Training data set: Who buys computer?

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

— N H N F N
— N H N E N

— N M N F N
— Y L Y F Y
— Y M Y E Y

ขั้น

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

ขั้นแต่ละตัวที่แบ่งด้วย (≤30, 31…40, >40)

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

$$Info_{income}(D) = \frac{4}{14}I(3,1) + \frac{6}{14}I(4,2) + \frac{4}{14}I(2,2)$$
$$= \frac{4}{14}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) + \frac{6}{14}\left(-\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right) + \frac{4}{14}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right)\right)$$
$$= 0.911$$

$$Info_{student}(D) = \frac{7}{14}I(4,3) + \frac{7}{14}I(6,1)$$
$$= \frac{7}{14}\left(-\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right)\right) + \frac{7}{14}\left(-\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right)$$
$$= 0.788$$

$$Info_{credit\_rating}(D) = \frac{8}{14}I(6,2) + \frac{6}{14}I(3,3)$$
$$= \frac{8}{14}\left(-\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right) + \frac{6}{14}\left(-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right)$$
$$= 0.892$$

ขั้น Gain

$$Gain(age) = info(D) - info_{age}(D) = 0.940 - 0.694 = 0.246$$
$$Gain(income) = info(D) - info_{income}(D) = 0.940 - 0.911 = 0.029$$
$$Gain(student) = info(D) - info_{student}(D) = 0.940 - 0.788 = 0.152$$
$$Gain(credit\_rating) = info(D) - info_{credit\_rating}(D) = 0.940 - 0.892 = 0.048$$

ดังนั้น เลือกในแอ Gain age มีค่า 0.246

age : <= 30

$Info_{age:<=30}(D) = I(\overset{Y}{2},\overset{N}{3}) = -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}) = 0.971$

$Info_{income}(D) = \frac{1}{5}I(\overset{low}{1},\overset{Y}{0},\overset{N}{0}) + \frac{2}{5}I(\overset{M}{1},\overset{Y}{1},\overset{N}{1}) + \frac{2}{5}I(\overset{H}{0},\overset{Y}{2},\overset{N}{0})$

$= \frac{1}{5}(-\frac{1}{1}\log_2(\frac{1}{1}) - 0) + \frac{2}{5}(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})) + \frac{2}{5}(-\frac{2}{2}\log_2(\frac{2}{2}) - 0)$

$= 0.4$

$Info_{student}(D) = \frac{3}{5}(0,3) + \frac{2}{5}(2,0)$

$= \frac{3}{5}(-\frac{3}{3}\log_2(\frac{3}{3})) + \frac{2}{5}(-\frac{2}{2}\log_2(\frac{2}{2}))$

$= 0$

F N E N
F N E Y
F Y

$Info_{credit\_rating}(D) = \frac{3}{5}(\overset{F}{1},\overset{E}{2}) + \frac{2}{5}(1,1)$

$= \frac{3}{5}(-\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3})) + \frac{2}{5}(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log(\frac{1}{2}))$

$= 0.951$

$Gain(income) = info_{age:<30}(D) - info_{income}(D) = 0.971 - 0.400 = 0.571$

$Gain(Student) = info_{age:<30}(D) - info_{student}(D) = 0.971 - 0 = 0.971$

$Gain(credit\_rating) = info_{age:<30}(D) - info_{credit\_rating}(D) = 0.971 - 0.951 = 0.020$

ดังนั้นเราเลือก Node student เพราะมี Gain สูงสุด

age : 31 ... 40

$Info_{age:31...40}(D) = I(4,0)$

$= -\frac{4}{4}\log_2(\frac{4}{4}) - \frac{0}{4}\log_2(\frac{0}{4})$

$= 0$

เพราะตรงว่า จะซื้อยังไงก็ซื้อเพราะมี Yes 4 อัน

age : 40

$Info_{age:>40}(D) = I(3,2)$

$= -\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5})$

$= 0.971$

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

$= \frac{2}{5}(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})) + \frac{3}{5}(-\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}))$

$= 0.951$

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

$= \frac{3}{5}(-\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2\frac{1}{3}) + \frac{2}{5}(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}))$

$Info_{credit\_rating}(D) = \frac{3}{5}I(3,0) + \frac{2}{5}I(0,2)$

$= \frac{3}{5}(-\frac{3}{3}\log_2(\frac{3}{3})) + \frac{2}{5}(-\frac{2}{2}\log_2(\frac{2}{2}))$

$= 0$

$Gain(income) = info_{age:>40}(D) - info_{income}(D) = 0.971 - 0.951 = 0.02$

$Gain(student) = info_{age:>40}(D) - info_{student}(D) = 0.971 - 0.951 = 0.02$

$Gain(credit\_rating) = info_{age:>40}(D) - info_{credit\_rating}(D) = 0.971 - 0 = 0.971$

เลือก Credit_rating เพราะ Gain สูงสุด