Decision Tree Regressor

DT Classifier
→ Entropy / Gini
→ Information Gain

DT Regressor
→ Variance.
↳ Variance reduction

Experience	Career Gap	Salary
2	Yes	40K
2.5	Yes	42K
3	No	52K
4	No	60K
4.5	yes	56K

$\bar{y} = 50k$

① [40K, 42K, 52K, 60K, 56K]
≤ 2
Yes → 40K
No → 42K, 52K, 60K, 56K

* final aim is variance reduction.
* you will split whichever feature / feature value gives the highest feature reduction.

② ≤ 2.5
40K, 42K 52K 60K, 56K
40, 42
52, 60, 56

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \quad \text{[mean square Error]}$$

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

$$= \frac{1}{5}(100 + 64 + 4 + 100 + 36)$$

$$= 60.8$$

$$\text{Variance}(L\cdot C) = \frac{1}{1}\left((40-50)^2\right)$$

$$= 100$$

$$\text{Variance}(R\cdot C) = \frac{1}{4}\left(8^2 + 2^2 + 10^2 + 6^2\right)$$

$$= \frac{1}{4}(64 + 4 + 100 + 36)$$

$$= 51$$

$\text{Var}(P(\text{root})) = 60.8$

$\text{Var}(\text{left child}) = \frac{1}{2}(100 + 64)$

$= 82$

$\text{Var}(\text{right child}) = \frac{1}{3}(4 + 36 + 100)$

$= 46.66$

Variance Reduction = Var$_p$ − Var$_{child\ combined}$

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

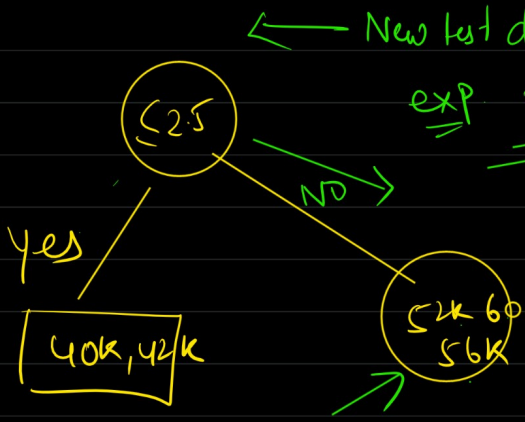
$= 0$

$60.8 - \left(\frac{2}{5} \times 82 + \frac{3}{5} \times 46.66\right)$

$= 0.004$

Variance reduction for 2 < Variance Reduction for 2.5

Select this split of the feature.



Pred for 3 years = $\frac{52 + 60 + 56}{3}$

$= 56k.$