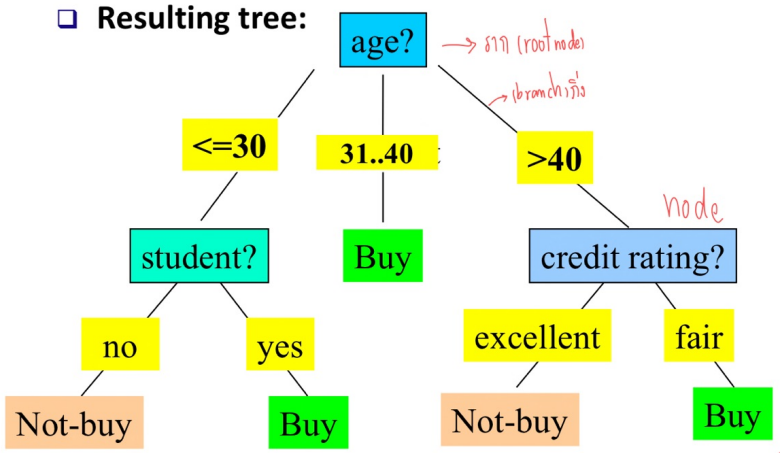
วิธีคำนวณ Decision-Tree

- **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" = 9
- Class N: buys_computer = "no" = 5

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

$= \frac{5}{14}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right] + \frac{4}{14}\left[-\frac{4}{4}\log_2\left(\frac{4}{4}\right)\right] + \frac{5}{14}\left[-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right] = 0.694$

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

$= \frac{4}{14}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] + \frac{6}{14}\left[-\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right] + \frac{4}{14}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right] = 0.911$

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

$= \frac{7}{14}\left[-\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right] + \frac{7}{14}\left[-\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right] = 0.789$

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

$= \frac{8}{14}\left[-\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right] + \frac{6}{14}\left[-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right] = 0.892$

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

3. คำนวณ Information Grain โดย ค่า Grain ที่สูงที่สุดจะเป็น ราก (root node)

3.1 Grain (age) = 0.940 - 0.694 = 0.246

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) เป็น ราก (root node) เพราะมีค่าสูงที่สุด

4. แยกกลุ่มของ feature ตามค่าใน ราก (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

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

$Info_{income}(D) = \frac{2}{5} I(0,2) + \frac{2}{5} I(1,1) + \frac{1}{5} I(1,0) = 0.4$

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

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

คำนวณ Information Grain

Grain (Income) = 0.971 - 0.4 = 0.571
Grain (Student) = 0.971 - 0 = 0.971
Grain (Credit_rating) = 0.971 - 0.951 = 0.02

เลือก Grain(student) เป็น node ในอายุ <= 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 yes = 4
no = 0
เมื่ออายุ 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

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

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

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

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

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 ในอายุ >40

5. สามารถสร้าง Decision Tree ได้ดังนี้