DECISION TREE EXAMPLE

S.No.	Age	Income	Student	Credit Rating	Buys Computer
1	Youth	high	no	fair	no
2	Youth	high	no	excellent	no
3	Middle-age	high	no	fair	yes
4	Senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle-age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle-age	medium	no	excellent	yes
13	middle-age	high	yes	fair	yes
14	senior	medium	no	excellent	no

Attribute Selection :

Information Gain

Let us consider class : buys.Computer as Decision criteria D.

① Calculate information

$$-P_y \log_2 (P_y) - P_n \log_2 (P_n)$$

where P_y : Propability of 'yes'
P_n : probability of 'no'

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

$$= 0.940 \text{ bits}$$

② Calculate entropy for '<u>Youth</u>' for attribute age.

$$\text{Entropy 'youth'} = \underbrace{-\frac{2}{5}\log_2\frac{2}{5}}_{\downarrow \text{ yes}} - \underbrace{\frac{3}{5}\log_2\frac{3}{5}}_{\downarrow \text{ no}}$$

③ Calculate entropy for 'middle-age' for attribute age

$$\text{Entropy 'middle-age'} = \underbrace{-\frac{4}{4}\log_2\frac{4}{4}}_{\downarrow \text{ yes}} - \underbrace{\frac{0}{4}\log_2\frac{0}{4}}_{\swarrow \text{no}}$$

④ Similarly

$$\text{Entropy 'senior'} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

⑤ The expected information needed to classify a tuple in D if the tuples are partitioned according to age is

$$\begin{aligned} \text{info}_{age}(D) = {} & \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \\ & \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) + \\ & \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) \end{aligned}$$

$$\Rightarrow 0.694$$

⑥ Now,

Gain of Age : $\text{info}(D) - \text{info}_{age}(D)$

$$\Rightarrow 0.940 - 0.694$$

$$\Rightarrow \boxed{0.246}$$

⑦ Similarly,

$$\begin{aligned} \text{info}_{income}(D) = {} & \frac{4}{14} \times \left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \\ & \frac{6}{14} \times \left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) + \\ & \frac{4}{14} \times \left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right) \end{aligned}$$

$$\Rightarrow 0.911$$

Gain of income : $\text{info}(D) - \text{info}_{income}(D)$

$$\Rightarrow 0.940 - 0.911 = \boxed{0.029}$$

⑧ $$\text{info}_{\text{student}}(D) = \frac{7}{14} \times \left(-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right) + \frac{7}{14} \times \left(-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right)$$

$\Rightarrow 0.789$

Gain of student : $\text{info}(D) - \text{info}_{\text{student}}(D)$

$\Rightarrow 0.940 - 0.789$

$\Rightarrow \boxed{0.151}$

⑨ $$\text{info}_{\text{credit\_rating}}(D) = \frac{8}{14} \times \left(-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right) + \frac{6}{14} \times \left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right)$$

$\Rightarrow 0.892$

Gain of credit_rating : $\text{info}(D) - \text{info}_{\text{credit\_rating}}(D)$

$\Rightarrow 0.94 - 0.892$

$\Rightarrow \boxed{0.048}$

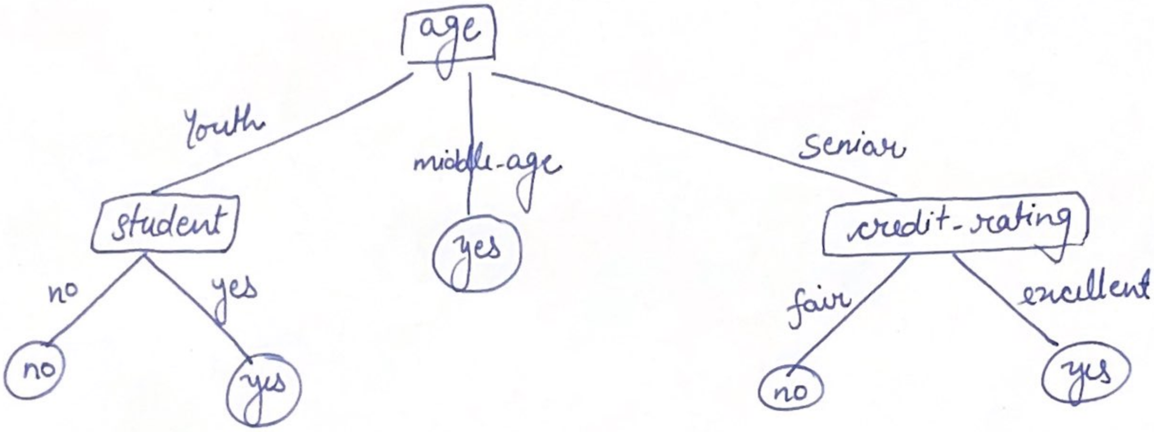
⑩ At last,

Independent variable	Information Gain
Age	0.246
Income	0.029
Student	0.151
Credit_rating	0.048

Because age has highest information gain among the attributes, it is selected as the splitting attribute.

- Tuples falling into the partition for age = middle-age all belong to the same class i.e. yes. Therefore, a leaf should be created at the end of the branch and labelled with "yes".

- Final Decision Tree


age
Youth
middle-age
senior
student
yes
credit-rating
no
yes
fair
excellent
no
yes
no
yes