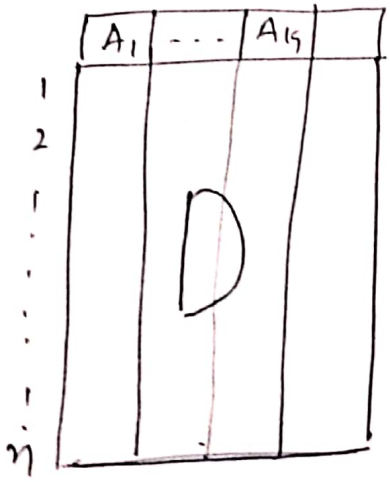
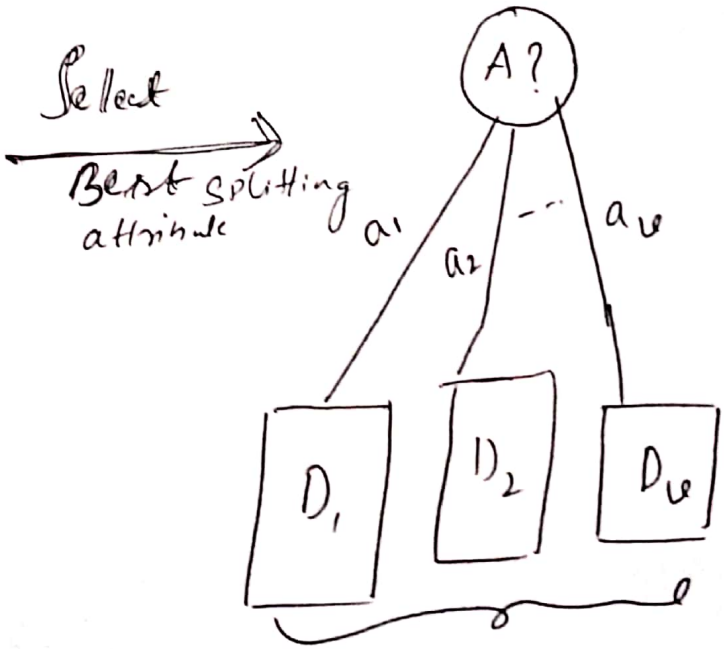
Attribute Selection Measures

An attribute selection Measure is a heuristic for selecting criterion that best seperates a given Data Partition, D of class-labeled training tuples in individual classes.

Selecting Criterion → Best Splitting attribute or Best Splitting point





Ideally, we Except resultant partitions to be pure i.e., after Splitting.

→ The attribute Selecting Measure Provides a Ranking for Lach attribute Describing given training tuples.

→ The attribute Selection measure provides a ranking for Lach attribute

→ The attribute having the best Score is choosen as the Splitting attribute for the given tuples.

→ The tree node Created for partition D is labelled with Splitting Criterion & branches grown for Lach outcome of Criterion & tuples are partitioned accordingly.

Attribute Selection Measures

(1) Information Gain
(2) Gain ratio
(3) Gini index.

Notation :

Let D, Data partition be a training Set of class-labeled tuples

Let m be the no. of Distinct classes.

C_i $(i = 1 \cdots m)$

Let C_{iD} be the Set of tuples of class C_i in D.

Let $|D|$ be the no. of tuples in D

Let $|C_{iD}|$ be no. of tuples in C_{iD}

Information Gain

→ ID3 Algorithm Uses Information Gain as its Attribute Selecting Measure.

→ This Measure is based on Shannon's Information Theory.

→ The attribute with the highest information gain is choosen as the Spliting attribute for Node N.

This attribute minimizes the information needed to classify tuples in the resulting partitions & reflects ~~resulting~~ These ~~base~~ partitions. least impurity in

The Expected information needed to classify a tuple in D is given by

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

Where

P_i = probability of a tuple in D belongs to class C_i

i.e.,

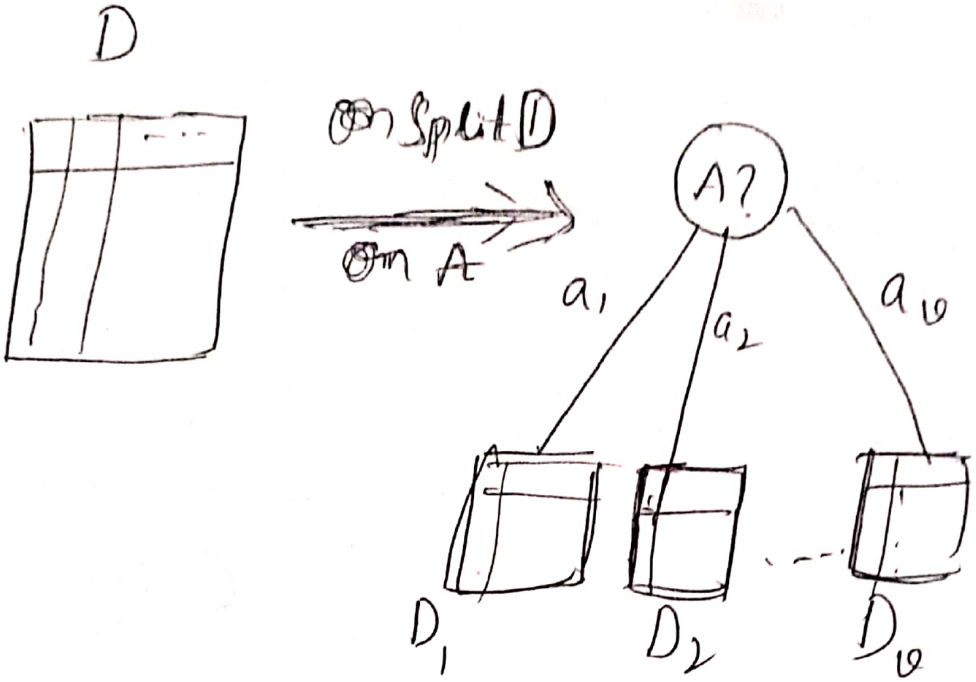
$$P_i = \frac{|C_i D|}{|D|} \quad \rightarrow (2)$$

Then, partition the tuples in D on some attribute A having v distinct values, $\{a_1, a_2 \cdots a_v\}$, as observed from the training Data.

Attribute A can be used to split D into v partitions $\{D_1, D_2 \cdots D_v\}$

where D_j = contains tuples in D that have outcome a_j of A.

i.e.



Then Information required after partitioning is given by on A

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

Where D_j = no. of ⊕ tuples in D_j

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

$$p_i = \frac{|C_i D_j|}{|D_j|}$$

Then

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

Where Gain(A) is Expected reduction in the information requirement on partition D by attribute (A)

Note: The attribute A with highest information gain is choosen as the splitting attribute at node N.

Ex:-

TID	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

$$\text{Info}(D) = -\sum_{i=1}^{2} p_i \log_2 p_i$$

$$= -\left[p_1 \log_2 p_1 + p_2 \log_2 p_2\right]$$

where $p_1 = \frac{|C_1 D|}{|D|}$ $\quad p_2 = \frac{|C_2 D|}{|D|}$

$= \frac{9}{14}$ $\quad p_2 = \frac{5}{14}$

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

$$\boxed{\text{Info}(D) = 0.940 \text{ bits}}$$

(x) Let partition D on attribute age

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

$$= \frac{|D_{age=youth}|}{|D|} Info\left(D_{age=youth}\right)$$

$$+ \frac{|D_{age=middle\_aged}|}{|D|} Info\left(D_{age=middle\_aged}\right)$$

$$+ \frac{|D_{age=Senior}|}{|D|} Info\left(D_{age=Senior}\right)$$

$$= \frac{5}{14} \times \left[-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right]$$

$$+ \frac{4}{14} \times \left[-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right]$$

$$+ \frac{5}{14} \times \left[-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right]$$

$$= 0.694 \text{ bits}$$

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

$$= 0.940 - 0.694$$

$$= 0.246 \text{ bits}$$

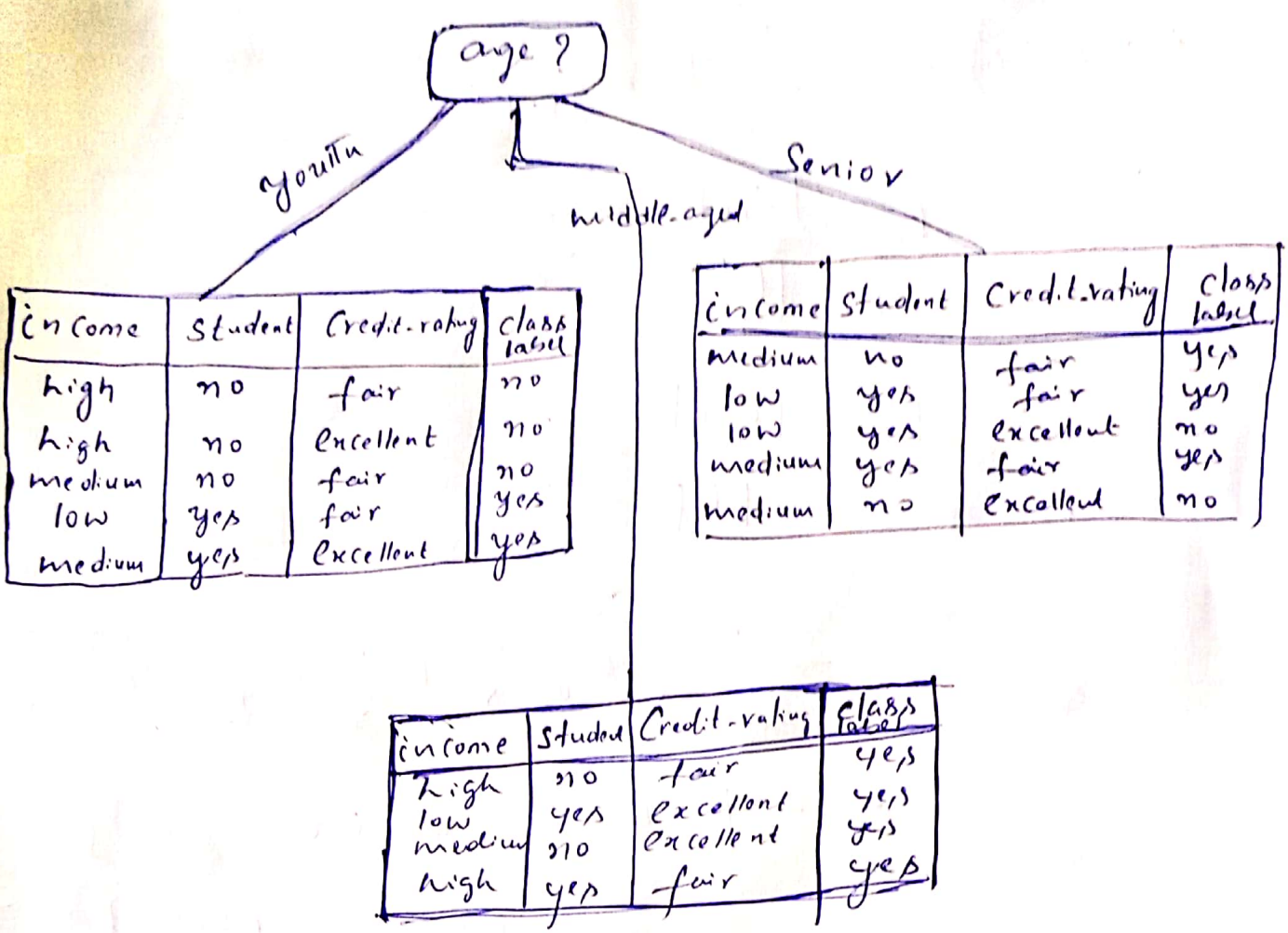
IIIly

Gain (income) = 0.029 bits

Gain (student) = 0.151 bits

Gain (credit-rating) = 0.048 bits.

∴ attribute age is selected as splitty attribute



income	Student	Credit-rating	class label
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	Excellent	yes

income	Student	Credit-rating	Class label
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	Student	Credit-rating	Class label
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

resultant first

Since The ↑ partition is impure,

Split partition by best Attribute

from remaining attribut list

= {income, Student, Credit-rating}

Then ⓓ find Information gain ⇒

for each remain attributes.

& Select ⓢ attribute with highest

Information gain.

$$Info(D) = -\sum_{i=1}^{2} p_i \log_2 p_i = -[P_1 \log_2 P_1 + P_2 \log_2 P_2]$$

$$P_1 = \frac{|C_1 D|}{|D|} = \frac{2}{5} = 0.4$$

$$P_2 = \frac{|C_2 D|}{|D|} = \frac{3}{5} = 0.6$$

$$= -\left[\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right]$$

$$= -[0.4(-1.321) + 0.6(-0.7369)]$$

$$= -[-0.5284 - 0.4421]$$

$$= \underline{0.9705} \text{ bits}$$

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

$$= \frac{|D_{high}|}{|D|} \times Info(D_{high})$$

$$+ \frac{|D_{medium}|}{|D|} \times Info(D_{medium})$$

$$+ \frac{|D_{low}|}{|D|} \times Info(D_{low})$$

$$= \frac{2}{5} \times \left[-\frac{0}{2} \log_2\left(\frac{0}{2}\right) - \frac{2}{2} \log_2\left(\frac{2}{2}\right) \right]$$

$$+ \frac{2}{5} \times \left[-\frac{1}{2} \log_2\left(\frac{1}{2}\right) - \frac{1}{2} \log_2\left(\frac{1}{2}\right) \right]$$

$$+ \frac{1}{5} \times \left[-\frac{1}{1} \log_2\left(\frac{1}{1}\right) - \frac{0}{1} \log_2\left(\frac{0}{1}\right) \right]$$

$$= 0.4 \times \left[-\frac{1}{2}(-1) - \frac{1}{2}(-1) \right]$$

$$= 0.4 \times \left[\frac{1}{2} + \frac{1}{2} \right]$$

$$= 0.4 [1]$$

$$= \underline{0.4} \text{ bits}$$

Gain(income)

$$= \text{Info}(D) - \text{Info}_{\text{income}}(D)$$

$$= 0.9705 - 0.4$$

Gain(income) = 0.5705 bits

$$\underset{\text{student}}{Info(D)} = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$$

$$= \frac{|D_{yes}|}{|D|} Info(D_{yes}) + \frac{|D_{no}|}{|D|} Info(D_{no})$$

$$= \frac{2}{5} Info(D_{yes}) + \frac{3}{5} Info(D_{no})$$

$$= \frac{2}{5}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right] + \frac{3}{5}\left[-\frac{0}{3}\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\log_2\left(\frac{3}{3}\right)\right]$$

$$= 0 \text{ bits}$$

Gain (student)

$$= Info(D) - \underset{\text{student}}{Info}(D)$$

$$= 0.9705 - 0$$

Gain(student) = 0.9705 bits

$$\underset{\text{Credit\_rating}}{Info(D)} = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$$

$$= \frac{|D_{fair}|}{|D|} Info(D_{fair})$$

$$+ \frac{|D_{excellent}|}{|D|} Info(D_{excellent})$$

$$= \frac{3}{5} Info(D_{fair})$$

$$+ \frac{2}{5} Info(D_{excellent})$$

$$= \frac{3}{5}\left[-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right]$$

$$+ \frac{2}{5}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]$$

$$= \frac{3}{5}\left[-(-0.5331) - (-0.399)\right]$$

$$= \frac{3}{5}\left[0.9321\right) + \frac{2}{5}(1)$$

$$= 0.5592 \text{ bits} + 0.4 = 0.9592$$

$$Gain(credit\_rating) = Info(D) - \underset{\text{credit\_rating}}{Info(D)}$$

$$= 0.9705 - 0.9592$$

$$Gain(credit\_rating) = 0.0113 \text{ bits}$$

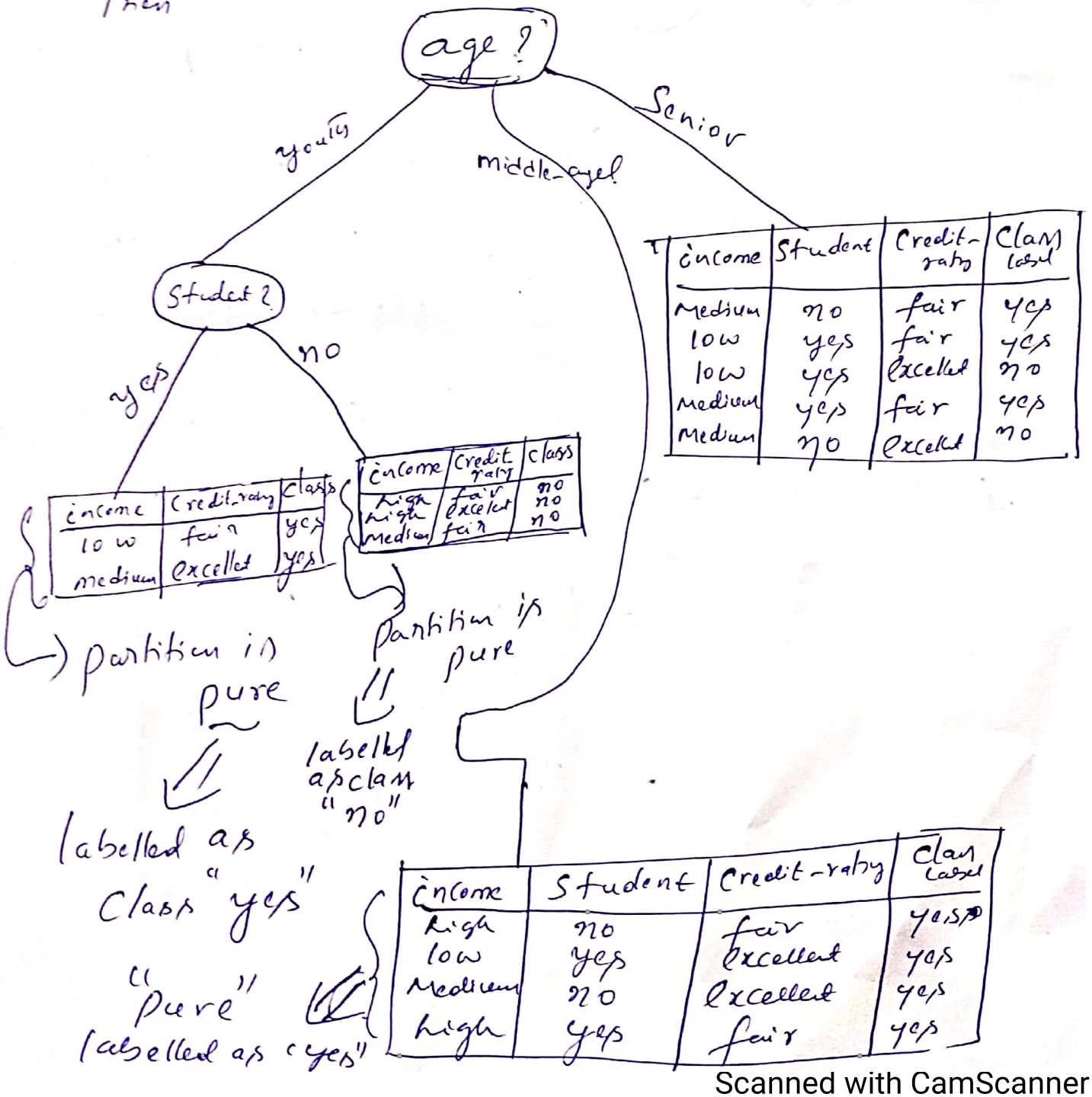
Since, Gain (income) = 0.5705 bits

Gain (student) = 0.9705 bits

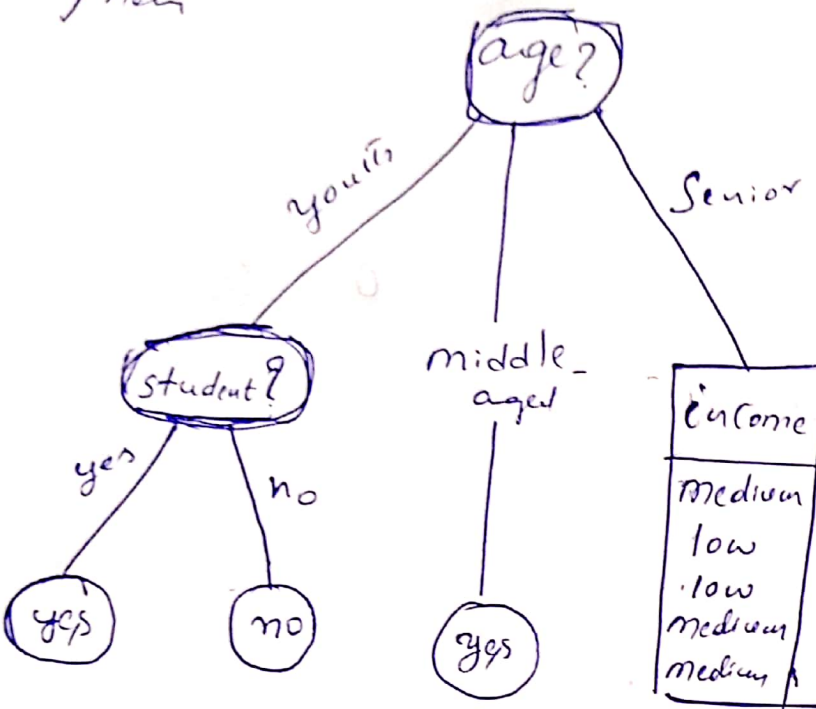
Gain (Credit.rating) = 0.0113 bits

Therefore student attribute will be selected for splitting partition.

Then



Then



income	Student	Credit-rating	Class:buys
Medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

Since impure partition Split the partition using next attribute from remaining attribute list

= { income, Credit_rating }

$$Info(D) = -\sum_{i=1}^{2} p_i \log_2 p_i$$

$$= -\left[P_1 \log_2 P_1 + P_2 \log_2 P_2\right]$$

$$= -\left[\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5}\right]$$

$$= -\left[0.6 \log_2 0.6 + 0.4 \log_2 0.4\right]$$

$$= 0.9705 \text{ bits}$$

$$\underset{\text{income}}{Info(D)} = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$$

$$= \frac{|D_{low}|}{|D|} Info(D_{low})$$

$$+ \frac{|D_{medium}|}{|D|} Info(D_{medium})$$

$$= \frac{2}{5}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]$$

$$+ \frac{3}{5}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$$

$$= \frac{2}{5}[1]$$

$$+ 3/5\,[0.9321]$$

$$= 0.9592$$

Gain(income)

$$= Info(D) - \underset{\text{income}}{Info(D)}$$

$$= 0.9705 - 0.9592$$

Gain(income) = 0.0113 bits

$$Info_{Credit\_rating}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$$

$$= \frac{|D_{fair}|}{|D|} Info(D_{fair})$$

$$+ \frac{|D_{excellent}|}{|D|} Info(D_{excellent})$$

$$= \frac{3}{5}\left[-\frac{3}{3}\log_2(3/3) - \frac{0}{3}\log_2(0/3)\right]$$

$$+ \frac{2}{5}\left[-\frac{0}{2}\log_2(0/2) - \frac{2}{2}\log_2(2/2)\right]$$

$= 0$ bits.

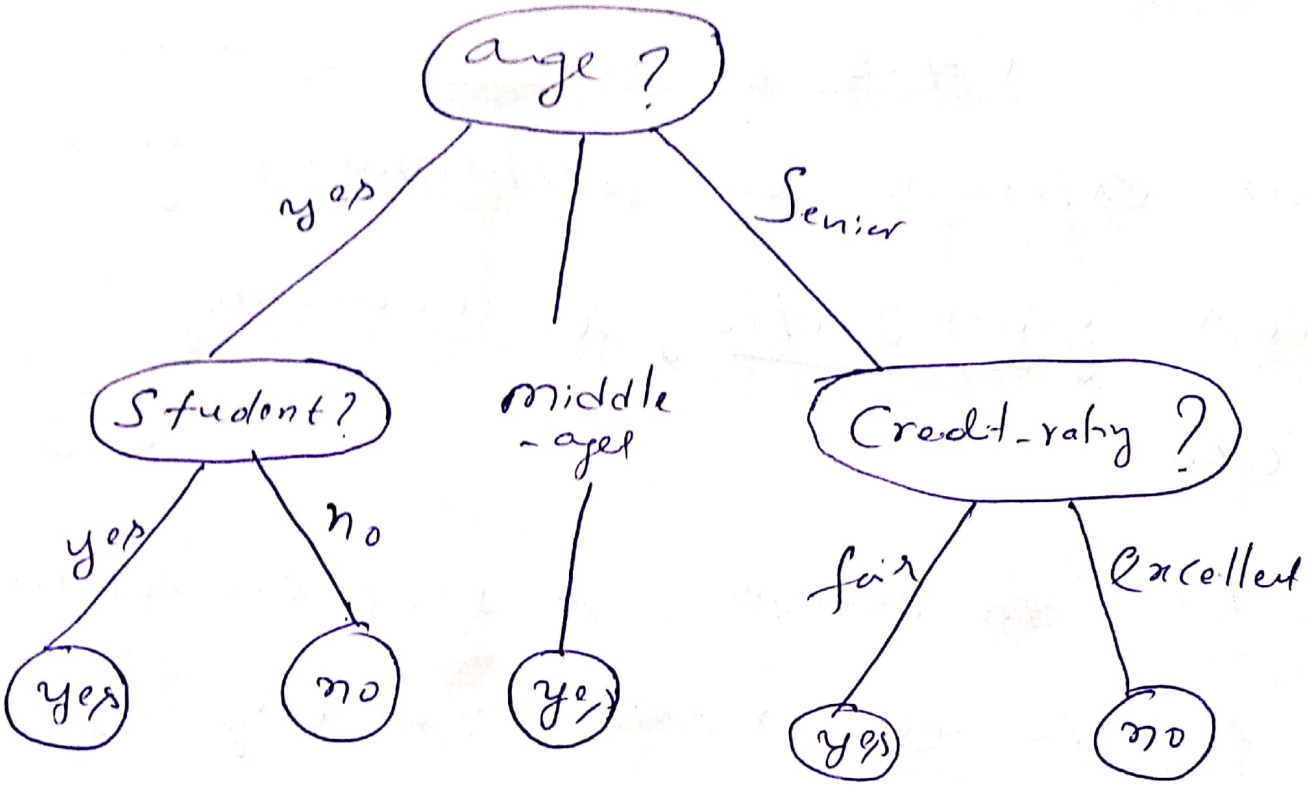
$Gain(Credit\_rating)$

$= 0.9705 - 0$

$= \underline{0.9705}$

Therefore, Attribute Credit_rating is selected for splitting a partition.

Tran



This value represents the potential information generated by splitting the training Data set D into v partitions, corresponding to the v outcomes of a test on attribute A.

Note:- The attribute with the maximum gain ratio is selected as the Splitting attribute.

Ex:- Consider attribute income,

$$\text{GainRatio}(\text{income}) = \frac{\text{Gain}(\text{income})}{\text{Split}_{\text{income}}(D)}$$

W.K.T, $\text{Gain}(\text{income}) = 0.029$

$$\text{Split}_{\text{income}}(D) = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \left(\frac{4}{14}\right) \times \log_2\left(\frac{4}{14}\right)$$

$$= 1.557$$

$$\text{Gain Ratio}(\text{income}) = \frac{0.029}{1.557} = 0.019$$

Next Consider attribute age

$$\text{Gainratio}(\text{Age}) = \frac{\text{Gain}(\text{Age})}{\text{Splitinfo}_{\text{Age}}(D)}$$

W.K.T $\text{Gain}(\text{Age}) = 0.246$

$$\text{Splitinfo}_{\text{age}}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

$$= -\frac{5}{14}\log_2\left(\frac{5}{14}\right) - \frac{4}{14}\log_2\left(\frac{4}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= -0.35\log_2(0.35) - 0.28\log_2(0.28) - 0.35\log_2(0.35)$$

$$= -[-0.5301 - 0.5142 - 0.5301]$$

$$= -[-1.5744]$$

$$= 1.5744$$

$$\text{Gainratio}(\text{Age}) = \frac{0.246}{1.5744} = 0.1562$$

Next Consider attribute → Student

$$\text{Gain ratio(student)} = \frac{\text{Gain(student)}}{\text{Split info}_{\text{student}}(D)}$$

WKT

$$\text{Gain(student)} = 0.151$$

$$\text{Split info}_{\text{student}}(D) = -\frac{7}{14}\log_2\left(\frac{7}{14}\right) - \frac{7}{14}\log_2\left(\frac{7}{14}\right)$$

$$= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$= -\frac{1}{2}(-1) - \frac{1}{2}(-1)$$

$$= \frac{1}{2} + \frac{1}{2}$$

$$= 1$$

$$\text{Gain ratio(student)} = \frac{0.151}{1}$$

$$= 0.151$$

Next Consider attribute Credit-rating

$$\text{Gainratio(Credit-rating)} = \frac{\text{Gain(Credit-rating)}}{\text{SplitInfo}_{\text{Credit-rating}}(D)}$$

WKT Gain(Credit-rating) = 0.048

$$\text{SplitInfo}_{\text{Credit-rating}}(D) = -\frac{8}{14}\log_2 8/14 - \frac{6}{14}\log_2$$

$$= -0.57 \log_2(0.57) - 0.42 \log_2(0.42)$$

$$= -[-0.4622 - 0.5256]$$

$$= 0.9878$$

$$\text{Gainratio(credit-rating)} = \frac{0.048}{0.9878}$$

$$= 0.0485$$

∴ ~~Gainratio~~ = Gainratio(age) = 0.1562
Gainratio(Student) = 0.151
Gainratio(income) = 0.019
Gainratio(Credit-rating) = 0.0485