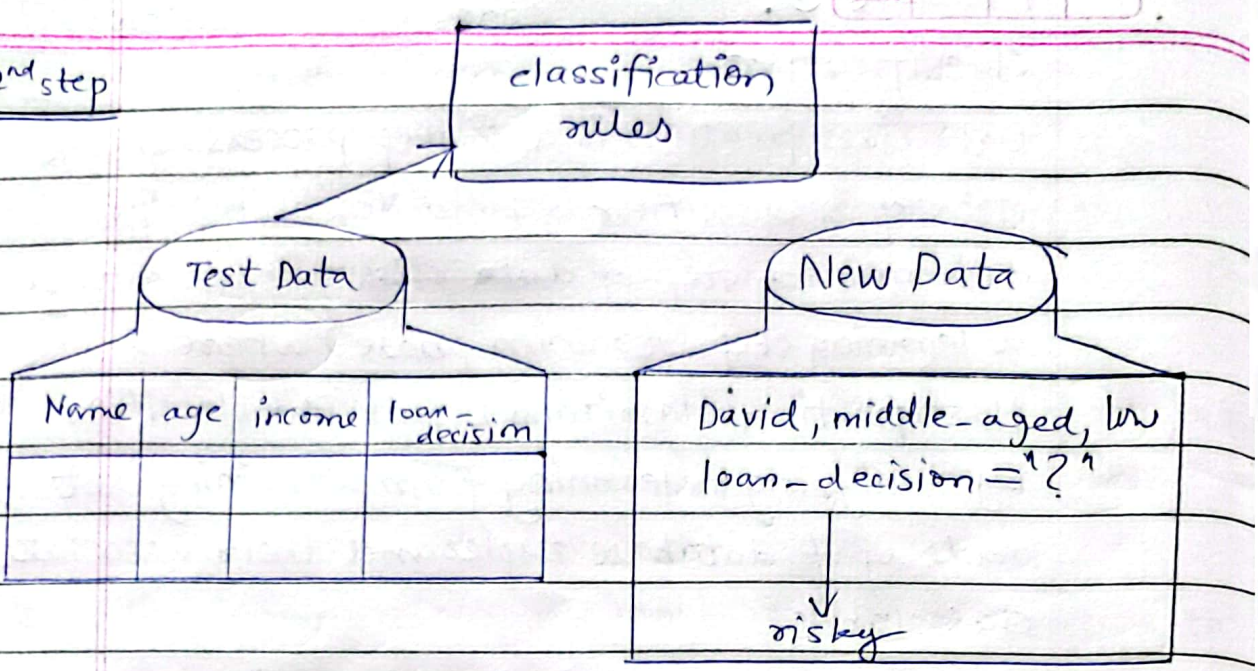
2nd step



Decision Tree

A tree is constructed to model the classification process. once the tree is build, it is applied to each tuple in the data base and results in the classification for that tuple.

Attribute Selection measure

It is selecting the best splitting criteria that separates the given data partition 'D' of class labelled training tuples into individual classes. They determined how the tuples at the given node are to be split. There are 3 popular attribute selection measures -

1) Information Gain
2) Gain Ratio
3) Gini index

1) Information Gain

let node N represent or hold the tuples of partition 'D'. The information to classify the tuple in 'D' is given by -

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

where, p_i is the probability that an arbitary tuple in D belongs to class C_i.

Info(D) is also known as entropy of D

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

$Info_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A.

→ Information gain is defined as the ~~requ~~ difference between the original information and the new requirement.

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

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

Database name

All electronics customer database.

Q1.

Rid	age	income	student	credit rating	class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	mid	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	mid	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	mid	medium	no	excellent	yes
13	mid	high	yes	fair	yes
14	senior	medium	no	excellent	no

First determine classes.

class C1 buys_computer = "yes"

class C2 buys_computer = "no"

no. of tuples with class C1 ("yes") = 9

no. of tuples with class C2 ("no") = 5

$$\text{Info}(D) = -\frac{c1}{c1+c2}\log_2\left(\frac{c_1}{c_1+c_2}\right) - \frac{c_2}{c_1+c_2}\log_2\left(\frac{c_2}{c_1+c_2}\right)$$

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

$$= 0.940$$

1st iter (I)

start with attribute age

Age category youth there are 2 yes tuples and 3 no tuples.

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

$$= 0.971$$

Age category middle aged there are 4 yes tuples and 0 no tuples.

$$I(4,0) = 0$$

Age category senior there are 3 yes tuples and 2 no tuples.

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

$$= 0.971$$

Calculate Entropy

$$E(age) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$= \frac{5}{14} \times I(2,3) + \frac{4}{14} \times I(4,0) + \frac{5}{14} \times I(3,2)$$

$$= 0.694$$

Information gain of age is

$$Gain(age) = Info(D) - E(age)$$

$$= 0.940 - 0.694$$

$$= 0.246$$

start with income attribute

Income attribute there a high, 2 yes tuples and 2 no tuples

$$I(2,2) = \frac{-2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right)$$

$$= -\log_2\left(\frac{2}{4}\right) = 1$$

Income category medium there are 4 yes tuples and 2 no tuples

$$I(4,2) = \frac{-4}{6}\log\left(\frac{4}{6}\right) - \frac{2}{6}\log\left(\frac{2}{6}\right)$$

$$= 0.918$$

Income category low there are 3 yes tuples and 1 no tuple.

$$I(3,1) = \frac{-3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right)$$

$$= 0.811$$

Calculate Entropy

$$E(\text{income}) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

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

$$= 0.911$$

$$\text{Gain (income)} = \text{Info}(D) - E(\text{Income})$$

$$= 0.940 - 0.911$$

$$= 0.029$$

start with student attribute.

Student category yes there are 6 yes tuples and 1 no tuple

$$I(6,1) = \frac{-6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)$$

$$= 0.592$$

student category no there are 3 yes tuples and 4 no tuples.

$$I(3,4) = \frac{-3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)$$

$$= 0.985$$

Calculate entropy

$$E(\text{student}) = \sum_{j=1}^{v} \frac{p_j + n_j}{p+n} I(p_j, n_j)$$

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

$$= 0.7885$$

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

$$= 0.940 - 0.7885$$

$$= 0.1515$$

Start with credit rating attribute.

credit rating category fair there are 6 yes tuples and 2 no tuples.

$$I(6,2) = \frac{-6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$$

$$= 0.811$$

credit rating category excellent there are 3 yes tuples and 3 no tuples

$$I(3,3) = \frac{-3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$$

$$= 1$$

Calculate entropy,

$$E(\text{credit rating}) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$= \frac{8}{14} I(6,2) + \frac{6}{14} I(3,3)$$

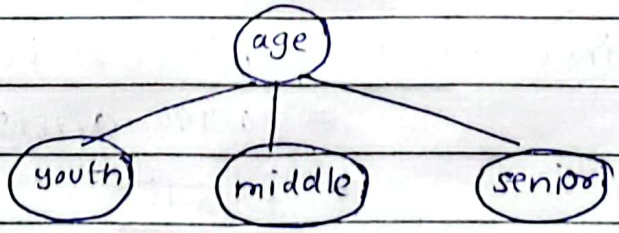
$$= 0.892$$

$$\text{Gain}(\text{credit rating}) = \text{Info}(D) - E(\text{credit rating})$$

$$= 0.940 - 0.892$$

$$= 0.048$$

Since, Age attribute has the highest information gain among the attributes. It is selected as the next attribute.



2nd Iter (II) Now age = youth total tuples = 5

There are 2 yes and 3 no.

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

Income category high ⟶ 0 yes and 2 no

$I(0,2) = 0$

Income = medium → 1 yes & 1 no

$I(1,1) = 1$

Income = low → 1 yes & 0 no

$I(1,0) = 0$

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

$$= 0.4$$

$$\text{Gain(income)} = \text{Info}(D) - E(\text{income}) = 0.571$$

now, student = yes → 2 yes and 0 no

$I(2,0) = 0$

student = no → 0 yes and 3 no

$I(0,3) = 0$

$E(\text{student}) = 0$

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

$$= 0.971$$

now, credit rating = fair → 1 yes and 2 no

$$I(1,2) = \frac{-1}{3} \log_2\left(\frac{1}{3}\right) - \frac{2}{3} \log_2\left(\frac{2}{3}\right) = 0.918$$

credit rating = excelent → 1 yes and 1 no

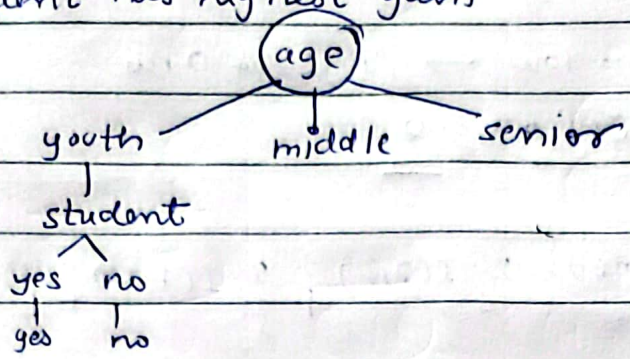
$I(1,1) = 1$

$$E(\text{credit rating}) = \frac{2}{5} I(1,1) + \frac{3}{5} I(1,2) = 0.951$$

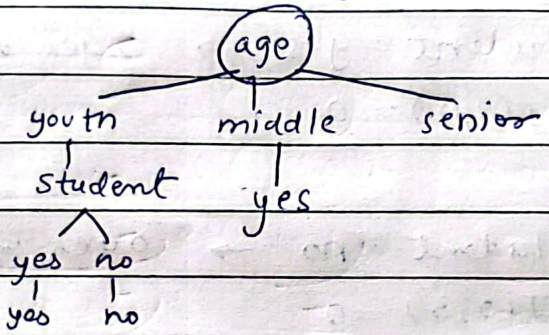
Gain (credit rating) = Info(D) − E(credit rating)

= 0.02

student has highest gain



Iter 3 (III) The no. of tuples for category middle age is 4 since the attribute's income & credit rating. The class : buys computer is yes. So you can assign the class yes to middle_aged.



Iter 4 (IV) senior age category age = senior tuples = 5

3 yes 2 no

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

Income = high → 0 yes & 0 no I(0,0) = 0

Income = medium → 2 yes & 1 no I(2,1) = 0.918

Income = low → 1 yes & 1 no I(1,1) = 1

$$E(\text{income}) = \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1) + \frac{0}{5} I(0,0)$$

= 0.951

Gain (Income) = Info (D) − E(income)

= 0.971 − 0.951 = 0.02

credit rating = fair → 3 Yes & 0 No

$I(3,0) = 0$

credit rating = excellent → 0 Yes & 2 No

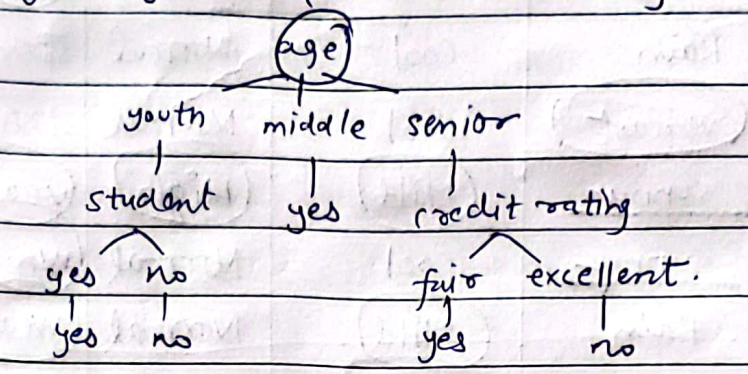
$I(0,2) = 0$

$E(\text{credit rating}) = 0$

$\text{Gain}(\text{credit rating}) = \text{Info}(D) - E(\text{credit rating})$

$= 0.971 - 0 = 0.971$

Highest gain is from credit rating



Classification Rules

→ If age = youth & student = yes then buys_computer = yes

→ If age = youth & student = no then buys_computer = no

→ If age = middle_aged then buys_computer = yes

→ If age = senior and credit_rating = fair then buys_computer = yes

→ If age = senior and credit_rating = excellent then buys_computer = no.

Classification

(I) Decision Tree

Q1. Table.

Day	Outlook	Temperature	Humidity	Wind	Class : play Football
D1	Sunny	Hot	High	weak	NO
D2	Sunny	Hot	High	strong	NO
D3	overcast	Hot	High	weak	YES.
D4	Rain	Mild	High	weak	YES.
D5	Rain	cool	Normal	Weak	YES
D6	Rain	Cool	Normal	strong	NO.
D7	overcast	cool	Normal	strong	YES
D8	sunny	Mild	High	weak	NO.
D9	sunny	cool	Normal	Weak	YES
D10	Rain	Mild	Normal	weak	YES
D11	Sunny	Mild	Normal	strong	YES.
D12	overcast	Mild	High	strong	YES.
D13	overcast	Hot	Normal	Weak	YES
D14	Rain	Mild	High	strong	NO

class C1 : plays football = YES

class C2 : plays football = NO.

Iteration 1

Total number of records = 14

Number of Tuples with class C1 = 9

Number of Tuples with class C2 = 5

$$\text{Info}(D) = -\left(\frac{C1}{C1+C2}\right)\log_2\left(\frac{C1}{C1+C2}\right) - \left(\frac{C2}{C1+C2}\right)\log_2\left(\frac{C2}{C1+C2}\right)$$

$$= 0.940$$

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

$$= 0.940$$

For Outlook Property Attribute.

outlook category sunny there are 2 YES tuples and 3 NO tuples.

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

$$= 0.971$$

outlook category overcast there are 4 YES tuples and 0 NO tuples.

$$I(4,0) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)$$

$$= 0$$

outlook category Rain there are 3 YES tuples and 2 NO tuples.

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

$$= 0.971$$

Calculate entropy

$$E(\text{outlook}) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$= \frac{5}{14} \underset{0.971}{I(2,3)} + \frac{4}{14} \underset{0}{I(4,0)} + \frac{5}{14} \underset{0.971}{I(3,2)}$$

$$= 0.6935.$$

Calculate Info Gain

Info Gain = Info(D) − E(outlook)

= 0.940 − 0.6935

≈ 0.2465

For Temperature Attribute,

Temp category Hot there are 2 YES tuples and 2 NO tuples.

$$I(2,2) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$

$$= 1$$

Temp category Mild there are 4 YES tuples and 2 NO tuples.

$$I(4,2) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)$$

$$= 0.9182$$

Temp category Cool there are 3 YES tuples and 1 NO tuple

$$I(3,1) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$$

$$= 0.8112$$

$$E(\text{temperature}) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

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

$$= \frac{4}{14}(1) + \frac{6}{14}(0.9182) + \frac{4}{14}(0.8112)$$

$$= 0.911$$

Info Gain = Info (D) − E (temperature)

= 0.940 − 0.911

= 0.029

For Humidity Attribute;

High → 3 YES and 4 NO tuples.

$$I(3,4) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)$$

$$= 0.9852$$

Normal → 6 YES and 1 NO tuples.

$$I(6,1) = -\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)$$

$$= 0.5916$$

$$E(\text{humidity}) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$= \frac{7}{14} \underset{0.9852}{I(3,4)} + \frac{7}{14} \underset{0.5916}{I(6,1)}$$

$$= 0.7884$$

Info Gain = Info (D) − E (humidity)

= 0.940 − 0.7884

= 0.1516

For Wind attribute,

weak → 6 YES and 2 NO tuples

$$I(6,2) = \frac{-6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$$

$$= 0.8112$$

strong → 3 YES and 3 NO tuples

$$I(3,3) = \frac{-3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$$

$$= 1.$$

$$E(\text{wind}) = \sum_{j=1}^{v} \frac{p_j + n_j}{p+n} I(p_i, n_i)$$

$$= \frac{8}{14} I(6,2) + \frac{6}{14} I(3,3)$$

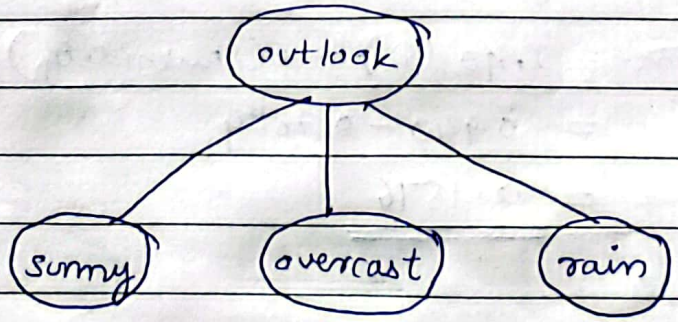
$$= \frac{8}{14}(0.8112) + \frac{6}{14}(1)$$

$$= 0.8921$$

$$\text{Info Gain} = \text{Info}(D) - E(\text{wind})$$

$$= 0.940 - 0.8921$$

$$= 0.0479$$



Iteration 2 - outlook = sunny.

For attribute outlook = sunny

Tuples = 5

There are 2 YES and 3 NO

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

Info(D) = 0.9709

Check for attribute temperature under outlook = sunny.

temperature = ~~high~~ hot → 0 YES and 2 NO.

$$I(0,2) = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)$$

$$= 0$$

temperature = mild → 1 YES and 1 NO

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

$$= 1$$

temperature = cool → 1 YES and 0 NO

$$I(1,0) = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)$$

$$= 0$$

$$E(\text{temperature}) = \frac{2}{5}I(0,2) + \frac{2}{5}I(1,1) + \frac{1}{5}I(1,0)$$

$$= 0.4$$

Gain (temperature) = Info(D) - E(temp)

$$= 0.9709 - 0.4 = 0.5709$$

check for attribute humidity under outlook = sunny

humidity = high → 0 yes and 3 No

$I(0,3) = 0$

humidity = normal → 2 Yes and 0 No

$I(2,0) = 0$

$E(\text{humidity}) = 0$

$\text{Gain}(\text{humidity}) = \text{Info}(D) - E(\text{humidity})$

$= 0.971 - 0 = 0.971$

check for attribute wind under outlook = sunny

wind = weak → 1 Yes and 2 No

$I(1,2) = \frac{-1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.918$

wind = strong → 1 Yes and 1 No.

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

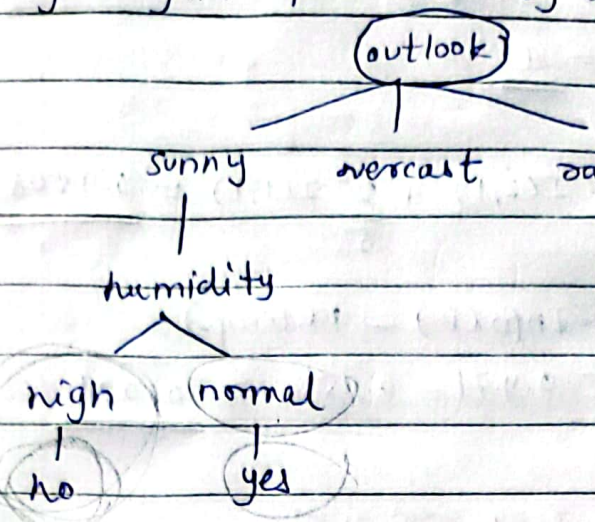
$E(\text{wind}) = \frac{3}{5} I(1,2) + \frac{2}{5} I(1,1)$

$= 0.9508$

$\text{Gain}(\text{wind}) = 0.971 - 0.9508$

$= 0.021$

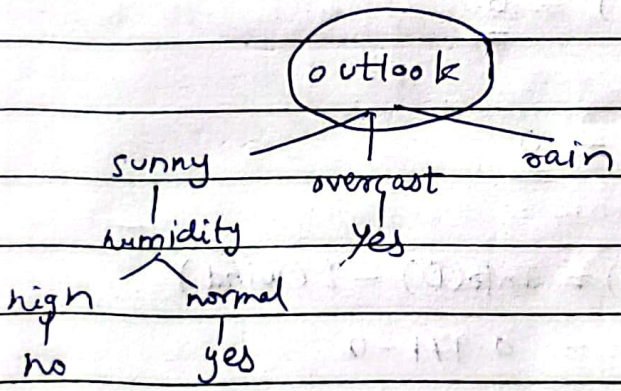
Highest gain is from humidity -



Iteration 3 - outlook = overcast

Total tuples = 4

there are 4 Yes & 0 No hence outcome will always be yes.



3 Yes & 2 No

Iteration 4 - outlook = rain Total tuples = 5

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

$$= 0.971$$

check for attribute temperature under outlook = rain

temp = hot → 0 Yes & 0 No

$I(0,0) = 0$

temp = mild → 2 Yes & 1 No

$$I(2,1) = \frac{-2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.918$$

temp = cool → 1 Yes & 1 No

$I(1,1) = \underline{1}$

$E(temp) = \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1) = \underline{0.9508}$

Gain (temp) = Info(D) − E(temp)

= 0.971 − 0.9508 = 0.021

check for attribute wind under outlook = rain

wind = weak ⟶ 3 Yes & 0 No

$I(3,0) = \underline{0}$

wind = strong → 0 Yes & 2 No

$I(0,2) = \underline{0}$

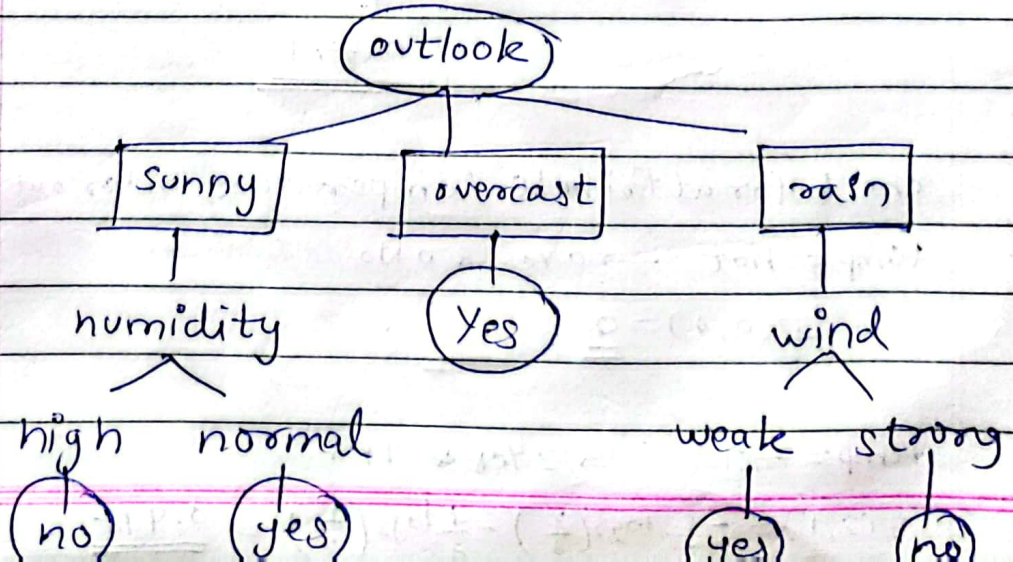
$E(wind) = \underline{0}$

Gain (wind) = Info(D) − E(wind)

= 0.971 − 0

= 0.971

Since, wind has highest gain consider wind



Classification Rules

→ If outlook = sunny and humidity = high then play_football = yes'

→ If outlook = sunny and humidity = normal then play_football = no

→ If outlook = overcast then play_football = yes

→ If outlook = rain and wind = weak then play_footall = yes

→ If outlook = rain and wind = strong then play_football = no