Information Gain

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

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

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

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

$I(2,2) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 0.5+0.5 = 1$

$I(4,2) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.389 + 0.528 = 0.917$

$I(3,1) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.311 + 0.5 = 0.811$

$I(3,4) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.523 + 0.461 = 0.984$

$I(6,1) = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.190 + 0.401 = 0.591$

$I(6,2) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.311 + 0.5 = 0.811$

$I(3,3) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 0.5 + 0.5 = 1$

$Info(D) = I(9,5) = -\frac{9}{14}\log_2\left(\frac{9}{4}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$

Age

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

$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = \mathbf{0.246}$ #

Income

$Info_{income}(D) = \frac{4}{14}I(2,2) + \frac{6}{14}I(4,2) + \frac{4}{14}I(3,1) = \frac{4}{14}(1) + \frac{6}{14}(0.917) + \frac{4}{14}(0.811)$

$= 0.286 + 0.393 + 0.232 = 0.911$

$Gain(income) = 0.940 - 0.911 = \mathbf{0.029}$ #

Student

$Info_{student}(D) = \frac{7}{14}I(3,4) + \frac{7}{14}I(6,1) = \frac{7}{14}(0.984) + \frac{7}{14}(0.591) = 0.492 + 0.296 = 0.788$

$Gain(student) = 0.940 - 0.788 = \mathbf{0.152}$ #

Credit_rating

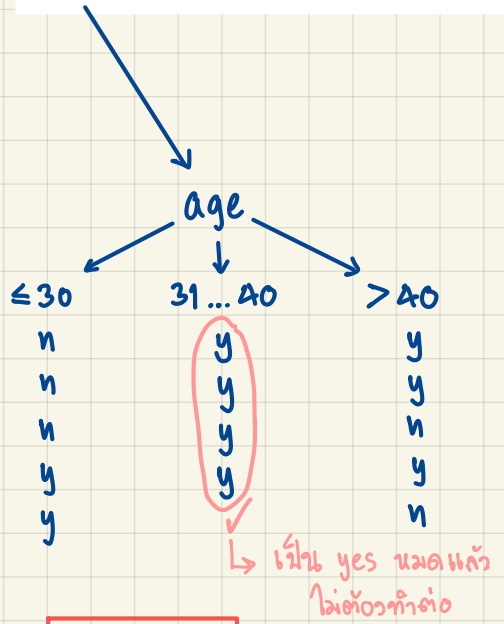
$Info_{credit}(D) = \frac{8}{14}I(6,2) + \frac{6}{14}I(3,3) = \frac{8}{14}(0.811) + \frac{6}{14}(1) = 0.463 + 0.429 = 0.892$

$Gain(credit) = 0.940 - 0.892 = \mathbf{0.048}$ #

เนื่องจาก Gain(age) มีค่ามากที่สุด จึงเลือก age เป็น root node

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



Age ≤ 30

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

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

$= 0.529 + 0.442 = 0.971$

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

$= \frac{2}{5}(0.5+0.5) = 0.4$

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

$= 0$

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

$= \frac{3}{5}(0.528+0.389) + \frac{2}{5}(0.5+0.5)$

$= 0.550 + 0.4 = 0.95$

$Gain(income) = 0.971 - 0.4 = 0.571$

$Gain(Student) = 0.971 - 0 = 0.971$

$Gain(credit) = 0.971 - 0.95 = 0.021$

เนื่องจาก Gain(Student) มีค่ามากที่สุด จึงเลือก Student เป็น node ต่อไปนี้

Age > 40

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

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

$= \frac{3}{5}(0.389+0.528) + \frac{2}{5}(1) = 0.95$

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

$= 0.4 + 0.550 = 0.95$

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

$= 0$

$Gain(income) = 0.971 - 0.95 = 0.021$

$Gain(Student) = 0.971 - 0.95 = 0.021$

$Gain(credit) = 0.971 - 0 = 0.971$

เนื่องจาก Gain(credit) มีค่ามากที่สุด จึงเลือก Credit_rating เป็น node ต่อไปนี้

จะได้ Decision Tree ออกมา ดังนี้