Agenda

1. Gradient Boost ✓
2. Xgboost ✓
3. Unsupervised ML
   - K Means ✓
   - Hierarchical Means ✓
   - DBScan
   - → Silhouette Score

→ CatBoost

4. Gradient Boosting

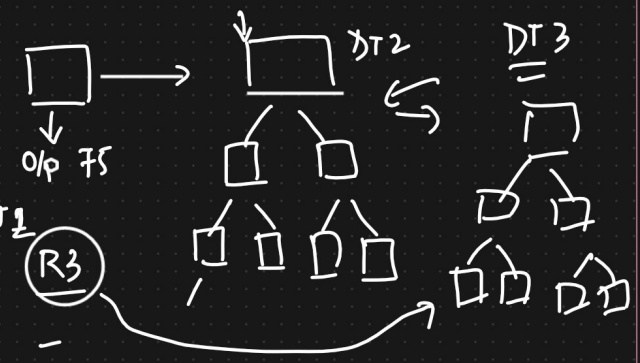
- → Classification
- → Regression

Exp	Degree	Salary ⇓ O/P	R1 → O/P (Error)	R2 (O/P)
2	BE	50K	-25	-23
3	Masters	70K	-5	-3
5	Masters	80K	5	3
6	PHD	100K	25	23

Test → 3.5 Masters

1. Base Model → O/p → Same O/p
   ⇓
   Constant Model

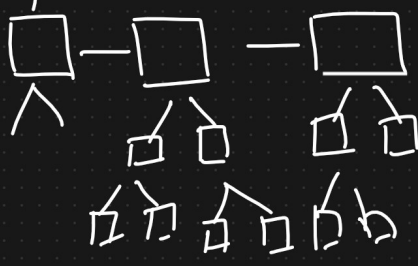
Base Model = $\frac{50+70+80+100}{4} = 75K$

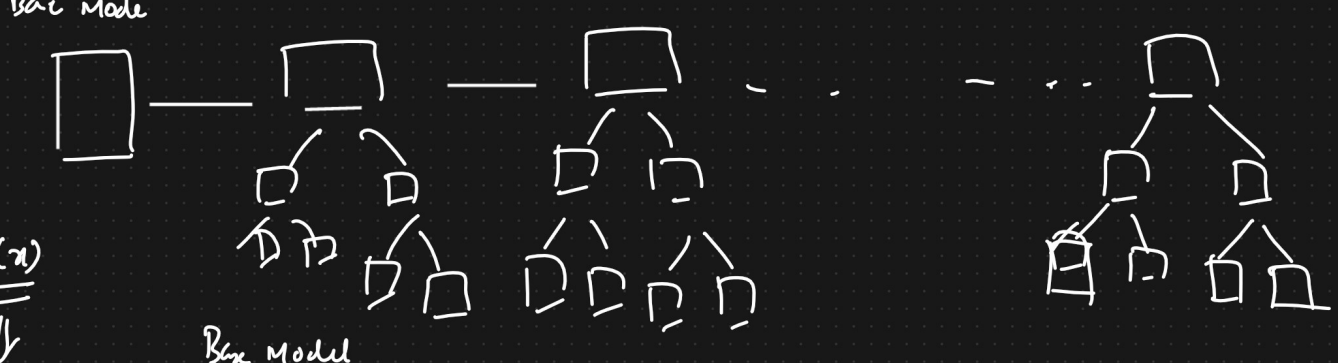


- 75 + (-3) = 72 → Overfitting
- $75 + \alpha_1(-3)$ ($\alpha_1 = 0.5$)
- $75 + (0.5)(-3) \Rightarrow 75 - 1.5 = 73.5$

Reduce $75 + \alpha 23 = 52$

Overfitting





Hypothesis ↑ ↑DT1 ↑DT2 ↑DT3

$$f(n) = h_0(x) + \alpha_1 h_1(x_1) + \alpha_2 h_2(x) + \cdots \alpha_n (h_n(x))$$

Hypothesis — 100 DT

Exp Degree → R1 (O/P) → Predicted O/P R2

$$f(n) = \text{Base Model} + \alpha_1 (DT1) + \alpha_2 (DT2) + \alpha_3 (DT3) + \cdots + \alpha_n (DTn)$$

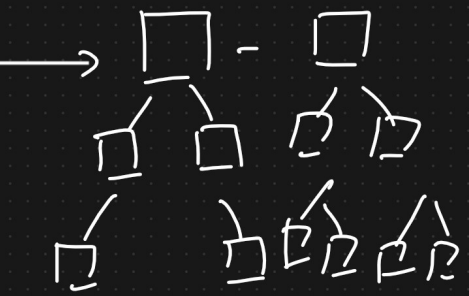
↓

To reduce overfitting.

f1	f2	O/p	R1	O/p	R2	O/p	R3
—	—	50	-15	-16	-7		
		60	-5	-4	-		
		70	5	7	-		
		80	15	13	-		

Base model

$\frac{50+60+70+80}{4} = 65$



$\alpha_1 = 0.5$

$65 + \alpha_1 (-16) = 65 + (10.5)(-16)$

$= 65 - 8$

$= 57$

② Xgboost Classifier

Base model → 0.5

Salary	Credit	Approval	Res
<=50K	B	0	-0.5
<=50K	G	1	0.5
<=50K	G	1	0.5
>50K	B	0	-0.5
>50K	G	1	0.5
>50K	N	1	0.5
<=50K	N	0	-0.5

Credit ✓

⇓

0.250

Similarity $= \frac{0.25}{1.75} = 0.142$

[-0.5, 0.5, 0.5, -0.5, 0.5, 0.5, -0.5]

① Construct Base Model →

⇓

O/p = 0.5

② Construct Tree with Root Node.

← Salary

<=50K >50K

[-0.5, 0.5, 0.5, 0.5] [-0.5, 0.5, 0.5]

③ Calculate Similarity Weight

$$= \frac{(\Sigma \text{ Residual})^2}{\Sigma Pr(1-Pr) + \lambda}$$

$\lambda = 0$

Similar Weight

$$= \frac{[-0.5+0.5+0.5-0.5]^2}{[0.5(0.5)+0.5(0.5)+0.5(0.5)+0.5(0.5)]} = \frac{[-0.5+0.5+0.5]^2}{[0.5(0.5)+0.5(0.5)+0.5(0.5)]}$$

$= 0 +$

$= \frac{0.25}{0.75} = 0.33$

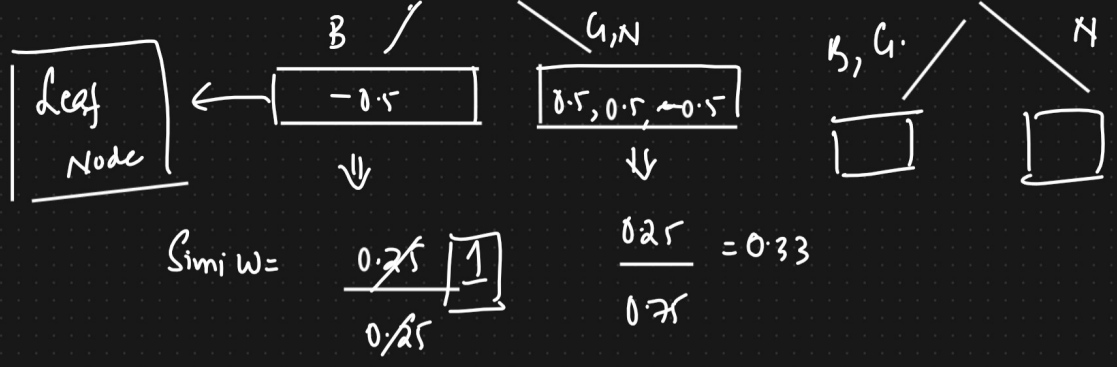
④ Gain

Gain = 0 + 0.33 − 0.142

Gain = 0.188

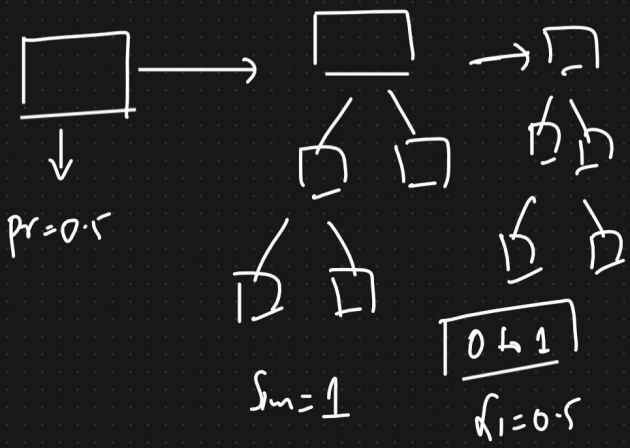
[-0.5, 0.5, 0.5, -0.5] [-0.5, 0.5, 0.5]

[Credit]



Simi W= $\frac{0.25}{0.25}$ [1] $\frac{0.25}{0.75} = 0.33$

Info gain = $1 + 0.33 - 0 = 1.33$



$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right) = \log\left(\frac{0.5}{0.5}\right) = 0$

Base model + $\gamma_1(DT1) + \gamma_2(DT2) + \gamma_3(DT3) + \cdots \gamma_n(DT_n)$

Xgboost = $\sigma(0 + \gamma_1(1))$

$= \sigma((0) + (0.5)(1))$

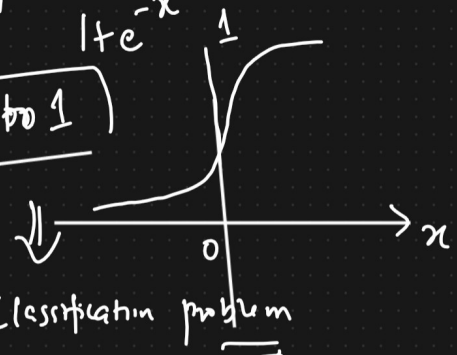
↳ Sigmoid Activation fn

Sigmoid $= \sigma(0.5) = \frac{1}{1 + e^{-0.5}}$

0.78

$y = \frac{1}{1+e^{-x}}$

= 0 to 1



Classification problem

Xgboost : Base model + $\sigma(\gamma_1(DT_1) + \gamma_2(DT_2) + \gamma_3(DT_3) + \cdots \gamma_n(DT_n))$

↓ log loss → $\log\left(\frac{p}{1-p}\right)$

↓ Sigmoid Activation function.

f1 f2 → O/p Categor → O/p → Regression Continuous

⑧ Xgboost Regressor

Exp	Gap	Salary	Res 1
→ 2	Yes	40K	-11
→ 2.5	Yes	42K	-9
3	No	52K	1
4	No	60K	9
4.5	Yes	62K	11

O/p

$$\frac{(-11-9+1+9+11)^2}{5+1} = \frac{1}{6} \text{ //.}$$

Base Model → 51K

↓

[-11, -9, 1, 9, 11]

$\frac{1}{6} = 0.166$

Exp

<=2 | >2

[-11] | [-9, 1, 9, 11] → O/p 3

$\lambda = 1$

Similarity weight = $\frac{121}{1+1} = \frac{121}{2} = 60.5$

$\frac{(-9+1+9+11)^2}{4+1} = \frac{144}{5} = 28.8$

② Similarity weight = $\frac{(\Sigma \text{Residual})^2}{\Sigma p_i(1-p_i) + \lambda}$

$= \frac{(\Sigma \text{Residual})^2}{\text{No. of residuals} + \lambda}$

Gain = 60.5 + 28.8 − 0.166

= 89.134

[-11, -9, 1, 9, 11]

Exp

0.166

≤2.5 | >2.5

O/p = -10 | O/p = 21/3 = 7

O/p = $\frac{-11-9}{2} = -10$ [-11, -9] | [1, 9, 11]

$\text{Sim(wt)} = \frac{(-11-9)^2}{2+1}$

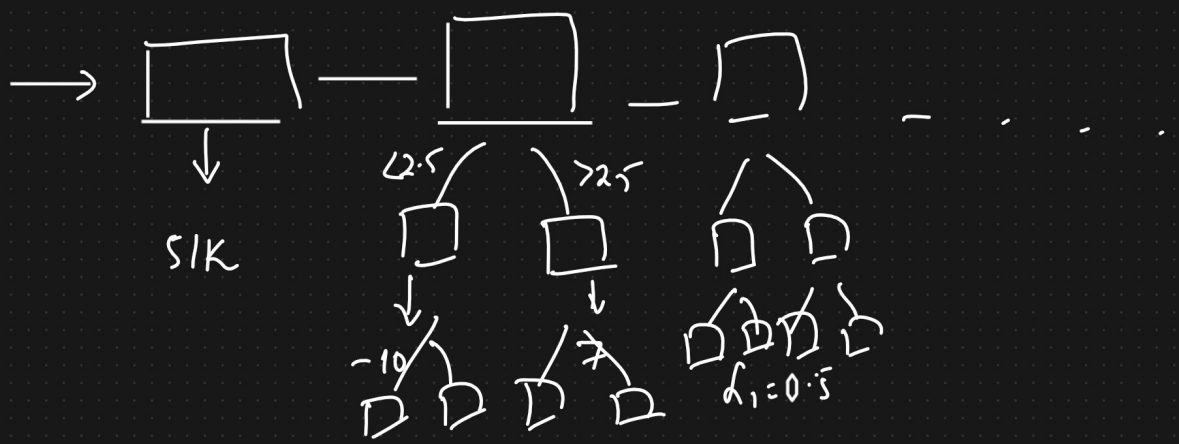
$\text{Simcor} = \frac{(1+9+11)^2}{4} = \frac{(21)^2}{4} = 110.23$

Yes | No

$= \frac{400}{3} = 133.3$

Info gain $= 133.3 + 110.25 - 0.166$

≈ 243



$$51 + \alpha_1(-10) + \alpha_2() + \alpha_3() + \alpha_4() + \cdots + \alpha_n(DTa)$$

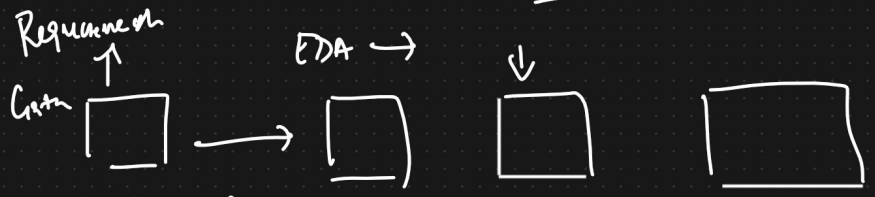
$51 + 0.5(-10) = 51 - 5 = 46K$

Interview → 2 interviews

Logistic

{ ① Projects ② Theoretical ③ Lifecycle of DS Project }.

Projects: ① Goal ⇒ 1% 3% Automating something } Business }.



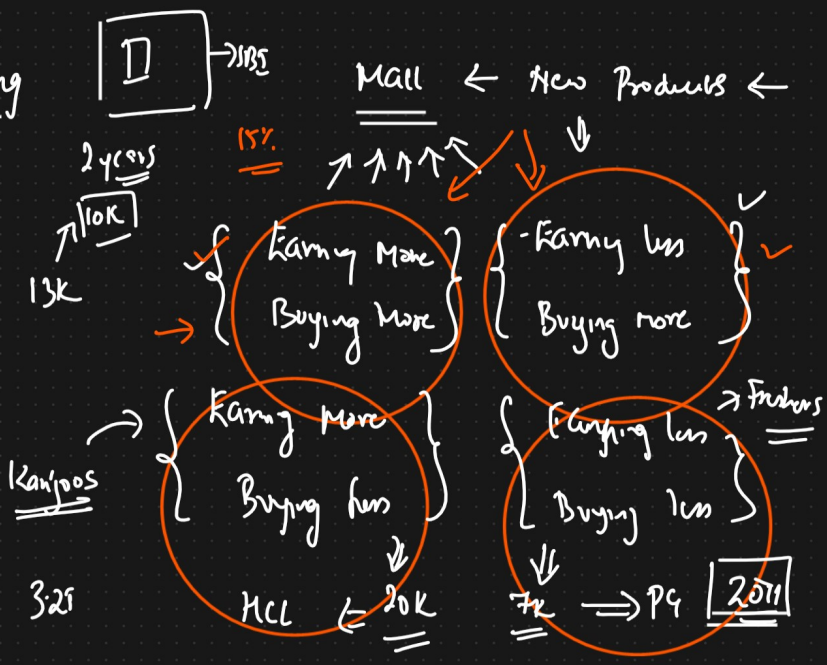
{CATBOOST} ←

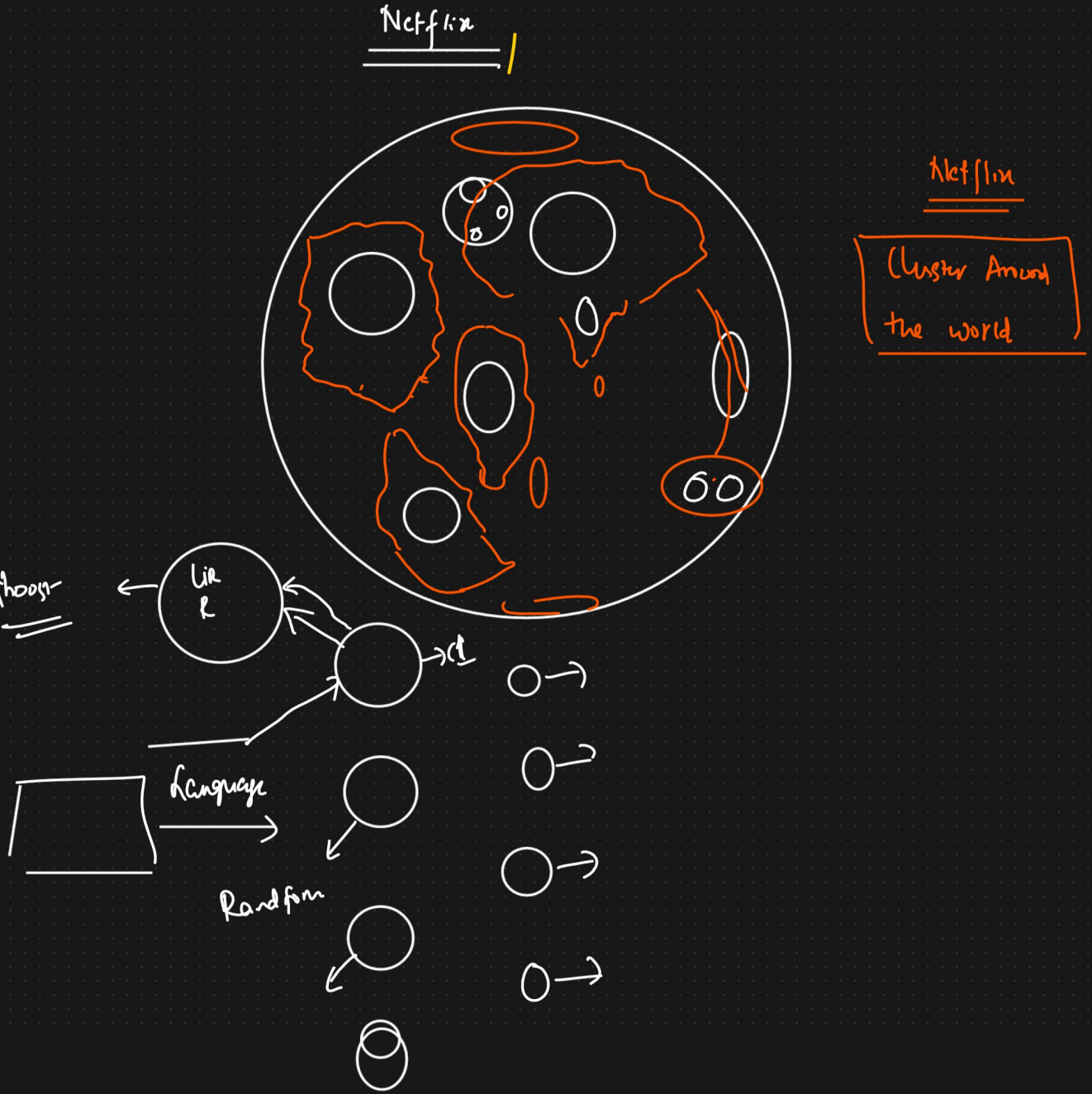
⑧ Unsupervised Machine Learning

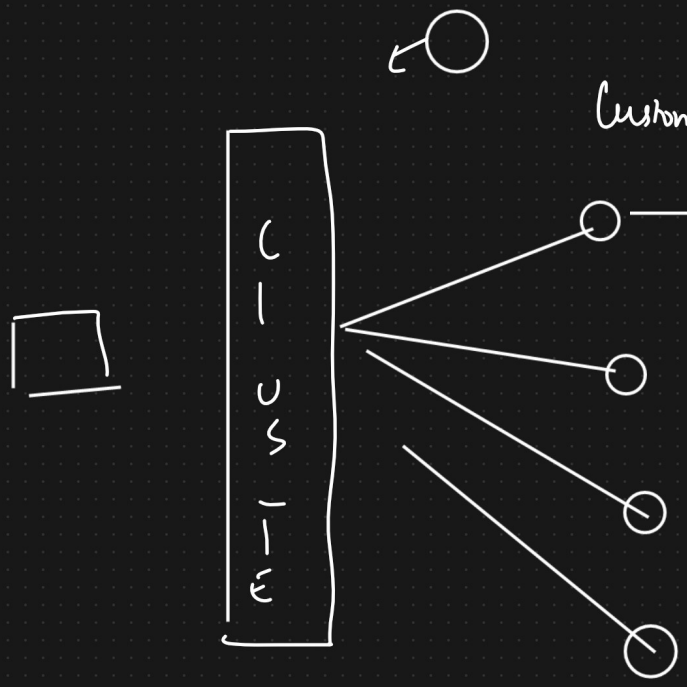
Salary	Spending Score (0-10)
100K	9
50K	4
120K	9
40K	8
70K	3

Clusters

O/P







Custom Ensemble Techniques

Salary	Spending Score
100K	9
40K	8
70K	4
120K	8
10K	1
50K	4

Discount to be sent

o/p

15%.

12%.

5%.

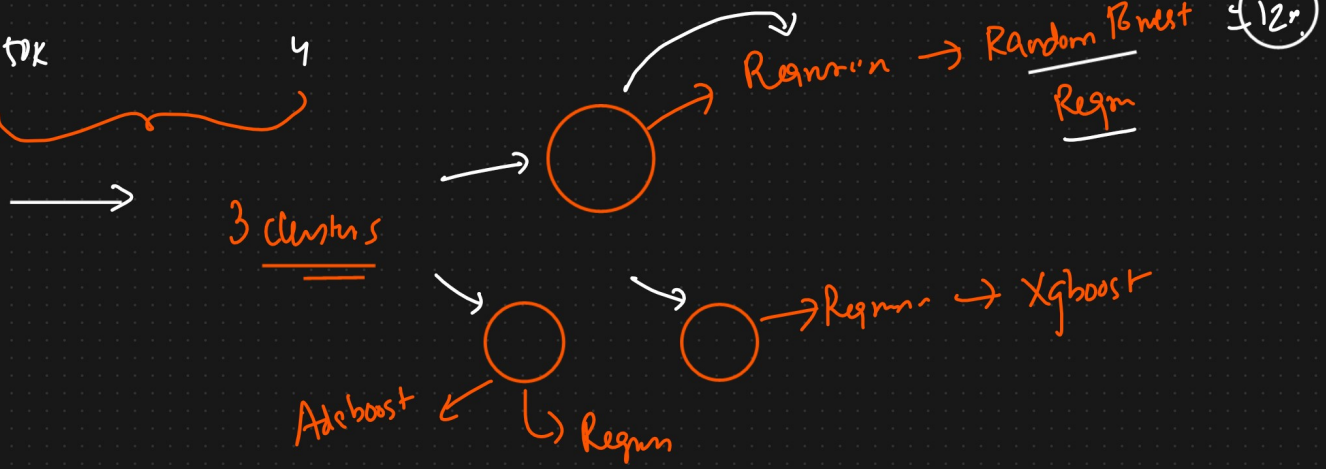
15%.

13%.

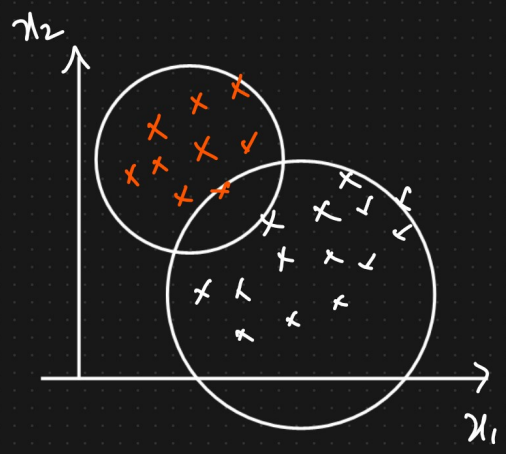
→ Regression problem statement

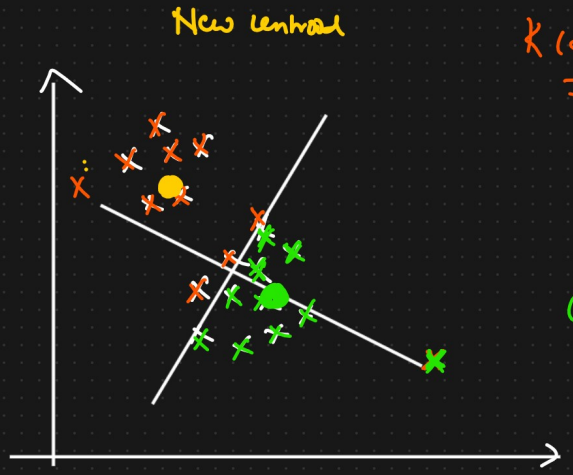
# of Actors

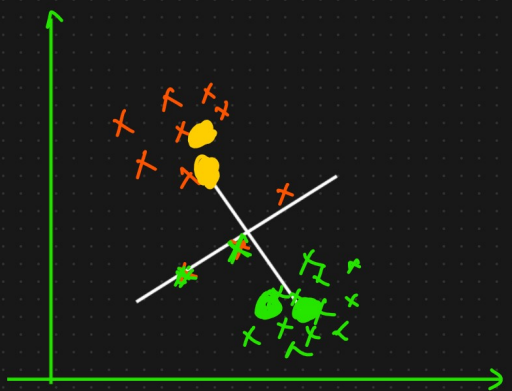
Netflix Notification

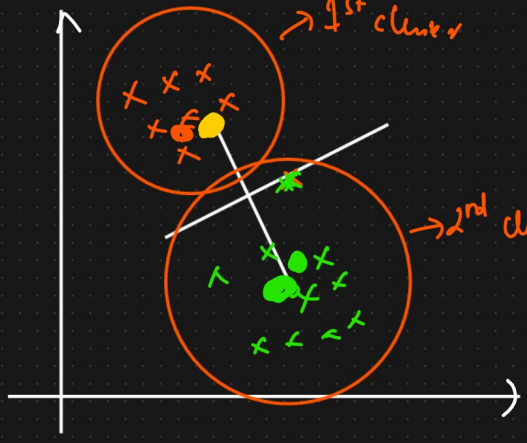


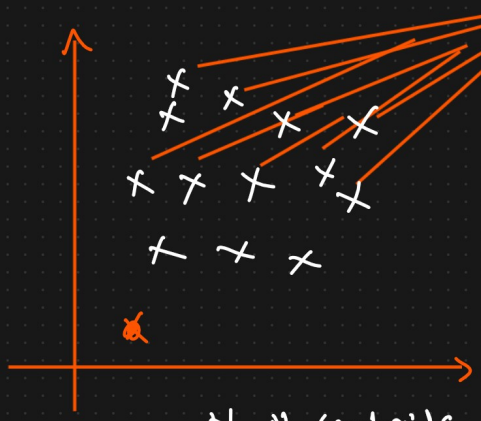
① K Means Clustering











K = 1 to 20

K=1

↑↑↑ wcss {within cluster sum of square}

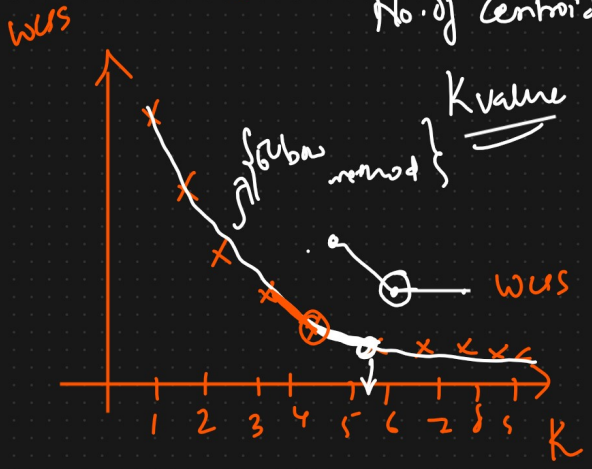
K=2

↓↓↓ wcss {will decrease}

↓↓↓ K=3

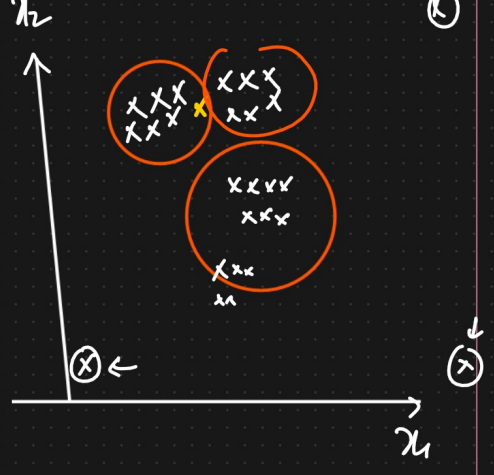
↓↓↓ K=4

↓↓↓



Kmeans ++

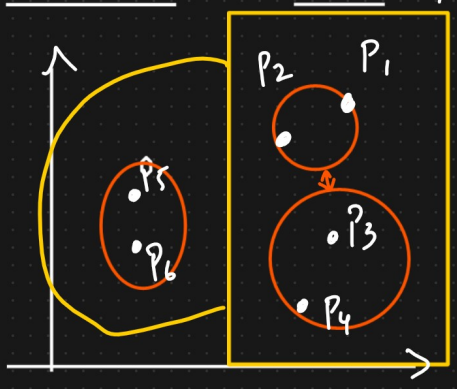
K=3

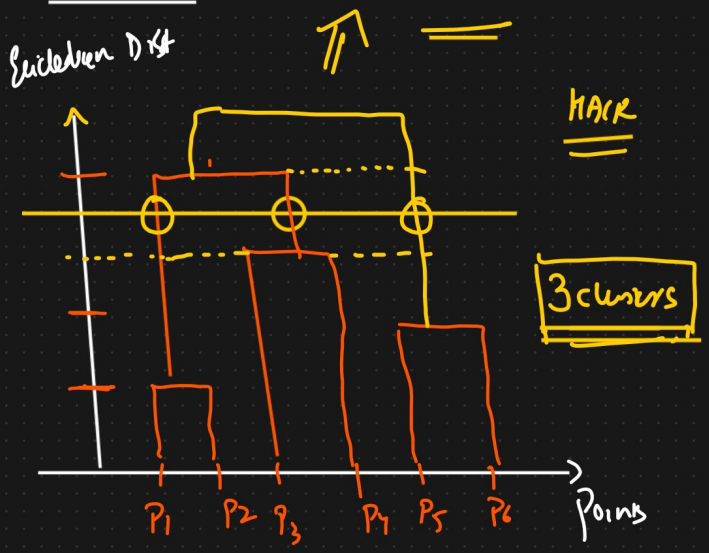


Kmeans ++

Dendogram

② Heirarchical Mean Clustering





{① We need to the longest vertical line such that none of the horizontal line passes through it. }