K-Nearest Neighbours

HOUSE   PRICE
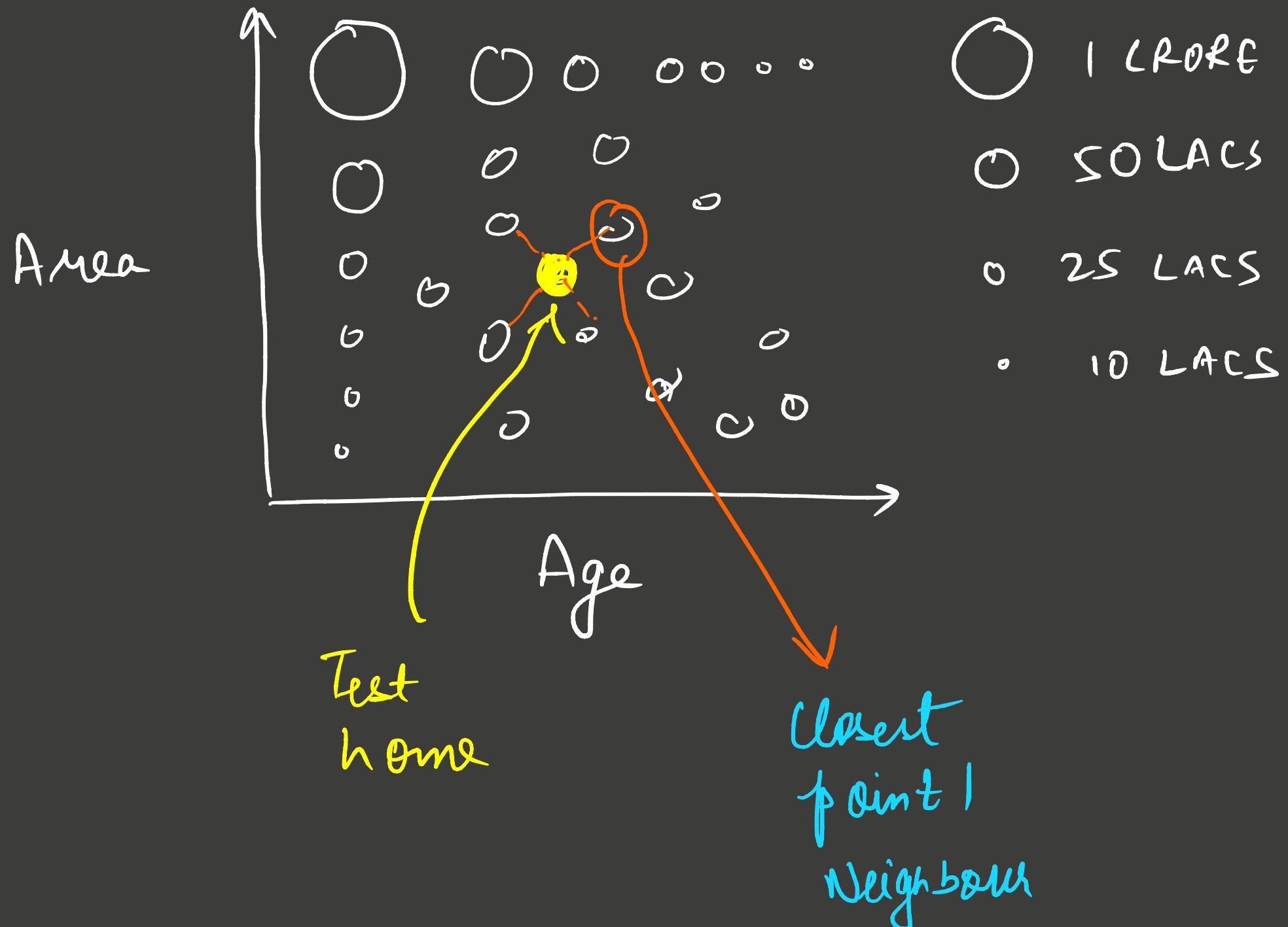
SIMILARITY IN:

* SIMILAR IN GEO-COORDINATE

* Area

* OLD

* AMENITIES IN THE HOME

↓↓ THEN

SIMILAR IN "PRICE"

1-NN or Nearest Neighbours.

Area

Age

○ — 1 CRORE

○ — 50 LACS

○ — 25 LACS

• — 10 LACS

Test home

Closest point 1 Neighbour
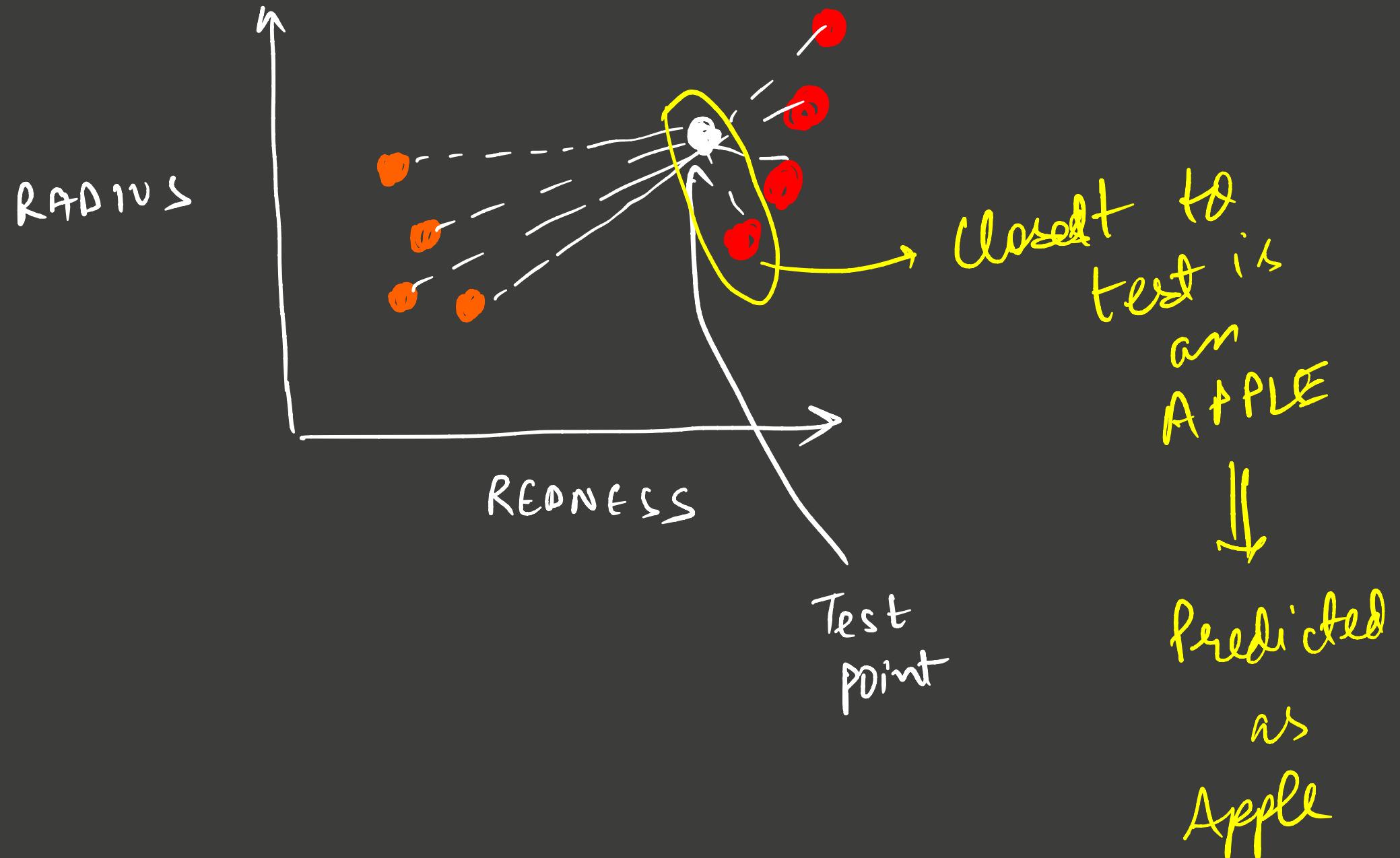
Price (Test home) = Price (Nearest Neighbour to Test home)

IMPORTANT QUESTIONS

1. What features?

2. # features

3. Distance Metric

4. COMPUTATIONAL COMPLEXITY | APPROXIMATE 'KNN

LNN FOR CLASSIFICATON

RADIUS

REDNESS

Test point

Closest to test is an APPLE

⇓

Predicted as Apple
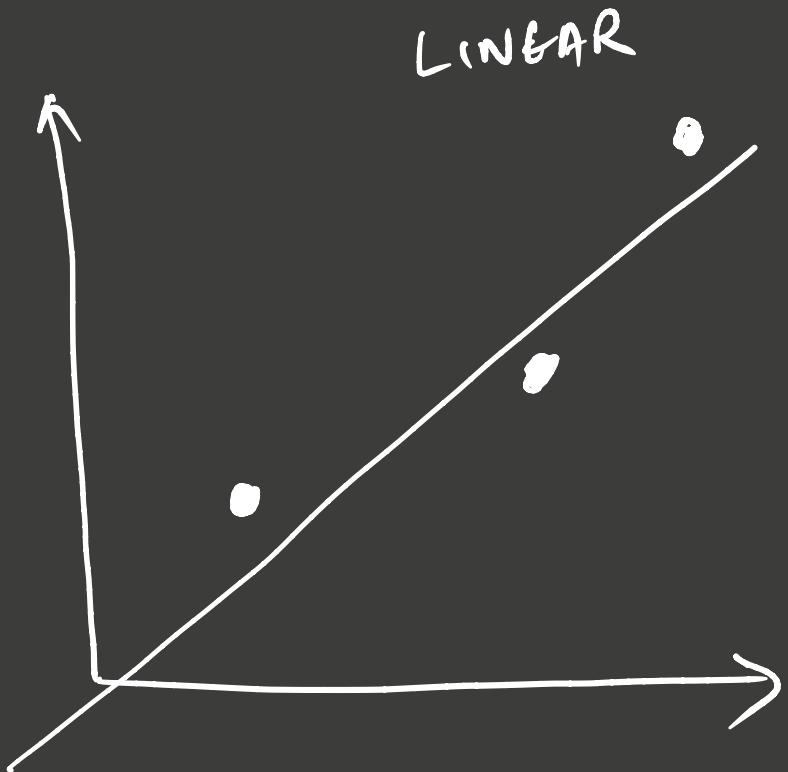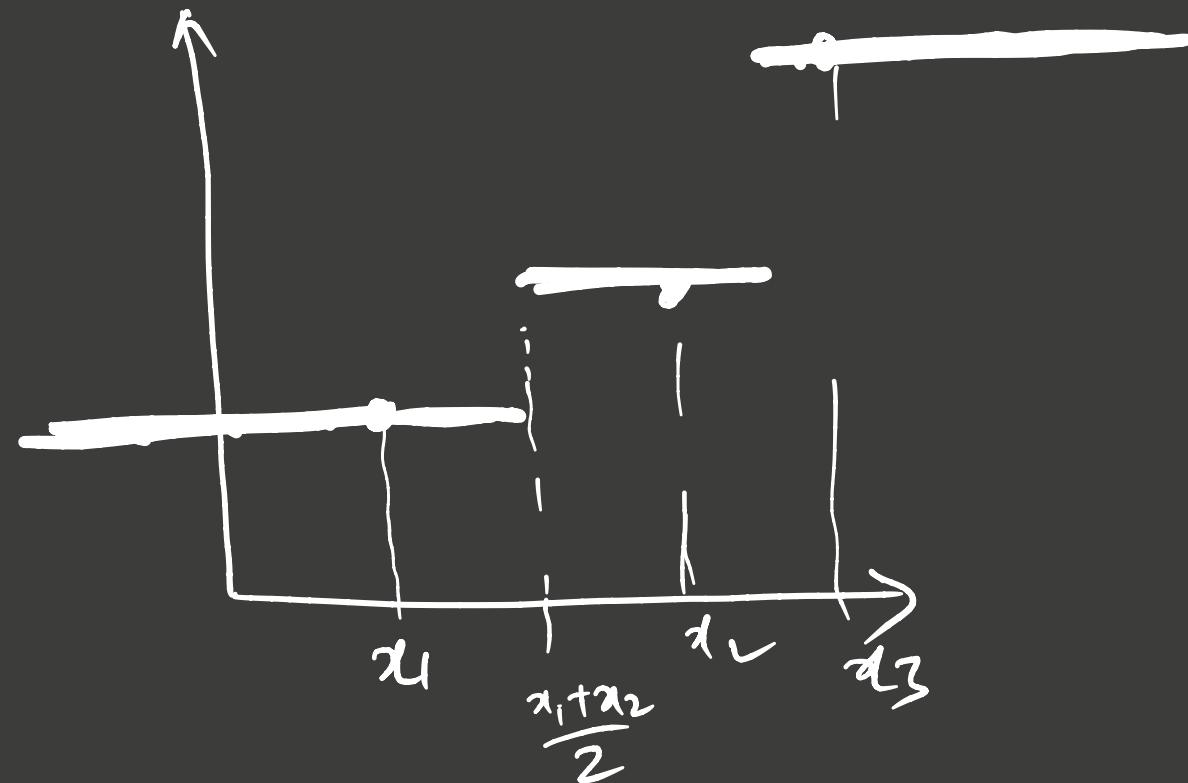
1-NN VIS LINEAR REGRESSION

LINEAR

KNN

FOR ALL TEST $x < x_1$; NN is (x_1, y_1)
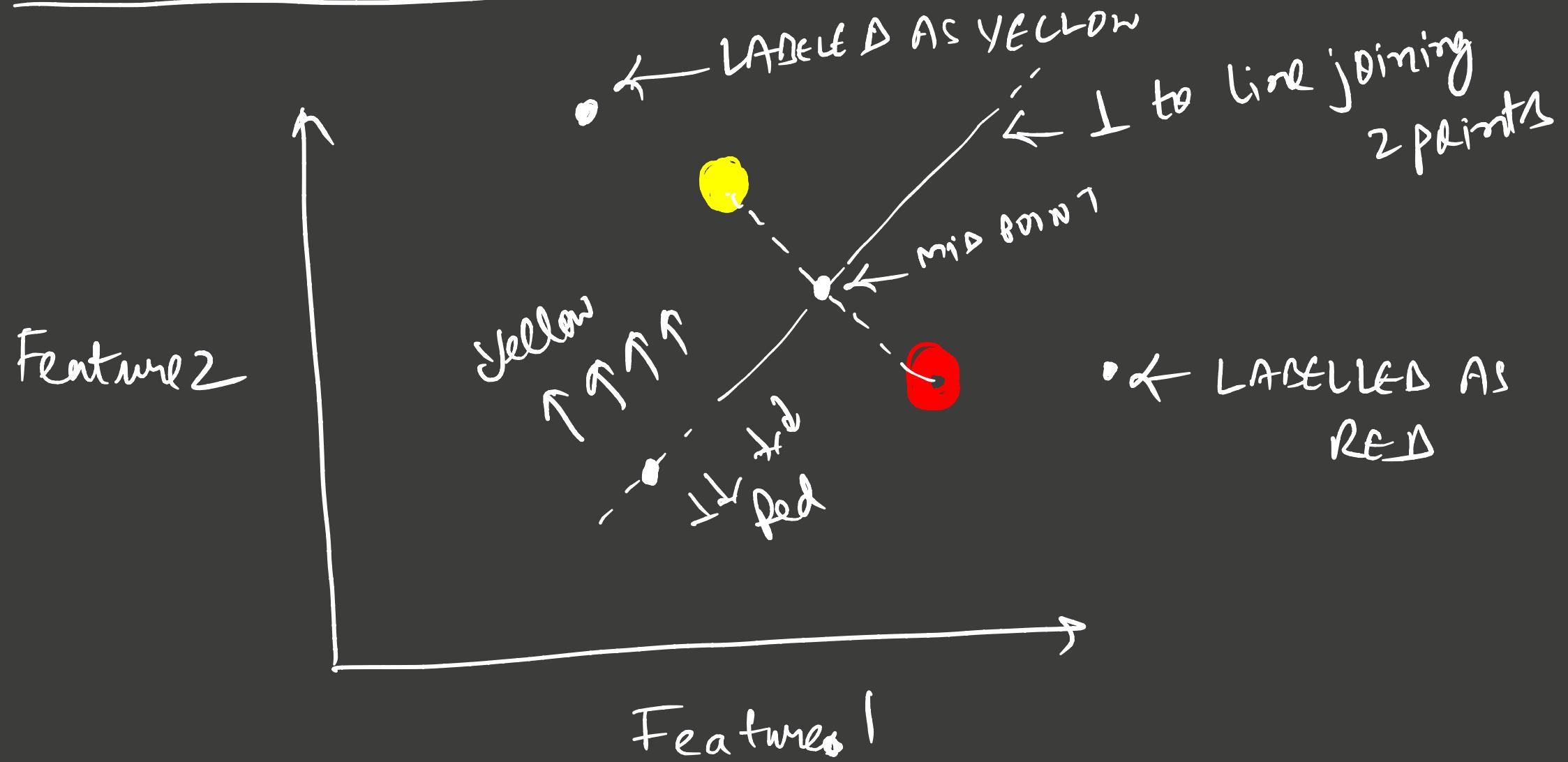
$x_1 < x < \dfrac{x_1 + x_2}{2}$; NN is (x_1, y_1)
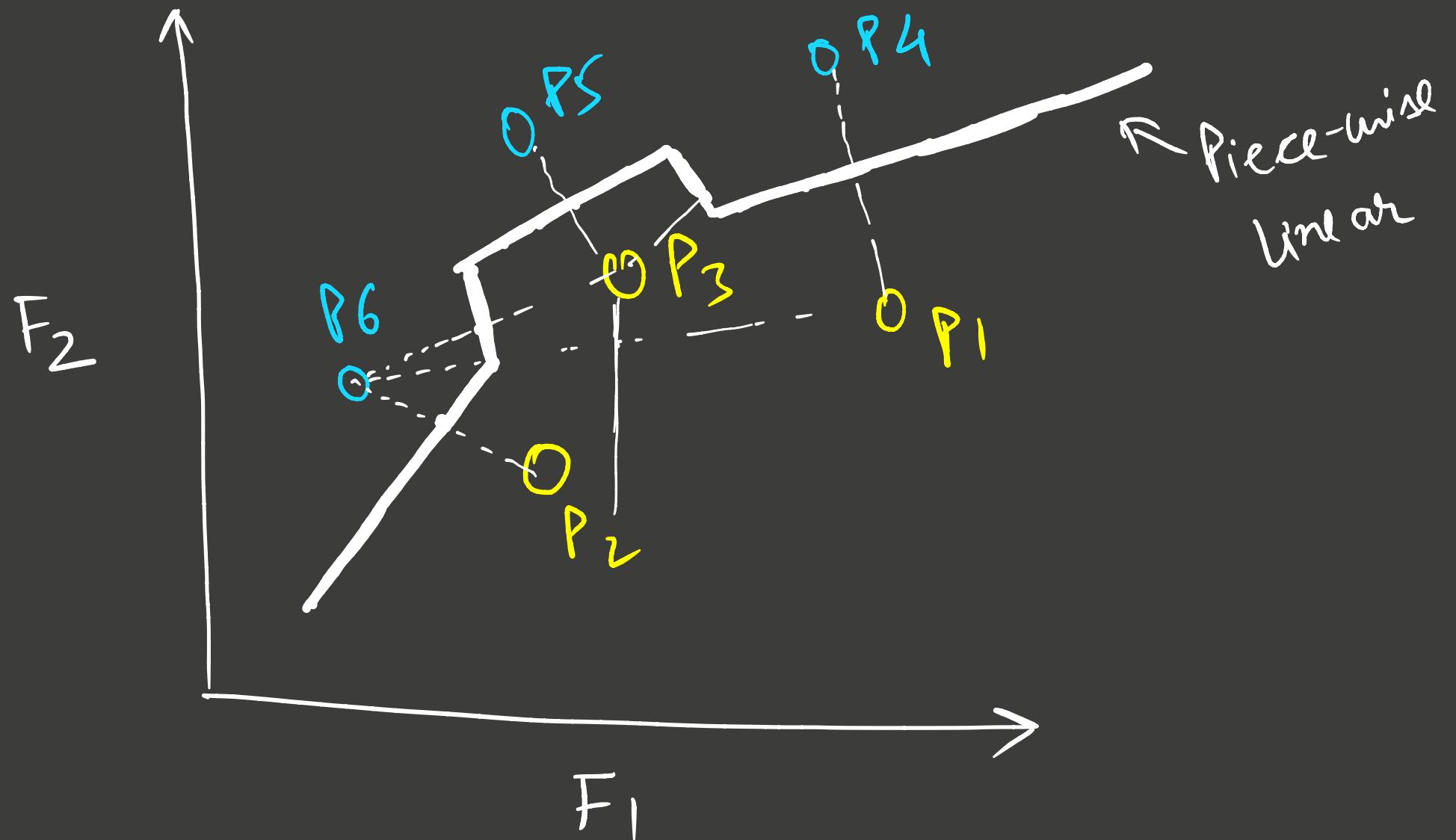
$\dfrac{x_1 + x_2}{2} < x < \dfrac{x_2 + x_3}{2}$ NN is (x_2, y_2)

NN NOT GOOD AT
EXTRAPOLATION

VORONOI DIAGRAM & 1-NN

LABELED AS YELLOW

⊥ to line joining 2 points

MID POINT

LABELLED AS RED

Yellow ↑↑↑↑

⊥ to Red

Feature 2

Feature 1

F_2

F_1

O P5

O P4

P6

O P3

O P1

↖ Piece-wise
linear

O P2

L_1: Blw P_2 & P_6

L_2: Blw P_6 & P_3
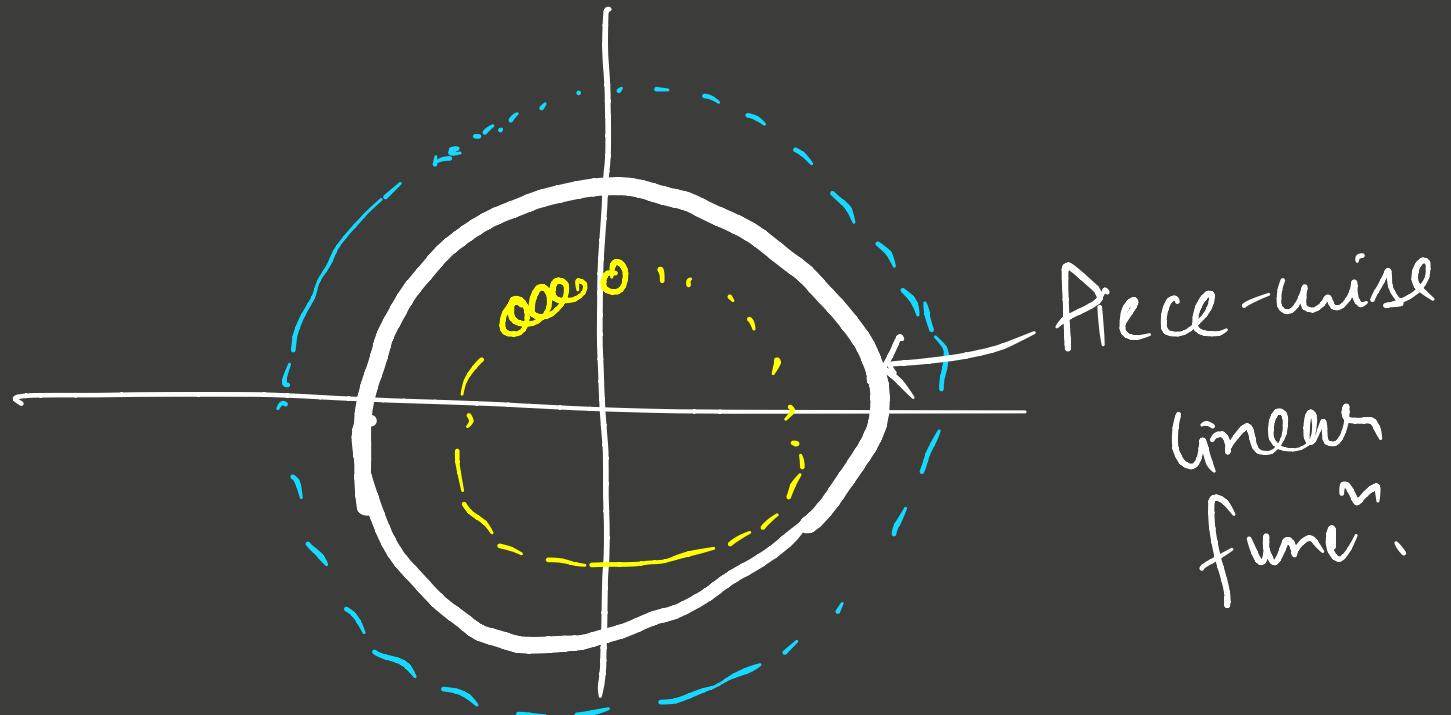
L_3: Blw P_6 & P_1 (DOESN'T ADD VALUE)

L_4: Blw P_5 & P_3

L_5:

Piece-wise linear funcⁿ.
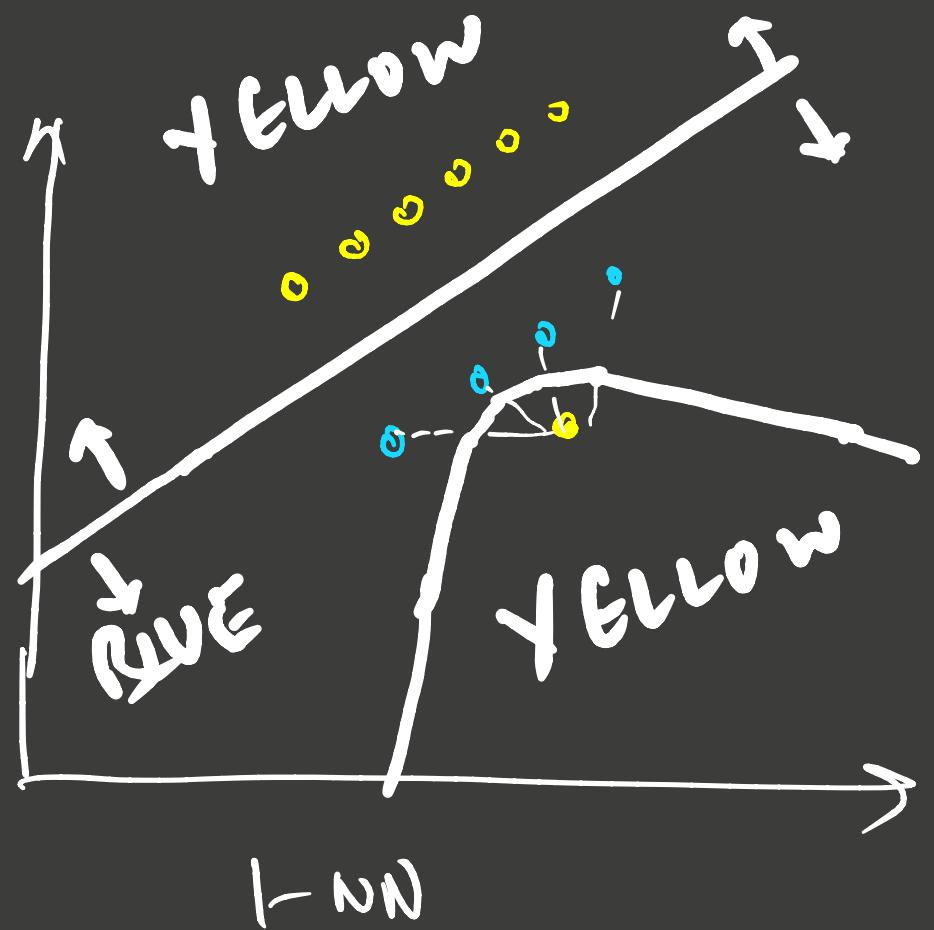
KNN
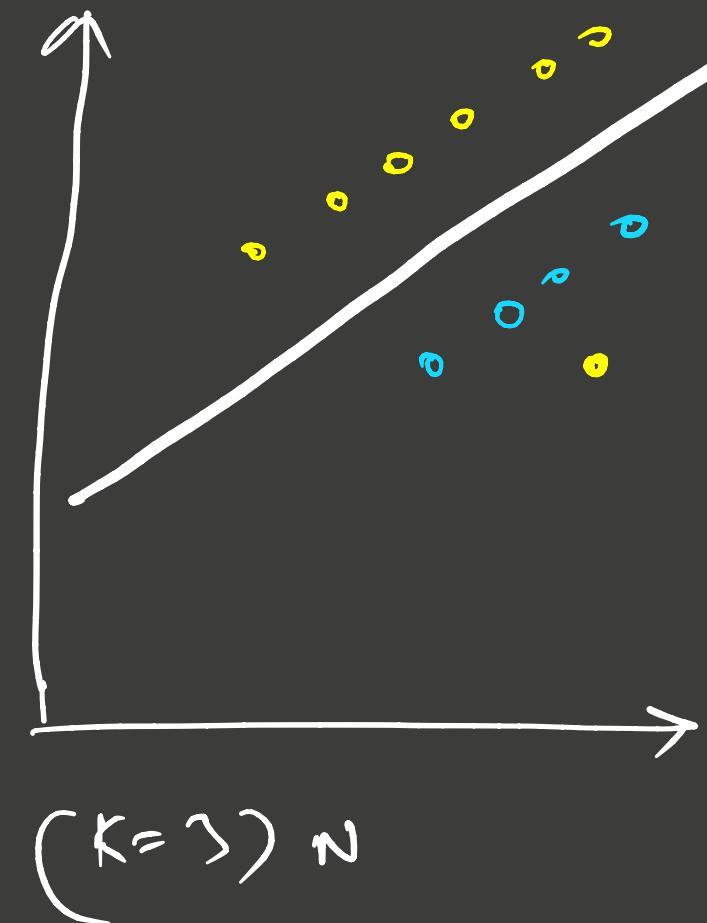
* Regression : $\hat{y} = $ Mean (y of 'k' Nearest pts.)

* Classificat" : $\hat{y} = $ MAJORITY VOTE (y of 'k'. NN)

Radius -NN

* All points within 'r' of query point.
  distance
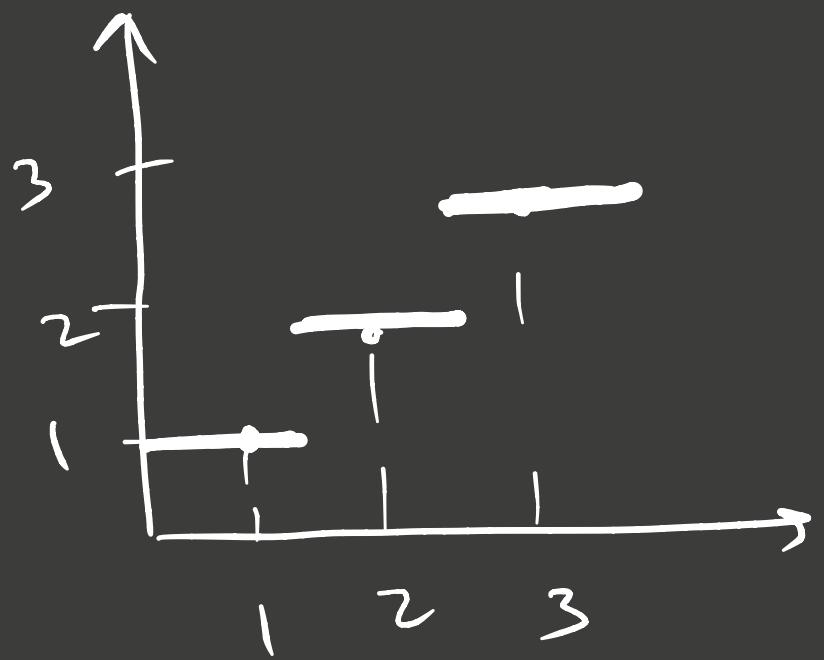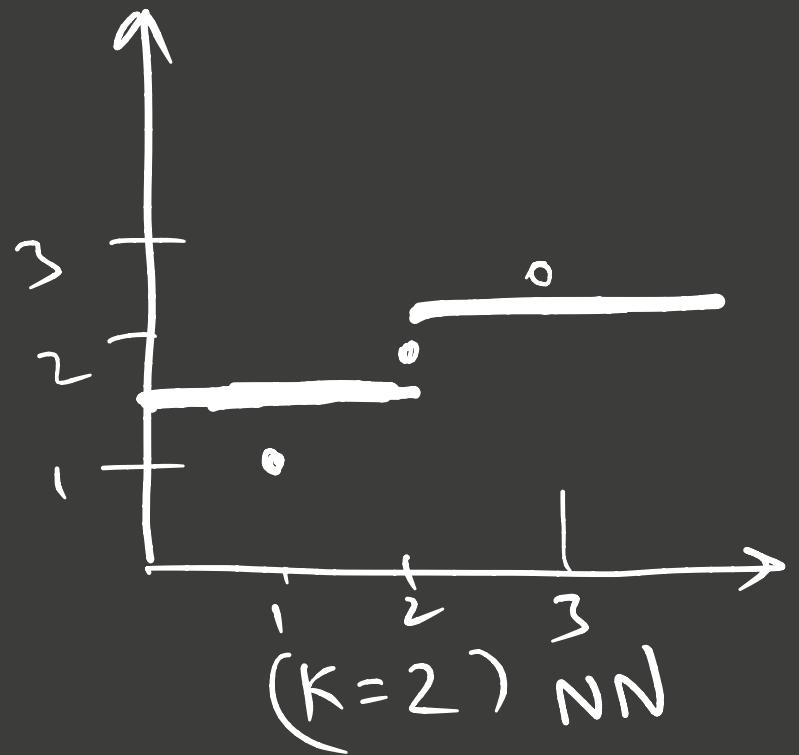
YELLOW

BLUE

YELLOW

1-NN

$(K=3)$ N

$K \uparrow$

SMOOTHNESS $\uparrow$

1-NN

(K=2) NN

$x < 1$ (2NN are $x=1$ &

$x=2$

$\hat{y} = (1+2)/2 = 1.5$

$1 < x < 2$ ($x=1$ & $x=2$

$\hat{y} = 1.5$

$x > 2$ (2NNs are $x=2$ &

$x=3$

$\hat{y} = 2.5$)

BIAS- VARIANCE

ERROR

TEST ERROR

TRAIN ERROR

K →
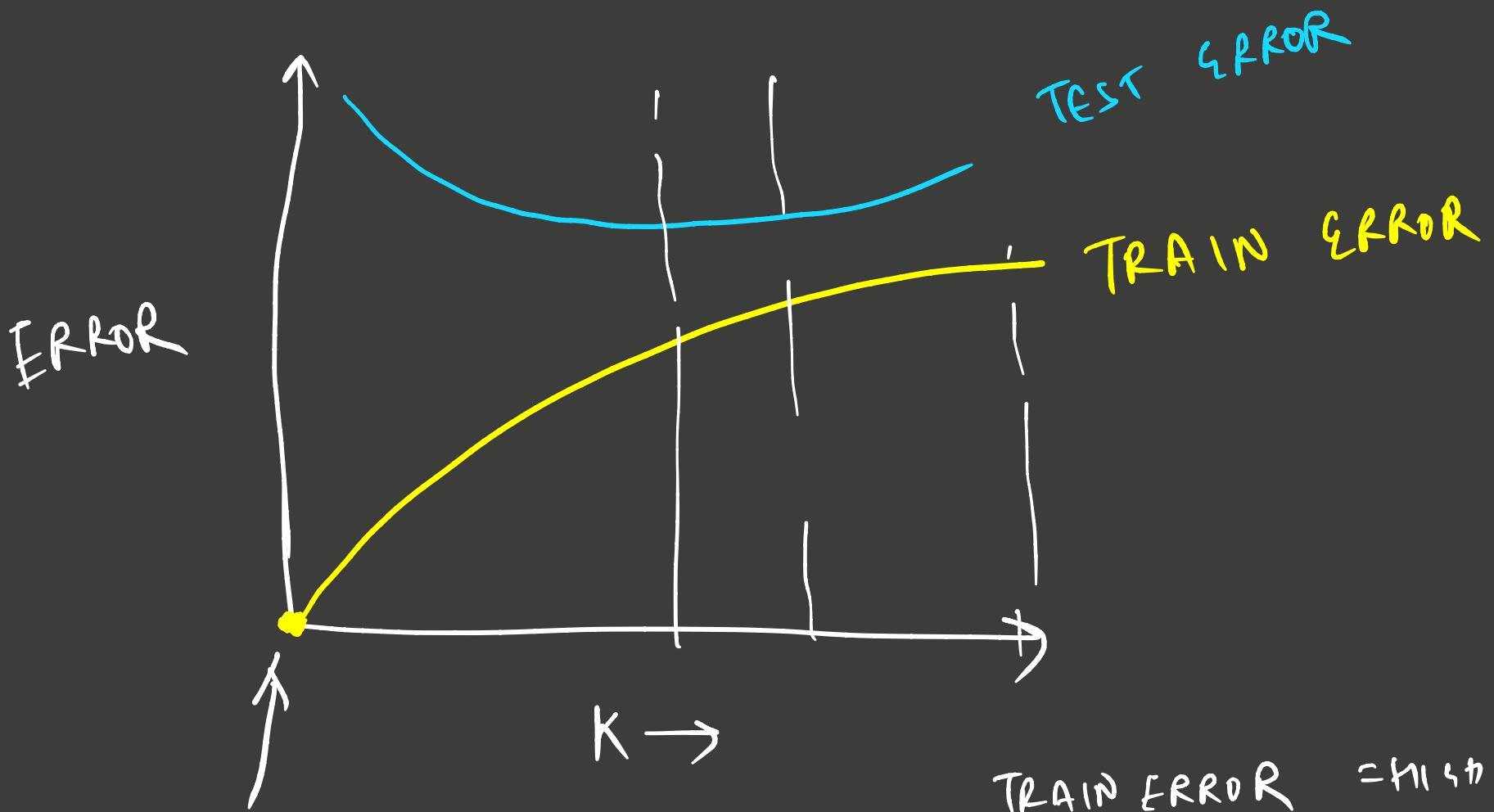
TRAIN
ERROR = 0

TEST ERROR
HIGH

VARIANCE
HIGH

OVERFITTING
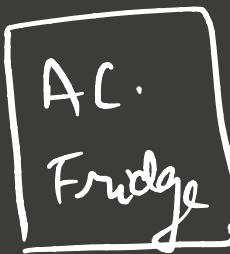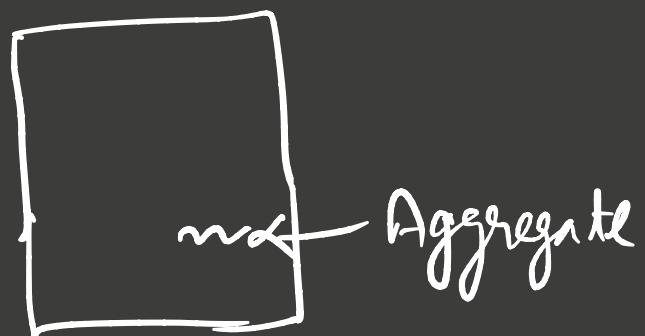
TRAIN ERROR = HIGH
BIAS = HIGH
UNDERFITTING
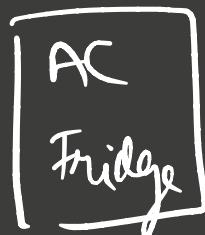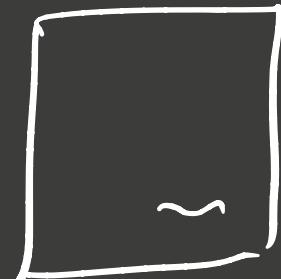
EXAMPLE APPLICATION (KDD2016 GEMELLO)

Jan

←— Aggregate

A.C.
Fridge

Feb
:

AC
Fridge
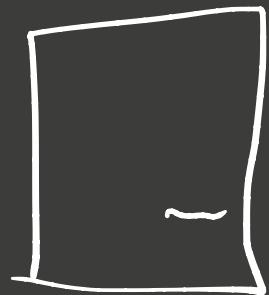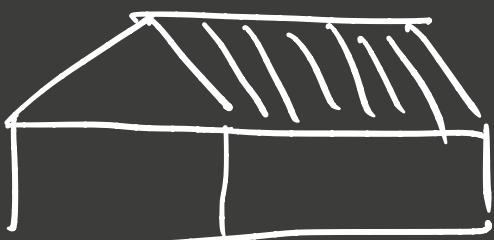
Dec
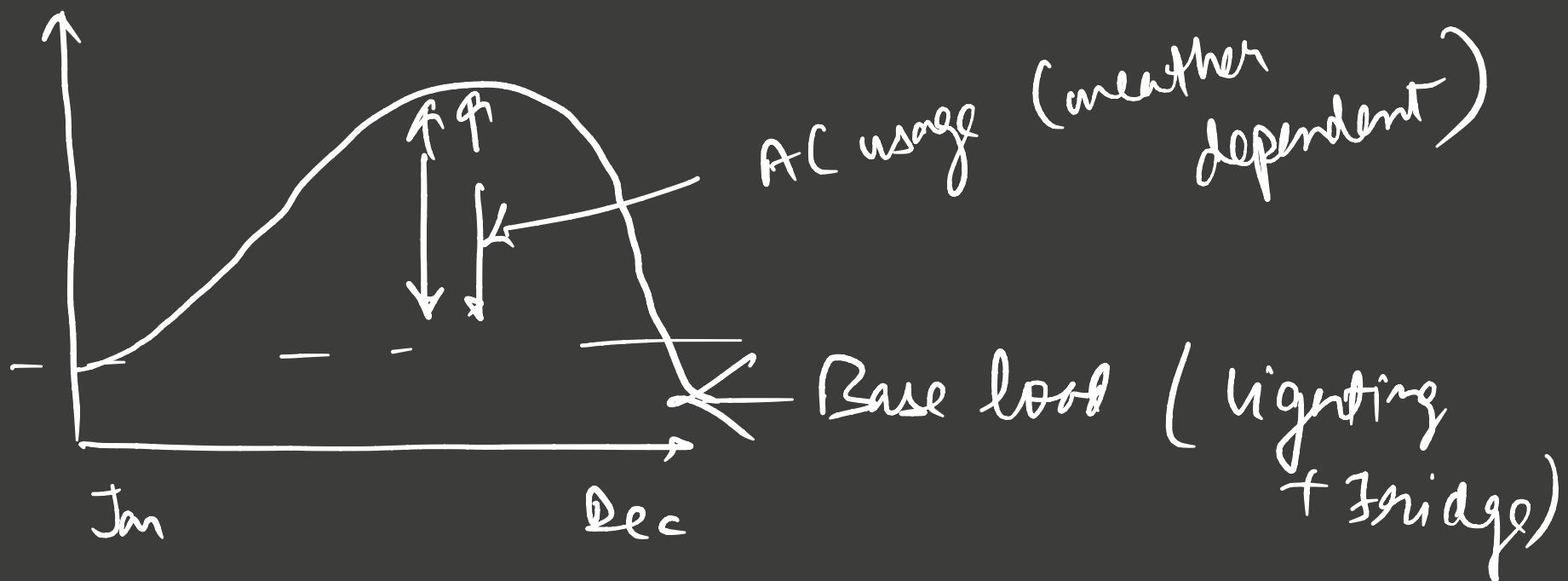
Area ; # Occupants

TRAIN SET

HOME	MONTH	Aggregate	AC	Fridge	
H_1	1	200	100	50	
H_1	2	✓	✓	✓	TRAIN
⋮	⋮				
H_m	12	✓	✓	✓	

HOME	AREA	# Occupants
H_1	100	3
⋮	⋮	⋮

HOMES   SIMILAR   IN

AREA   =>   SIMILAR   IN AC

AC usage (weather dependent)

Base load (lighting + Fridge)

Jan   Dec

AC usage

① Area

② Max (Aggregate) - Min (Aggregate)