Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	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_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Step:1: Entropy of D: Info(D)

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

where P_i is the nonzero probability that an arbitrary tuple in D belongs to class C_i and is estimated by $\frac{|C_{i,D}|}{|D|}$ = $\frac{\text{No. of tuples with same } i}{\text{Total No. of tuples}}$

for the given dataset

$$Info(D) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94.$$

Step:2: Consider the attributes in the dataset (input)
They are age, income, student, creditrating.
The attribute age has 3 categories 1) <=30 [2Y,3N]
2) 31..40 [4Y]
3) >40 [3Y,2N]
Calculate the Information gain for age. Find first the entropy.
[Y-Yes, N-No]

Because Gain(age) = Info(D) − Entropy(age) (or) $Info_{(age)}(D)$

Therefore, $Entropy(age) = Info_{Age}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$

$$Info_{Age}(D) = \frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{4}{14}\left(-\frac{4}{4}\log_2\frac{4}{4}\right) + \frac{5}{14}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$$

$$Info_{Age}(D) = \frac{5}{14}(0.9709) + \frac{4}{14} \times 0 + \frac{5}{14}(0.9709)$$

$Info_{Age}(D) = 0.6935$

$Gain(age) = Info(D) - Info_{Age}(D) = 0.94 - 0.6935$

$= 0.2465$

Step 3: Consider Attribute: Income.

This attribute has 3 categories such as high, medium and low.

high [2 Y, 2 N] [Y-Yes, N-No].

medium [4 Y, 2 N]

low [3 Y, 1 N]

$Entropy(income) = Info_{income}(D)$

$= \frac{4}{14}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) +$

$\frac{6}{14}\left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) +$

$\frac{4}{14}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right)$

$= \frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.811)$

$= 0.285 + 0.393 + 0.231$

$= 0.9108$

$Gain(income) = Info(D) - Info_{income}(D)$

$= 0.94 - 0.9108$

$= 0.0292.$

Step: 4 Consider the next attribute: Student

It is revealing whether the customer is a student or not. Categories: Yes [6Y 1N]

No [3Y 4N]

Class labels: Y-Yes, N-No.

$$\text{Entropy(Student)} = \text{Info}_{\text{Stud}}(D)$$

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

$$= \frac{7}{14}(0.5916) + \frac{7}{14}(0.9852)$$

$$= 0.2958 + 0.4926 = 0.7884$$

$$\text{Gain(Student)} = \text{Info}(D) - \text{Info}_{\text{Stud}}(D)$$

$$= 0.94 - 0.7884 = 0.1516$$

Step: 5: Consider the next attribute: Credit-Rating which has only 2 categories like, fair & Excellent

fair: [6Y, 2N]

excellent: [3Y, 3N]

$$\text{Entropy(C\_R)} = \text{Info}_{\text{C\_R}}(D)$$

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

$$= \frac{8}{14}(0.8112) + \frac{6}{14}(1)$$

$$= 0.4635 + 0.4285 = 0.8920.$$

$$\text{Gain(C\_R)} = \text{Info}(D) - \text{Info}_{\text{C\_R}}(D)$$

$$= 0.94 - 0.8920$$

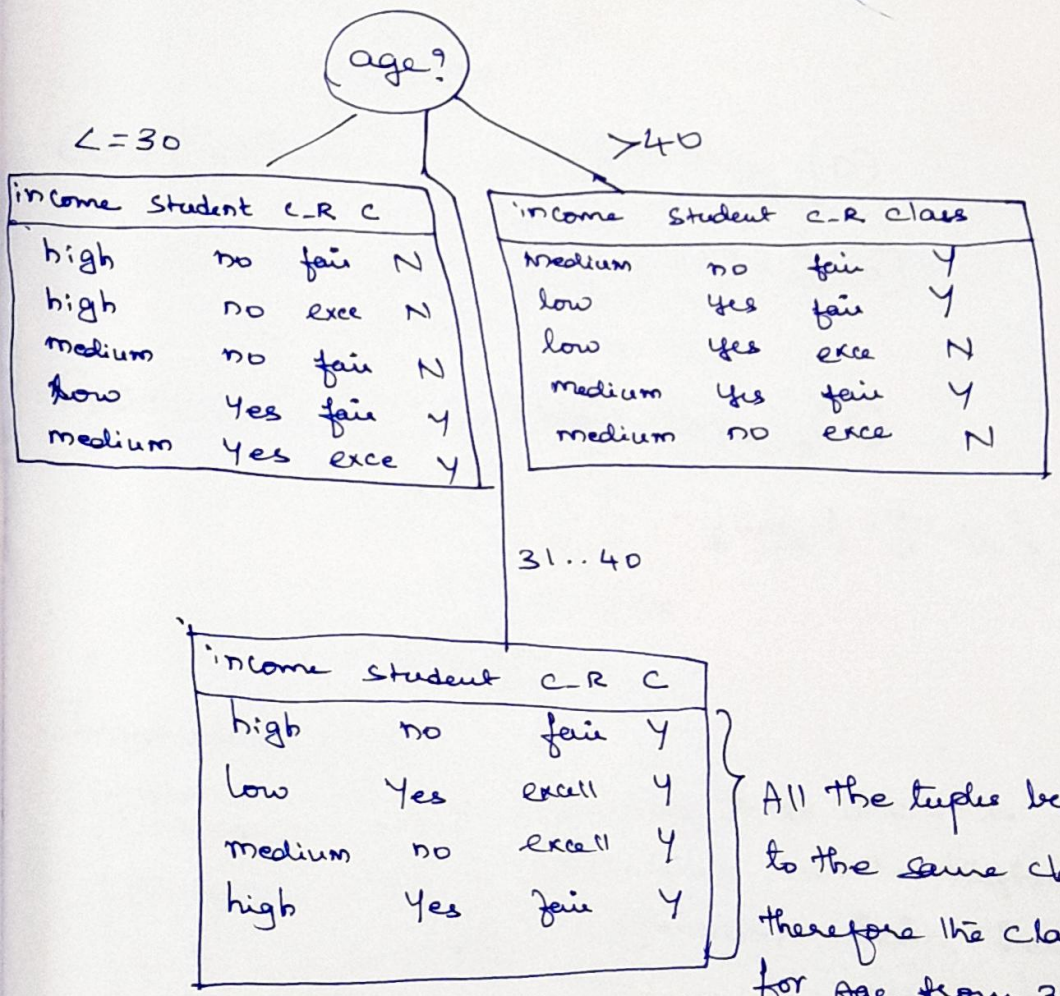
$$= 0.048$$

Step: 6

Information Gain

Age	0.2465
Income	0.0292
Student	0.1516
Credit Rtg	0.048

Attribute Age has the highest information Gain when compared to other attributes. Thus the Selected attribute for the root node of the desision tree is Age.



All the tuples belong to the same class and therefore the class label for Age from 31 to 40 is **Yes**.

Step: 7: Consider the table for age <=30

Income	Student	C_R	Class
high	no	fair	No
high	no	exce	No
medium	no	fair	No
low	Yes	fair	Yes
medium	Yes	excell	Yes

Attribute: Income

$$Info(D): Info(Age) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$$

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

$$Gain(income) = Info(Age) - Info_{income}(D)$$
$$= 0.97 - 0.4$$
$$= 0.57.$$

$$Gain(student) = Info(Age) - Info_{Stud}(D)$$
$$= 0.97 - \left[\frac{2}{5}(0) + \frac{3}{5}(0)\right]$$
$$= 0.97 - 0 = 0.97$$

$$Gain(C\_R) = Info(Age) - Info_{C\_R}(D)$$
$$= 0.97 - \left[\frac{3}{5}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{2}{5}(1)\right]$$
$$= 0.97 - 0.9508$$
$$= 0.0192.$$

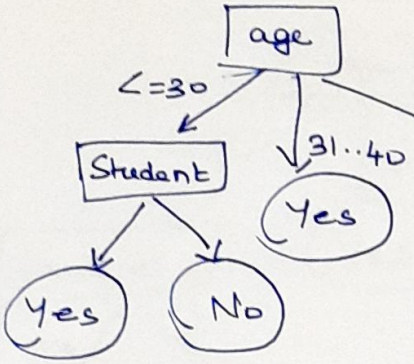
Information Gain:

Income	= 0.57
Student	= 0.97
C_R	= 0.0192

Among these attributes Student has the highest information Gain. Therefore the next attribute to be selected for the decision tree is Student.

Step: 8

Decision Tree at this level:



income	student	Credit	class
medium	no	fair	Yes
low	yes	fair	Yes
low	Yes	excell	No
Medium	yes	fair	Yes
Medium	no	excell	No

Now find the Information gain of the attributes again with respect to age >40.

$$\text{Info}(\text{Age}>40) = E(3,2) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= 0.97.$$

Find the Information Gain for income, Student, CR.

$$\text{Gain}(\text{income}) = 0.97 - \left[\frac{3}{5}\left(\left(-\frac{2}{3}\log\frac{2}{3}\right) - \left(\frac{1}{3}\log_2\frac{1}{3}\right)\right) + \frac{2}{5}(1)\right]$$

$$= 0.97 - (0.55 + 0.4)$$

$$= 0.97 - 0.95$$

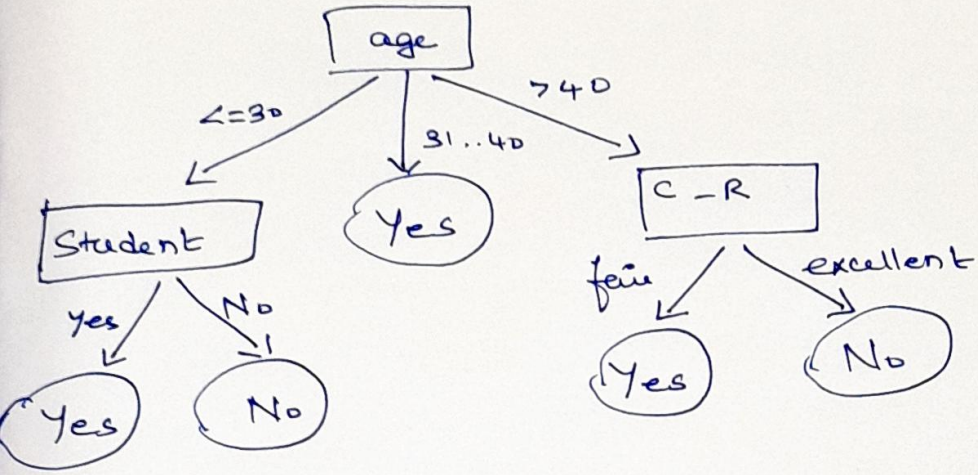
$$= 0.02.$$

$$\text{Gain}(\text{Student}) = 0.97 - 0.95 = 0.02$$

$$\text{Gain}(\text{C\_R}) = 0.97 - 0 = 0.97$$

Among the calculated Information Gain C_R has the highest gain. The Credit-rating is the selected attribute

The final Decision Tree (ID3 Algorithm)



For the new record the class label is:

age <=30, income = medium, Student = Yes

CR = fair. **Result : Yes**