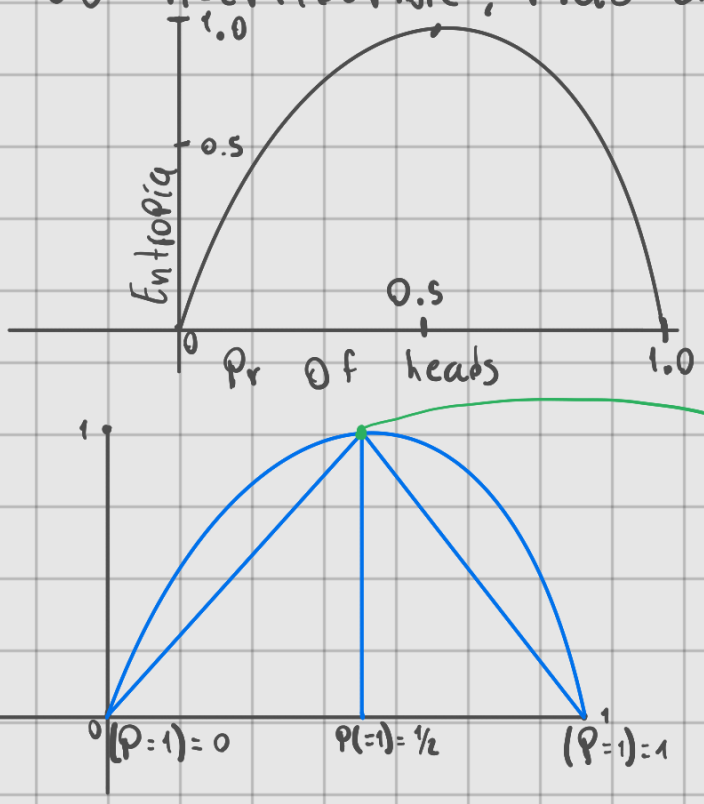
# Midiendo el desorden

Entropía $H(y)$ de una variable $y$

$$H(y) = -\sum_{i=1}^{k} P(y = y_i) \log_2 P(y = y_i)$$

Más incertidumbre, más entropía!




Datos completamente desordenados → Entropía máxima

Tenemos dos clases

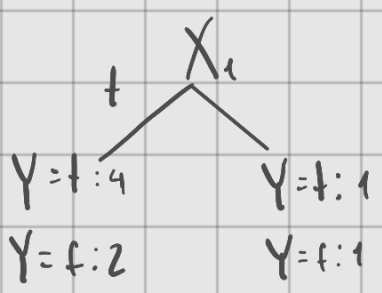
$P(=1) = \frac{N_1}{N}$ $P(=0) = \frac{N_0}{N} = 1 - \frac{N_1}{N}$

$$H(\langle N_0, N_1 \rangle) = \frac{N_0}{N} \log_2 \left(\frac{N_0}{N}\right) - \frac{N_1}{N} \log_2 \left(\frac{N_1}{N}\right)$$

$$\log_2 X = \frac{\ln X}{\ln 2}$$

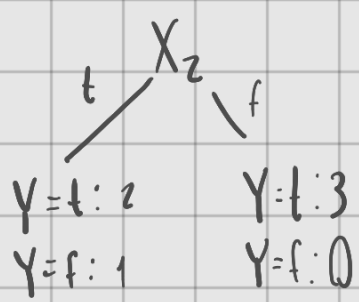
$P(X_1 = t) = 4/6$

$P(X_1 = f) = 2/6$




$$H(\langle X_1, y \rangle) = -\frac{4}{6}(1 \log_2 1 + 0 \log_2 0) - \frac{2}{6}(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2}) = 2/6$$

$X_2$




$$H(\langle Y, X_2 \rangle) = \frac{3}{6}\left[-\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3}\right] + \frac{3}{6}[0 \log_2 0 - 1 \log_2 1]$$

$$3/6\,[0.39 + 0.53] = 0.46$$

## Ganancia De Información

Entropía → $H(y)$ ; Entropía condicional → $H(y|x)$

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

$$H(y) - H(y|x_i)$$

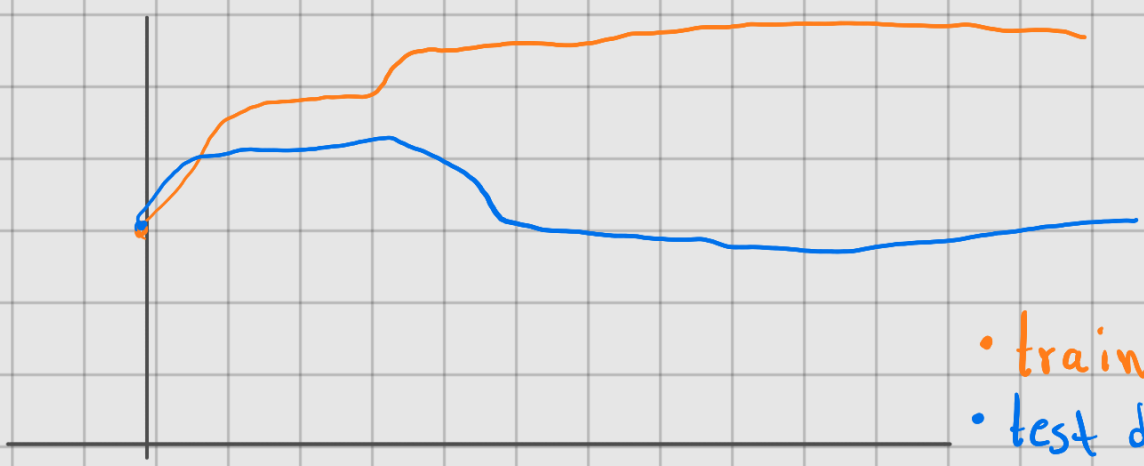
$$0.65 - 0.33 \simeq 0.32$$

- Empieza con un DT vacío
- Divide en el sig mejor atributo

$$\arg\max IG(x_i) = \arg\max H(y) - H(y|x_i)$$

- Recursividad

Si limitas mucho el árbol, pierdes mucha precisión



- training data
- test data