Decision Tree Regression

Dataset

Exp	Gap	Salary (O/P)
2	Yes	40 K
25	Yes	42 K
3	NO	52 K
4	NO	60 K
4.5	Yes	56 K

here the O/P feature is continous, hence we will have to use regression

In decision tree we cannot use

- → Entropy
- → Gini Impurity
- → Information gain

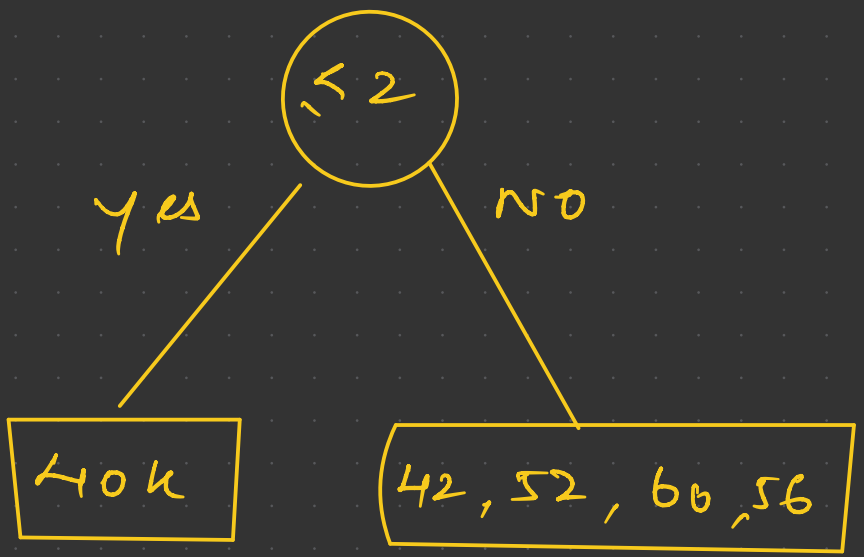
↓

In order to use this output has to be fixed number of categories, but in case of regression output will be continous.

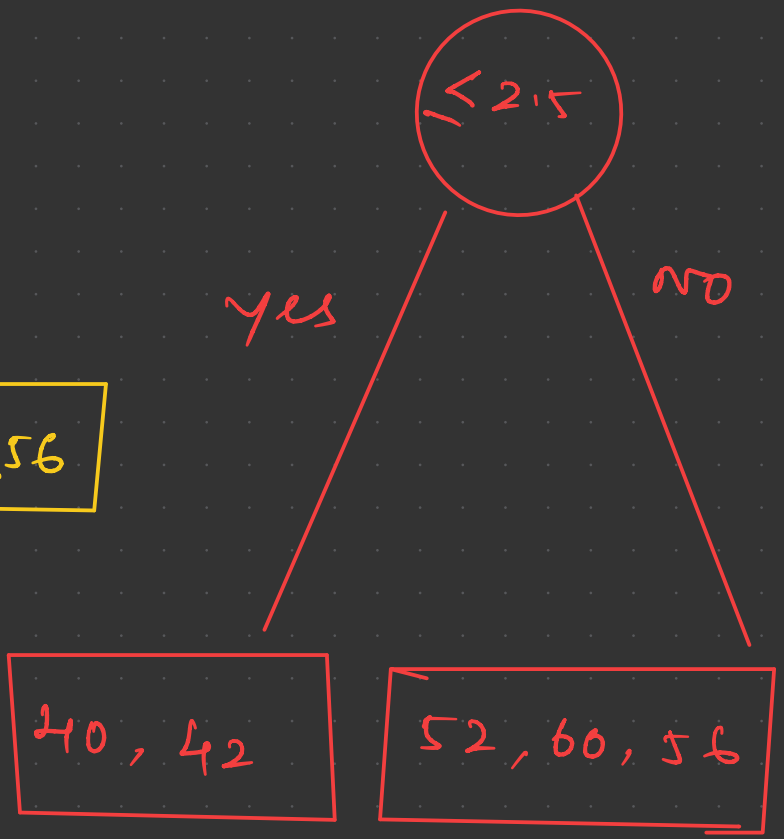
Dataset

Exp	Gap	Salary (o/p)
2	yes	40 K
2.5	yes	42 K
3	NO	52 K
4	NO	60 K
4.5	yes	56 K

Threshold : 2



Threshold : 2.5



how we select, which split needs to be selected.?

→ we will be using Variance Reduction to select the best split for the regression Problem.

In Variance Reduction, we need to

① compute the variance.

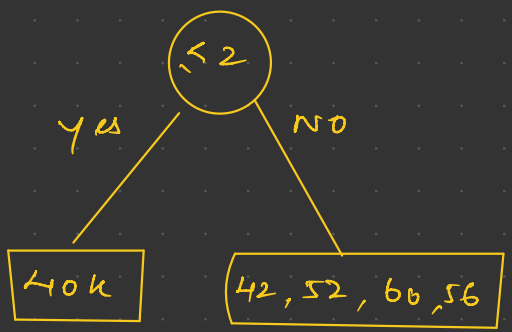
$$\text{Variance} = \frac{1}{n} \sum_{i=1}^{n} (y - \bar{y})^2$$

↳ average.

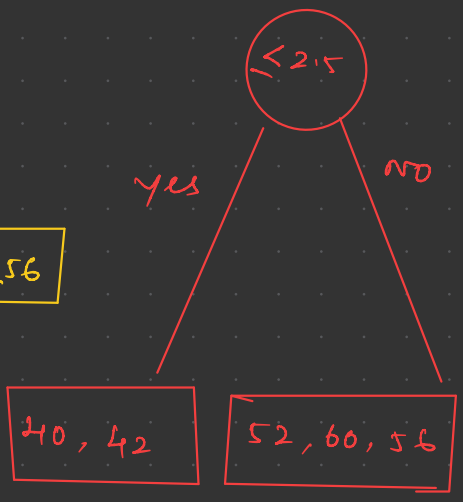
② we will be finding the mean square error for each and every node and combine them to find out the variance reduction

③ Select the split which has maximum variance reduction.

[40, 42, 52, 60, 56]



[40, 42, 52, 60, 56]



$\bar{Y} = 50\,k$

⇓

mean of o/p

Variance of root:

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

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

$$= 60.8$$

Variance of child 1

$$\text{Variance}(C_1) = \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2$$

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

$$\text{Variance}(C_1) = 100$$

Variance of child 2

$$\text{Variance}(C_2) = \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2$$

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

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

$$\text{Variance}(C_2) = 51$$

Variance Reduction

$$VR = Var(root) - \sum \omega_i \, var(child)$$

$\omega_i \longrightarrow$ Ratio of the Split

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

$$= 60.8 - \left[20 + 40.8\right]$$

$= 0 \longrightarrow$ Variance reduction of first split

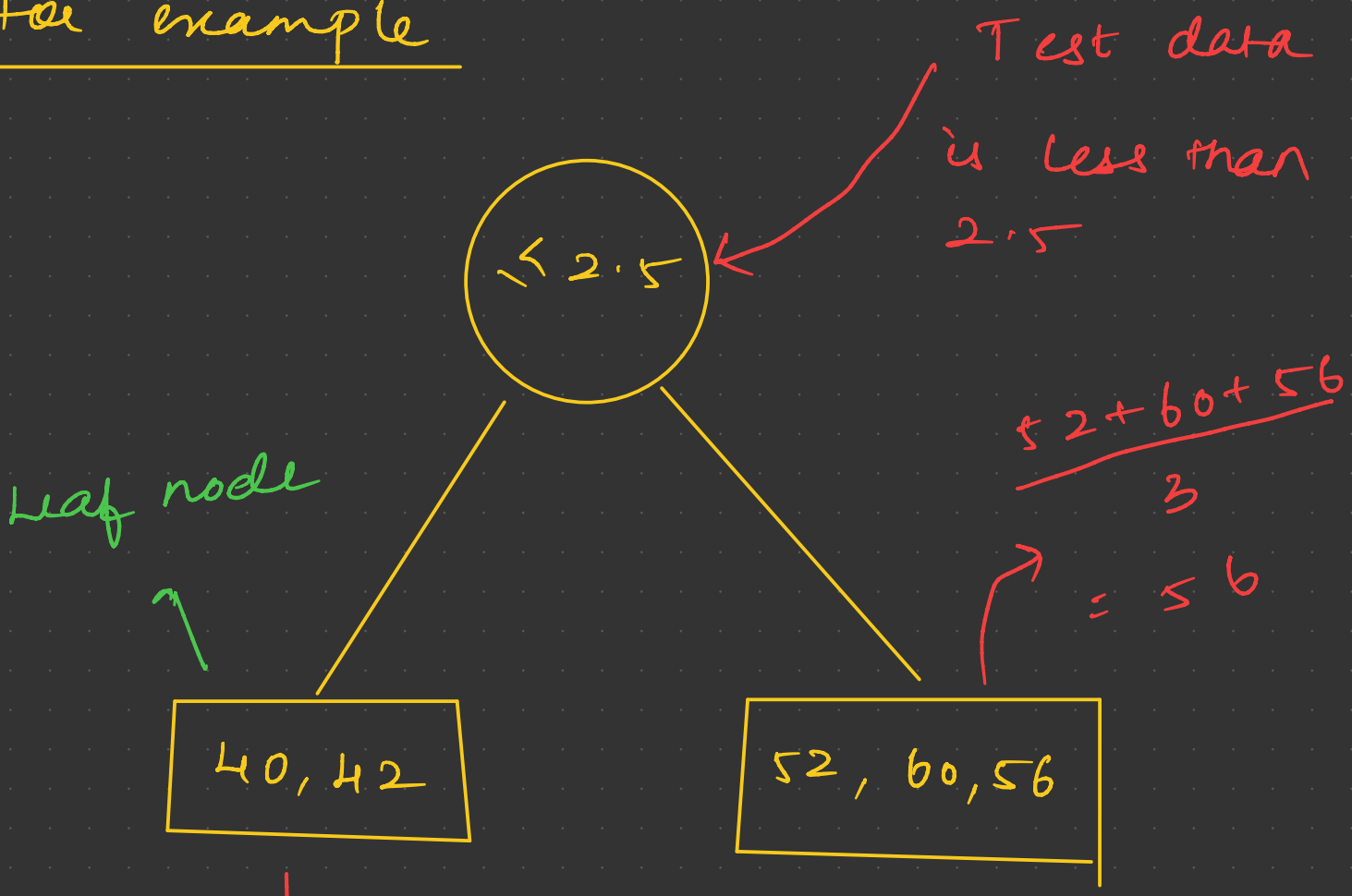
Similarly, after computing the Variance reduction of the second split

Variance Reduction of split 2 is

0.004.

hence we will be choosing the split 2 for Decision Tree Regression.

For example



Output = $\left[\frac{40 + 42}{2}\right]$

= 41

output will be the average of the leaf node.