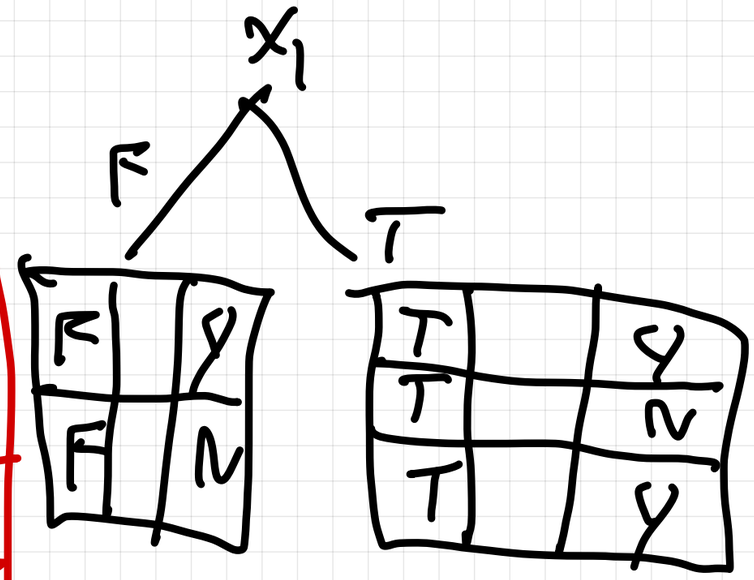
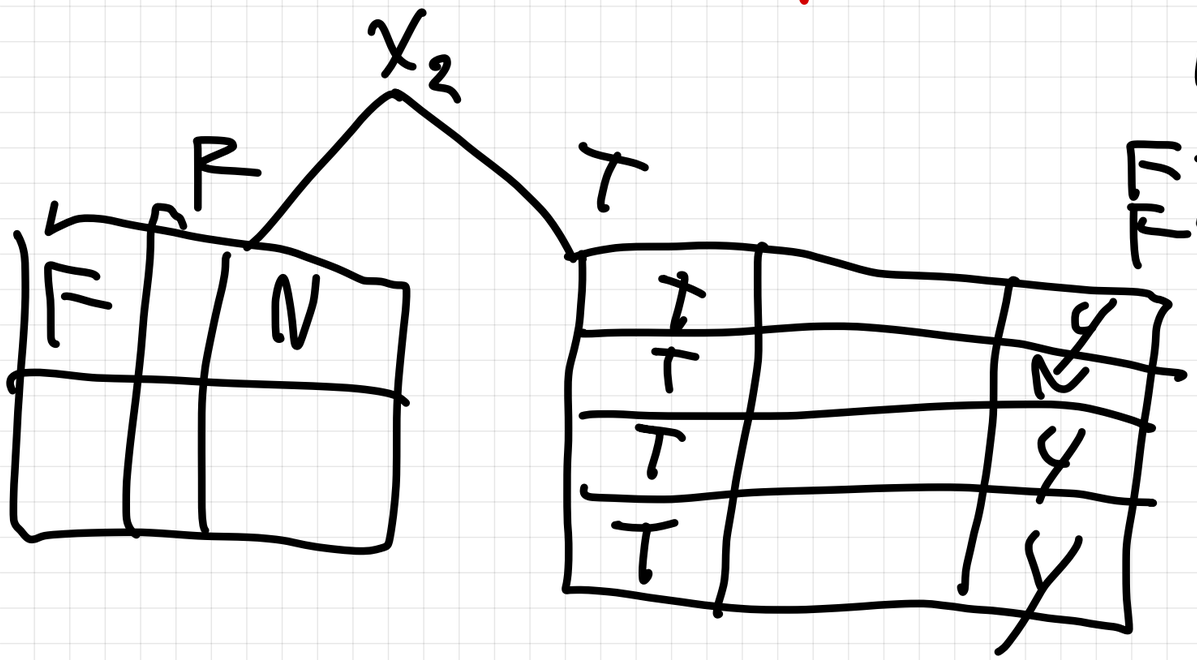
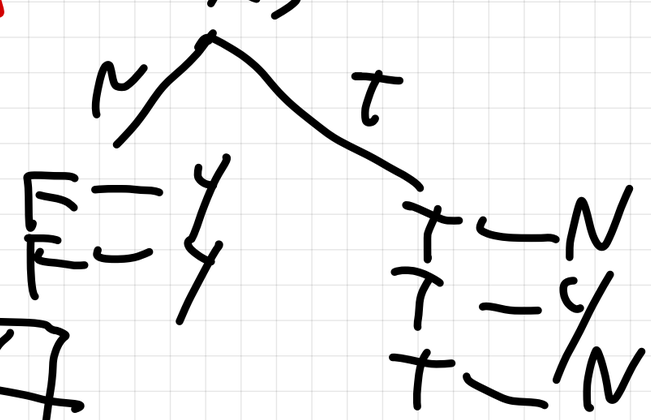
X_1	X_2	X_3	Y
T	T	F	Y
T	T	T	N
T	F	F	Y
F	T	T	Y
F	F	T	N







Example: Attribute Selection with Information Gain

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31...40	4	0	0
>40	3	2	0.971

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

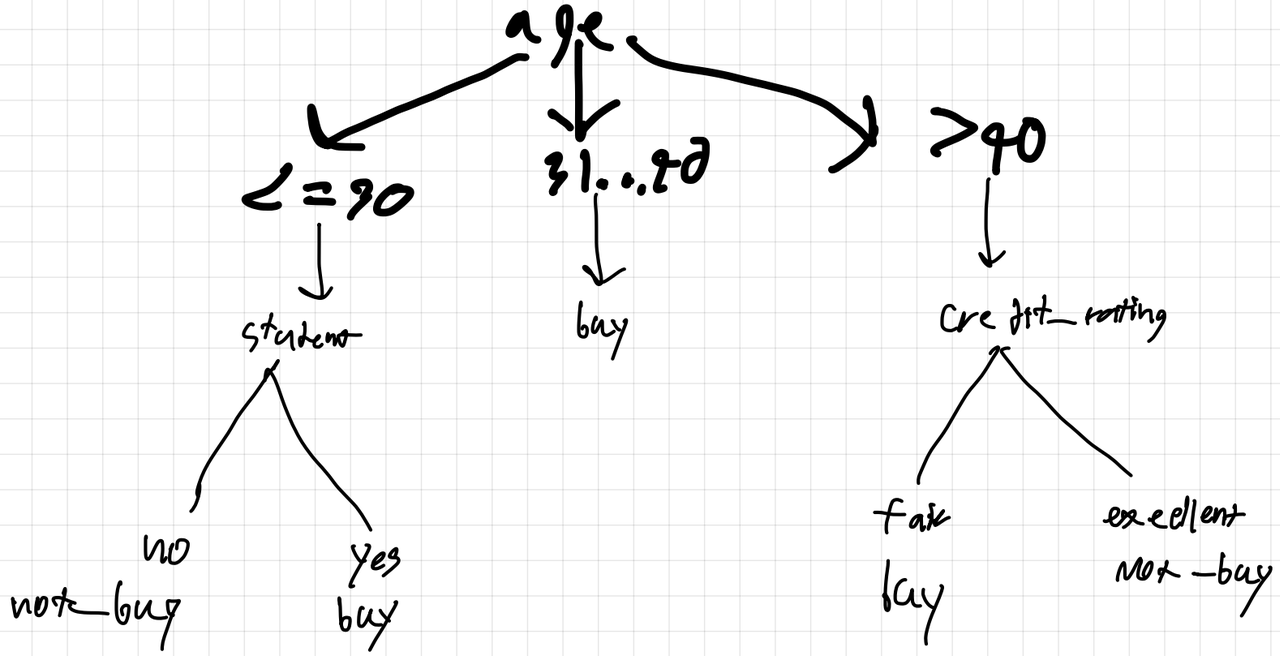
$\frac{5}{14}I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's.

Hence

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly, we can get

$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$



1. คำนวณ Info(D)

$$Info(D) = I(8,4) = -\frac{8}{12}\log_2\frac{8}{12} - \frac{4}{12}\log_2\frac{4}{12}$$

$$= 0.31 + 0.5285$$

$$\therefore Info(D) = 0.9183$$

2. คำนวณ $Info_{age,income,student,credit}(D)$

$$Info_{age}(D) = \frac{4}{12}i(2,2) + \frac{3}{12}i(3,0) + \frac{5}{12}i(3,2)$$

$$= \frac{4}{12}\left[-\frac{2}{4}\log(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4})\right] + \frac{3}{12}\left[-\frac{3}{3}\log_2(\frac{3}{3}) - \frac{0}{3}\log_2(\frac{0}{3})\right]$$

$$+ \frac{5}{12}\left[-\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5})\right]$$

$$= \frac{4}{12}(1) + \frac{3}{12}(0) + \frac{5}{12}(0.9710)$$

$$Info(D)_{age} = 0.6904$$

$$Info_{(income)}(D) = \frac{4}{12}I(2,2) + \frac{5}{12}I(4,1) + \frac{3}{12}I(2,1)$$

$$= \frac{4}{12}\left[-\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4})\right] + \frac{5}{12}\left[-\frac{4}{5}\log_2(\frac{4}{5}) - \frac{1}{5}\log_2(\frac{1}{5})\right] + \frac{3}{12}\left[-\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3})\right]$$

$$= \frac{4}{12}(1) + \frac{5}{12}(0.7219) + \frac{3}{12}(0.9183)$$

$Info_{(income)}(D) = 0.8637$

$$Info_{(student)}(D) = \frac{6}{12} I(5,1) + \frac{6}{12} I(3,3)$$

$$= \frac{6}{12}\left[-\frac{5}{6}\log_2\left(\frac{5}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right)\right] + \frac{6}{12}\left[-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right]$$

$$= \frac{6}{12}(0.6500) + \frac{6}{12}(1)$$

$Info_{(student)}(D) = 0.825$

$$Info_{(credit)}(D) = \frac{7}{12} I(6,1) + \frac{5}{12} I(2,3)$$

$$= \frac{7}{12}\left[-\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right] + \frac{5}{12}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]$$

$$= \frac{7}{12}(0.5917) + \frac{5}{12}(0.9710)$$

$Info_{(credit)}(D) = 0.7497$

3. คำนวณ Gain

Gain(age) = 0.9183 − 0.6409 = 0.2779

Gain(Income) = 0.9183 − 0.8637 = 0.0546

Gain(Student) = 0.9183 − 0.8250 = 0.0933

Gain(Credit rating) = 0.9183 − 0.7497 = 0.1686

ได้ Gain(Age) ค่ามากที่สุด ดังนั้น Age เป็น root node