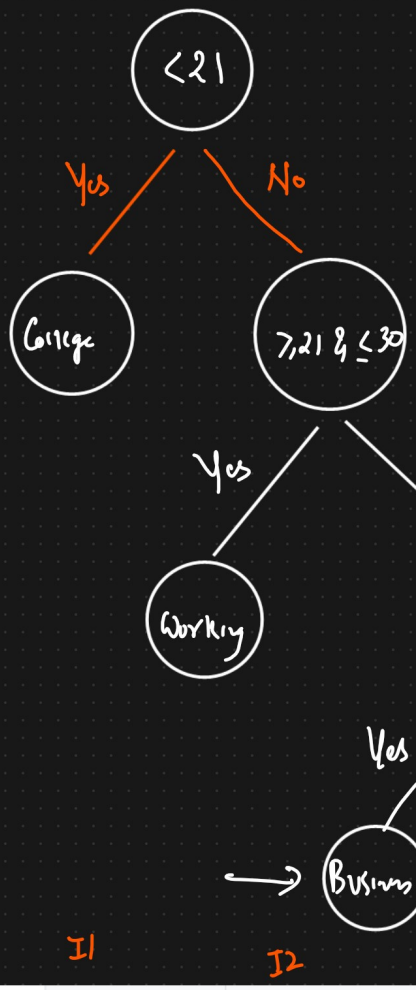
Decision Tree Classifier And Decision Tree Regressor

① Classification
② Regression Problem

age = 34



I1 I2 I3 I4

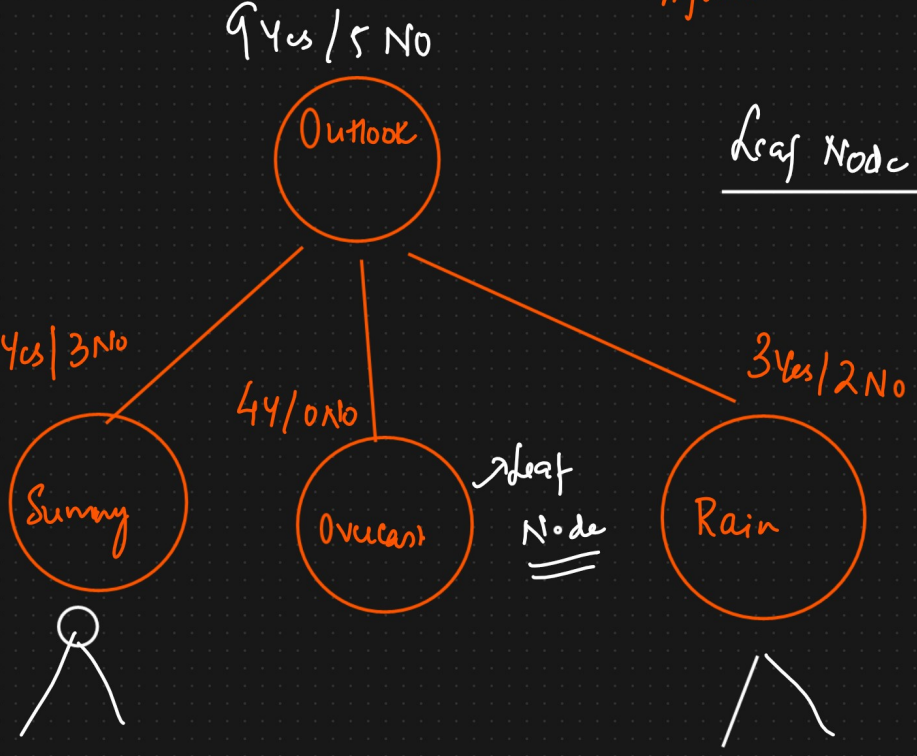
Nested if-else Clause

```
if (p_age < 21):
    Print ("He should be in College")
elif (p_age >,21 and p_age ≤ 30):
    Print ("He should be working")
elif (p_age > 30 and p_age ≤ 35):
    Print ("He owns a business")
else:
    Print ("He has retired")
```

	Outlook	Temperature	Humidity	Wind	Play Tennis
2	Sunny	Hot	High	Weak	No
3	Sunny	Hot	High	Strong	No
4	Overcast	Hot	High	Weak	Yes
5	Rain	Mild	High	Weak	Yes
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Cool	Normal	Strong	No
8	Overcast	Cool	Normal	Strong	Yes
9	Sunny	Mild	High	Weak	No
10	Sunny	Cool	Normal	Weak	Yes
11	Rain	Mild	Normal	Weak	Yes
12	Sunny	Mild	Normal	Strong	Yes
13	Overcast	Mild	High	Strong	Yes
14	Overcast	Hot	Normal	Weak	Yes
15	Rain	Mild	High	Strong	No

Decision Tree : ① ID3 Algorithm ② CART (Sklearn)

Classification and Regression Tree Algorithm



① Purity → Pure Split ??

↳ Entropy
↳ Gini Index

② How the features are selected

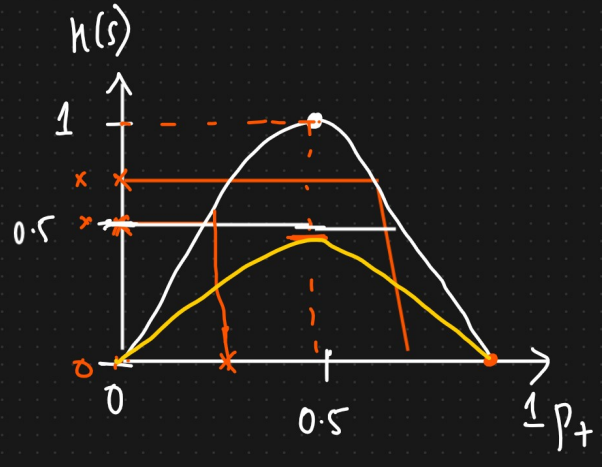
⇒ Information Gain?

① Entropy {Multiclassification} {Binary classification}

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$$H(S) = -P_{c_1} \log_2 P_{c_1} - P_{c_2} \log_2 P_{c_2} - P_{c_3} \log_2 P_{c_3}$$

② Gini Index



3Y|3N C_1 C_2 3Y|0N

$$H(S) = -\frac{3}{6}\log_2(3/6) - \left(\frac{3}{6}\right)\log_2(3/6)$$

$H(S) = \boxed{1}$ → Very Impure Split

3Y|0N

$$H(S) = -\frac{3}{3}\log_2(3/3) - \frac{0}{3}\log_2(0/3)$$

$$= 0 - 0$$

$= \boxed{0}$ → Pure Split.

Pr P-
3Yes|0No → 3/3 = 1
H(S) = 0

0Yes|3No → 0/3 = 0

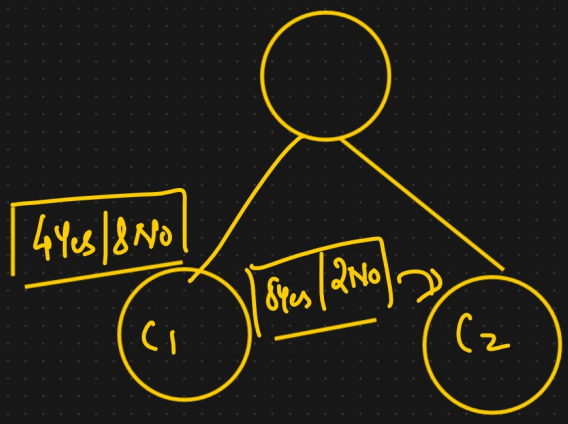
H(S) = 0

Pure Split

$\boxed{2Y|3N}$ → Impure Split.

$$H(S) = -\frac{2}{5}\log_2(2/5) - (3/5)\log_2(3/5)$$

$$= 0.97$$



⑥ Gini Index

$$G.I = 1 - \sum_{i=1}^{n} (p)^2 = 1 - \left[(p_+)^2 + (p_-)^2\right]$$

3/6 = 0.5

$$= 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$$

3Yes|3No ⇒ 0.5

$$G.I = 1 - \frac{1}{2} = 0.5$$

$$[4Yes/8No] = 1 - \sum_{i=1}^{n} (p)^2 = 1 - \left[\left(\frac{4}{12}\right)^2 + \left(\frac{8}{12}\right)^2\right]$$

$$= 1 - \left[\frac{1}{9} + \frac{4}{9}\right]$$

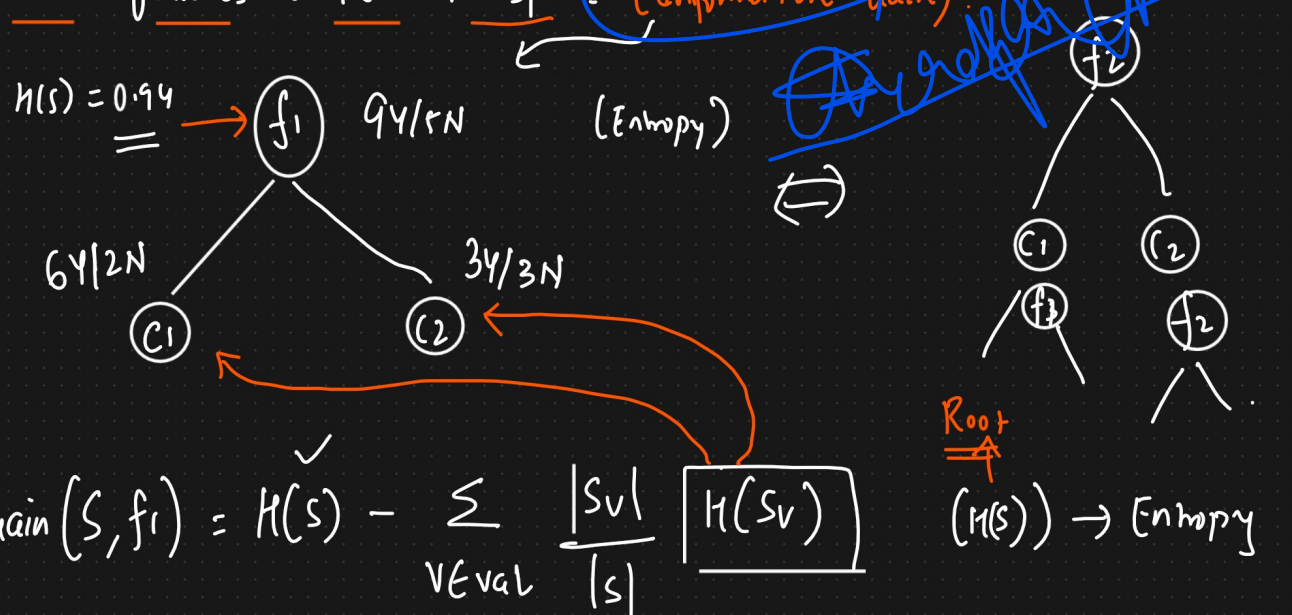
Gini Index

$$= 1 - \frac{5}{9} = \boxed{4/9} \approx 0.444$$

$$[8Yes/2No] = 1 - \left[\left(\frac{8}{10}\right)^2 + \left(\frac{2}{10}\right)^2\right]$$

$$= 1 - \left[\frac{16}{25} + \frac{1}{25}\right] = 1 - \frac{17}{25} = \frac{8}{25} \approx 0.32$$

(2) Which features to take to split? (Information Gain)



$$Gain(S, f_1) = H(S) - \sum_{V \in Val} \frac{|S_V|}{|S|} H(S_V)$$

(H(S)) → Entropy

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$$= -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) \approx \boxed{0.94}$$

(6Y/2N)

$$H(C_1) = \frac{-6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) = 0$$

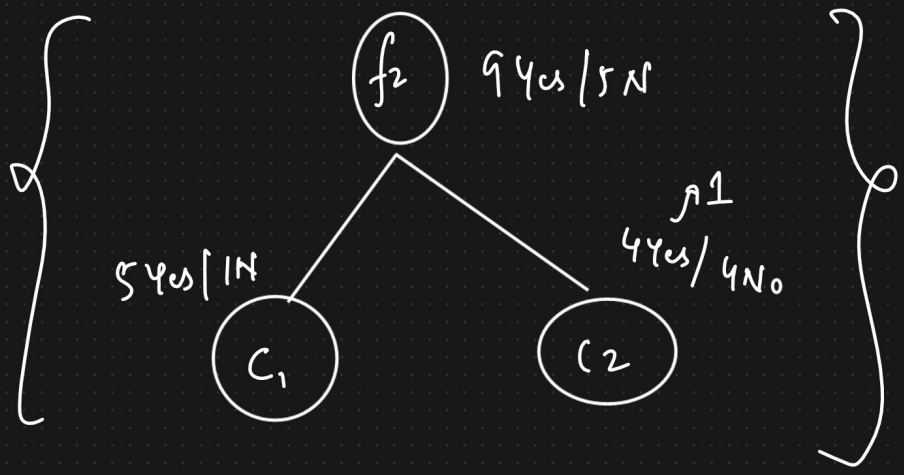
$$H(C_2) = 1$$

(3+/3-)

$$\text{Gain}(S, f_1) = 0.94 - \left[\frac{8}{14}(0.81) + \frac{6}{14} \times 1\right]$$

$$\frac{|S_v|}{|S|} = \frac{6}{14} \quad \boxed{\frac{8}{14}}$$

$$= 0.94 - [0.462 + 0.42] = 0.049$$



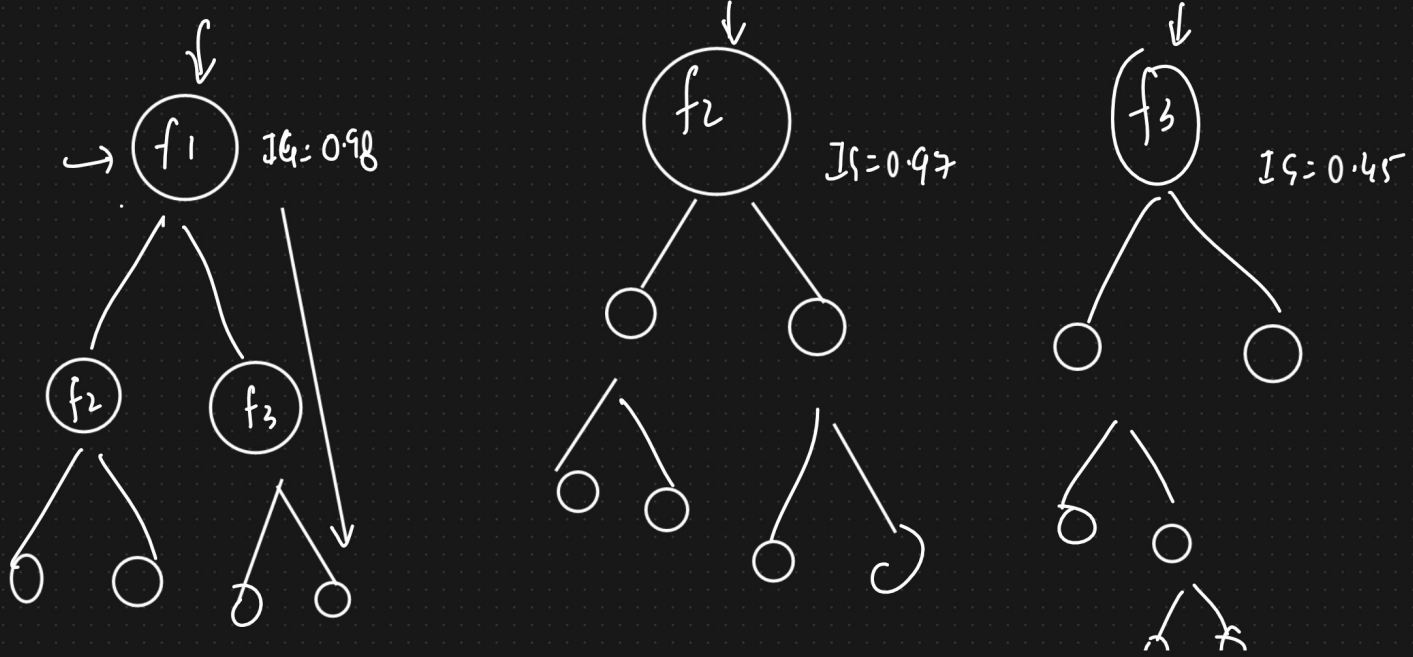
⇒ Information gain.

= 0.09

Gain = =0.94 -(6/14)*0.65 - (8/14)*1 = 0.09

feature 2

$IG(f_2) > IG(f_1)$

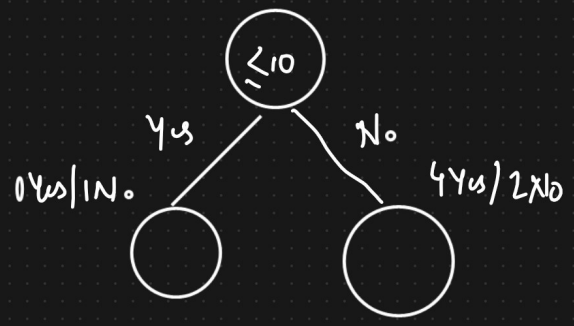


Numerical feature = CART

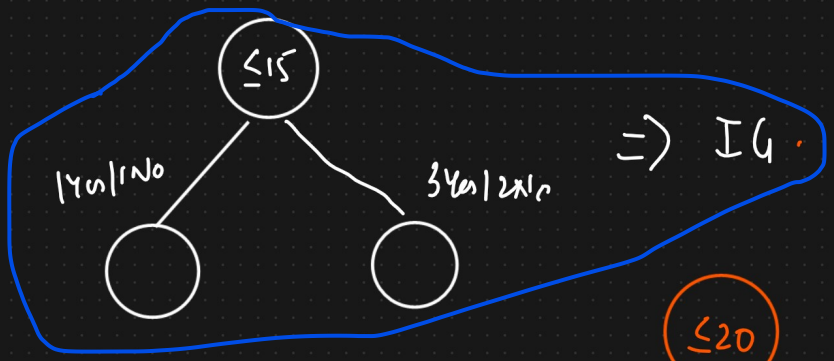
{Independent feature}

f1	f2	Op
10	–	0 ✓
15	–	1 ✓
20	–	1
25	–	0
30	–	1
35	–	0
40	–	1

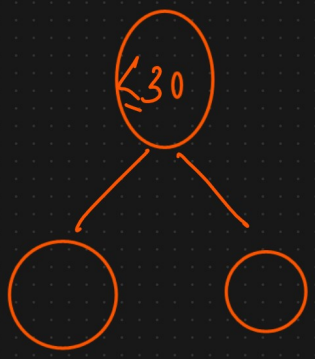
Binary split

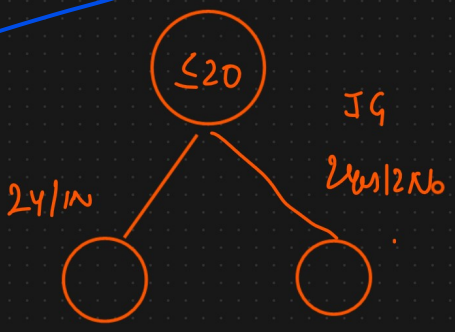


IG =



=> IG.





IG

Decision Tree Regressor

Op

$MSE = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ Continuous value

↓ Mean

2 Op feature

f1	f2	Price
–	–	20
–	–	24
–	–	28
–	–	14
–	–	16
–	–	20

Mean

Size of House

Price of the House.

Op

↓ Mean( )

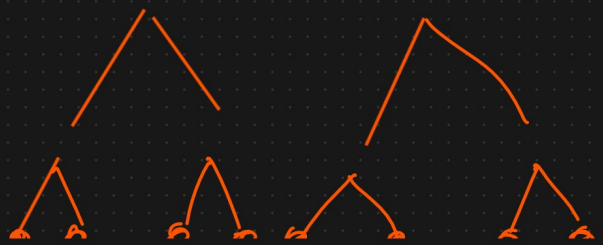
MSE ↑ Mean(20, 24, 28)

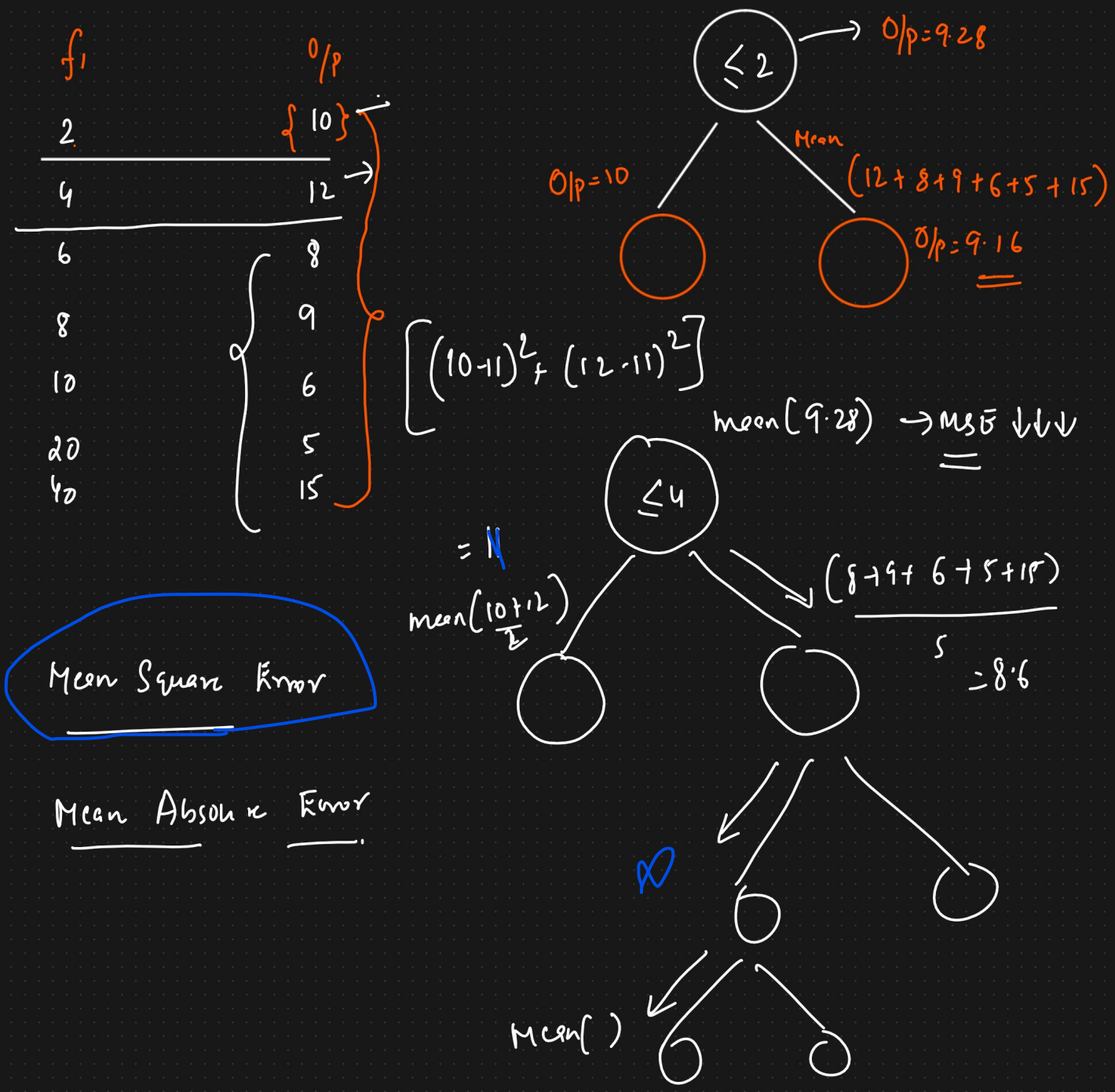
f1

(MSE or MAE)

Mean(14, 16, 20)

C1 C2





Mean Square Error

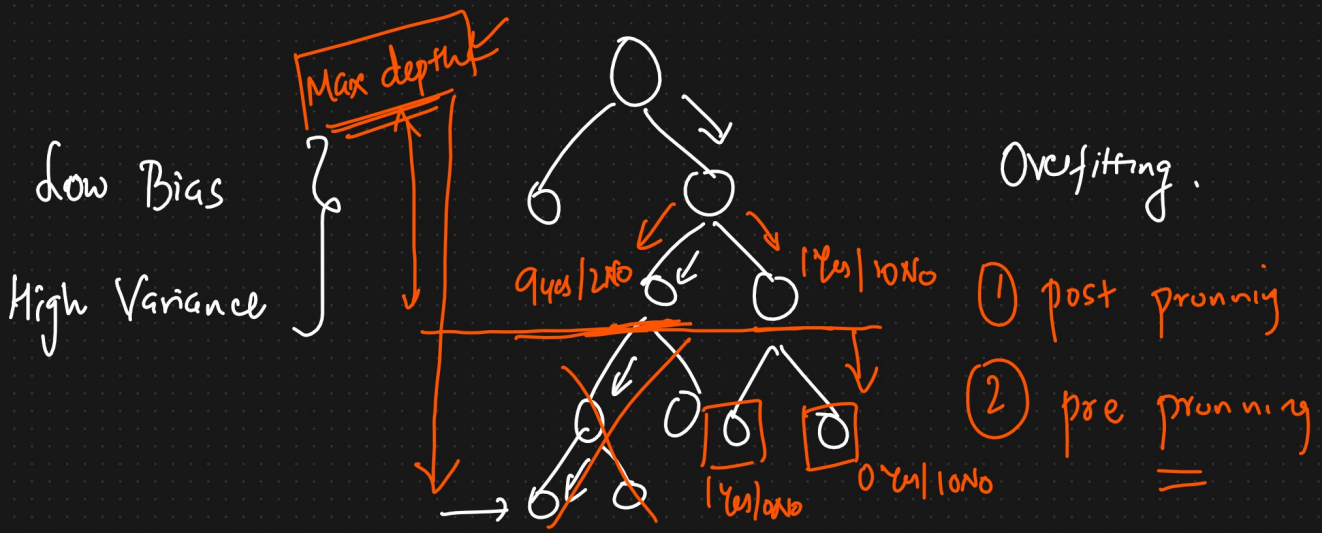
Mean Absolute Error

Decision Tree

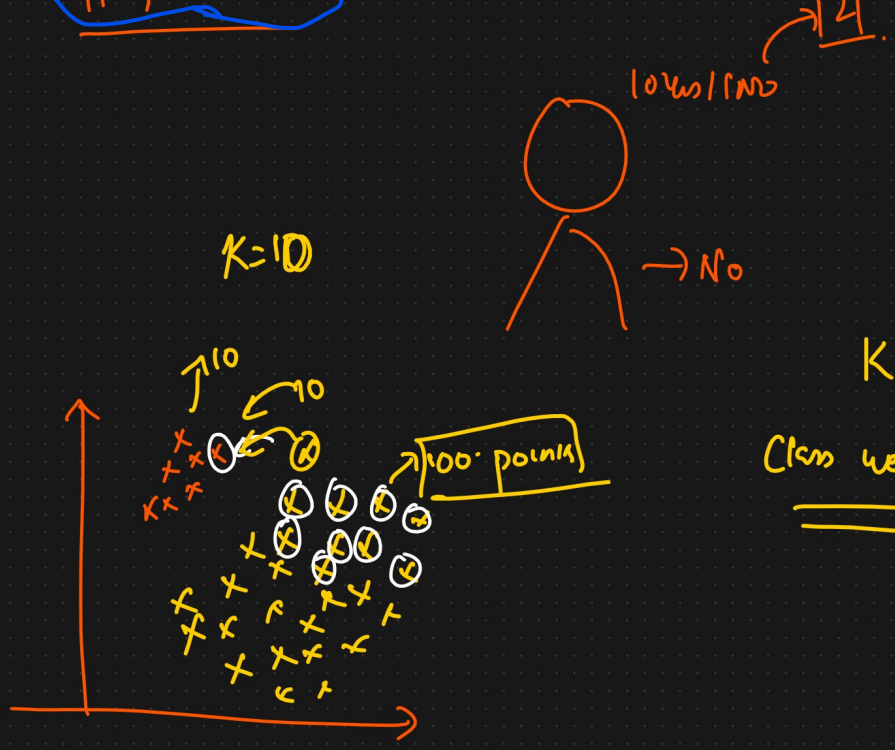
Overfitting

Training Accuracy ↑↑ {Low bias}

Test Accuracy low {High Variance}



Hyperparameter



K Nearest Neigh

Class weight

f1 f2 f3 O/P

Regression { ↔ }

{what is the approach to fill these NAN values?}.

↓

70 - 30%