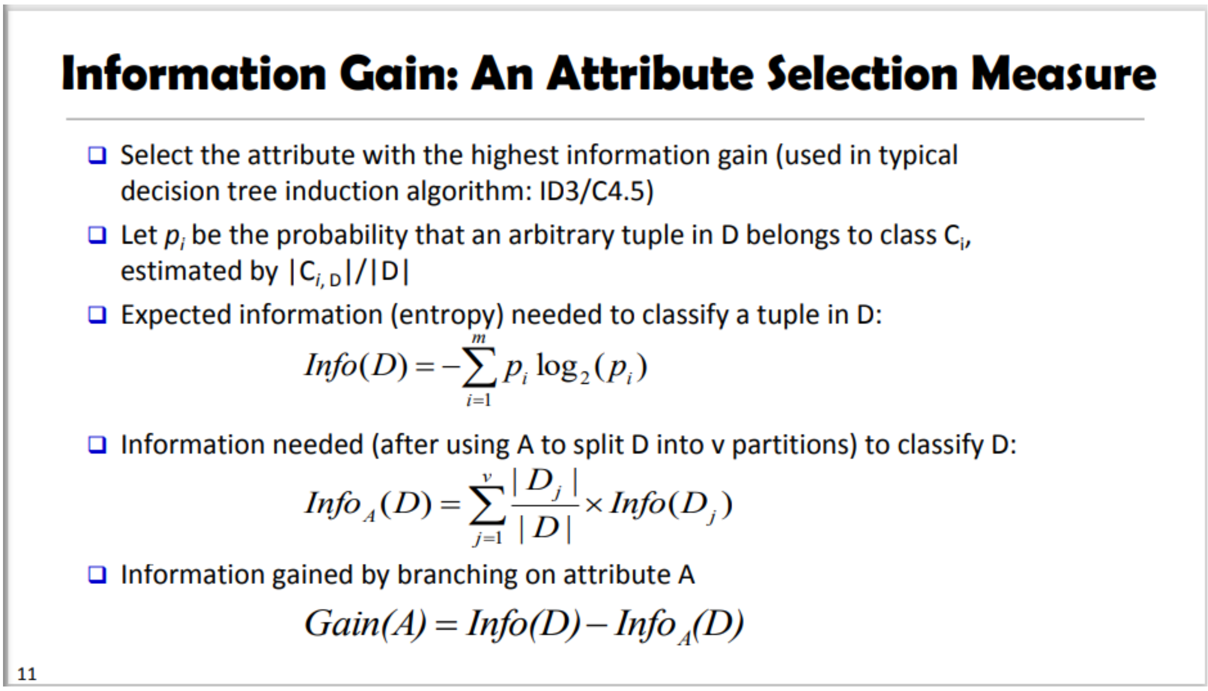
Decision Tree Induction



Information Gain: An Attribute Selection Measure

- ❑ Select the attribute with the highest information gain (used in typical decision tree induction algorithm: ID3/C4.5)
- ❑ Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- ❑ Expected information (entropy) needed to classify a tuple in D:

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

- ❑ Information needed (after using A to split D into v partitions) to classify D:

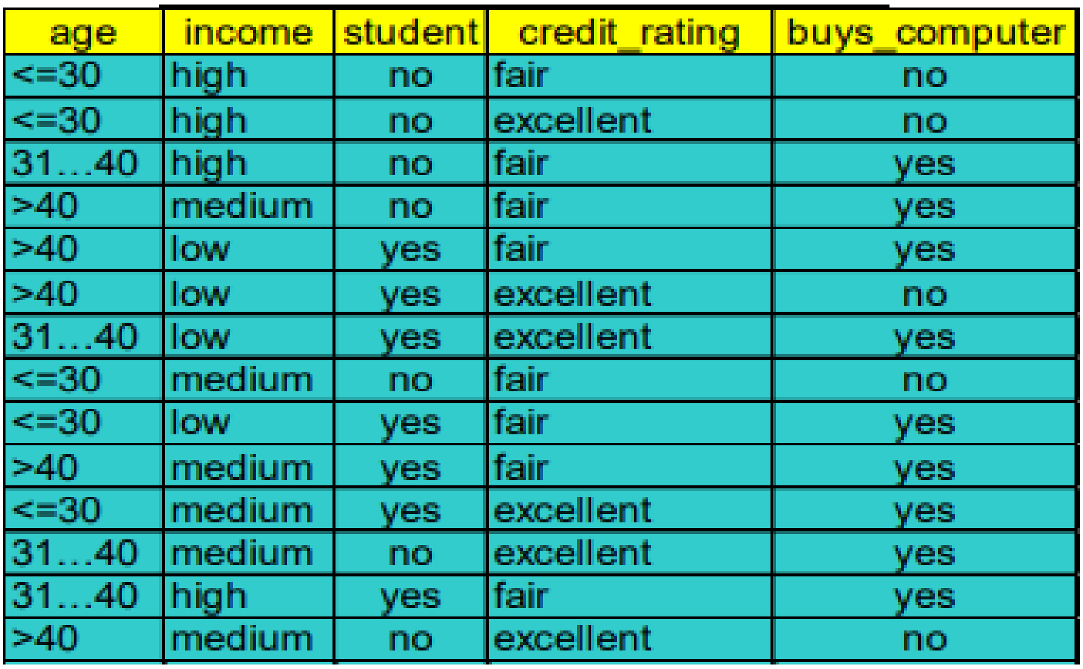
$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- ❑ Information gained by branching on attribute A

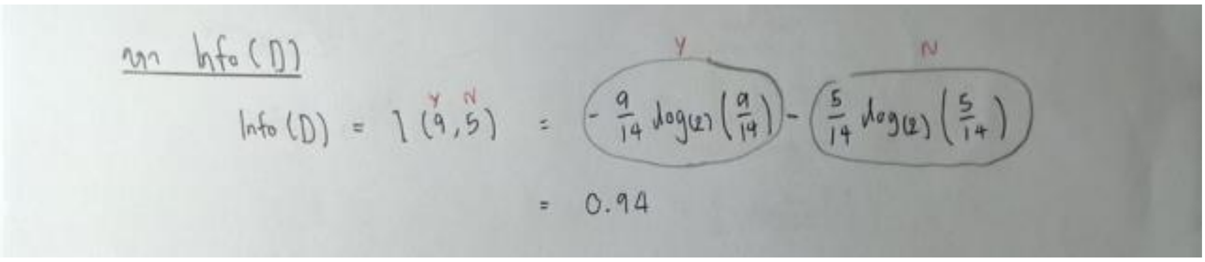
$$Gain(A) = Info(D) - Info_A(D)$$

11

Example:



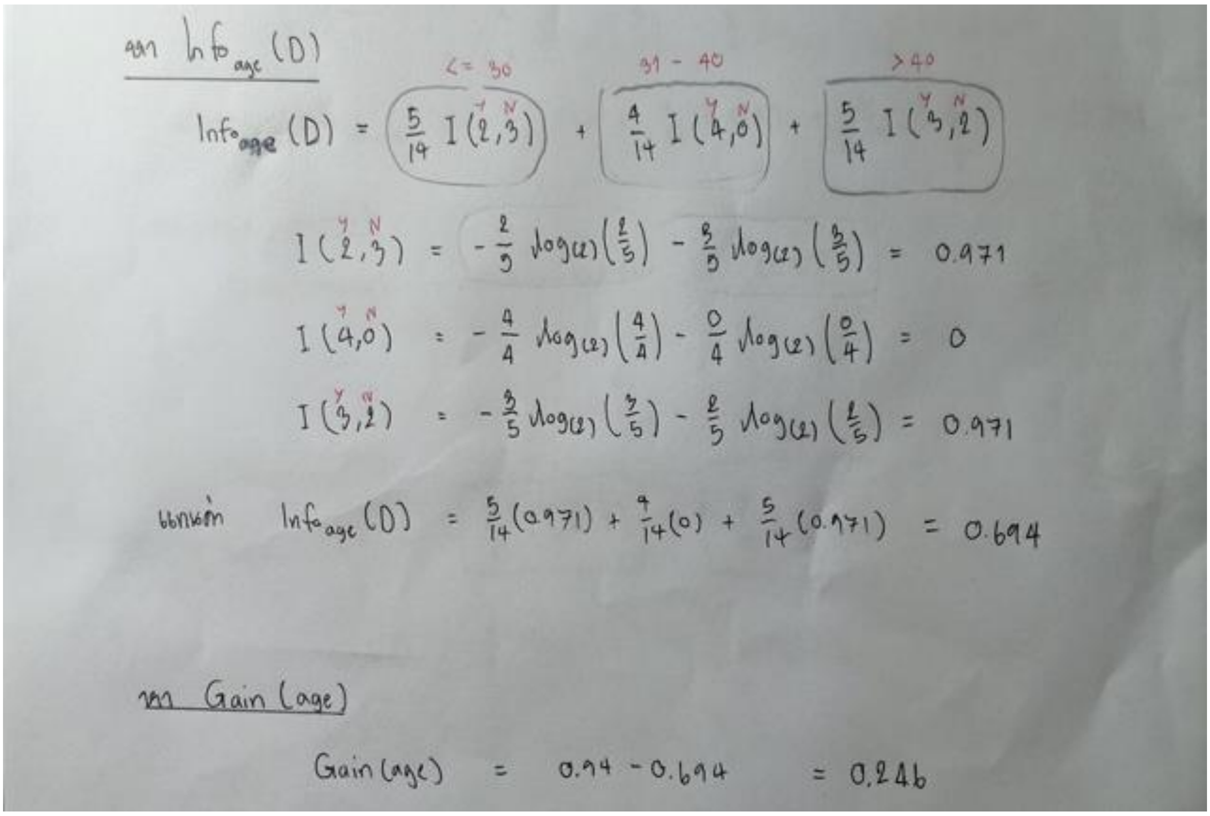
age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent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



หา Info(D)

$$Info(D) = I(\overset{Y}{9}, \overset{N}{5}) = \overset{Y}{\left(-\frac{9}{14}\log(2)\left(\frac{9}{14}\right)\right)} - \overset{N}{\left(\frac{5}{14}\log(2)\left(\frac{5}{14}\right)\right)}$$

$$= 0.94$$



หา $\text{Info}_{age}(D)$

$$\text{Info}_{age}(D) = \overset{\text{<= 30}}{\boxed{\frac{5}{14} I(\overset{Y}{2},\overset{N}{3})}} + \overset{\text{31 - 40}}{\boxed{\frac{4}{14} I(\overset{Y}{4},\overset{N}{0})}} + \overset{\text{>40}}{\boxed{\frac{5}{14} I(\overset{Y}{3},\overset{N}{2})}}$$

$$I(\overset{Y}{2},\overset{N}{3}) = -\frac{2}{5}\log_{(2)}\left(\frac{2}{5}\right) - \frac{3}{5}\log_{(2)}\left(\frac{3}{5}\right) = 0.971$$

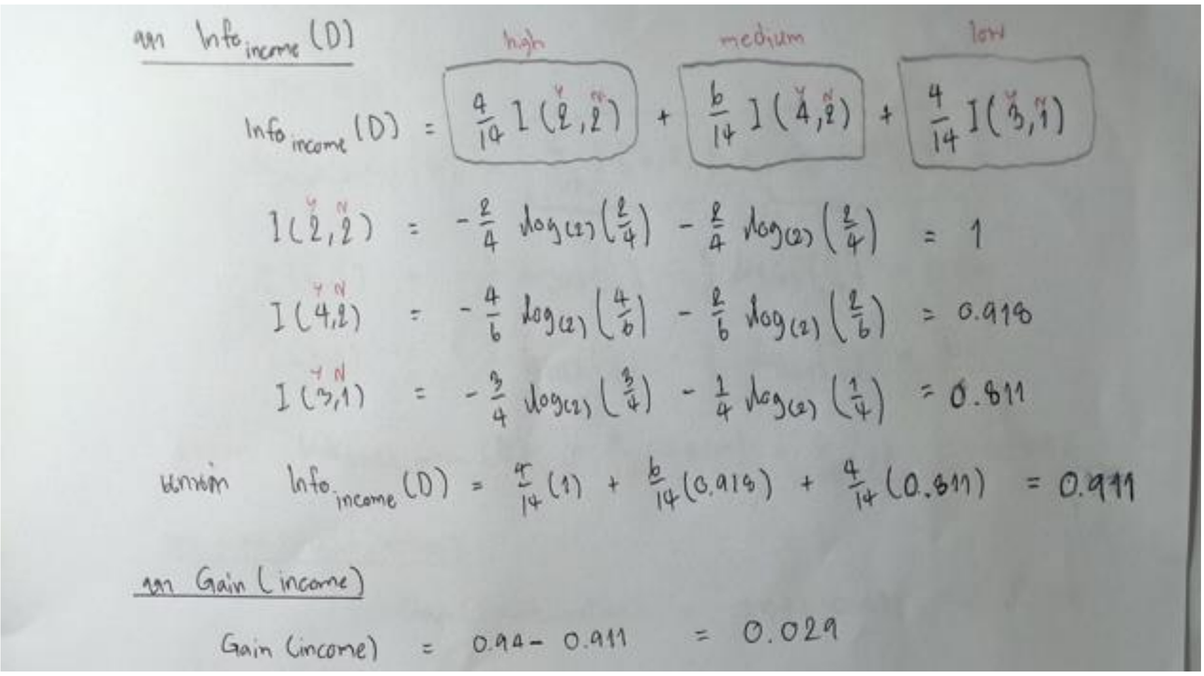
$$I(\overset{Y}{4},\overset{N}{0}) = -\frac{4}{4}\log_{(2)}\left(\frac{4}{4}\right) - \frac{0}{4}\log_{(2)}\left(\frac{0}{4}\right) = 0$$

$$I(\overset{Y}{3},\overset{N}{2}) = -\frac{3}{5}\log_{(2)}\left(\frac{3}{5}\right) - \frac{2}{5}\log_{(2)}\left(\frac{2}{5}\right) = 0.971$$

แทนค่า $\text{Info}_{age}(D) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = 0.694$

หา Gain(age)

$$\text{Gain}(age) = 0.94 - 0.694 = 0.246$$



หา $\text{Info}_{income}(D)$

$$\text{Info}_{income}(D) = \overset{\text{high}}{\boxed{\frac{4}{14} I(\overset{Y}{2},\overset{N}{2})}} + \overset{\text{medium}}{\boxed{\frac{6}{14} I(\overset{Y}{4},\overset{N}{2})}} + \overset{\text{low}}{\boxed{\frac{4}{14} I(\overset{Y}{3},\overset{N}{1})}}$$

$$I(\overset{Y}{2},\overset{N}{2}) = -\frac{2}{4}\log_{(2)}\left(\frac{2}{4}\right) - \frac{2}{4}\log_{(2)}\left(\frac{2}{4}\right) = 1$$

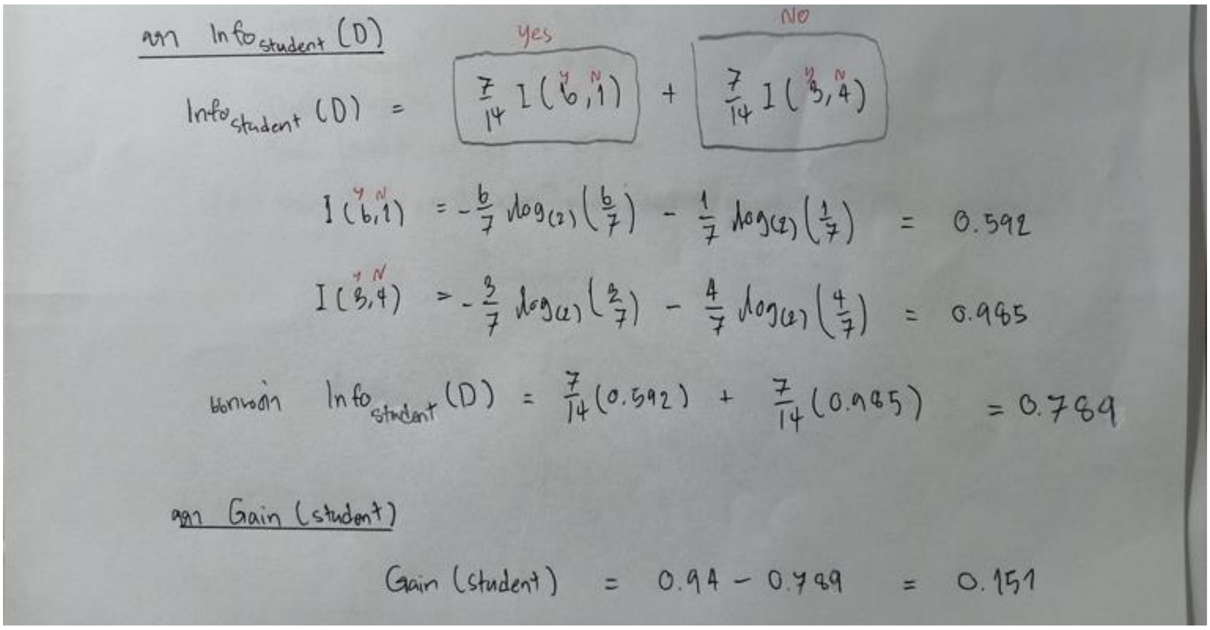
$$I(\overset{Y}{4},\overset{N}{2}) = -\frac{4}{6}\log_{(2)}\left(\frac{4}{6}\right) - \frac{2}{6}\log_{(2)}\left(\frac{2}{6}\right) = 0.918$$

$$I(\overset{Y}{3},\overset{N}{1}) = -\frac{3}{4}\log_{(2)}\left(\frac{3}{4}\right) - \frac{1}{4}\log_{(2)}\left(\frac{1}{4}\right) = 0.811$$

แทนค่า $\text{Info}_{income}(D) = \frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.811) = 0.911$

หา Gain(income)

$$\text{Gain}(income) = 0.94 - 0.911 = 0.029$$



หา $\text{Info}_{student}(D)$

$$\text{Info}_{student}(D) = \overset{\text{yes}}{\boxed{\frac{7}{14} I(\overset{Y}{6},\overset{N}{1})}} + \overset{\text{No}}{\boxed{\frac{7}{14} I(\overset{Y}{3},\overset{N}{4})}}$$

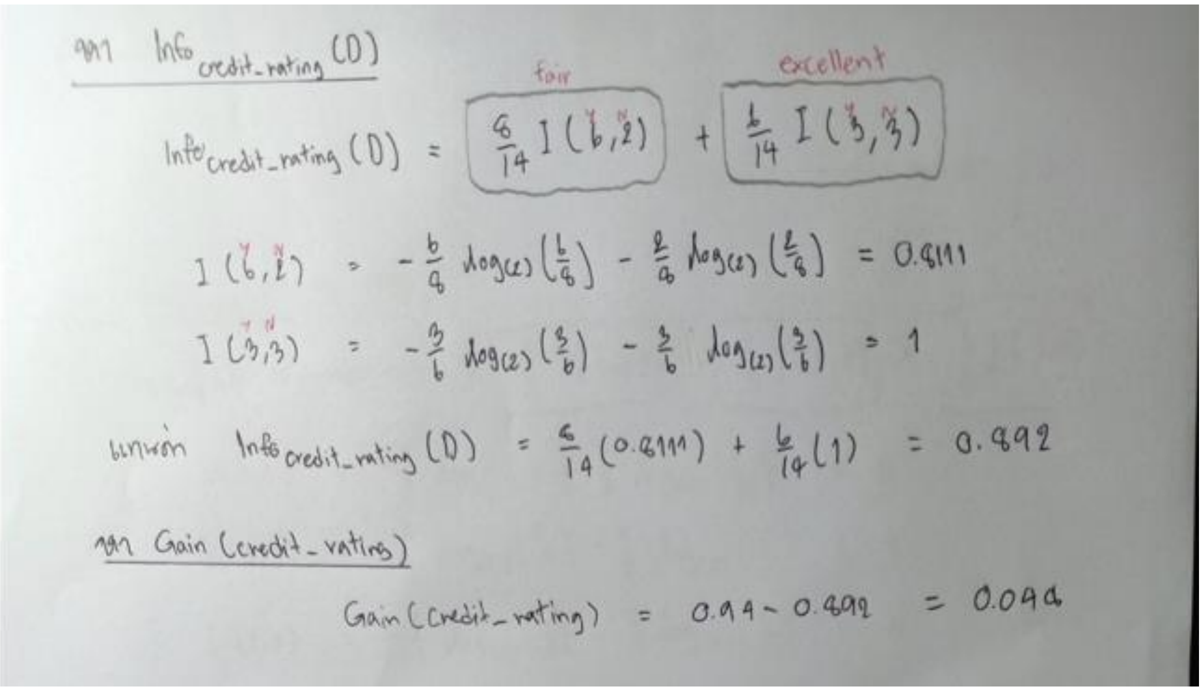
$$I(\overset{Y}{6},\overset{N}{1}) = -\frac{6}{7}\log_{(2)}\left(\frac{6}{7}\right) - \frac{1}{7}\log_{(2)}\left(\frac{1}{7}\right) = 0.592$$

$$I(\overset{Y}{3},\overset{N}{4}) = -\frac{3}{7}\log_{(2)}\left(\frac{3}{7}\right) - \frac{4}{7}\log_{(2)}\left(\frac{4}{7}\right) = 0.985$$

แทนค่า $\text{Info}_{student}(D) = \frac{7}{14}(0.592) + \frac{7}{14}(0.985) = 0.789$

หา Gain(student)

$$\text{Gain}(student) = 0.94 - 0.789 = 0.151$$



หา $\text{Info}_{credit\_rating}(D)$

$$\text{Info}_{credit\_rating}(D) = \underbrace{\boxed{\frac{8}{14} I(\overset{Y}{6},\overset{N}{2})}}_{} + \boxed{\frac{6}{14} I(\overset{Y}{3},\overset{N}{3})}$$

fair excellent

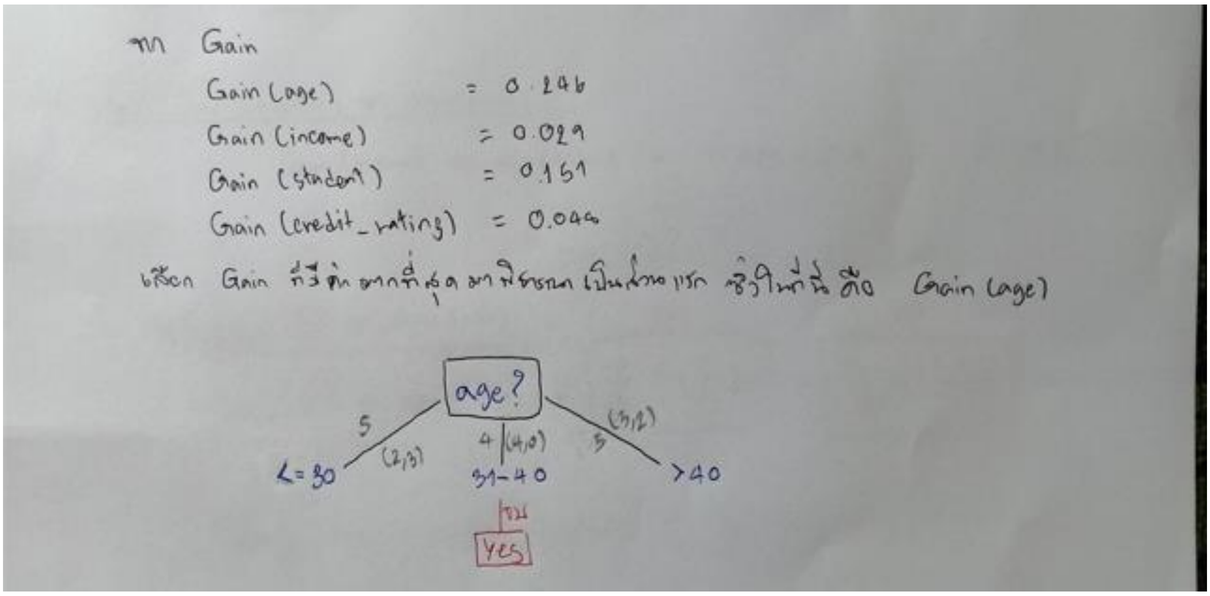
$$I(\overset{Y}{6},\overset{N}{2}) = -\frac{6}{8}\log_{(2)}\left(\frac{6}{8}\right) - \frac{2}{8}\log_{(2)}\left(\frac{2}{8}\right) = 0.8111$$

$$I(\overset{Y}{3},\overset{N}{3}) = -\frac{3}{6}\log_{(2)}\left(\frac{3}{6}\right) - \frac{3}{6}\log_{(2)}\left(\frac{3}{6}\right) = 1$$

แทนค่า $\text{Info}_{credit\_rating}(D) = \frac{8}{14}(0.8111) + \frac{6}{14}(1) = 0.892$

หา Gain (credit_rating)

$$\text{Gain}(credit\_rating) = 0.94 - 0.892 = 0.048$$



หา Gain

Gain (age) = 0.246

Gain (income) = 0.029

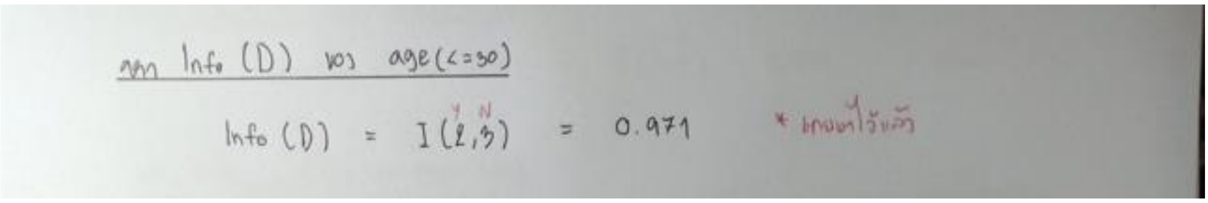
Gain (student) = 0.151

Gain (credit_rating) = 0.048

เลือก Gain ที่มีค่ามากที่สุด มาพิจารณาเป็นตัวแรก ซึ่งในที่นี้ คือ Gain (age)

age?
5
(2,3)
<=30
4
(4,0)
31-40
ใช่
Yes
5
(3,2)
>40

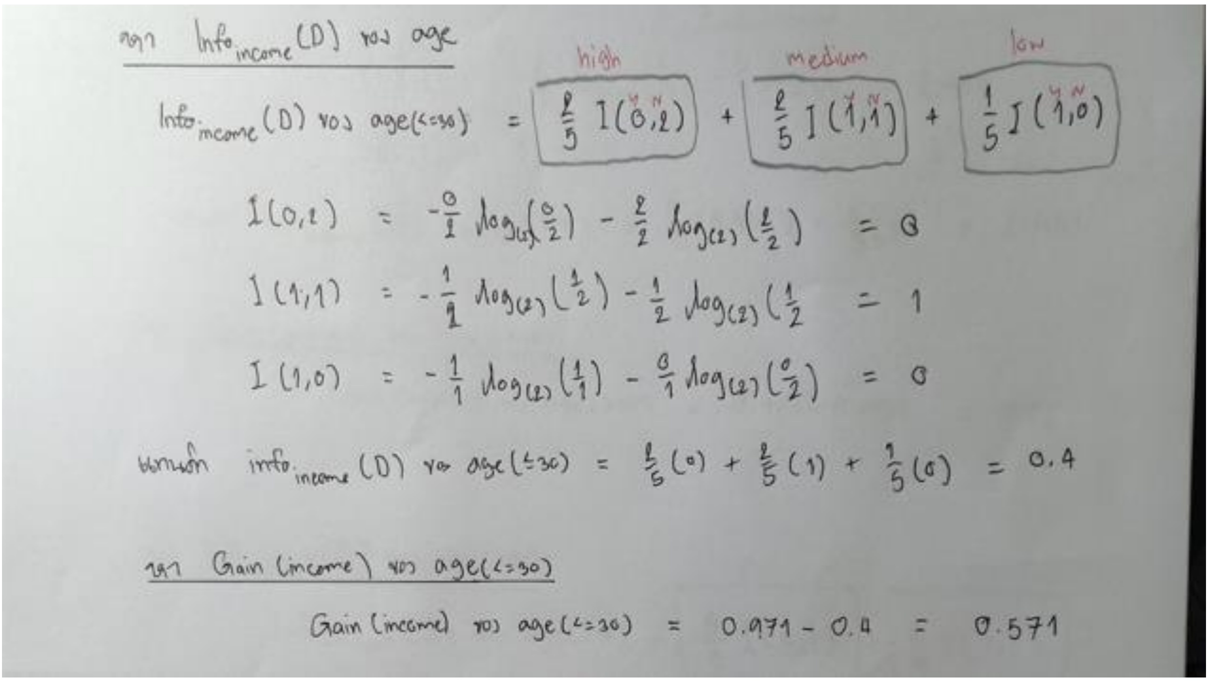
age(<=30)



หา Info (D) ของ age(<=30)

$$\text{Info}(D) = I(\overset{Y}{2},\overset{N}{3}) = 0.971$$

* เก็บค่าไว้ก่อน



หา $Info_{income}(D)$ ของ age

$Info_{income}(D)$ ของ $age(<=30)$ $= \overset{high}{\boxed{\frac{2}{5} I(\overset{Y}{0},\overset{N}{2})}} + \overset{medium}{\boxed{\frac{2}{5} I(\overset{Y}{1},\overset{N}{1})}} + \overset{low}{\boxed{\frac{1}{5} I(\overset{Y}{1},\overset{N}{0})}}$

$I(0,2) = -\frac{0}{2} \log_{(2)}(\frac{0}{2}) - \frac{2}{2} \log_{(2)}(\frac{2}{2}) = 0$

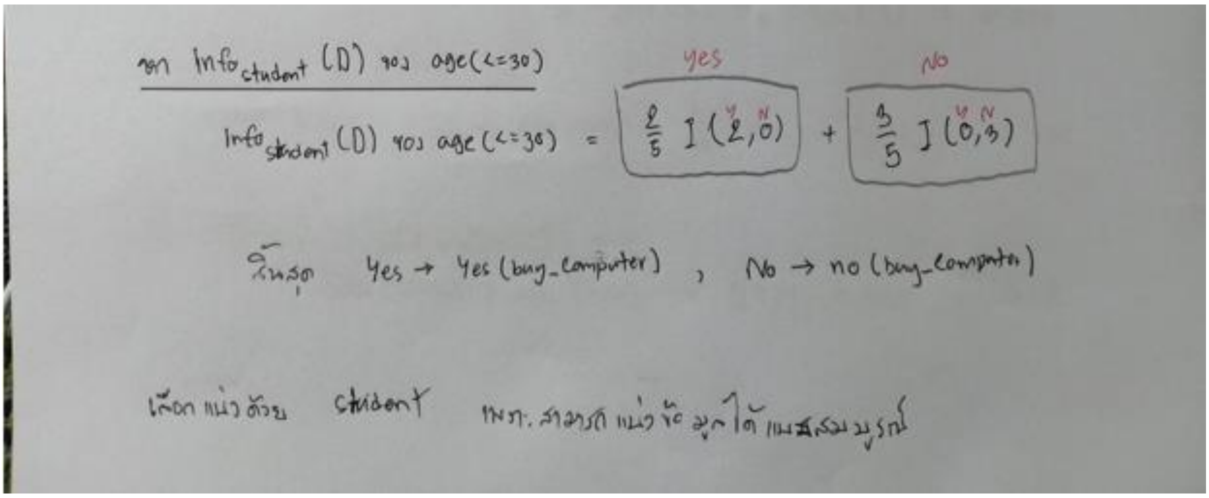
$I(1,1) = -\frac{1}{2} \log_{(2)}(\frac{1}{2}) - \frac{1}{2} \log_{(2)}(\frac{1}{2} = 1$

$I(1,0) = -\frac{1}{1} \log_{(2)}(\frac{1}{1}) - \frac{0}{1} \log_{(2)}(\frac{0}{2}) = 0$

แทนค่า $info_{income}(D)$ ของ $age(\leq 30) = \frac{2}{5}(0) + \frac{2}{5}(1) + \frac{1}{5}(0) = 0.4$

หา Gain (income) ของ age(<=30)

Gain (income) ของ age(<=30) = 0.971 - 0.4 = 0.571



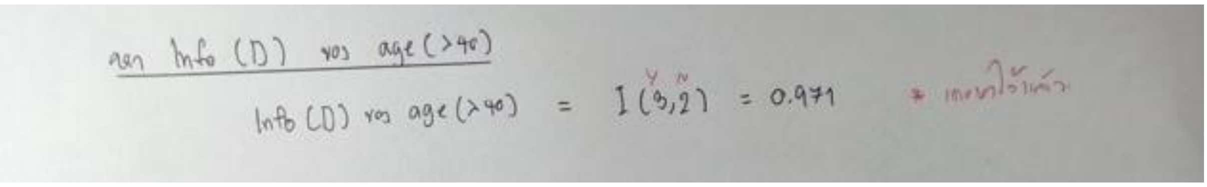
หา $Info_{student}(D)$ ของ $age(<=30)$

$Info_{student}(D)$ ของ $age(<=30) = \overset{yes}{\boxed{\frac{2}{5} I(\overset{Y}{2},\overset{N}{0})}} + \overset{No}{\boxed{\frac{3}{5} I(\overset{Y}{0},\overset{N}{3})}}$

สิ้นสุด Yes → Yes (buy_computer) , No → no (buy_computer)

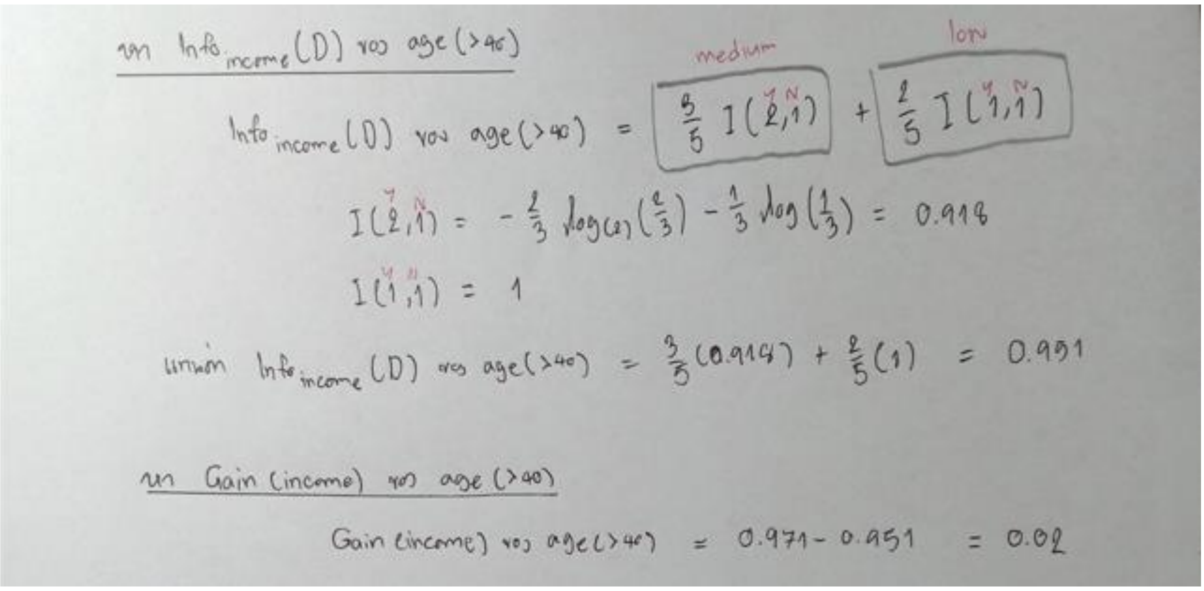
เลือกแบ่งด้วย student เพราะ สามารถ แบ่งข้อมูลได้แบบสมบูรณ์

age(>40)



หา Info (D) ของ age (>40)

Info (D) ของ age (>40) $= I(\overset{Y}{3},\overset{N}{2}) = 0.971$ * เคยหาไว้แล้ว



หา $Info_{income}(D)$ ของ $age(>40)$

$Info_{income}(D)$ ของ $age(>40) = \overset{medium}{\boxed{\frac{3}{5} I(\overset{Y}{2},\overset{N}{1})}} + \overset{low}{\boxed{\frac{2}{5} I(\overset{Y}{1},\overset{N}{1})}}$

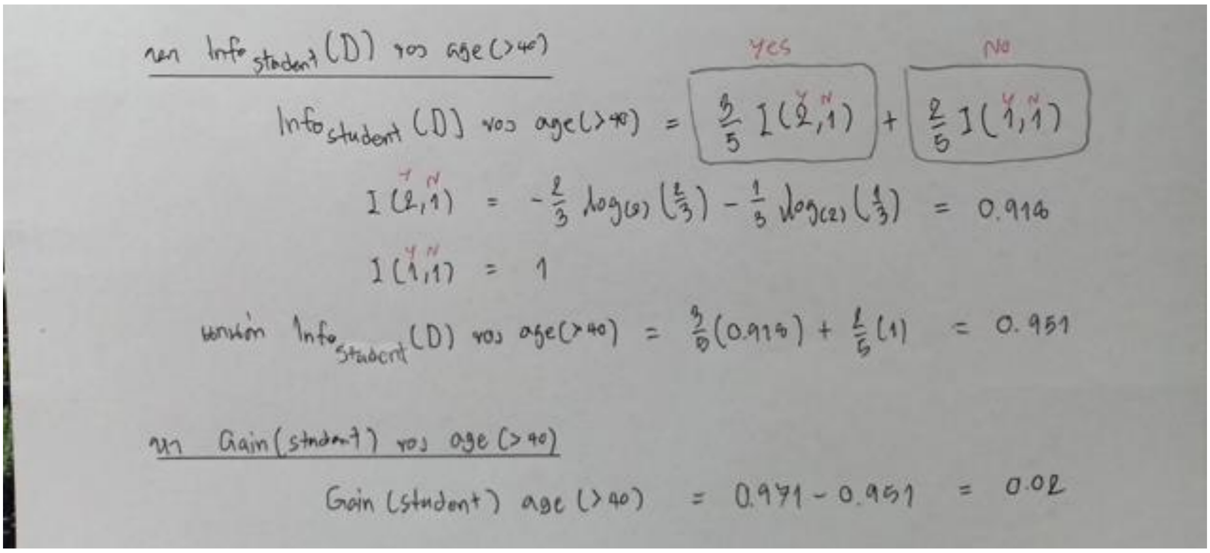
$I(\overset{Y}{2},\overset{N}{1}) = -\frac{2}{3} \log_{(2)}(\frac{2}{3}) - \frac{1}{3} \log(\frac{1}{3}) = 0.918$

$I(\overset{Y}{1},\overset{N}{1}) = 1$

แทนค่า $Info_{income}(D)$ ของ $age(>40) = \frac{3}{5}(0.918) + \frac{2}{5}(1) = 0.951$

หา Gain (income) ของ age (>40)

Gain (income) ของ age(>40) = 0.971 - 0.951 = 0.02



หา $\text{Info}_{student}(D)$ ของ age(>40)

$$\text{Info}_{student}(D) \text{ ของ } age(>40) = \overset{yes}{\boxed{\frac{3}{5} I(\overset{Y}{2},\overset{N}{1})}} + \overset{No}{\boxed{\frac{2}{5} I(\overset{Y}{1},\overset{N}{1})}}$$

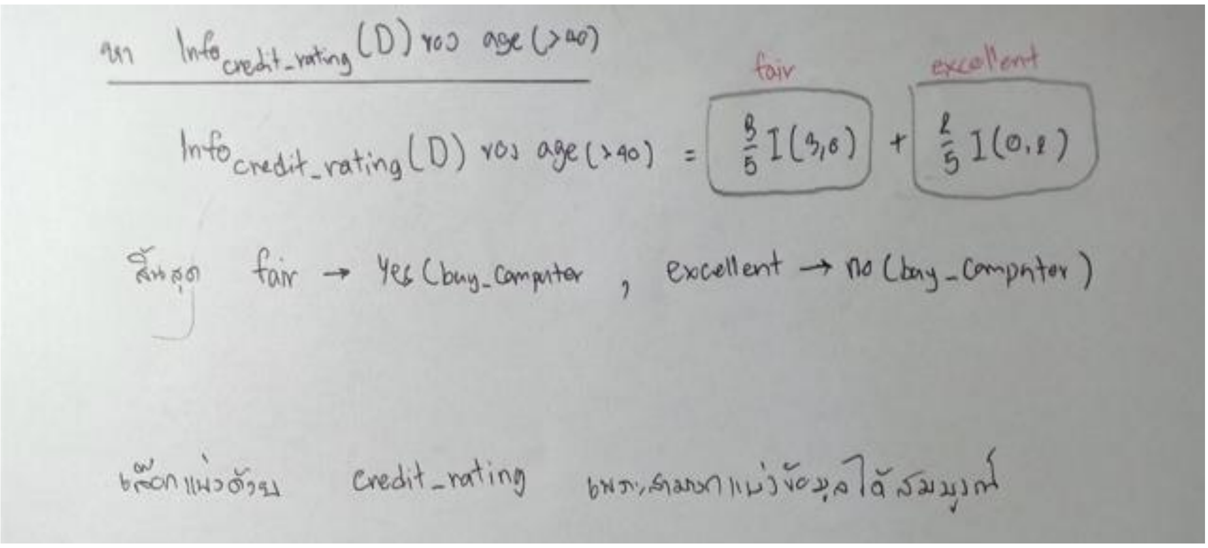
$$I(\overset{Y}{2},\overset{N}{1}) = -\frac{2}{3}\log_{(2)}\left(\frac{2}{3}\right) - \frac{1}{3}\log_{(2)}\left(\frac{1}{3}\right) = 0.916$$

$$I(\overset{Y}{1},\overset{N}{1}) = 1$$

แทนค่า $\text{Info}_{student}(D)$ ของ $age(>40) = \frac{3}{5}(0.916) + \frac{2}{5}(1) = 0.951$

หา Gain(student) ของ age(>40)

$$\text{Gain}(student)\ age(>40) = 0.971 - 0.951 = 0.02$$



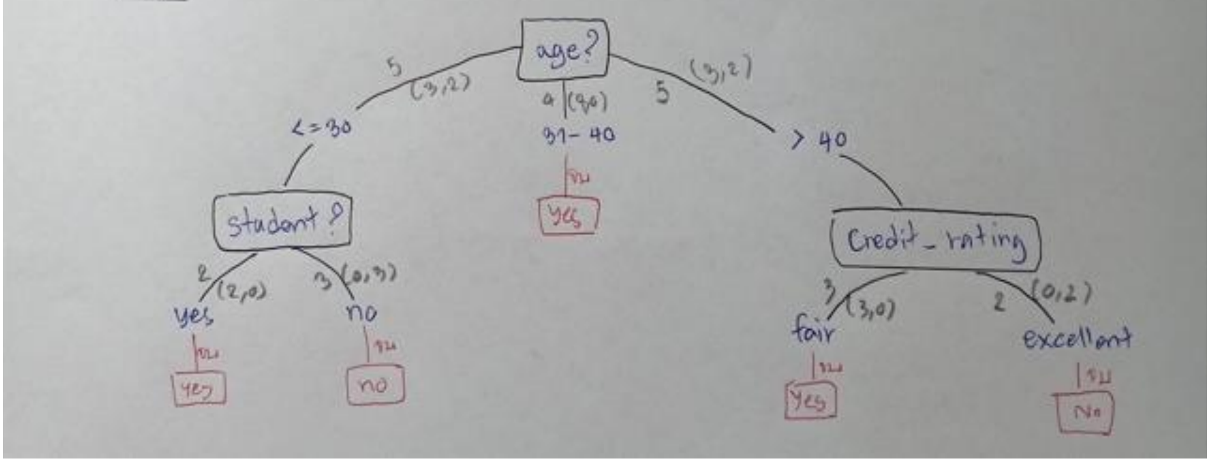
หา $\text{Info}_{credit\_rating}(D)$ ของ age(>40)

$$\text{Info}_{credit\_rating}(D) \text{ ของ } age(>40) = \overset{fair}{\boxed{\frac{3}{5} I(3,0)}} + \overset{excellent}{\boxed{\frac{2}{5} I(0,2)}}$$

สังเกต fair → yes (buy_Computer , excellent → no (buy_Computer)

เลือกแบ่งด้วย credit_rating เพราะสามารถแบ่งข้อมูลได้สมบูรณ์

สรุป


age?
5
(3,2)
<=30
4
(4,0)
31-40
5
(3,2)
> 40
student?
2
(2,0)
yes
3
(0,3)
no
จบ
yes
จบ
no
จบ
yes
Credit_rating
3
(3,0)
fair
2
(0,2)
excellent
จบ
Yes
จบ
No