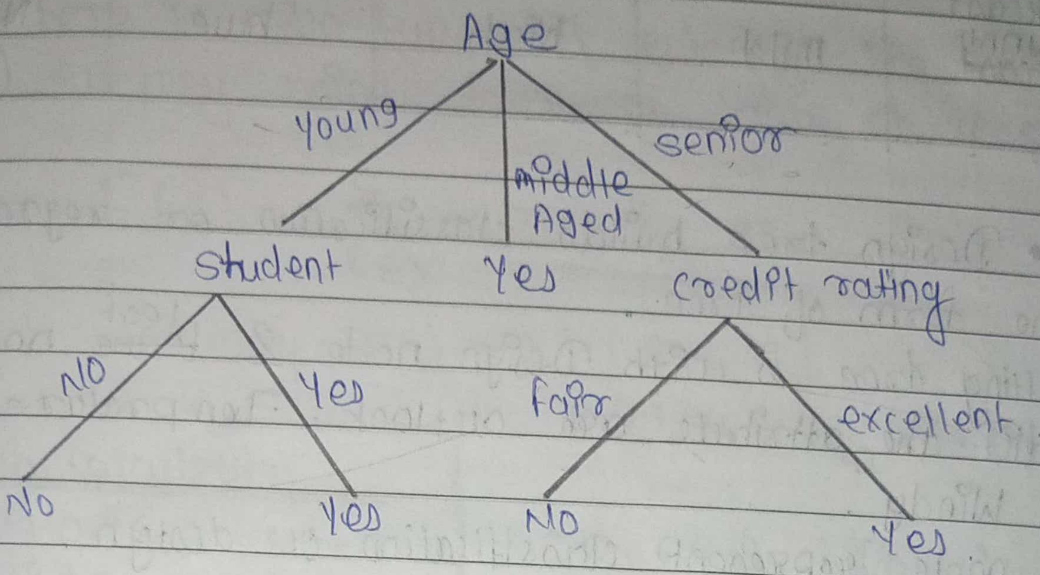
Page No. | Date

11 - may - 2022

Decision Tree



Imp Que. Consider following dataset

Sr.no.	Outlook	Temp	Humidity	windy	Play golf
1)	Rainy	Hot	High	false	No
2)	Rainy	Hot	High	True	No
3)	overcast	HOT	High	false	Yes
4)	Sunny	mild	High	false	yes
5)	Sunny	Cool	Normal	false	Yes
6)	Sunny	Cool	Normal	True	No
7)	overcast	cool	Normal	True	Yes
8)	Rainy	mild	High	false	No
9)	Rainy	Cool	Normal	false	Yes
10)	sunny	mild	Normal	false	Yes
11)	Rainy	mild	Normal	True	Yes

12)	overcast	mild	High	True	Yes
13)	Overcast	hot	Normal	false	Yes
14)	~~Overcast~~ sunny	mild	high	True	NO.

- Design tree builds classification or regression model in the form of tree.
- Resulting tree is with Design node & ~~leave~~ leaf node.
- predict the attribute are outlook, Tempreture, Humidity and Windy.
- Leaf node represent classification or design.
-

* Algorithm.

- The core algorithm for building design tree is called ID3 (Induction Diplomizer) which was proposed by J. R. quinlan.
- ID3 uses entropy and information gain to construct tree.

* entropy

- Design tree is build in top-down approch ~~the~~ and involves partitioning the data into subsets that contains instances with similar values.
- ID3 uses entropy to calculate the homogenity / Homogeneous of the sample
- If sample is homogenous then entropy is 0.
- If sample is not homogenous then entropy is different from '0'.

- (If sample is equaly divided then it has entropy of 1)

There is two types of entropy :

1) entropy using frequency table of a one attribute.

$$E(s) = \sum_{c \in X}^{c} - p_i \log_2 p_i$$

$E(\text{play golf}) = E(9,5)$

$= -P(9/14)\log_2 P(9/14) - P(5/14)\log_2 P(5/14)$

$= -0.64 \log_2(0.64) - 0.36 \log_2 0.36$

$= -0.64 \times (-0.63) - 0.36(-1.46)$

$= 0.40\cancel{32} + 0.53$

$= 0.93$

Rough calculation

$0.36 = \frac{\log 0.36}{\log 2} = \frac{-0.44}{0.30}$

$= -1.46$

$0.64 = \frac{\log 0.64}{\log 2} = \frac{-0.19}{0.30}$

$= -0.63$

2) entropy by using frequency table of Two attribute.

$$E(T, X) = \sum_{c \in X} P(c)\, E(c)$$

$E(\text{playgolf, outlook}) = P(\text{sunny})\, E(\text{sunny}) + P(\text{Rainy})\, E(\text{Rainy}) + P(\text{overcast})\, E(\text{overcast})$

$= \cancel{P(5/14)}\ \cancel{E(3,2)}\ (3,2) + P(5/14)\ \cancel{E(2,3)}\ (2,3) + \cancel{P(4/14)}$

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i .$$

Page No.
Date

		Playgolf		Total
		yes	No	
Outlook	Rainy	2	3	5
	Sunny	3	2	5
	over-cast	4	0	4

Total = 14

$$= P(5/14)\, \underline{E(3,2)} + P(5/14)\, \underline{E(2,3)} + P(4/14)\, \underline{E(4,0)}$$

$$= (0.35)\left(-(3/5)\log_2(3/5) - (2/5)\log_2(2/5)\right) +$$

$$(0.36)\left(-(2/5)\log_2(2/5) - (3/5)\log_2(3/5)\right) +$$

$$0.28\left(-(4/4)\log_2(4/4) - (0/4)\log_2(0/4)\right)$$

$$= (0.36 \times 0.97) + (0.36 \times 0.97) + (0.28 \times 0)$$

$$= 0.349 + 0.349 + 0$$

E(playgolf, outlook) = 0.698

3) Information gain is based upon decrease in entropy after a dataset is split on an attribute. constructing a design tree is all about finding an attribute that returns ~~a~~ highest information gain. formula to calculate the gain is :

Page No.
Date

$$\text{Gain}(\overset{\text{Target}}{T}, \overset{\text{Selected}}{X}) = \text{Entropy}(T) - E(T;X)$$

Gain (Playgolf, outlook) = Entropy (Playgolf) - F (Playgolf, Outlook).

= 0.93 - 0.698

Gain (Playgolf, outlook) = 0.24.

Que. construct the disision tree for a given dataset by using identity

→ Step 1 :

To compute information given for attributes (playgolf, outlook)

(a) To calculate entropy of target attribute (playgolf).

$E(T) = E(\text{Playgolf}) = \underline{0.93}$

(b) Information gain between a target attribute is (Playgolf, outlook)

$$E(T, X) = \sum_{C \in X} P(C)\, E(C)$$

$E(\text{playgolf}, \text{outlook}) = \underline{0.693}$

(c) Gain (T,x) = Gain (Playgolf, outlook)

= E(T) - E(T, x)

= E (Playgolf) - E (playgolf, outlook)

$= 0.93 - 0.693$
$= 0.247$

$Gain(Playgolf, outlook) = \underline{0.247}$

Step 2 :

To complete information gain for attribute. (Playgolf, temp).

		Playgolf		Total
		Yes	No.	
Temp	Hot	2	2	4
	cool	3	1	4
	mild	4	2	6

Total = 14

$E(T, X) = E(Playgolf, \cancel{outlook}\ temp)$

$= \sum_{c \in X} P(c)\ E(c)$

$= P(Hot)\ E((2/4), (2/4)) + P(cool)\ E((3/4), (1/4)) + P(mild)\ E((4/6), (2/6))$

$= (4/14)\ E(0.5, 0.5) + (4/14)\ E(0.75, 0.25) + (6/14)\ E(0.67, 0.33)$

$= 0.29(-0.5 \log_2 0.5 - 0.5 \log_2 0.5) +$

$0.29(-0.75 \log_2 0.75 - 0.25 \log_2 0.25) +$

$0.43(-0.67 \log_2 0.67 - 0.33 \log_2 0.33)$

$= 0.29(1) + 0.29(0.81) + 0.43(0.92)$

$= 0.29 + 0.23 + 0.39$

$= 0.91$

E (Playgolf, ~~Outlook~~ Temp) = 0.91

Gain (Playgolf, temp) = E (Playgolf) − E(Playgolf, temp)

$= 0.93 - 0.91$

$= 0.02$

Gain (Playgolf, temp) = 0.02

Step 3 :

To compute information gain for attribute (Playgolf, Humidity)

		Playgolf		Total
		Yes	No	
Humidity	High	3	4	7
	Normal	6	1	7

Total = 14

E(T, x) = E (Playgolf, Humidity)

$= \sum_{c \in x} P(c) E(c)$

$= P(\text{High}) E((3/7), (4/7)) + P(\text{Normal}) E((6/7), (1/7))$

Page No.
Date

$= P(7/14)\ E(0.43, 0.57) + P(7/14)\ E(0.88, 0.14)$

$= 0.5(-0.43 \log_2 0.43 - 0.57 \log_2 0.57) +$

$0.5(-0.88 \log_2 0.88 - 0.14 \log_2 0.14)$

$= 0.5(0.99) + 0.5(0.58)$

$= 0.49 + 0.29$

$= 0.79$

E (playgolf, Humidity) = 0.79

Gain (playgolf, humidity) = E (playgolf) − E (playgolf, Humidity)

$= 0.93 - 0.79$
$= 0.14$

Gain (playgolf, humidity) = 0.14

Step 4 :

To compute information gain for attribute (playgolf, windy)

		Playgolf Yes	Playgolf No	Total	
windy	True	3	3	6	Total = 14
	false	6	2	8	

Page No.
Date

$E(T, X) = E(\text{Playgolf}, \text{windy})$

$= \sum_{c \in X} P(c) E(c)$

$= P(\text{True}) E((3/6), (3/6)) + P(\text{false}) E((6/8), (2/8))$

$= p(6/14) E(0.5, 0.5) + P(8/14) E(0.75, 0.25)$

$= 0.43 (-0.5 \log_2 0.5 - 0.5 \log_2 0.5) +$

$0.57 (-0.75 \log_2 0.75 - 0.25 \log_2 0.25)$

$= 0.43 (1) + 0.57 (0.81)$

$= 0.43 + 0.46$

$= 0.89$

$E(\text{Playgolf}, \text{windy}) = \underline{0.89}$

$\text{Gain}(\text{playgolf}, \text{windy}) = E(\text{Playgolf}) - E(\text{Playgolf}, \text{windy})$

$= 0.93 - 0.89$

$= 0.04$

$\text{Gain}(\text{Playgolf}, \text{windy}) = \underline{0.04}$

Step 5 :

choose the attribute with largest information gain as design node and divide the dataset by its branches repeat the same process on every branch.

Outlook

Rainy

	Temp	Humidity	windy	Playgolf
1)	Hot	High	false	NO
2)	Hot	High	True	NO
3)	mild	High	false	NO
4)	cool	Normal	false	yes
5)	mild	Normal	True	yes

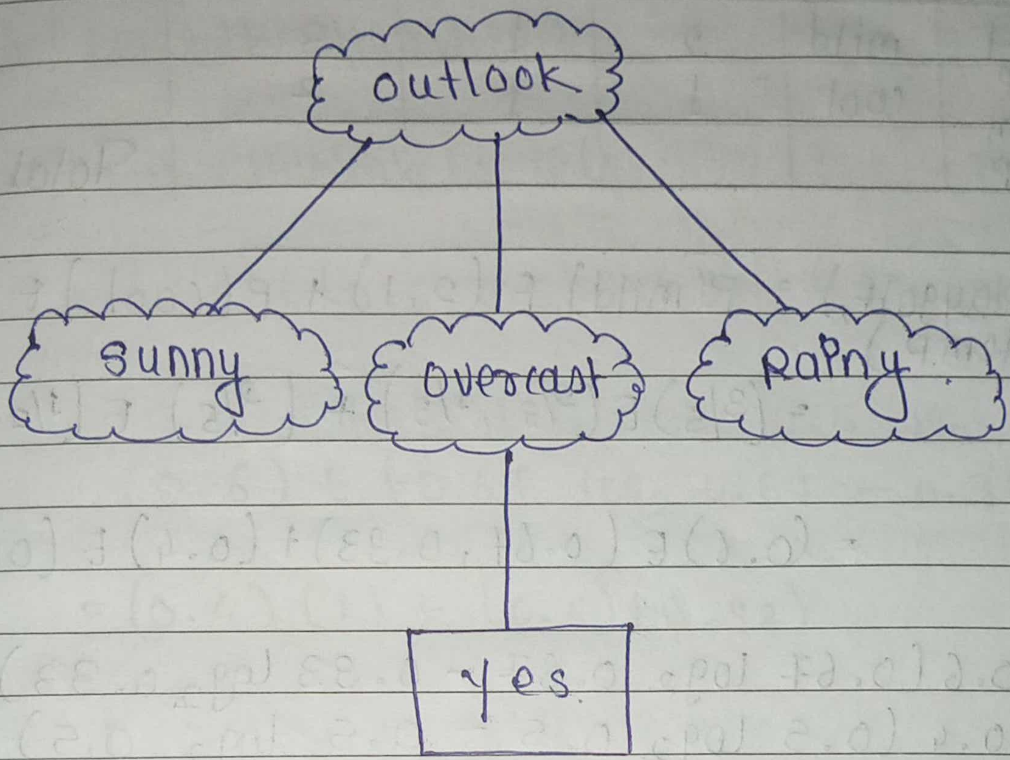
Sunny

1)	mild	High	false	yes
2)	cool	normal	false	yes
3)	cool	normal	True	No
4)	mild	normal	False	yes
5)	mild	high	True	No

overcast

1)	Hot	High	false	yes
2)	cool	Normal	True	yes
3)	mild	High	True	yes
4)	Hot	Normal	false	yes

∴ The part of the ~~design~~ dicision tree is :-



Step 6 :

A branch with entropy with more than 0 needs to be split up.

$$E(B)\ E(\text{playgolf}) = E(3|2)$$

$$= E(3/5, 2/5)$$

$$= E(0.6, 0.4)$$

$$= E(0.6 \log_2 0.6) - (0.4 \log_2 0.4)$$

$$E(\text{playgolf}) = \underline{0.97}$$

		Playgolf		Total
		Yes	NO	
Temp	mild	2	1	3
	cool	1	1	2

Total = 5

$E(\text{Playgolf, temp}) = P(\text{mild})\, E(2,1) + P(\text{cool})\, E(1,1)$

$= (3/5)\, E(2/3, 1/3) + (2/5)\, E(1/2, 1/2)$

$= (0.6)\, E(0.67, 0.33) + (0.4)\, E(0.5, 0.5)$

$= 0.6(0.67 \log_2 0.67 - 0.33 \log_2 0.33) + 0.4(0.5 \log_2 0.5 - 0.5 \log_2 0.5)$

$= 0.6(0.92) + 0.4(1)$

$= 0.55 + 0.4$

$= 0.95$

$E(\text{playgolf, temp}) = \underline{0.95}$

information gain (playgolf, temp) = E(Playgolf) − E(playgolf, temp)

$= 0.97 - 0.95$

$= \underline{0.02}$

Page No.
Date

Step 7 :

Select attribute humidity.

E (playgolf , Humidity) :

$E(\text{playgolf}, \text{Humidity}) = P(\text{High})E(1,1) + P(\text{normal})E(2,1)$

$= (2/5)E(0.5, 0.5) + (3/5)E(0.67, 0.33)$

$= (0.4)E(-0.5\log_2 0.5 - 0.5\log_2 0.5) +$
$(0.6)E(-0.67\log_2 0.67 - 0.33\log_2 0.33)$

$= (0.4)(1) + (0.6)(0.92)$

$= 0.4 + 0.552$

$= 0.95$

$E(\text{Playgolf}, \text{Humidity}) = \underline{\underline{0.95}}$

info gain (play golf, humidity) = E (playgolf) − E (Playgolf, Humidity)

$= 0.97 - 0.95$

info gain (playgolf, Humidity) = $\underline{\underline{0.02}}$

Step 8 :

attribute windy.

E (playgolf , windy).

		Playgolf		Total
		Yes	No	
windy	True	1	2	3
	false	2	0	2

Total = 5

$$E(\text{playgolf}, \text{windy}) = P(\text{True})\, E(0,2) + P(\text{False})\, E(3,0)$$

$$= (3/5)\, E(0, 2) + (2/5)\, E(3,0)$$

$$= \cancel{(0.6)\, E(0,1) + (0.4)\, E(1,0)}$$

$$= \cancel{(0.6)\, E(0, 2/5) + (2/5)\, E(3/5, 0)}$$

$$= (0.6)\, E(0,1) + (0.4)\, E(1,0)$$

$$= 0.6(-0 \log_2 - 1 \log_1) + 0.4(-1 \log_2 1 - 0 \log_2 0)$$

$$= 0.6(0) + 0.4(0)$$

$$= 0 + 0$$

$$= 0$$

$$E(\text{playgolf}, \text{windy}) = \underline{0}$$

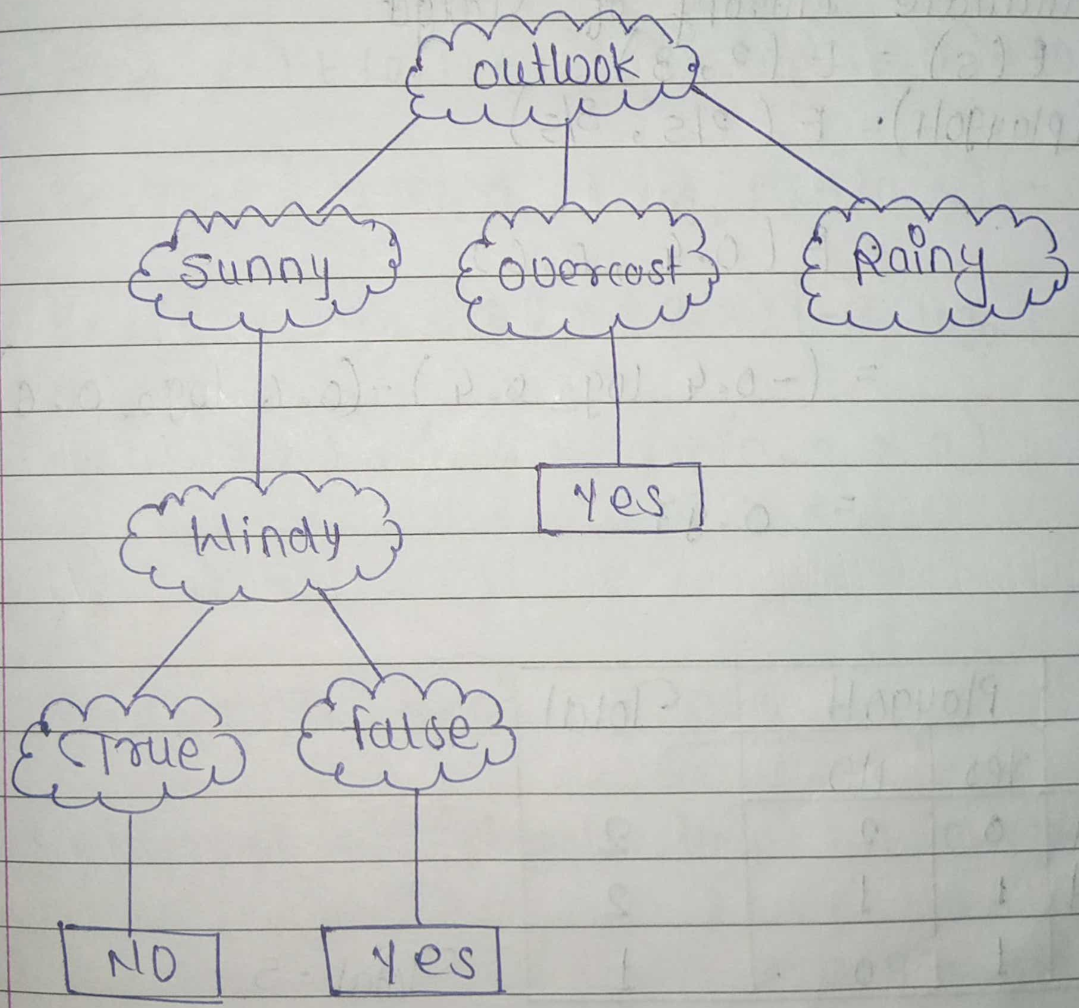
Page No.
Date

infogain (Playgolf, windy) = E.(Playgolf) − E (Playgolf , windy)

= 0.97 − 0

infogain (Playgolf, windy) = 0.97

Step 9 :
The attribute windy have largest Hence, tree structure is as below.



Page No.
Date

Outlook	Temp	Humidity	Windy	Playgolf
Rainy	Hot	High	false	No
Rainy	Hot	High	True	No
Rainy	mild	High	false	No
Rainy	cool	Normal	false	yes
Rainy	mild	Normal	True	yes

Step 10 : selected attribute is temprature.
Calculate entropy of Target

$E(s) = E(2, 3)$

$E(\text{playgolf}) = E(2/5, 3/5)$

$= E(0.4, 0.6)$

$= (-0.4 \log_2 0.4) - (0.6 \log_2 0.6)$

$= 0.97$

Temp.	Playgolf Yes	Playgolf No	Total
Hot	0	2	2
mild	1	1	2
cool	1	0	1

Total = 5

Page No.			
Date			

~~outlook~~ ~~temp~~ ~~humidity~~

$$E(T, x) = \sum_{c \in X} P(c)\, E(c)$$

$$E(\text{playgolf}, \text{temp}) = P(\text{Hot})\, E(0, 2) + P(\text{mild})\, E(1, 1) + P(\text{cool})\, E(1, 0)$$

$$= P(2/5)\, E(0/2, 2/2) + P(2/5)\, E(1/2, 1/2) + P(1/5)\, E(1/1, 0/1)$$

$$= (0.4)\, E(0, 1) + (0.4)\, E(0.5, 0.5) + (0.2)\, E(1, 0)$$

$$= (0.4)(-0 \log_2 0 - 1 \log_2 1) + (0.4)(-0.5 \log_2 0.5 - 0.5 \log_2 0.5) + (0.2)(-1 \log_2 1 - 0 \log_2 0)$$

$$= (0.4 \times 0) + (0.4 \times 1) + (0.2 \times 0)$$

$$= 0.4$$

$$E(\text{playgolf}, \text{temp}) = \underline{\underline{0.4}}$$

$$\text{Information gain}(\text{Playgolf}, \text{temp}) = E(\text{Playgolf}) - E(\text{playgolf}, \text{temp})$$

$$= 0.97 - 0.4$$

$$\text{Information gain}(\text{playgolf}, \text{temp}) = \underline{\underline{0.57}}$$

Page No.	
Date	

Step 11 : selected attribute is humidity.

		Playgolf		Total
		Yes	NO	
Humidity	High	3	0	3
	Normal	2	0	2

Total = 5

$$E(\text{playgolf}, \text{Humidity}) = \sum_{c \in X} p(c)\, E(c)$$

$= p(\text{High})\, E(0,3) + p(\text{Normal})\, E(2,0)$

$= p(3/5)\, E(0/3, 3/3) + p(2/5)\, E(2/2, 0/2)$

$= (0.6)\, E(0,1) + (0.4)\, E(1,0)$

$= (0.6)(-0 \log_2 0 - 1 \log_2 1) + (0.4)(-1 \log_2 1 - 0 \log_2 0)$

$= 0.6 \times 0 + 0.4 \times 0$

$= 0$

$E(\text{playgolf}, \text{Humidity}) = \underline{0}$

Information gain (playgolf, Humidity) = E (playgolf) − E(playgolf, Humidity)

$= 0.97 - 0$

Information gain (playgolf, Humidity) = $\underline{0.97}$

Step 12 : Selected attribute is Windy.

		Playgolf		Total
		Yes	NO	
Windy	True	1	1	2
	false	1	2	3

Total = 5.

$E(\text{Playgolf}, \text{windy}) = \sum_{c \in X} p(c) E(c)$.

$= p(\text{True}) E(1, 1) + p(\text{false}) E(1, 2)$

$= p(2/5) E(1/2, 1/2) + p(3/5) E(1/3, 2/3)$.

$= (0.4) E(0.5, 0.5) + (0.6) E(0.33, 0.67)$

$= 0.4 \times 1 + 0.6 (0.33, 0.67)$.

$= 0.4 + 0.6 (-0.33 \log_2 0.3 - 0.67 \log_2 0.67)$

$= 0.4 + 0.6 (0.91)$

$= 0.95$

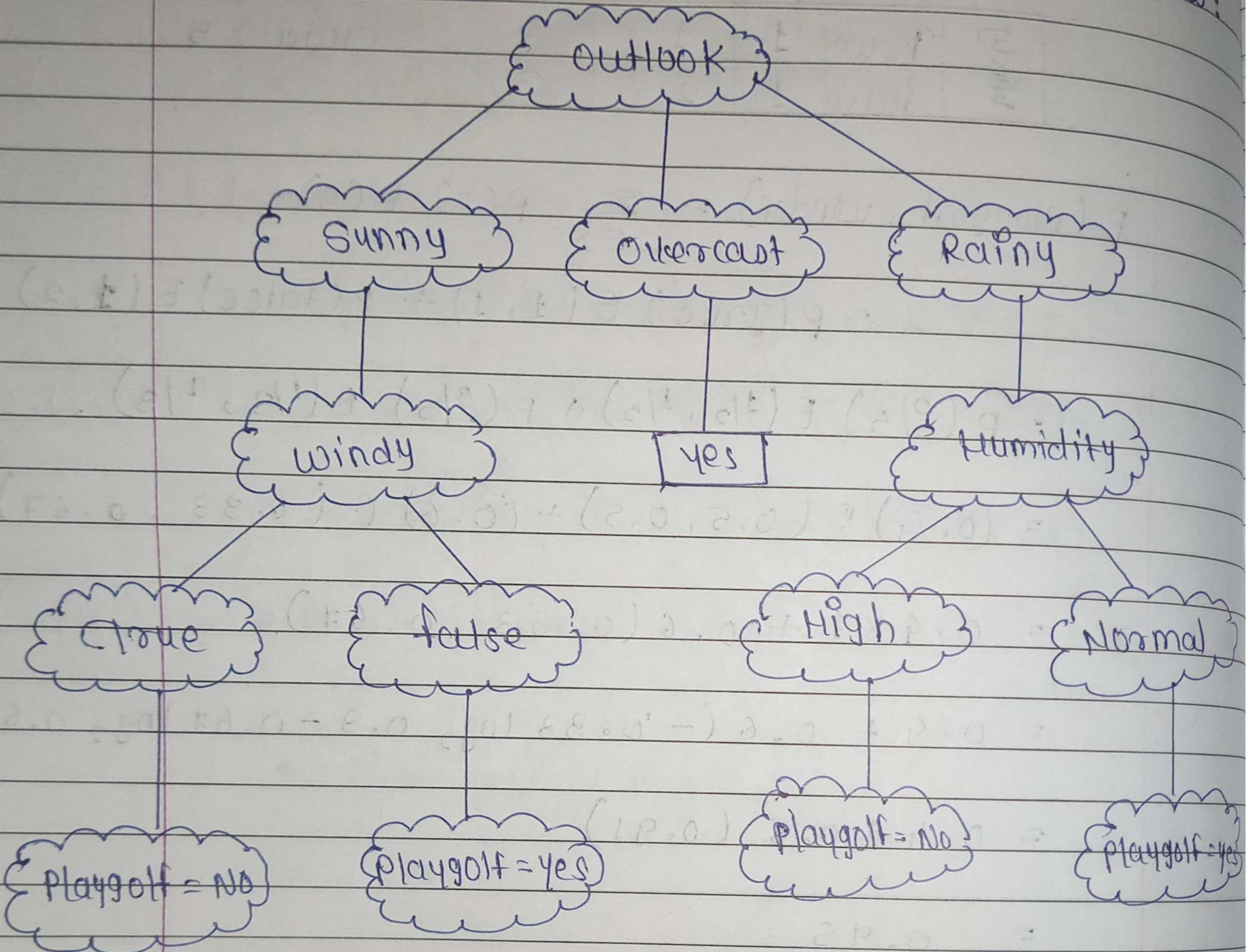
$E(\text{Playgolf}, \text{windy}) = \underline{0.95}$

Information gain (Playgolf, windy) = E(Playgolf) − E(Playgolf, windy)

$= 0.97 - 0.95$

Information gain (Playgolf, windy) = $\underline{0.02}$

The attribute combination (playgolf, Humidity) is having highest information gain.
Hence, Resulting tree structure is As below:



Page No.______

Date___/___/___

Qu.

Rec	Age	Income	Student	Credit Rating	Buys computer
R1	<=30	High	No	fair	NO
R2	<=30	High	No	Excellent	No
R3	31-40	High	No	fair	yes
R4	>=40	medium	No	fair	yes
R5	>=40	low	yes	fair	yes
R6	>=40	low	yes	Excellent	No
R7	31-40	low	yes	Excellent	yes
R8	<=30	medium	No	fair	No
R9	<=30	low	yes	fair	yes
R10	>=30	medium	yes	fair	yes
R11	<=30	medium	yes	Excellent	yes
R12	31-40	medium	No	Excellent	yes
R13	31-40	High	yes	fair	yes
R14	>=40	medium	NO	Excellent	NO

Step 1 :

for buys computer = yes

p(buy computer) = (9/14) = 0.64

for buys computer = NO

p(buy computer) = (5/14) = 0.36

Step 2 :

$$E(S) = \sum_{i=1} - p_i \log_2 p_i$$

$$E(\text{buycomputer}) = p(\text{yes}) \log_2 p(\text{yes}) - p(\text{NO}) \log_2 p(\text{NO})$$

$= p(9/14) \log_2 p(9/14) - p(5/14) \log_2 p(5/14)$

$= -0.41 - 0.53$

$= 0.94$

E(buys computer) = $\underline{0.94}$

Step 3 : Selected attribute is Age.

		buys computer		Total
		Yes	NO	
Age	<=30	2	3	5
	31-40	5	0	5
	>=40	2	2	4

Total = 14.

$E(\text{buys computer, Age}) = \sum_{c \in X} p(c)\, E(c)$

$= p(<=30)\, E(2/5, 3/5) + p(31\text{-}40)\, E(5/5, 0/5) + p(>=40)\, E(2/4, 2/4)$

$= p(5/14)\, E(2/5, 3/5) + p(5/14)\, E(5/5, 0/5) + p(4/14)\, E(2/4, 2/4).$

$= (0.36)\, E(0.4, 0.6) + (0.36)\, E(1, 0) + p(0.29)\, E(0.5, 0.5)$

Page No.
Date / /

$= (0.36)(-0.4 \log_2 0.4 - 0.6 \log_2 0.6) +$

$(0.36)(-1 \log_2 1 - 0 \log_2 0) +$

$(0.29)(-0.5 \log_2 0.5 - 0.5 \log_2 0.5).$

$= (0.36 \times 0.97) + (0.36 \times 0) + (0.29 \times 1).$

$= 0.35 + 0 + 0.29$

$= 0.64$

E(buys computer, Age) = 0.64

Information gain (buys computer, Age) = E(buys computer) - E(buys computer ; Age)

$= 0.94 - 0.64$

$= 0.30$

Information gain (buys computer, Age) = 0.30

Step 4 : Selected Attribute is Income.

		buys computer		Total
		Yes	NO	
Income	Low	3	1	4
	medium	4	2	6
	High	2	2	4

Total = 14

Shrikrupa

Page No.
Date / /

$E(\text{buys computer, Income}) = \sum_{c \in X} -p_i \log_2 p_i$

$= p(low)\, E(3, 1) + p(medium)\, E(4, 2) + p(High)\, E(2, 2)$

$= p(4/14)\, E(3/4, 1/4) + p(6/14)\, E(4/6, 2/6) + p(4/14)\, E(2/4, 2/4)$

$= (4/14)\, E(0.75, 0.25) + (6/14)\, E(0.67, 0.33) + (4/14)\, E(0.5, 0.5)$

$= (0.29)(-0.75 \log_2 0.75 - 0.25 \log_2 0.25) +$

$(0.43)(-0.67 \log_2 0.67 - 0.33 \log_2 0.33) +$

$(0.29)(-0.5 \log_2 0.5 - 0.5 \log_2 0.5)$

$= (0.29 \times 0.81) + (0.43 \times 0.92) + (0.29 \times 1)$

$= 0.23 + 0.\overset{40}{\cancel{39}} + 0.29$

$= 0.92$

$E(\text{buys computer, Income}) = \underline{\underline{0.92}}$

Information gain (buys computer, Income) = E (buys com) - E (buys computer, Income)

$= 0.94 - 0.92$

Information gain (buys com, Income) = $\underline{0.02}$

Shrikrupa

E - entropy
P - Probability

Date / /

Step 5 : Selected Attribute is student

		buys computer		Total
		Yes	No	
Student	Yes	6	1	7
	No	3	4	7

Total = 14

$E(\text{buys computer, student}) = \sum_{c \in X} - p_i \log_2 p_i$

$= P(\text{yes}) E(6, 1) + P(\text{NO}) E(3, 4)$

$= P(7/14) E(6/7, 1/7) + P(7/14) E(3/7, 4/7)$

$= P(7/14) E(0.86, 0.14) + P(7/14) E(0.43, 0.57)$

$= (7/14)(-0.86 \log_2 0.86 - 0.14 \log_2 0.14) +$
$(7/14)(-0.43 \log_2 0.43 - 0.57 \log_2 0.57)$

$= 0.5(0.58) + 0.5(0.98)$

$= 0.29 + 0.49$

$= 0.78$

$E(\text{buys computer, student}) = \underline{0.78}$

Information gain (buys computer, student) =
E (buys computer) - E (buy computer, student)

$= 0.94 - 0.78$

$= 0.16$

Information gain (buys computer, Student) = <u>0.16</u>

Step 6 : selected an attribute credit rating

		buys computer		Total
		Yes	No	
credit rating	Fair	6	2	8
	Excellent	3	3	6

Total = 14

$$E(\text{buys computer, credit rating}) = \sum_{c \in X} -p_i \log_2 p_i$$

$$= p(\text{fair})\, E(6, 2) + p(\text{Excellent})\, E(3, 3)$$

$$= p(8/14)\, E(6/8, 2/8) + p(6/14)\, E(3/6, 3/6)$$

$$= p(0.57)\, E(0.75, 0.25) + p(0.43)\, E(0.5, 0.5)$$

$$= (0.57)(-0.75 \log_2 0.75 - 0.25 \log_2 0.25) + (0.43)(-0.5 \log_2 0.5 - 0.5 \log_2 0.5)$$

$$= 0.57(0.81) + 0.43(1)$$

$$= 0.46 + 0.43$$

= 0.89

E(buys computer, credit rating) = 0.89

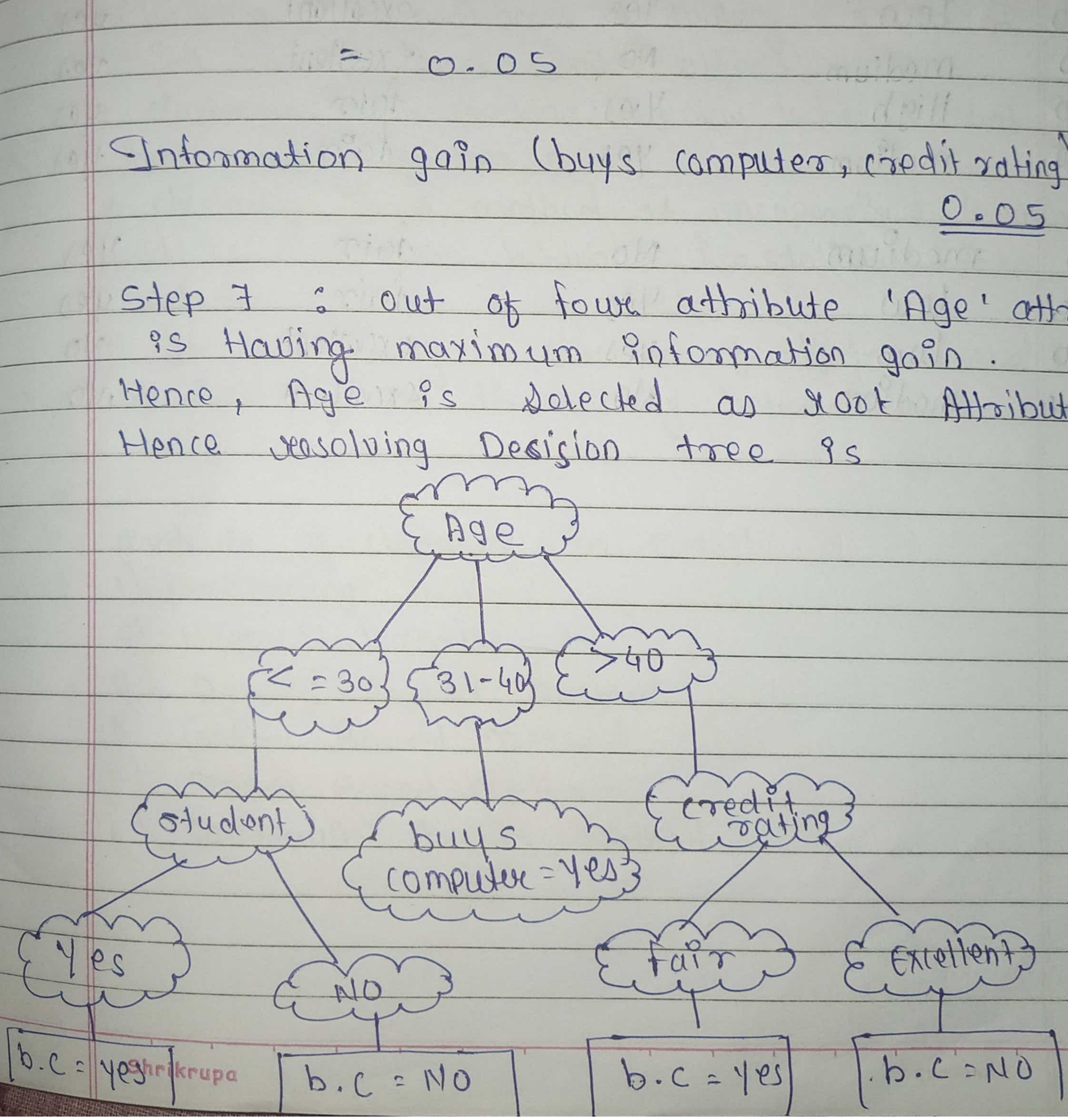
Information gain (buys computer, credit rating) = E(buys computer) − E(buys co., credit)

= 0.94 − 0.89

= 0.05

Information gain (buys computer, credit rating) = 0.05

Step 7 : out of four attribute 'Age' attribute is Having maximum information gain. Hence, Age is selected as root Attribute. Hence resolving Decision tree is



Age	Income	student	credit rating	buys computer
<=30	High	NO	fair	NO
<=30	High	NO	excellent	NO
<=30	~~High~~ medium	No	~~excellent~~ fair	NO
<=30	~~medium~~ low	Yes	fair	Yes
<=30	~~to~~ medium	yes	excellent	Yes
31-40	High	NO	fair	Yes
31-40	low	Yes	excellent	yes
31-40	medium	NO	excellent	yes
31-40	High	Yes	fair	yes
31-40	medium	yes	fair	yes.
>=40	medium	NO	fair	Yes
>=40	low	Yes	fair	yes
>=40	low	yes	excellent	NO.
>=40	medium	NO	excellent	NO.