DECISION TREE (CLASSIFICATION & REGRESSION)

↳ NESTED IF-ELSE.

- **Pure Node** : Node in Decision Tree where all datapoints belong to the same class. or have same value (regression)
  --→ leaf node are Pure Node

- **Impure Node** : where all datapoints don't belong to same (a) class or same value

AIM OF DT : Start of Impure node to reach to Pure Node

★ In some cases, it can happen we can't reach to pure state

Q How do we calculate to take which feature next in order to reach pure state?

→ ~~we use Loss for or cost FXN~~ INFORMATION GAIN

Loss Function ⇒ i) ENTROPY
in D.T ii) or GINI INDEX

LOSS FUNCTION

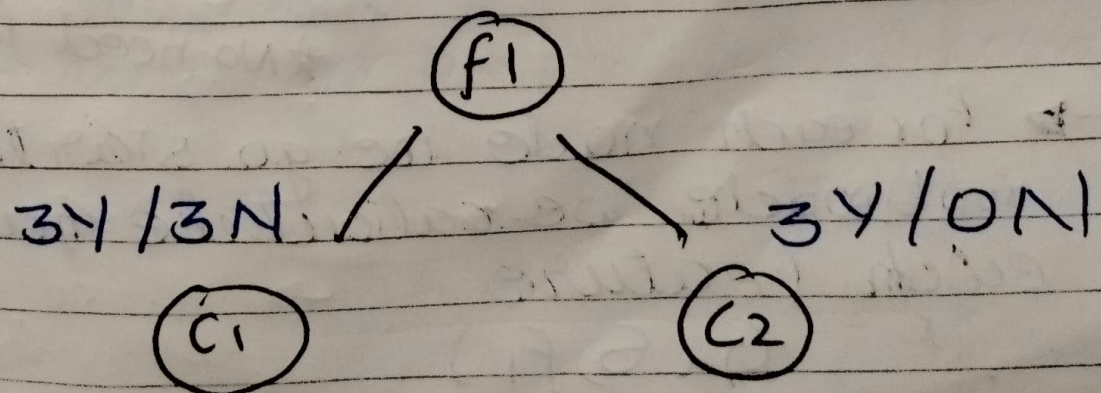
① Entropy (0 to 1)

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

p_+ ⇒ probability of Yes
p_- ⇒ probability of No.

★ Pure Split ⇒ Entropy ⇒ 0
Impure Split ⇒ Entropy ⇒ 1

❗ Check which node is Pure and which is Impure? (C1 & C2)



For C_1

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

$$= 1$$

$H(S) = 1$

↳ Impure split

For C_2

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

$H(S) = 0$

↳ Pure split.

FEATURE TO TAKE NEXT

↳ INFORMATION GAIN

$$Gain(S, f_x) = H(S) - \sum_{v \in val} \frac{|S_v|}{|S|} H(S_v)$$

feature

★No need to understand formula

---→ For each node we go starting from root node we calculate Gain for each feature

$G(S, f_1)$

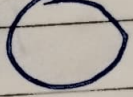
$G(S, f_2)$ ----

then select that feature there which has the highest Gain.

LOSS FUNCTION ⇒ GINI INDEX (DEFAULT)

$$G.I = 1 - \sum_{i=1}^{n} (p)^2$$

$$p^2 = (p+)^2 + (p-)^2$$



$$G.I = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = \frac{1}{2}$$

G.I ⇒ 0.5 ⇒ Impure Split
G.I ⇒ 0 ⇒ Pure Split.

Q In D.T which loss fxn to use when?
--→ when you have a large D.T with many parameter use Gini Index as loss fxn as it has less T.C as compare to Entropy which has log fxn in it.

R&D Nuclear

★ Decision Tree are very good at which column to include or not

Procedure

1. Choose the Best feature (Based on Gain) feature
2. Split the dataset of that category in category.
3. Repeat.
4. Stop when the leaf node is pure.

Q How to select the Best Feature? at each node?

↳ Two Way

(ID3) → If you using Entropy as loss fxn
↳ Calculate I.G for each feature & select the one with highest I.G

(CART Algo) → G.I (Default)
↳ Select feature with Gini Index

DECISION TREE REGRESSOR

↳ Loss Fxn ⇒ Mean Square Error.

Q How we select the best feature in Regression?

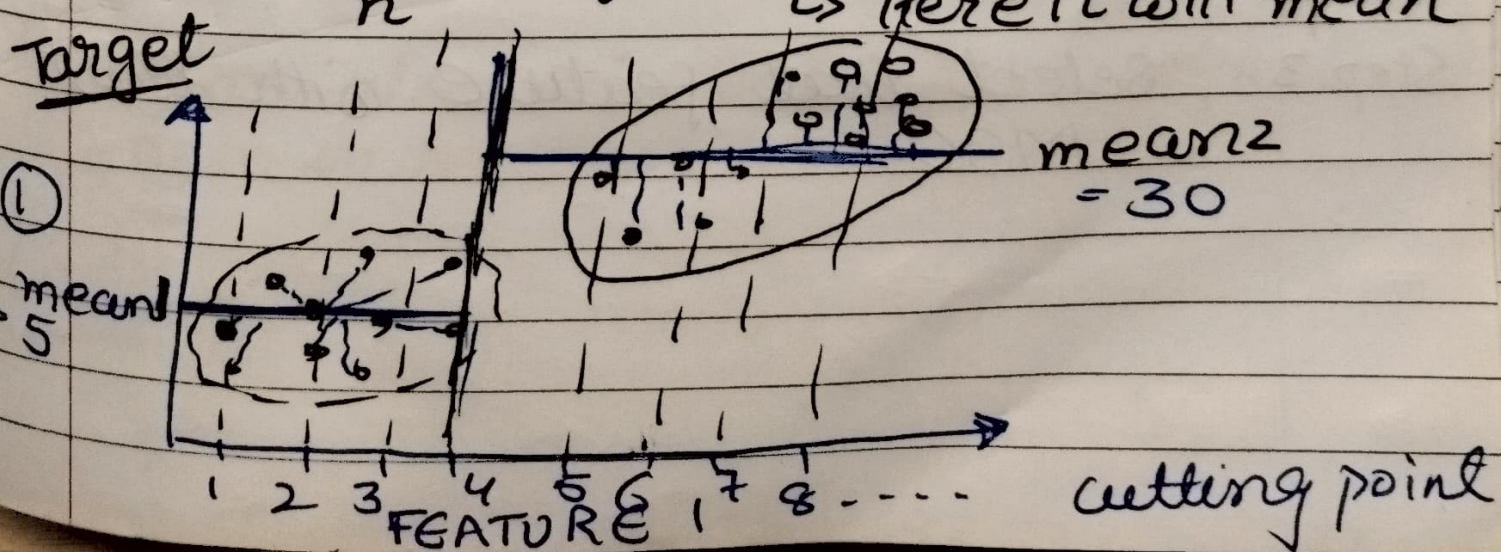
① First we calculate least MSE of each feature

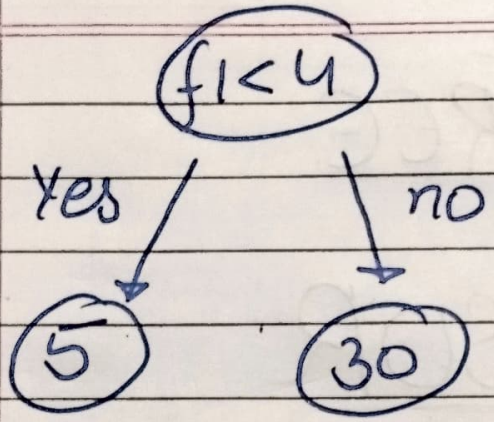
② Then select that feature among them which have the least MSE

MSE

↳ $\frac{1}{n} \Sigma (y - pred)^2$

↳ Here it will mean





---> Error = MSE1 (mean1) f1<4 + MSE2 (mean2) f>4

---> Error

Step (1) Calculating least MSE for a feature

i) Calculate MSE for each cutting point
ii) Select the cutting point with least MSE.

Step 2: Do This for all feature

Step 3: Select that feature with least MSE.

Problem with D.T

↳ ① Overfitting

↳ i) Tree Pruning
ii) Use Random Forest.
(Ensemble Method)

★ from sklearn.tree import DecisionTreeClass
-ifier

★
---> To See Graph (D.T)

↳
```
from sklearn import tree
plt.figure(figsize=(15,10))
tree.plot_tree(model, filled=True)
```

★ You can also check feature import
-ance in Decision Tree

Code