HW2 CLASSIFICATION : DECISION TREE

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

age	<30	31…40	>40
yes :	2	4	3
no :	3	0	2

income	high	med	low
yes :	2	4	3
no :	2	2	1

student	yes	no
yes :	6	3
no :	1	5

credit	fair	ex-
yes :	6	3
no :	2	3

① class $\quad Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$

$= I(Y,N) = I(9,5)$

$= -9/14 \log_2(9/14) - 5/14 \log_2(5/14) = 0.940$

② feature $\quad Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$

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

$= 5/14 \left(-2/5 \log_2(2/5) - 3/5 \log_2(3/5)\right) + 4/14 \left(-4/4 \log(4/4) - 0/4 \log_2(0/4)\right)$

$+ 5/14 \left(-3/5 \log_2(3/5) - 2/5 \log_2(2/5)\right)$

$= 0.694$

$Info_{income}(D) = 4/14\, I(2,2) + 6/14\, I(4,2) + 4/14\, I(3,1)$

$= 4/14 \left(-2/4 \log_2(4/4) - 2/4 \log_2(2/4)\right) + 6/14 \left(-4/6 \log_2(4/6) - 2/6 \log_2(2/6)\right)$

$+ 4/14 \left(-3/4 \log_2(3/4) - 1/4 \log_2(1/4)\right) = 0.911$

$Info_{student}(D) = 7/14\, I(6,1) + 7/14\, I(3,4)$

$= 7/14 \left(-6/7 \log_2(6/7) - 1/7 \log_2(1/7)\right) + 7/14 \left(-3/7 \log_2(3/7) - 4/7 \log_2(4/7)\right)$

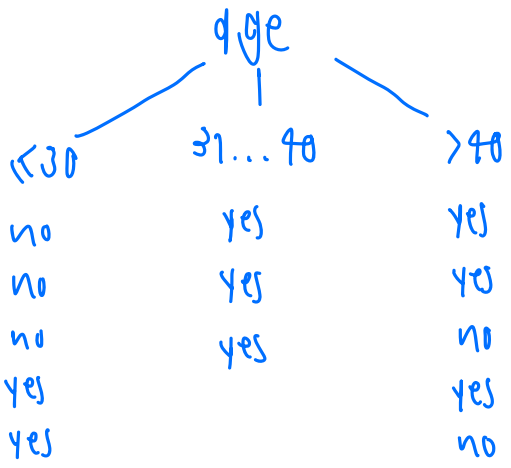
$= 0.788$

$Info_{credit}(D) = 6/14\, I(3,3) + 8/14\, I(6,2)$

$= 6/14 \left(-3/6 \log_2(3/6) - 3/6 \log(3/6)\right) + 8/14 \left(-6/8 \log_2(6/8) - 2/8 \log_2(2/8)\right)$

$= 0.892$

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



$Gain(age) = Info(D) - Info_{age}(D) = 0.94 - 0.694 = 0.246$ → เลือก Gain มากๆ

$Gain(income) = Info(D) - Info_{income}(D) = 0.94 - 0.911 = 0.029$

$Gain(student) = Info(D) - Info_{student}(D) = 0.94 - 0.788 = 0.152$

$Gain(credit) = Info(D) - Info_{credit}(D) = 0.94 - 0.892 = 0.048$

age <30

$Info_{age<30}(D) = I(2,3)$

$= -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}) = 0.971$

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

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

$= 0.4$

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

$= \frac{3}{5}(-\frac{3}{3}\log_2(\frac{3}{3})) + \frac{2}{5}(-\frac{2}{2}\log_2(\frac{2}{2})) = 0$

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

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

$= 0.951$