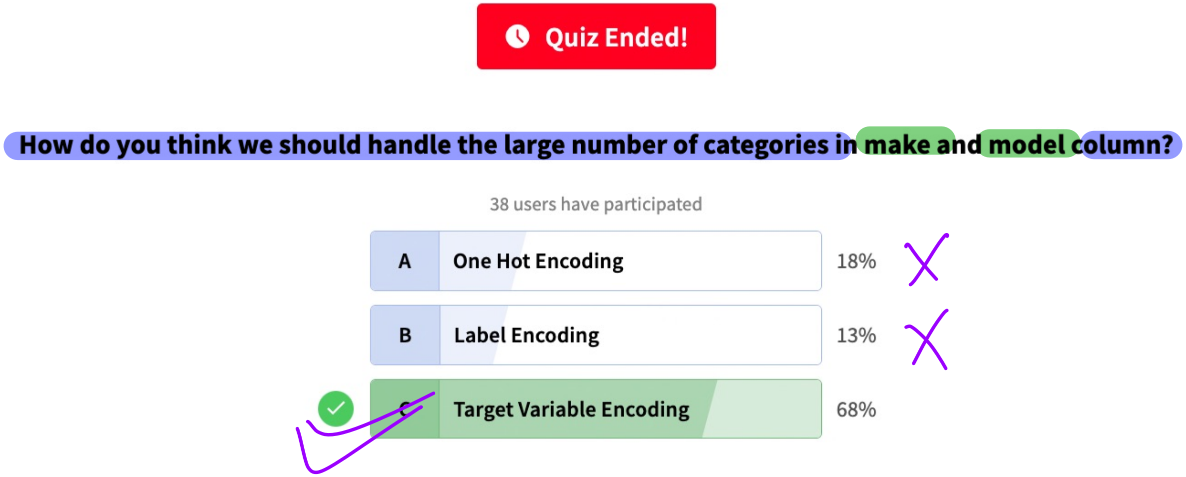
Target (HIT / FLOP)
(1) (0)

$P[T=1 \mid \text{Category}]$

0	→ Rom	H	-	0.5
1	drama	H	-	0.66
2	Thriller	H	-	1
0	Rom	F	-	0.5
1	drama	H	-	0.66
1	drama	F	-	6.66

Rom	H	$1/2 = 0.5$
Rom	F	
drama	H	$2/3 = 0.66$
drama	H	
drama	F	
Thriller	H	$1/1 = 1$

1000 classes → TE



Quiz Ended!

How do you think we should handle the large number of categories in make and model column?

38 users have participated

A	One Hot Encoding	18%	✗
B	Label Encoding	13%	✗
✓ C	Target Variable Encoding	68%	✓

1000

1000's Car → C1

→ C2

⋮

1 column

C1000

C1	C2	C3	- .	C1000
1	0	0	0 0	0
0	1	0	0	0

Input → Models → Output

high

① Scale

Standardisation

$$\frac{x-\mu}{\sigma}$$

Normalisation

$$\frac{x-\min}{\max-\min}$$

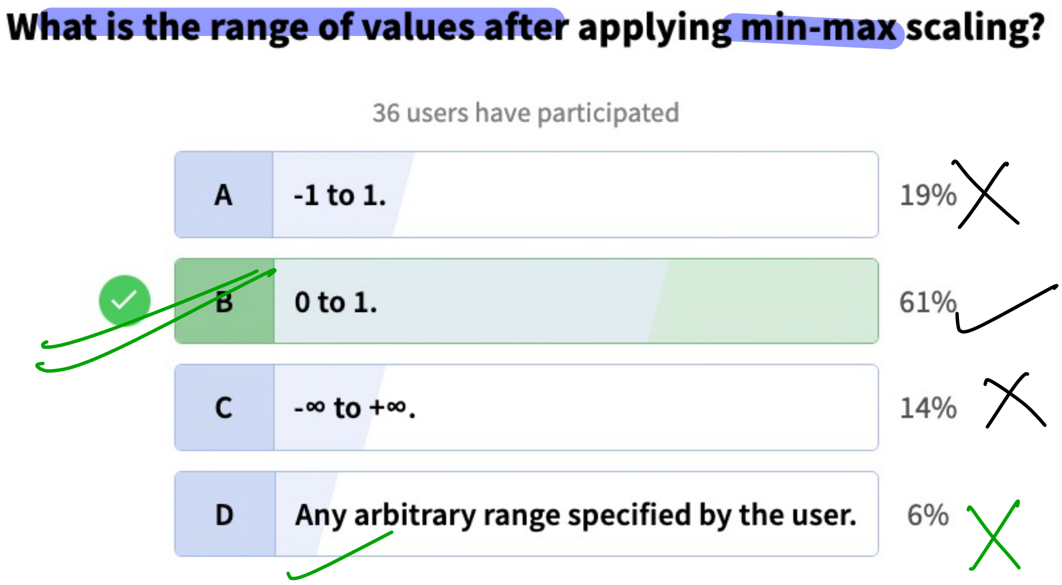
[0, 1]

99.7%

-3 —— 3

- -ve values
- $\mu = 0$

- I know Range
- +ve values



What is the range of values after applying min-max scaling?

36 users have participated

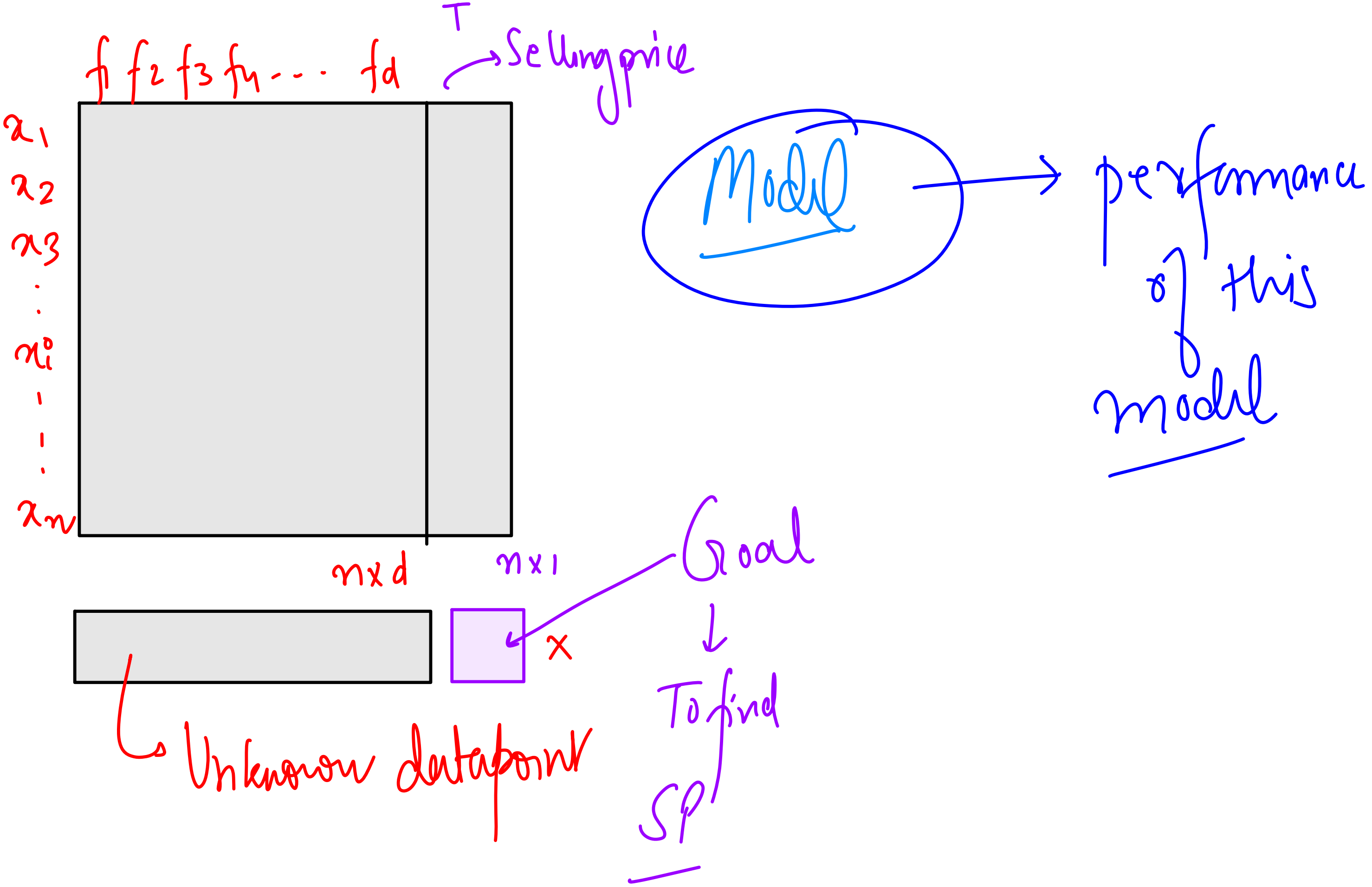
A	-1 to 1.	19%	✗
✓ B	0 to 1.	61%	✓
C	-∞ to +∞.	14%	✗
D	Any arbitrary range specified by the user.	6%	✗

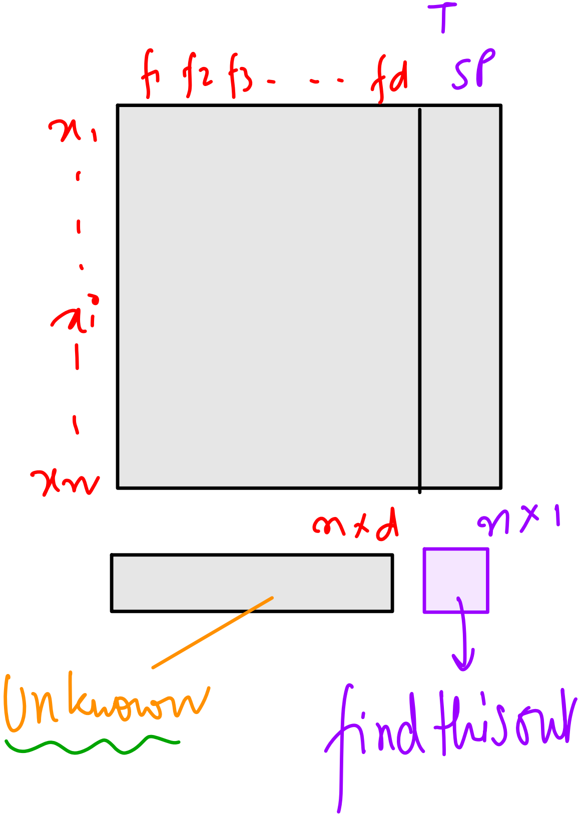
Structured data

Table

Unstructured data

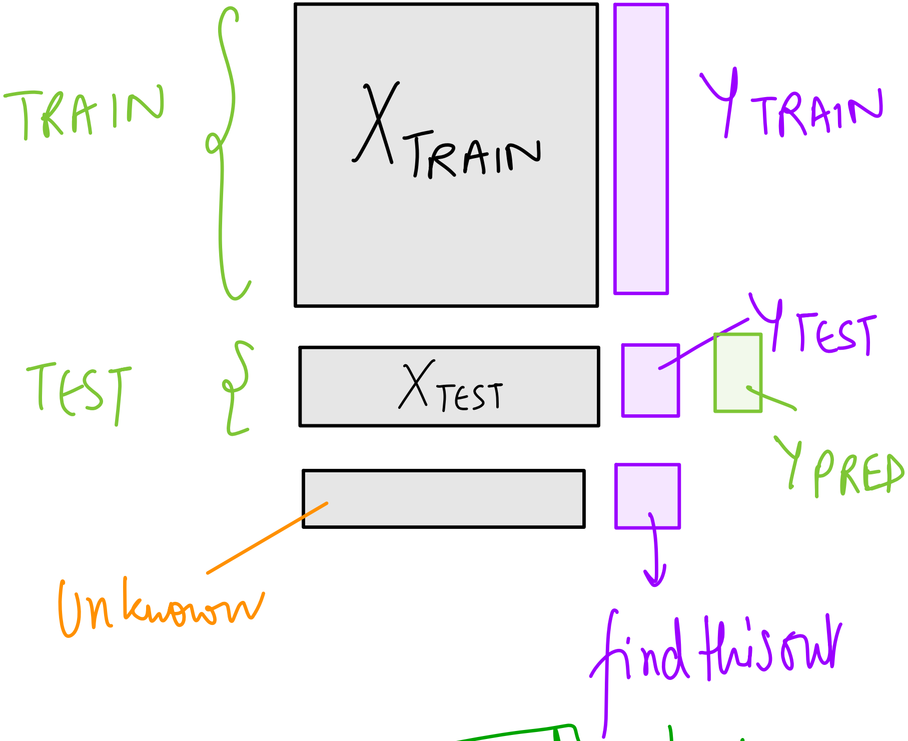
Audio, Video, Text, Images


f1 f2 f3 f4 . . . fd
T
Sellingprice
x1
x2
x3
xi
xn
n x d
n x 1
Unknown datapoint
x
Goal
To find
SP
Model
performance of this model



Dataset → Model → predict

unknown



X TRAN → Train Model → prediction

prediction → XTest → YPRED

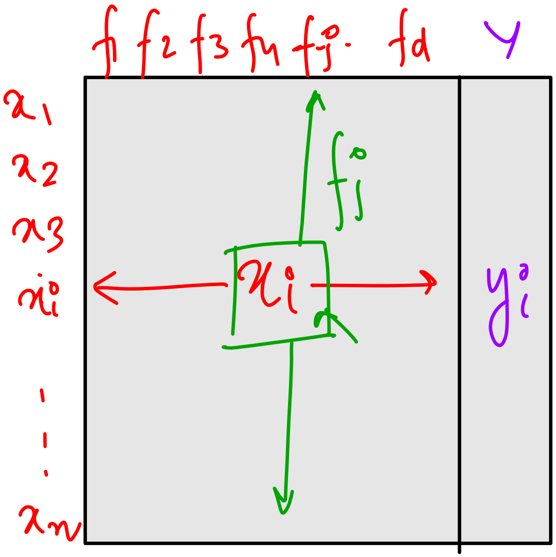
* Model performance good

⇒ YPRED ≈ YTest

* Model performing badly

⇒ YPRED ≉ YTEST

Retrain Model



n = no. of datapoints

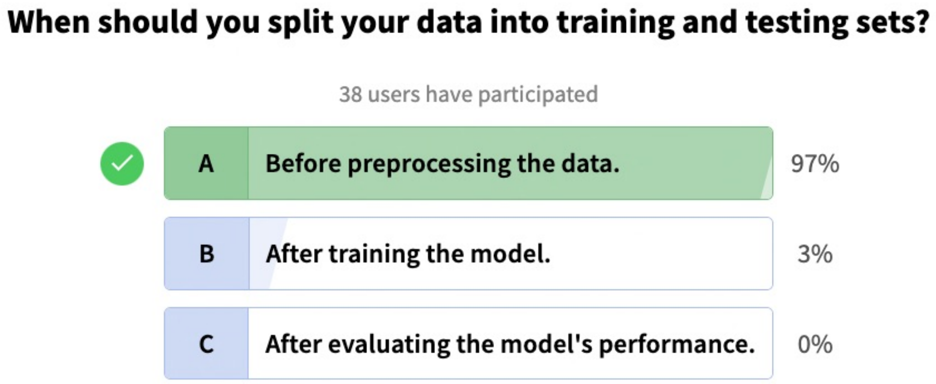
d = no. of features

True = y_i

Predicted = $\hat{y}_i$

x_{ij}

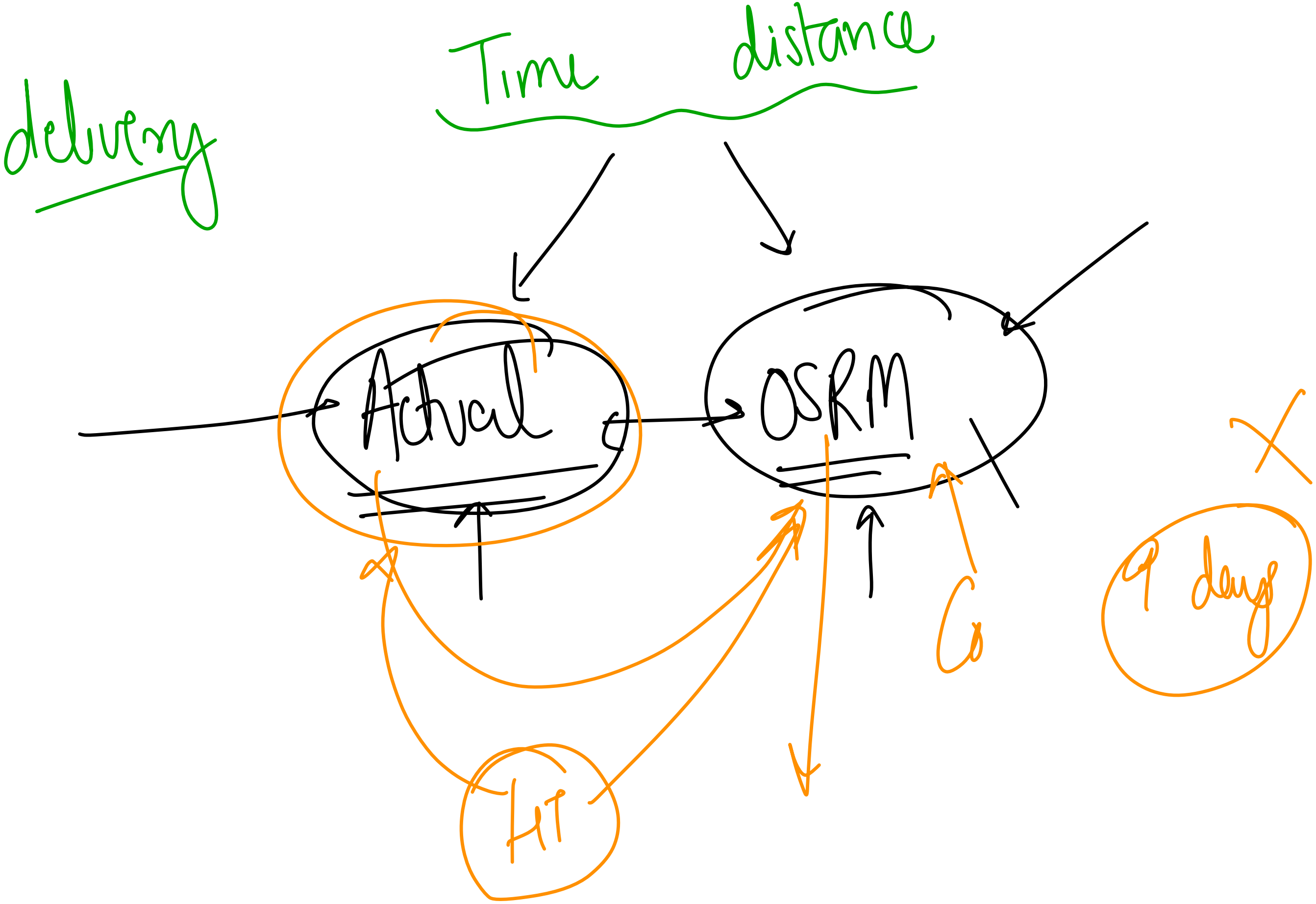
ith Sample →

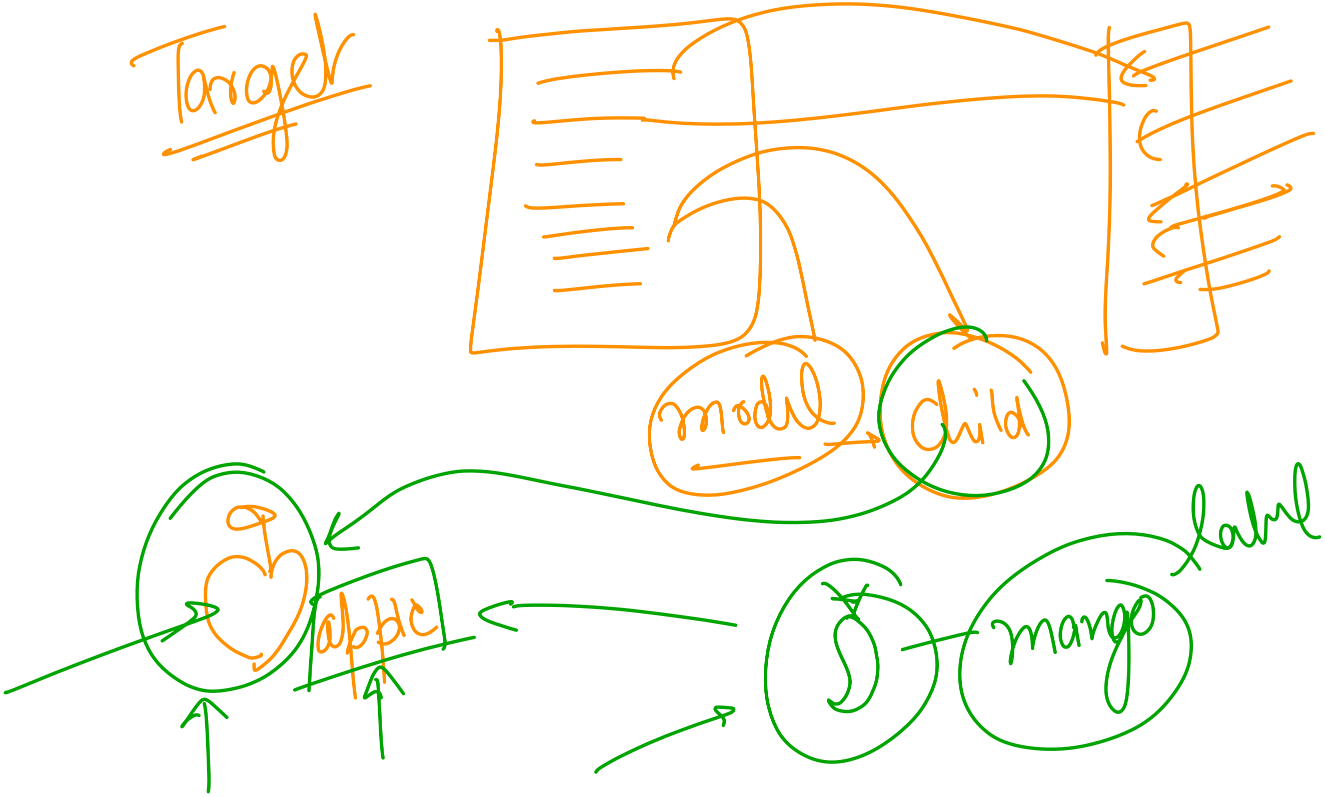


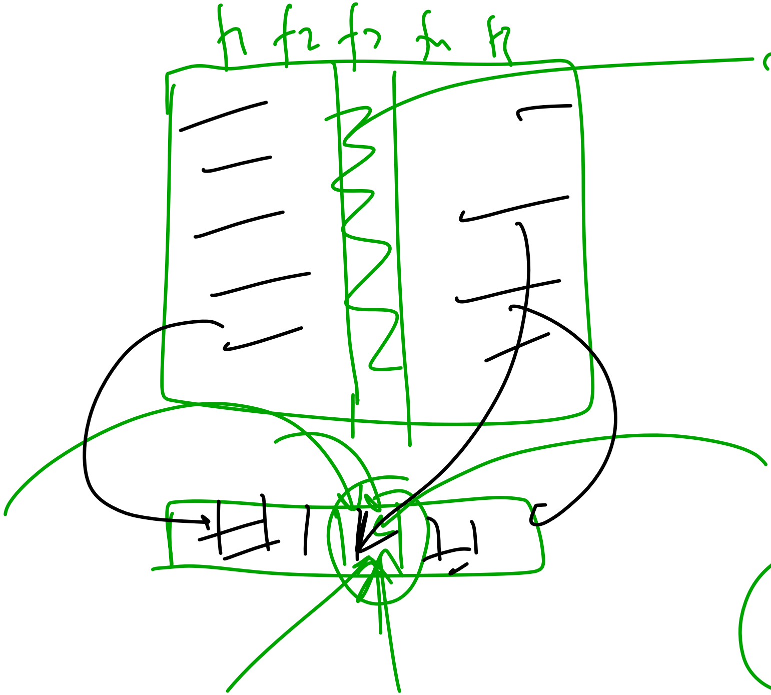
When should you split your data into training and testing sets?

38 users have participated

✓	A	Before preprocessing the data.	97%
	B	After training the model.	3%
	C	After evaluating the model's performance.	0%


delivery
Time
distance
Actual
OSRM
HT
6
9 days


Target
model
child
apple
mango
label

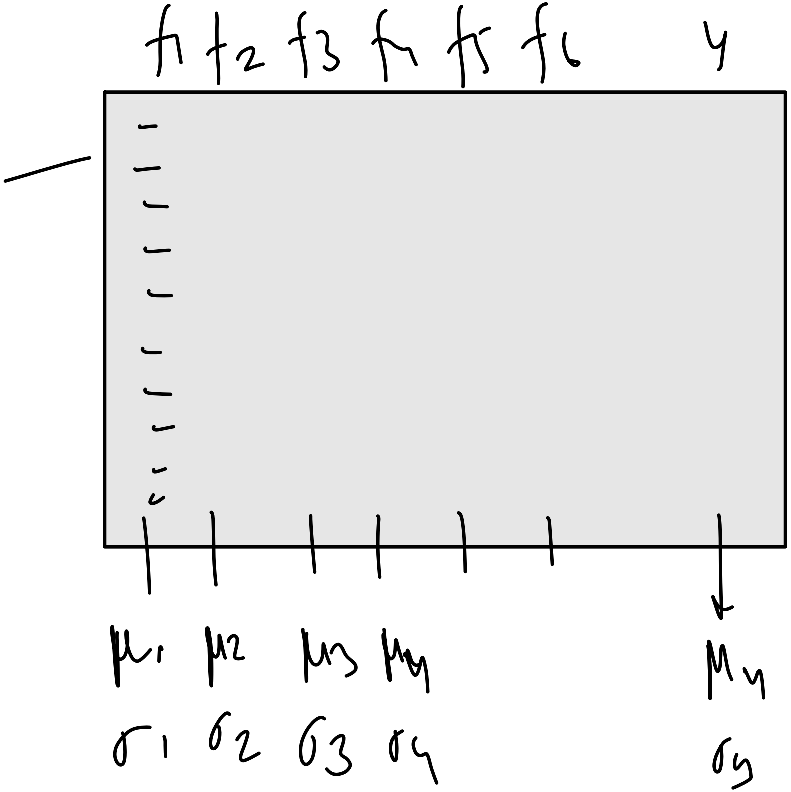


min max scaling

$w_1 f_1 + w_2 f_2 + \boxed{w_3} f_3 \dots$

$w_n f_n$

KNN


f1 f2 f3 f4 f5 f6 y
μ_1 μ_2 μ_3 μ_4 μ_y
σ_1 σ_2 σ_3 σ_4 σ_y

$$\frac{x_{i1} - \mu_1}{\sigma_1}$$