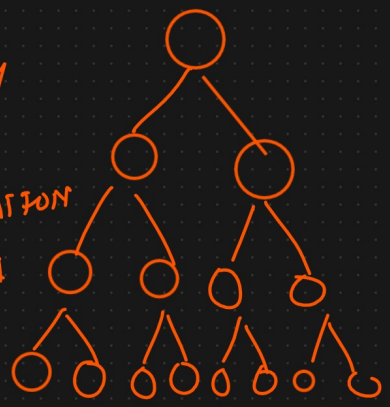
Decision Tree Regression :-

O/P is continuous feature.

Entropy
GINI
INFORMATION GAIN

CART



Dataset

Exp	Carcer Gap	Salary
2	Yes	40K
2.5	Yes	42K
3	No	52K
4	No	60K
4.5	Yes	56K

op

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

50K

60.8

≤2

Yes

No

40K

42, 52, 60, 50

V(C1) = 100

V(C2) = 51

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

≤2.5

60.8

Decision Tree Regressor

40, 42

52, 60, 56

Variance Reduction (Regression Problem)

It determines which value should select to split.

① Variance $= \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2$ → Average

Variance & Mean Squared Error is same in regression.

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

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

$$= 60.8$$

$$\text{Variance}(C1) = \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2$$

$$\text{Var}(C1) = \frac{1}{2}\left[(40-50)^2+(42-50)^2\right]$$

$$= \frac{1}{2}\left[100+64\right]$$

$$= \frac{164}{2} = 82$$

$$\text{Var(child2)} = \frac{1}{3}\left[4+100+36\right]$$

$$= \frac{140}{3} = 46.66$$

Variance Reduction

$$= 60.8 - \left[\frac{2}{5}\times 82 + \frac{3}{5}\times 46.66\right]$$

$$= 60.8 - \left[32.8 + 27.996\right]$$

$$= +0.004$$

$$= \frac{1}{1} (40-50)^2$$

$$= 100$$

$$\text{Variance (c2)} = \frac{1}{n} \sum_{i=1}^{n} (y-\bar{y})^2$$

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

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

$$= 51$$

Number of samples in child node.
Total number of samples in parent node.

Variance Reduction

$$= \text{Var(Root)} - \sum_{i=1}^{k} w_i \, \text{Var(child)}$$

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

$$= 60.8 - 20 - 40.8$$

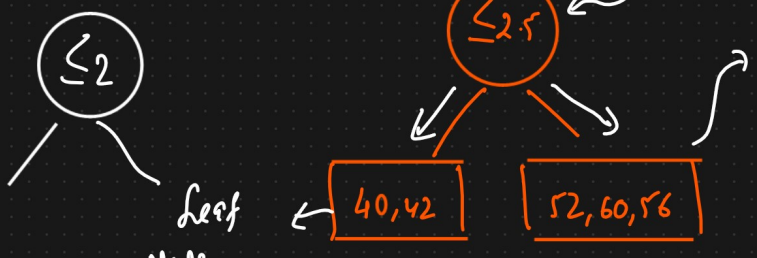
Variance Reduction = 0

0 0.004

Variance Reduction (Left Split) < VR (Right Split)

⇓

Left Tree



Decision Tree O/P Reg.

$\frac{52+60+56}{3} = 56$

$$\text{O/P} = \left[\frac{40+42}{2} \right]$$

$$= 41$$