2.4.1.5 Gain Ratio

Due to gain tends to favor test with many classes

C4.5 normalizes gain with a split information

$$\text{Split Info}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2\left(\frac{|D_j|}{|D|}\right)$$

$$\text{Gain Ratio}_A(D) = \frac{\text{Gain}(A)}{\text{Split Info}_A(D)}$$

Q: compute the Gain Ratio(D) of 4 attribute.

PID	Fever	Cough	Sore Throat	Tiredness	Flu
1	no	yes	no	yes	−
2	no	yes	no	no	−
3	mild	yes	no	yes	+
4	yes	mild	no	yes	+
5	yes	no	yes	yes	+
6	yes	no	yes	no	−
7	mild	no	yes	no	+
8	no	mild	no	yes	−
9	no	no	yes	yes	+
10	yes	mild	yes	yes	+
11	no	mild	yes	no	+
12	mild	mild	no	no	+
13	mild	yes	yes	yes	+
14	yes	mild	no	no	−

From 2.4.1.4.1

Gain(Fever) = 0.246

Gain (Cough) = 0.029

Gain (Sore Throat) = 0.151

Gain (Tiredness) = 0.048

Solution : ① Fever

$$SplitInfo_{Fever}(D) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{4}{14}\log_2\frac{4}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.5305 \times 2 + 0.5164$$

$$= 1.5774$$

$$GainRatio_{Fever}(D) = \frac{0.246}{1.5774}$$

$$= 0.156$$

② cough

$$SplitInfo_{cough}(D) = -\frac{4}{14}\log_2\frac{4}{14} - \frac{6}{14}\log_2\frac{6}{14} - \frac{4}{14}\log_2\frac{4}{14}$$

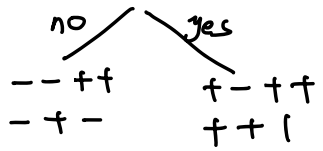
$$= 1.557$$

$$GainRatio_{cough}(D) = \frac{0.029}{1.557}$$

$$= 0.0186$$

$$= 0.019$$

③ Sore Throat

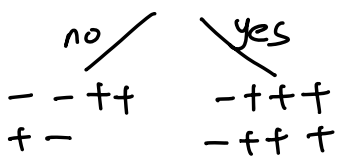


$$\text{Split Info}_{\text{Sore Throat}}(D) = -\frac{7}{14}\log_2\frac{7}{14} - \frac{7}{14}\log_2\frac{7}{14}$$

$$= 1$$

Gain Ratio (Sore Throat) = 0.151

④ Tiredness



$$\text{Split Info}_{\text{Tiredness}}(D) = -\frac{6}{14}\log_2\frac{6}{14} - \frac{8}{14}\log_2\frac{8}{14}$$

$$= 0.9852$$

$$\text{Gain Ratio Tiredness}(D) = \frac{0.048}{0.9852}$$

$$= 0.0487$$

⑤ Due to

Fever	cough	S.T.	T.
0.156	0.019	0.151	0.0487

Fever still got highese Gain Ratio.