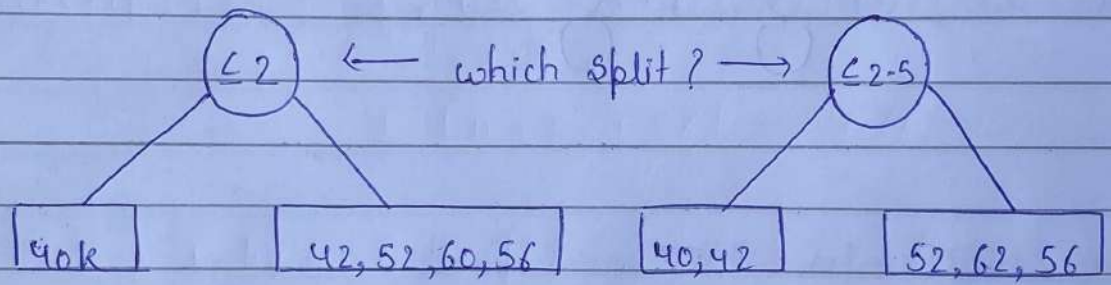
Decision Tree Regressor

→ Dataset :

	I/P		O/P
	Exp	Gap	Salary
→	2	Yes	40K
	2-5	Yes	42
	3	No	52
	4	No	60
	4-5	Yes	56

Sorted

$\hat{y} = 50k$



Note: In case of continuous feature, variance reduction is used to select split to choose.

* Variance Reduction :

$$\text{variance} = \frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2 \Rightarrow \text{MSE}$$

$\hat{y}$ → Average of O/P

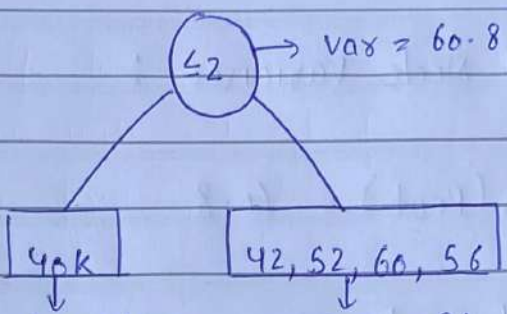
$$VR = Var(Root) - \sum w_i \, Var(child)$$

where, Var → Variance

w → weightage → $\frac{\text{child Node count}}{\text{Root Node count}}$

→ Selecting the split :

Split 1 :



• Root Node Variance :

$$Var(Root) = \frac{1}{n} \Sigma (y - \hat{y})^2$$

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

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

$$= \frac{304}{5} = 60.8$$

• Child Node Variance :

$$Var(child\ 1) = \frac{1}{1} \times \left[(40-50)^2 \right] = 100$$

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

$$= \frac{1}{4} \times \left[64 + 4 + 100 + 36 \right] = 51$$

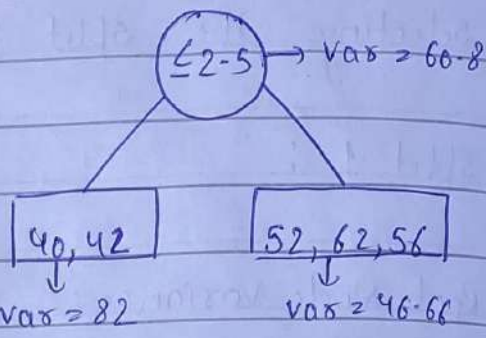
• Variance Reduction of split 1 :

$$VR_1 = Var(Root) - \Sigma w_i\, Var(child)$$

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

$$VR_1 = 0$$

Split 2 :



- Root Node Variance :

$Var(Root) = 60.8$

- child Node Variance :

$$Var(child\ 1) = \frac{1}{2} \times [(40-50)^2 + (42-50)^2]$$

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

$$var(child\ 2) = \frac{1}{3} \times [(52-50)^2 + (62-50)^2 + (56-50)^2]$$

$$= \frac{1}{3} \times [4 + 100 + 36] = 46.66$$

- variance Reduction of split 2 :

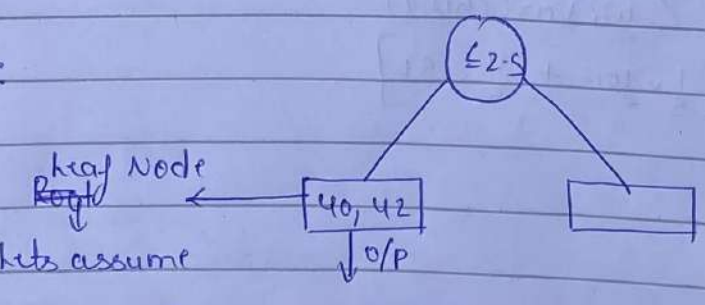
$$VR_2 = 60.8 - \left[\frac{2}{5} \times 82 + \frac{3}{5} \times 46.66\right]$$

$$VR_2 = 0.304.$$

$VR_2 > VR_1$

∴ split 2 will be selected for splitting.

Note :



Average → $\frac{40+42}{2} = 41.$