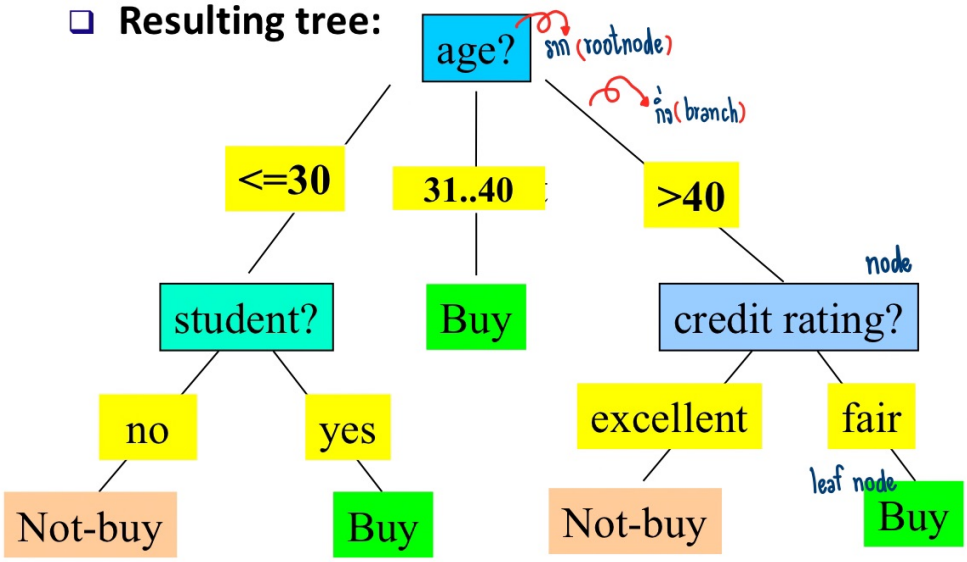
วิธีคำนวณ Decision - Tree

- **Resulting tree:**



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

- Class P: buys_computer = “yes”
- Class N: buys_computer = “no”

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

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

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

1. $$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2)$$
$$= \frac{5}{14}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right] + \frac{4}{14}\left[-\frac{4}{4}\log_2\left(\frac{4}{4}\right)\right] + \frac{5}{14}\left[-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$
(Y, N, Y, Y, N)
$$= 0.694$$

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

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

$= \frac{7}{14}\left[\frac{-3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right] + \frac{7}{14}\left[\frac{-6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right]$ (Y, N, Y, N)

$= 0.789$

4. $Info_{credit\_rating}(D) = \frac{8}{14} I(6,2) + \frac{6}{14} I(3,3)$

$= \frac{8}{14}\left[\frac{-6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right] + \frac{6}{14}\left[\frac{-3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log\left(\frac{3}{6}\right)\right]$ (Y, N, Y, N)

$= 0.892$

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

3. คำนวณ Information Grain โดยเอาค่า Grain ที่มีค่าสูงที่สุดจะเป็นราก (root node)

3.1 Grain (age) = 0.940 - 0.694 = 0.296

3.2 Grain (income) = 0.940 - 0.911 = 0.029

3.3 Grain (student) = 0.940 - 0.789 = 0.151

3.4 Grain (credit_rating) = 0.940 - 0.892 = 0.048

∴ เลือก Grain (age) เป็นราก (root node) เพราะว่ามีค่าสูงที่สุดจากค่าที่นำมาทั้งหมด

4. แยกกลุ่มของ Feature ตามค่าในราก (root node)

4.1 <= 30

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

Info (D) = I(2,3) = 0.971

$Info_{income} = \frac{2}{5} I(0,2) + \frac{2}{5} I(1,1) + \frac{1}{5} I(1,0)$

$= 0.4$

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

$= 0$

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

$= 0.951$

คำนวณ Information Grain

Grain (income) = 0.971 - 0.400 = 0.571

Grain (Student) = 0.971 - 0 = 0.971

Grain (Credit_rating) = 0.971 - 0.951 = 0.02

∴ เลือก Grain (student) เป็น node ในกลุ่ม <= 30

4.2 31...40

age	income	student	credit_rating	buys_computer
31...40	high	no	fair	yes
31...40	low	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes

จาก
- Class P: buys_computer = "yes" yes = 4
- Class N: buys_computer = "no" no = 0

เมื่อกลุ่ม 31...40 สามารถตอบ yes ใน buys_computer ได้เลย

4.3 >40

age	income	student	credit_rating	buys_computer
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
>40	medium	yes	fair	yes
>40	medium	no	excellent	no

$$Info(D) = I(3,2) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$$
$$= 0.971$$

$$Info_{(income)} = \frac{3}{5}I(1,1) + \frac{2}{5}I(1,1)$$
$$= 0.951$$

$$Info_{(student)} = \frac{3}{5}I(1,1) + \frac{2}{5}I(1,1)$$
$$= 0.951$$

$$Info_{(credit\_rating)} = \frac{3}{6}I(3,0) + \frac{2}{5}I(0,2)$$
$$= 0$$

คำนวณ Information Grain

Grain (income) = 0.971 - 0.951 = 0.02

Grain (Student) = 0.971 - 0.951 = 0.02

Grain (Credit_rating) = 0.971 - 0 = 0.971

∴ เลือก Grain (Credit_rating) เป็น node ในชั้นที่ 1 ของ >40

5. สามารถสร้าง Decision Tree ได้ดังนี้