3.4 :

Def^n of Information Gain:

for feature x and target y,

$$IG(y|x) = H(y) - H(y|x)$$

$H(y)$ = Entropy, $H(y|x)$ is the conditional entropy of y given x

Given : splitting feature has non zero IG

$$\Rightarrow IG(y|x) > 0$$

$$\Rightarrow H(y) - H(y|x) > 0$$

$$\Rightarrow H(y) > H(y|x)$$

wkt $H(y|x) = \sum_{v \in val(x)} P(x=v) H(y|x=v)$

- for $IG(y|x) > 0$, there must be a reduction in uncertainity about y when x is known $\Rightarrow$ knowing x $\Rightarrow$ useful information on y.
- But, if all training samples are sent to only 1 child node, then x wouldnot provide any additional information i.e.

$$IG(y|x) = 0$$

$\therefore$ for $IG(y|x) \neq 0$ or > 0, atleast one training sample is to be sent to each of the child nodes which ensures split differentiates b/w different outcomes of y.

This differentiation reduces overall entropy & results in positive IG.