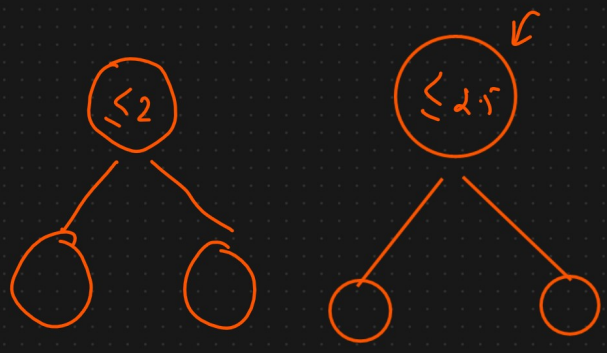
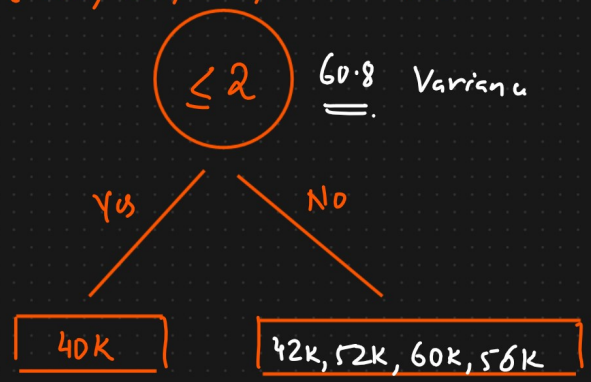
Decision Tree Regressor

Dataset

Exp	Career Gap	Salary
→ 2	Yes	40K
→ 2.5	Yes	42K
3	No	52K
4	No	60K
4.5	Yes	56K
		50K = $\bar{y}$



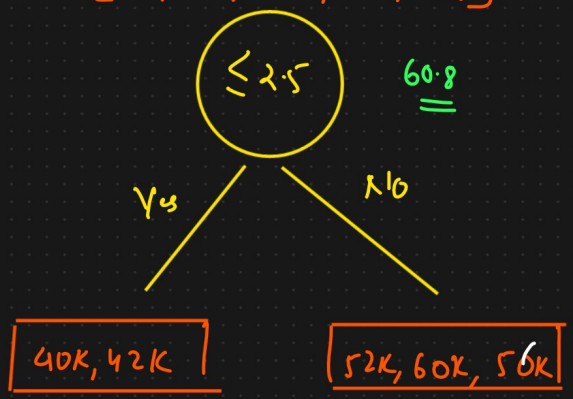
[40K, 42K, 52K, 60K, 50K]



100 (Variance)

Variance (51)

[40K, 42K, 52K, 60K, 50K]



Variance 82

Variance Reduction

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$ [Mean Squared Error].

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

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

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

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

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

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

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

$= 60.8$

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

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

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

$\text{Variance (left)} = 100$

$\text{Variance Reduction} = 60.9 - \left[\frac{2}{5} * 82 + \frac{3}{5} * 46.66\right]$

$= 0.004$

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

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

$= 51$

$Wi(L) = \frac{1}{5} \quad Wi(R) = 4/5$

$\text{Variance Reduction} = \text{Var(Root)} - \sum Wi\,\text{Var(child)}$

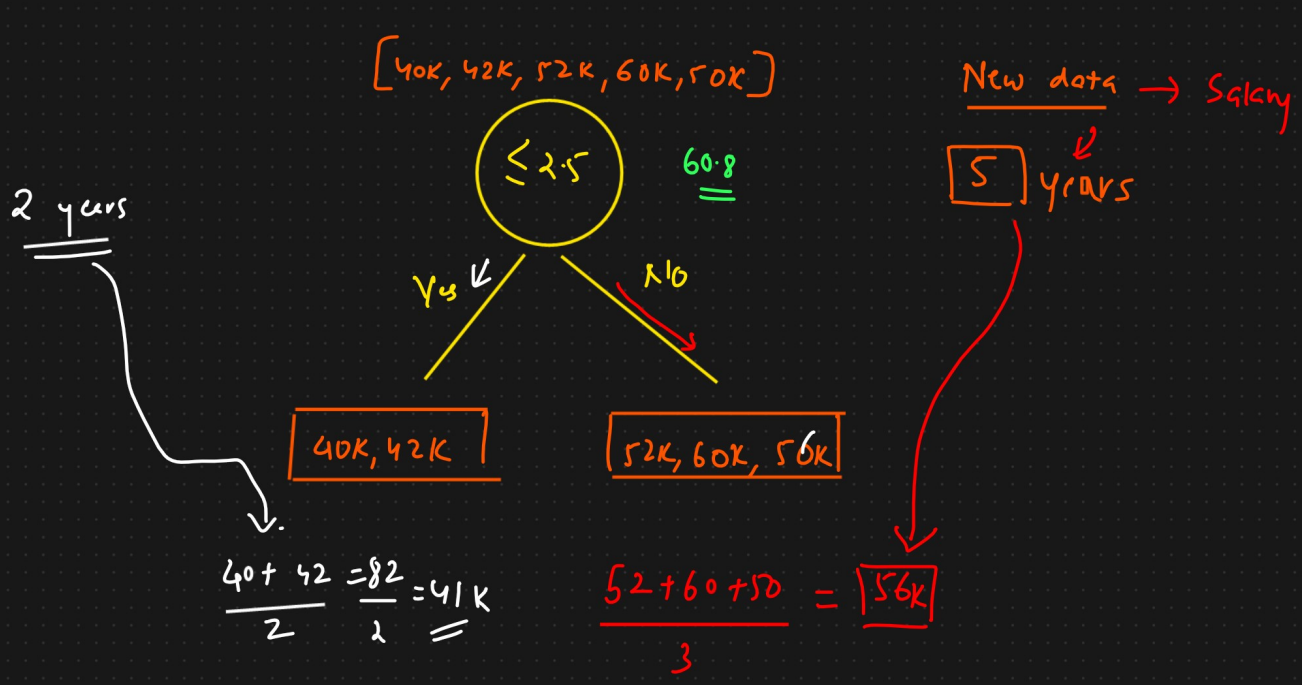
$= 60.8 - \left[\frac{1}{5} * 100 + \frac{4}{5} * 51\right]$

$= 60.8 - 20 - 40.8$

Variance Reduction $= 0$

0.004

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


[40K, 42K, 52K, 60K, 50K]
New data → Salary
≤ 2.5
60.8
5 years
Yes
No
2 years
40K, 42K
52K, 60K, 56K
40 + 42 / 2 = 82 / 2 = 41K
52 + 60 + 50 / 3 = 56K