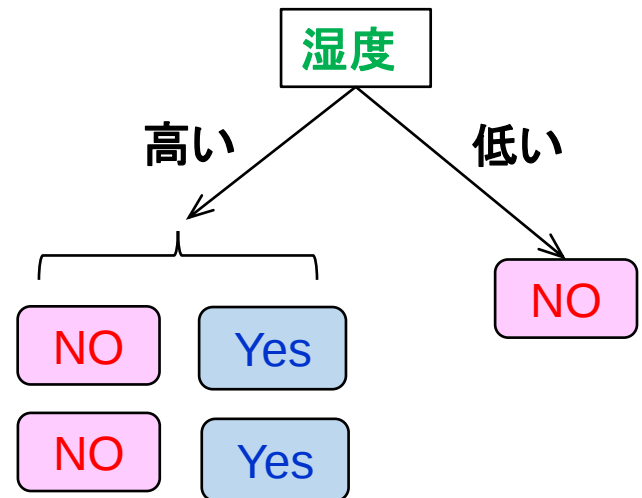
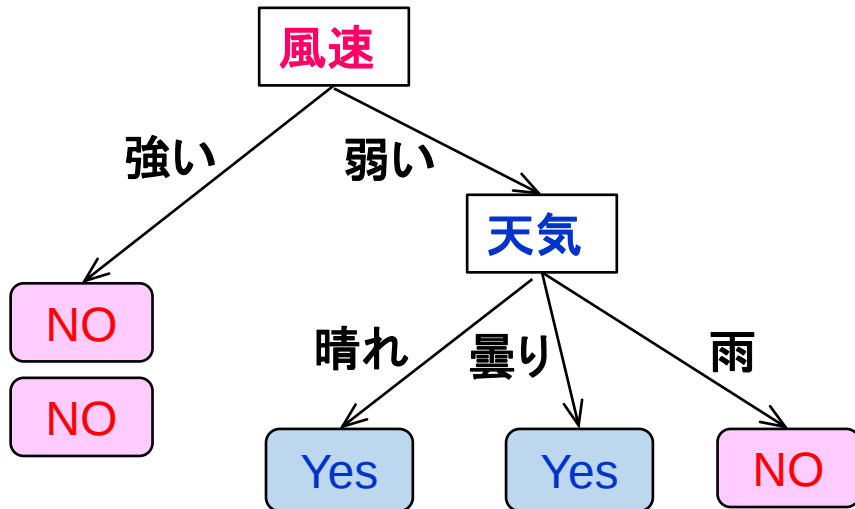
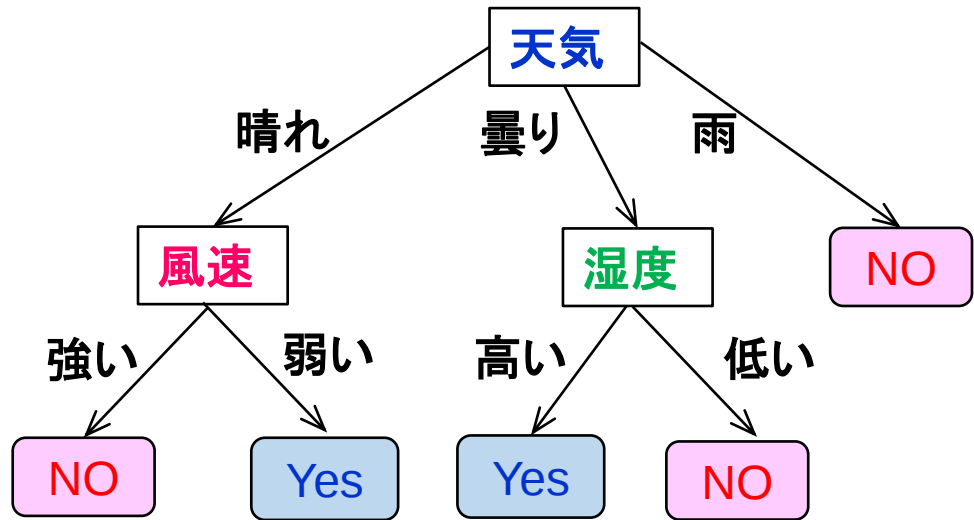


決定木

天気	風速	湿度	遠足
晴れ	強い	高い	No
曇り	弱い	高い	Yes
曇り	強い	低い	No
晴れ	弱い	高い	Yes
雨	弱い	高い	No



平均情報量(エントロピー)

各クラスのデータ数を、 m 個、 n 個とおく

$$I\left(\frac{m}{m+n}, \frac{n}{m+n}\right) = \frac{m}{m+n} \log_2 \frac{m+n}{m} + \frac{n}{m+n} \log_2 \frac{m+n}{n}$$

	分離良好	分離良好	まあまあ分離	まあまあ分離	分離不十分
Yes	●●●●		●●●	●	●●
No		●●●●	●	●●●	●●
I	0	0	0.81	0.81	1

$I\left(\frac{4}{4}, \frac{0}{4}\right) = \frac{4}{4} \log_2 \frac{4}{4} + \frac{0}{4} \log_2 \frac{4}{0} = 0$

$I\left(\frac{0}{4}, \frac{4}{4}\right) = \frac{0}{4} \log_2 \frac{4}{0} + \frac{4}{4} \log_2 \frac{4}{4} = 0$

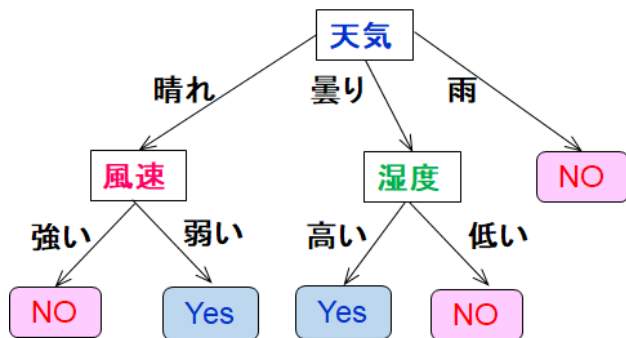
$I\left(\frac{2}{4}, \frac{2}{4}\right) = \frac{2}{4} \log_2 \frac{4}{2} + \frac{2}{4} \log_2 \frac{4}{2} = 1$

$I\left(\frac{3}{4}, \frac{1}{4}\right) = \frac{3}{4} \log_2 \frac{4}{3} + \frac{1}{4} \log_2 \frac{4}{1} = 0.81$

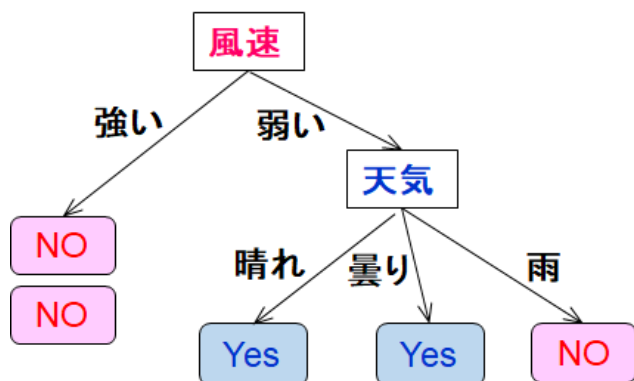
$I\left(\frac{1}{4}, \frac{3}{4}\right) = \frac{1}{4} \log_2 \frac{4}{1} + \frac{3}{4} \log_2 \frac{4}{3} = 0.81$

一般式

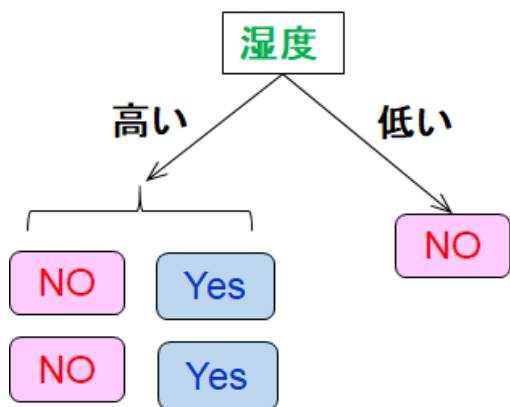
$$I(P_1, P_2, \cdots P_n) = \sum_{i=1}^n P_i \log_2 \frac{1}{P_i} = - \sum_{i=1}^n P_i \log_2 P_i$$



晴れ	曇り	雨	最初の状態
Yes=1 No=1	Yes=1 No=1	Yes=1 No=0	$I\left(\frac{2}{5}, \frac{3}{5}\right) = \frac{2}{5} \log_2 \frac{5}{2} + \frac{3}{5} \log_2 \frac{5}{3} = 0.972$
$I = 1$	$I = 1$	$I = 0$	
2/5	2/5	1/5	
$\frac{2}{5} \times 1 + \frac{2}{5} \times 1 + \frac{1}{5} \times 0 = 0.8$			$\text{ゲイン} = 0.972 - 0.8 = 0.172$



強い	弱い	最初の状態
Yes=0 No=2	Yes=2 No=1	$I\left(\frac{2}{5}, \frac{3}{5}\right) = \frac{2}{5} \log_2 \frac{5}{2} + \frac{3}{5} \log_2 \frac{5}{3} = 0.972$
$I = 0$	$I = 0.915$	
2/5	3/5	
$\frac{2}{5} \times 0 + \frac{3}{5} \times 0.915 = 0.549$		$\text{ゲイン} = 0.972 - 0.549 = 0.423$



高い	低い	最初の状態
Yes=2 No=2	Yes=0 No=1	$I\left(\frac{2}{5}, \frac{3}{5}\right) = \frac{2}{5} \log_2 \frac{5}{2} + \frac{3}{5} \log_2 \frac{5}{3} = 0.972$
$I = 1$	$I = 0$	
4/5	1/5	
$\frac{4}{5} \times 1 + \frac{1}{5} \times 0 = 0.8$		$\text{ゲイン} = 0.972 - 0.8 = 0.172$

最初の状態 (Yes=2、No=3)

$$I\left(\frac{2}{5}, \frac{3}{5}\right) = \frac{2}{5} \log_2 \frac{5}{2} + \frac{3}{5} \log_2 \frac{5}{3} = 0.972$$

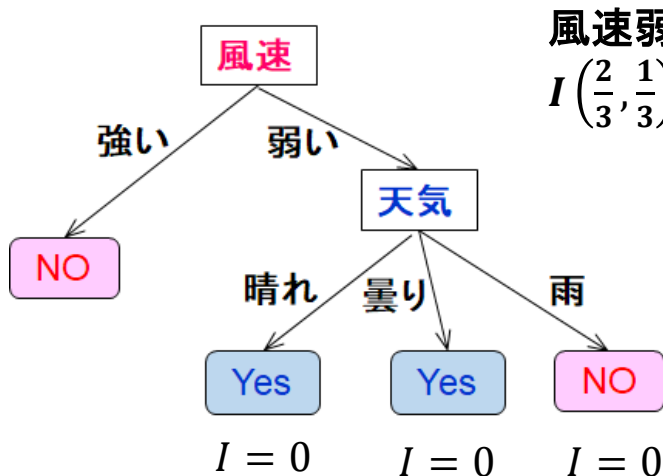
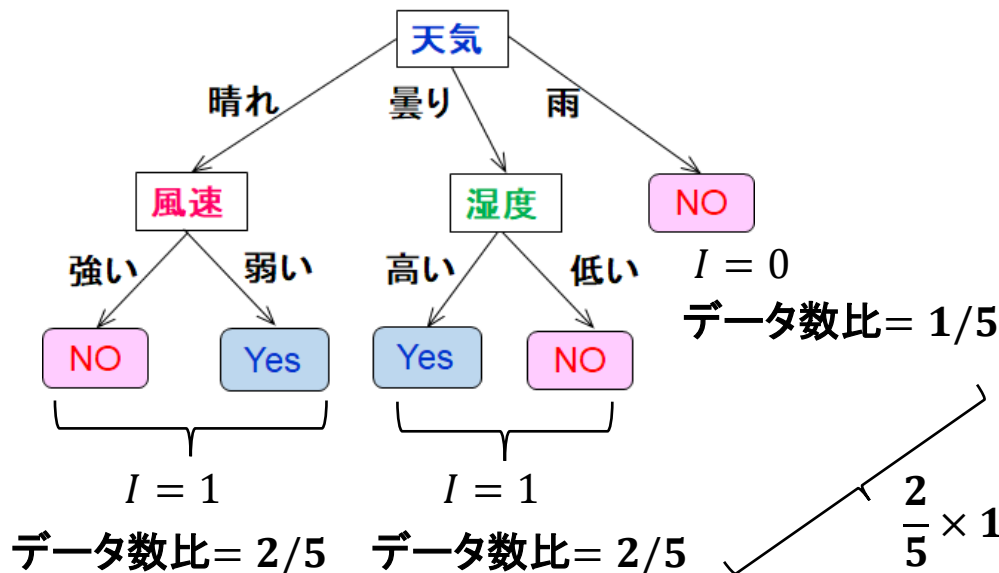
$$\text{天気}のゲイン = 0.972 - 0.8 = 0.172$$

$$\frac{2}{5} \times 1 + \frac{2}{5} \times 1 + \frac{1}{5} \times 0 = 0.8$$

最初に風速で分類した方がゲインが大きい

$$\text{天気}のゲイン = 0.918 - 0 = 0.918$$

$$\frac{1}{3} \times 0 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0 = 0$$



データ数比 = 1/3 データ数比 = 1/3 データ数比 = 1/3