Name: M. ASWIN REDDY

Date: 23/2/24

Roll No: 211121

Data Warehouse & Data Mining

Assignment - 1

Implementation of Decision Tree Classifier

Data Set:

Age	Income	Student	Credit-Rating	Buys Compute
≤30	high	no	fair	no
≤30	high	no	excellent	no
31-40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes.
>40	low	yes	excellent	no
31-40	low	yes	excellent	yes
≤30	medium	no	fair	no
≤30 ~~31-40~~	low ~~medium~~	yes	fair	yes
>40. ~~31-41~~	medium	yes	fair	yes
≤30	medium	yes	excellent	yes
31-40	medium.	no	excellent	yes
31-40	high	yes.	fair	yes.
>40	medium	no	excellent	no

First we need to calculate Information gain for every attribute other than target attribute which is Buys_Computer

From the attribute which has highest ~~for~~ information gain we can partition the dataset from that attribute and continue the Process.

Partition is stopped when the following conditions met:

i. All samples belong to the same class

ii. No samples left

iii. No Remaining Attributes for further Partitioning

~~Info~~ Entropy of Target $= -\sum^{m} P_i \log_2(P_i)$

Class P: buys_computer = "yes" = 9

Class N: buys_computer = "no" = 5

$E(\text{Buys computer}) = -\frac{9}{14}\log_2(9/14) - \frac{5}{14}\log_2(5/14)$

$E(\text{Target}) = 0.940$

$E(\text{Buys\_computer}, \text{Age}) = P(\leq 30) \cdot E(\text{target}, \leq 30) +$

$P(31\text{-}40) \cdot E(\text{target}, 31\text{-}40) +$

$P(>40) \cdot E(\text{target}, >40)$

		Buys_Computer yes	No
Age	≤30	2	3
	31-40	4	0
	>40	3	2

$$= P(\leq 30)\cdot E(2,3) + P(31\text{-}40)\cdot E(4,0) + P(>40)\cdot E(3,2)$$

$$= \frac{5}{14}\left[-\frac{2}{5}\cdot\log_2(2/5) - 3/5\log_2(3/5)\right] + \frac{4}{14}\left[-\frac{4}{4}\log(4/4) - \frac{0}{4}\log(0/4)\right]$$

$$+ \frac{5}{14}\left[-3/5\log_2(3/5) - 2/5\log(2/5)\right]$$

$$= \frac{5}{14}(0.970) + \frac{4}{14}\cdot 0 + \frac{5}{14}(0.970)$$

$E(\text{Buys comp, age})$

$$= 0.6935$$

$$\text{Information Gain}_{\text{Age}}(\text{Target}) = E(\text{Target}) - E(\text{Target, age})$$

$$= 0.940 - 0.6935$$

$$\text{Gain}_{\text{Age}}(\text{Target}) = 0.246$$

$$E(\text{Target, Income}) = P(\text{high})\cdot E[\text{Target, high}) + P(\text{medium})\cdot E(\text{Target, medium}) + P(\text{Low})\cdot E(\text{Target, Low})$$

		~~Bu~~ Buys_computer	
		yes	No
Income	High	2	2
	medium	4	2
	Low	3	1

$$= 4/14\cdot E(2,2) + \cancel{D}\,\frac{6}{14}E(4,2) + 4/14\,E(3,1)$$

$$= 4/14\left[-2/4\log_2(2/4) - 2/4\log_2(2/4)\right] +$$

$$6/14\left[-4/6\log_2(4/6) - 4/6\log_2(4/6)\right] +$$

$$4/14\left[-3/4\log_2(3/4) - 1/4\log_2(1/4)\right]$$

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

$E(\text{target, Income}) = 0.910$

$$\underset{\text{Income}}{\text{Gain}}(\text{Target}) = E(\text{Target}) - E(\text{Target, Income})$$

$$= 0.940 - 0.910$$

$$\underset{\text{income}}{\text{Gain}}(\text{Target}) = 0.03$$

$$E(\text{target, Student}) = P(\text{yes}) \cdot E(\text{target, yes}) + P(\text{No}) \cdot E(\text{target, no})$$

		Buys_computer	
		yes	No
Student	yes	06	01
	no	03	04

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

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

$$= \frac{7}{14} \cdot (0.591) + \frac{7}{14}(0.985)$$

$$E(\text{target, Student}) = 0.788$$

$$\underset{\text{Student}}{\text{Gain}}(\text{Target}) = E(\text{Target}) - E(\text{Target, Student})$$

$$= 0.940 - 0.788$$

$$= 0.152$$

$E(\text{target}, \text{Credit\_Rating}) = P(\text{fair}) \cdot E(\text{target}, \text{fair}) +$

$P(\text{excellent}) \cdot E(\text{target}, \text{Excellent})$

		Buys_Computer	
		yes	No
Credit_Rating	fair	3	3
	excellent	6	2

$= \frac{6}{14} \cdot E(3,3) + 8/14\ E(6,2)$

$= 6/14\,(1) + 8/14\,(0.811)$

~~Gain~~ $E(\text{target}, \text{credit}) = 0.892$

$\text{Gain}_{\text{Credit}}(\text{Target}) = E(\text{Target}) - E(\text{target}, \text{Credit})$

$= 0.940 - 0.892$

$= 0.048$

~~Eor~~

By Comparing Information Gain from 4 attributes

Age	Income	Student	Credit_Rating.
0.246	0.03	0.152	0.048

Age has highest Information Gain.

Age

≤30

Income	Student	Credit-rating	Class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes.

31-40

Income	Student	CreditRating	Class
high	no	fair	yes
Low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes.

>40

Income	Student	Credit-Rating	Class.
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

→ In the samples of 31-40, all belong to same class of the target so, 31-40 group buys computer.

Now consider ≤30 group/node.

Entropy of ≤30

$$, E(\text{target}_{,\le 30}) = E(2,3) \qquad \begin{array}{l}2\text{-yes's}\\3\text{- No's}\end{array}$$

$$= -\tfrac{2}{5}\log_2(\tfrac{2}{5}) - \tfrac{3}{5}\log_2(\tfrac{3}{5})$$

$$E(\text{target}_{\le 30}) = 0.970$$

$$E(\underbrace{\text{target}, \le 30}_{y}, \text{income}) = P(\text{high})E(y, \text{high}) + P(\text{medium})\,E(y, \text{med}) + P(\text{Low})E(y, \text{Low})$$

	yes	no
high	0	2
medium	1	1
Low	1	0

$$= \tfrac{2}{5}E(0,2) + \tfrac{2}{5}E(1,1) + \tfrac{1}{5}E(1,0)$$

$$= \tfrac{2}{5}(0) + \tfrac{2}{5} + \tfrac{1}{5}(0)$$

$$= 0.4$$

$$\text{Gain}\underset{\text{income}}{(\text{target}, \le 30)} = E(\text{target}, \le 30) - E(\text{target}, \le 30, \text{income})$$

$$= 0.970 - 0.4$$

$$= 0.570$$

$$E(\underbrace{\text{target}, \le 30}_{y}, \text{Student}) = P(\text{no}).E(y, \text{no}) + P(\text{yes}).E(y, \text{no})$$

$$= \tfrac{3}{5}E(0,3) + \tfrac{2}{5}E(2,0)$$

$$= \tfrac{3}{5}(0) + \tfrac{2}{5}(0)$$

$$= 0$$

$$\underset{\text{Student}}{\text{Gain}}(\text{target}, \le 30) = 0.970 - 0$$

$$= 0.970$$

$$E(\underbrace{\text{target}, \le 30}_{y}, \text{Credit\_Rating}) = P(\text{fair}).E(y, \text{fair}) + P(\text{excellent}).E(y, \text{excellent})$$

$$= \tfrac{3}{5} E(1,2) + \tfrac{2}{5} E(1,1)$$

$$= 0.6591$$

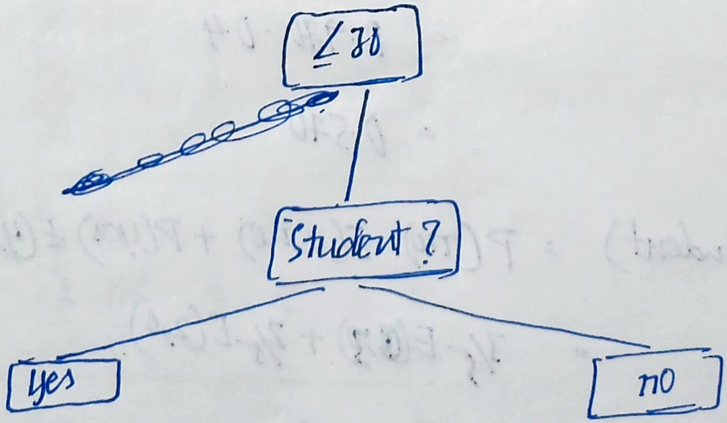
$$\underset{\text{Credit\_Rating}}{\text{Gain}}(\text{target}, \le 30) = 0.970 - 0.6591$$

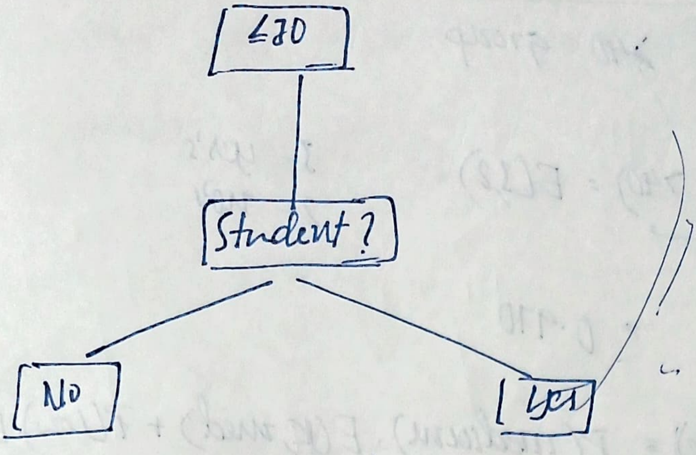
$$= 0.310$$

By comparing Information Gain from 3 Attributes in ≤30 group

Income	Student	Credit_Rating
0.570	~~0.6591~~ 0.970	0.310

Student has highest Information Gain.

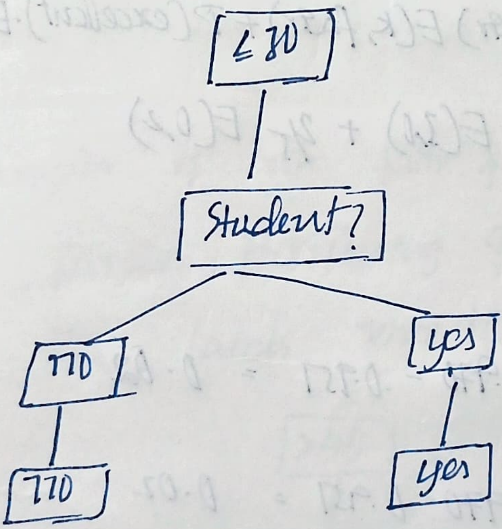




Income	Credit	Class
high	fair	no
high	excellent	no
~~high~~ med	fair	no

Income	Credit	class
Low	fair	yes
med.	excellent	yes
		~~yes~~

As samples of both yes & no have Pure class ~~classes~~ ~~#~~ ~~their~~ ~~own~~, further Partioning is not needed.



Now Consider >40 group

$E(\underbrace{\text{target}, >40}_{k}) = E(3,2)$ 3- yes's
2- no's

$= 0.970$

$E(k, \text{income}) = P(\text{medium}) . E(K, \text{med}) + P(\text{Low}). E(K, \text{Low})$

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

$= \cancel{0.950}\ 0.951$

$E(K, \text{Student}) = P(\text{yes}) \cdot E(K, \text{yes}) + P(\text{no}) \cdot E(K, \text{no})$

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

$= 0.951$

$E(K, \text{Credit\_Rating}) = P(\text{fair})\, E(K, \text{fair}) + P(\text{excellent}) \cdot E(K, \text{excellent})$

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

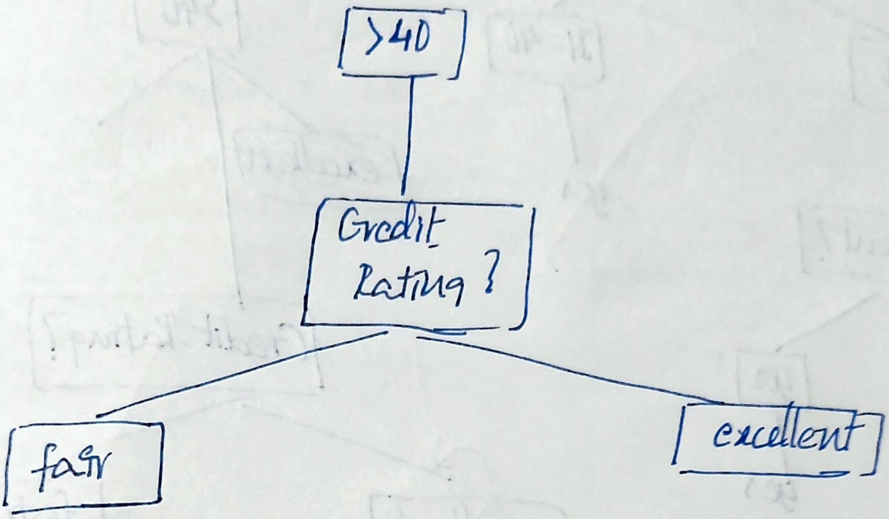
$= 0$

$\text{Gain}_{\text{income}}(\text{target} >40) = 0.970 - .0.951 = 0.02$

$\text{Gain}_{\text{Student}}(\text{target}, >40) = 0.970 - 0.951 = 0.02$

$\text{Gain}_{\text{credit}}(\text{target}, >40) = 0.970 - 0 = 0.970.$

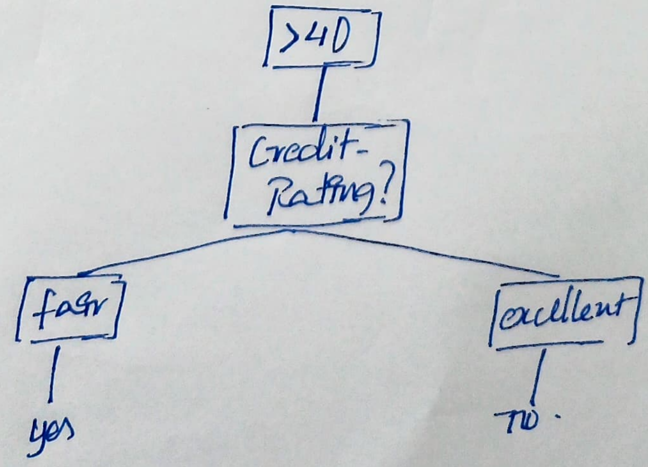
Credit_Rating has highest information gain in >40 group.

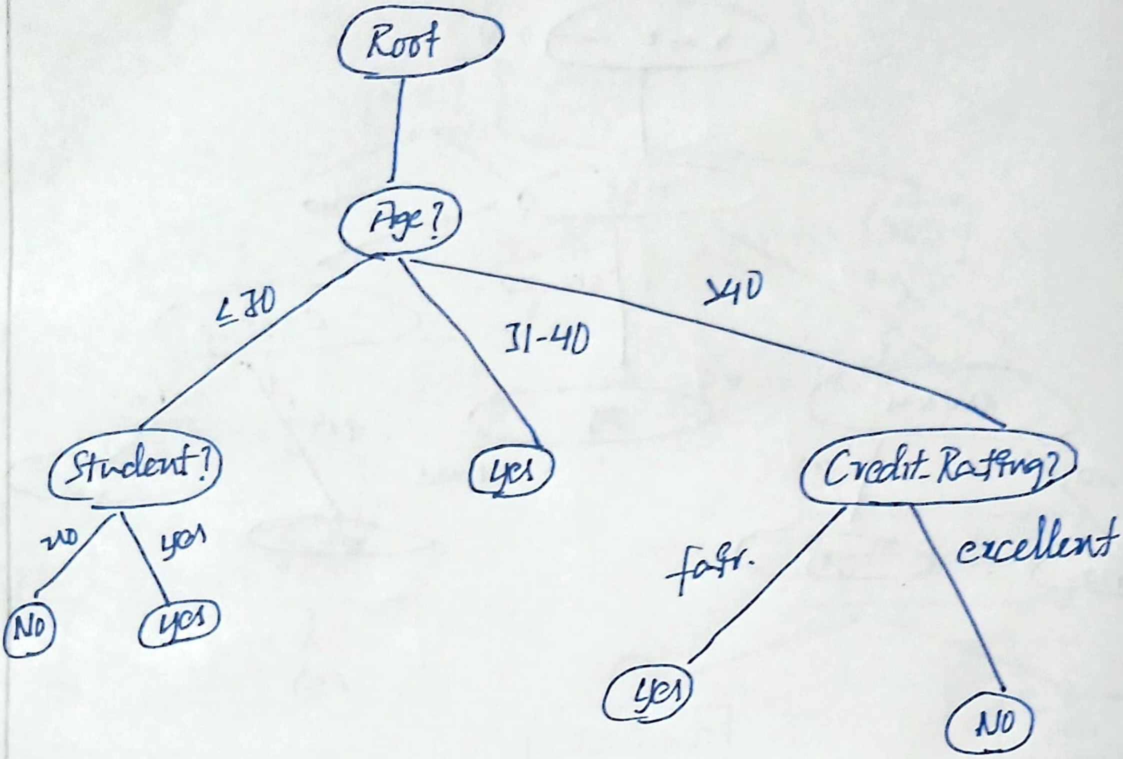


Income	Student	Class
med	no	yes
low	yes	yes
med	yes	yes

Income	Student	Class
low	yes	no
med	no	no.

As samples of both fair & excellent have Pure Classes, further Partitioning is not needed and assign the labels respectively.




Root
Age?
≤30
31-40
>40
Student?
yes
Credit Rating?
no
yes
fair.
excellent
NO
yes
yes
NO