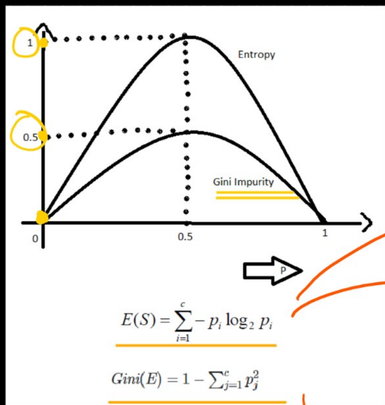
Gini vs Entropy

↳ Impurity



Highest Impure → 1

logarithmic ⟶ computationally expensive

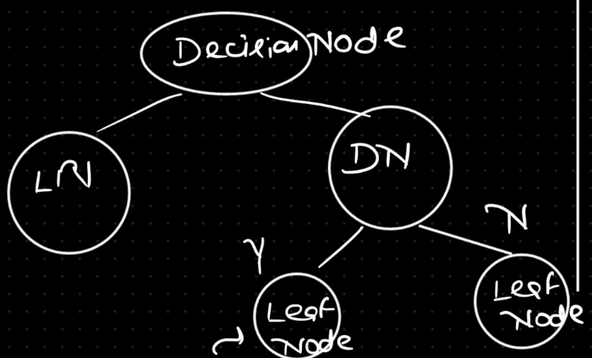
Highest Impure → 0.5

Default (Documentation)

Gini

Best case

Sample (Highest Pure)



$$\left\{ \begin{array}{l} P(Y) = 1 \\ P(N) = 0 \end{array} \right\} \longrightarrow \text{certain}$$

$$1 - \sum_{i=1}^{c} p_i^2 = 1 - 1 = 0 \text{ (Minimum)}$$

Worst case

Sample (Highest Impure)

$$\left. \begin{array}{l} P(Y) = 1/2 \\ P(N) = 1/2 \end{array} \right\} \text{Highly Confused}$$

$$1 - \sum_{i=1}^{c} p_i^2 = 1 - \left(\tfrac{1}{2}\right)^2 - \left(\tfrac{1}{2}\right)^2$$

$$= 1 - 1/4 - 1/4$$

$$= 1 - 1/2 = 0.5$$

Entropy

best case

worst case

Highest Purity	Highest Impure
$P(Y) = 1$	$P(Y) = \frac{1}{2}$
$P(N) = 0$	$P(N) = \frac{1}{2}$
$\mathcal{E} = -\sum_{i=1}^{c} P_i \log P_i$	$\mathcal{E} = -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2}$
$= -1 \log_2 1$	$\mathcal{E} = 1$
$\Rightarrow 0$	

Real time use-cases

- Gini Index → Slower training time
- Entropy → balanced decision tree

MSE → ↓↓ } Decision Tree Regressor

Leaf Node → Prediction of the Continuous Value