* Decision tree classifier

* Decision Tree classifier
  - ID3.
  - cart.

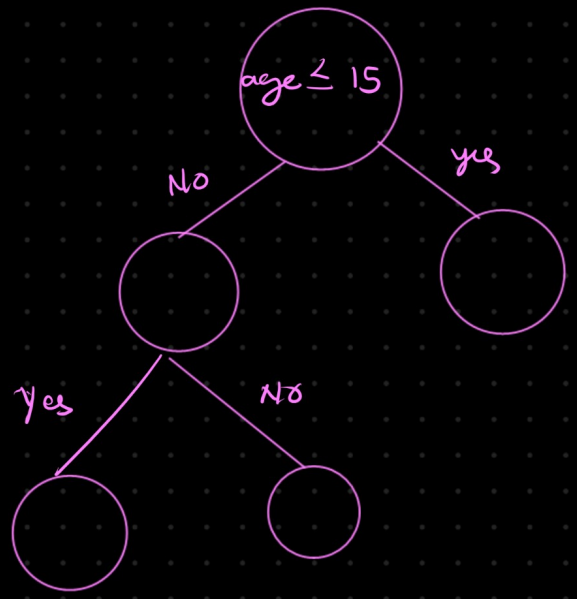
a) Entropy and giniindex → purity split.

b) Information gain → features to select for DT construction.

```
age=14
if (age ≤ 15)
    print ("sobaol")

elif (case<15 and age 21):
    print ("the person may ne college")

else:
    Print ("the person has passed")
```



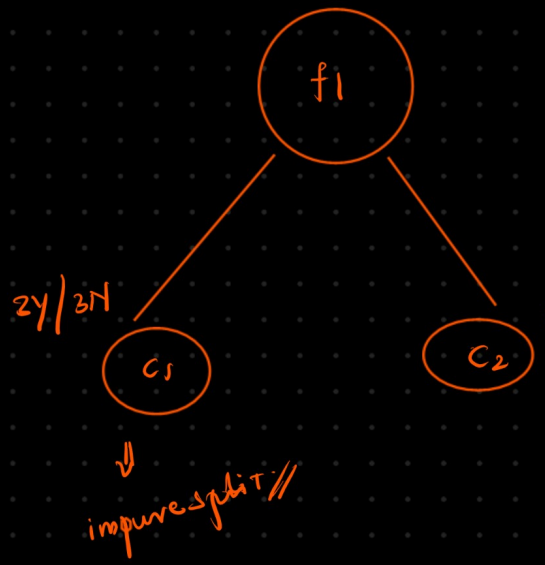
1) Entropy

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

{ binary classification }

one category → $-p_-$ = negative class

another category → $+p_+$ = positive class



$$H(C_2) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$$= -3/3 \log_2 3/3 - 0 \log_2 0$$

$$= -1 \log_2 1 = 0 = \text{pure split} //$$

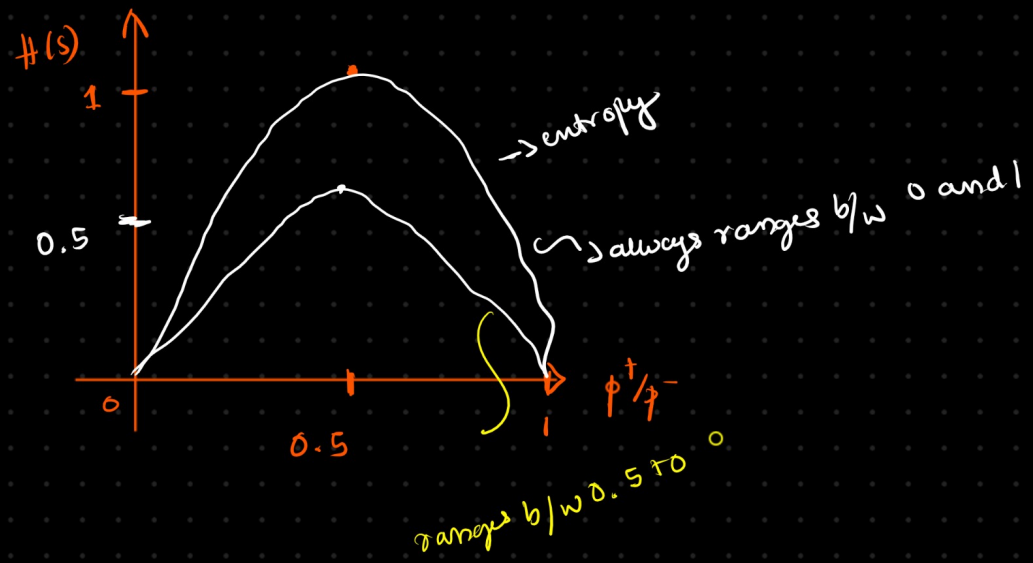
$$H(C_1) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$$= -3/6 \log_2 3/6 - 3/6 \log_2 3/6$$

$$= 1 // = \text{Impure Split.}$$

2) Gini Impurity

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



* Gini index

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

$$= 1 - ((p_+)^2 + (p_-)^2)$$

$$= 1 - ((1/2)^2 + (1/2)^2)$$

$$= 0.5 \Rightarrow \text{impure}$$

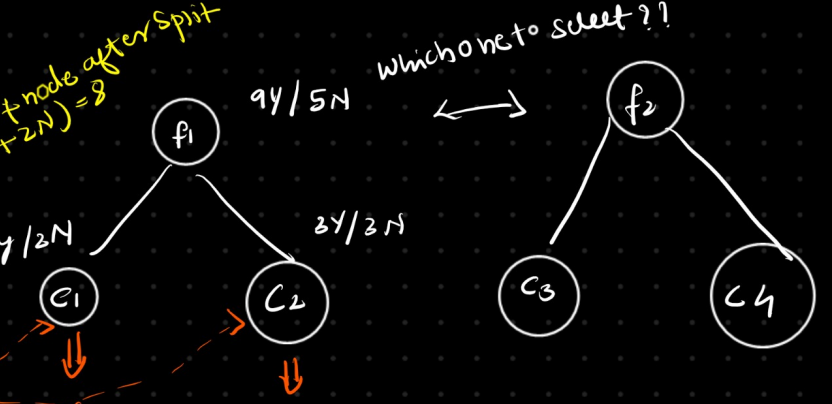
3Y/0N

$$= 1 - \left(\left(\frac{3}{3}\right)^2\right)$$

$$= 1 - 1$$

$$= 0 \;//\; \Rightarrow \text{pure}$$

Information gain



Entropy of root node

total number of node after split (6Y+2N) = 8

$$Gain(S, f_1) = H(S) - \sum_{v \in val} \frac{|S_v|}{|S|} H(S_v)$$

$|S|$ → no of dataset

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

entropy of these two nodes

$$= -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14}$$

$$\approx 0.94$$

$$H(C_1) = -\frac{6}{8} \log \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8}$$

$H(C_1) = 0.81$

$H(C_2) = 1$ //

$\frac{|S_v|}{|S|}$

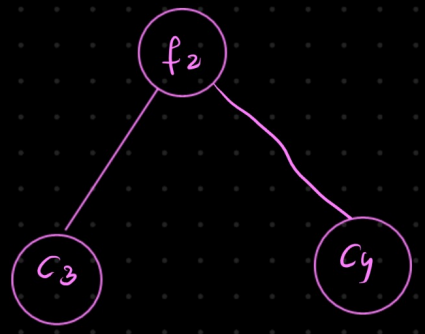
$$Gain(S, f_1) = 0.94 - \left[\frac{8}{14} * 0.8 + \frac{6}{14} * 1 \right]$$

$H(S_v)$

$IG(S, f_1) = 0.49$ // → with respect to f1

>>

$IG(S, f_2) = 0.51$ → with respect to f2 → it should start from f2 as root node.



When should we use Entropy and GINI index

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

$$\Downarrow$$

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

$$\longrightarrow \text{GiniIndex} = 1 - \sum_{i=1}^{n} (p_i)^2 \Rightarrow$$

* whenever the dataset is small use entropy"

* entropy → small dataset
* Gini index → large dataset

} → when we use decision tree classifier

Decision tree with numerical features

f1	o/p
2.3	Yes
3.6	Yes
4	No

threshold

Step - sort the values

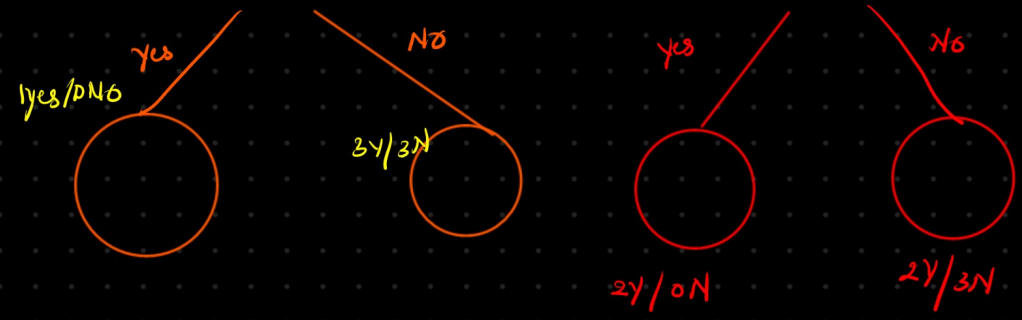
* threshold

≤ 2.3 | 4/4/3/4

* {information gain}

≤ 3.6

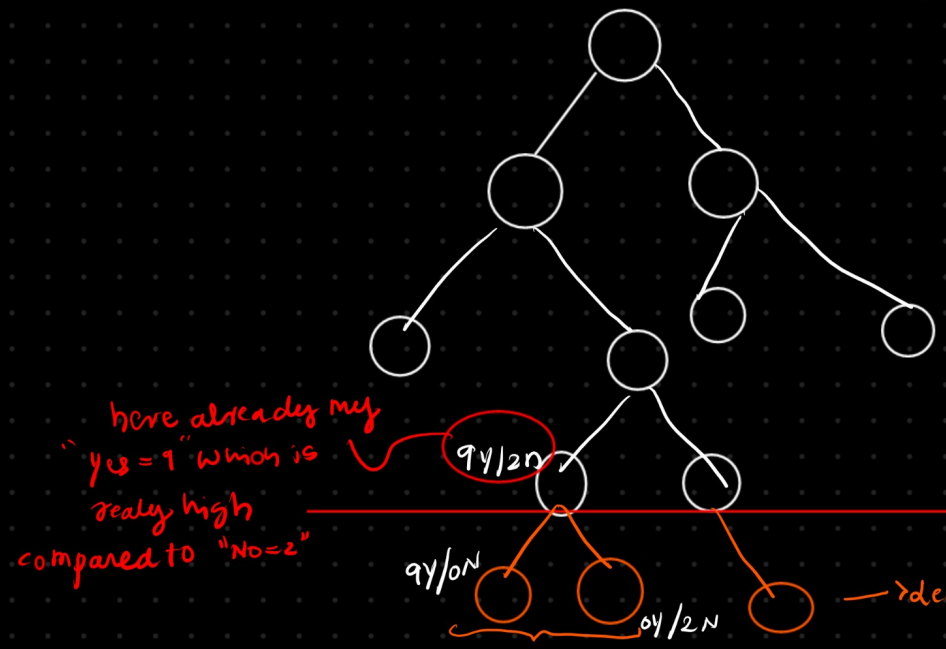
5.2	No
6.7	Yes
8.9	No
10.5	Yes



* there will be many splits like this which split will have larger information gain then we will consider that.

Post pruning & Pre pruning

* training dataset



- when we train the model until the leaf node our model will be "overfitting"
  - to reduce this we use
    - post-pruning.
    - pre-pruning.

→ post pruning
- ↳ construct DT
- ↳ prune it with respect to the
- ↳ for small length dataset

- even though i did the split these type of split will cause overfitting

① Post pruning

construction the DT → Prune it with respect to depth } used in smaller dataset.

② pre-pruning : tune with hyperparameters while constructing the decision tree

- max_depth = 3
- max_features
- split_ration
- criterion
- Splitter.

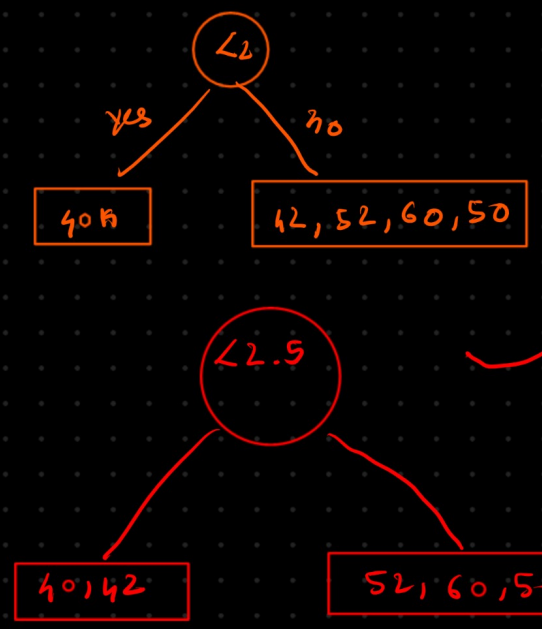
Decision Tree Regression

Dataset

Exp	gap	salary
2	Yes	40K
2.5	Yes	42K
3	No	52K
4	No	60K
4.5	Yes	50K

threshold

Continous value



between this which feature should i select. So we use "variance reduction"

① Variance $= \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2$ [mean Square Error]

$\bar{y}$ → Average.

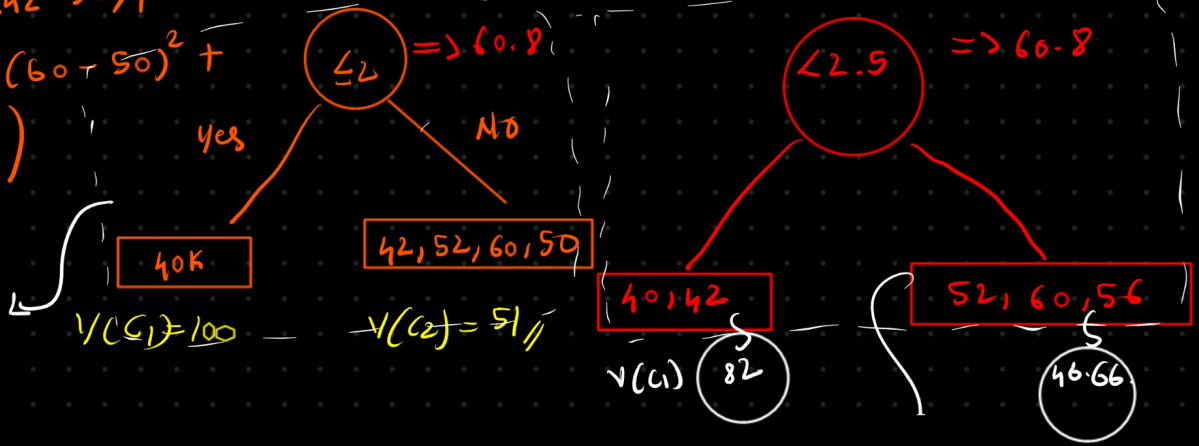
Variance (Root) $= \frac{1}{5}\left((40-50)^2+(42-50)^2+(52-50)^2+(60-50)^2+(56-50)^2\right)$

$= \frac{1}{5}[100+64+4+100+36]$

$= 60.8$ //

Variance reduction

[40K, 42K, 52K, 60K, 56K] [40K, 42K, 52K, 60K, 56K]



$$\text{Variance}(c_1) = \frac{1}{1}\left[(40-50)^2\right]$$

$$= 100 \;//$$

$$\text{Variance}(c_2) = \frac{1}{4}\left[(42-50)^2 + (52-50)^2 + (60-50)^2 + (56-50)^2\right]$$

$$= 51 \;//$$

$\sim 0.04 \;//$

Variance Reduction

$$= \text{Var}(\text{root}) - \left(w_j\right)\text{Var}(\text{child})$$

w_j → ratio of that particular child node to the total no.

$$= 60.8 - \left[\frac{1}{5} \times \frac{20}{100} + \frac{4}{5} \times 51\right]$$

$$= 60.8\left[20 + \frac{4}{5} \times 51\right]$$

$$= 60.8 - 20 - 40.8$$

Variance reduction = 0 < Variance Reduction = 0.04

↓ left split

↓ Right split