Name: Siddharth Betala
Roll: BE19B032
MA5710 MMI Assignment-3

Date
Page

(1) Training Data:

Index	Age	Salary	Class
1	30	65	G
2	23	15	B
3	40	75	G
4	55	40	B
5	55	100	G
6	45	60	G

This Column has been added by me for use ahead.

Usually we decide the root node condition & attribute based on Gini & the impurity value.

(i) Age & Salary as Attributes
Since 'age' & 'salary' are both numeric attributes, we need to presort the data.

Presorting based on age:

Age	Index
23	2
30	1
40	3
45	6
55	5
55	4

Presorting based on salary:

Salary	Index
15	2
40	4
60	6
65	1
75	3
100	5

for a potential nodes,

$$\text{Gini Impurity } (G_I) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2)$$

$$\text{Gini}(D_K) = 1 - \sum_{j=1}^{N} p_j^2$$

$K = 1, 2$

N = No. of Classes

p_j = Frequency / Probability of a class.

(A) Age as Attribute

a) Age < 23:

Age < 23

True | False

True: G	True: B	False: G	False: B
0	1	4	1

$|D_1| = 1$ $\quad$ $|D_2| = 2$.

$$\text{Gini}(D_1) = 1 - \left(\frac{0}{0+1}\right)^2 - \left(\frac{1}{0+1}\right)^2$$

$$= 1 - 1$$

$$= 0$$

$$\text{Gini}(D_2) = 1 - \left(\frac{4}{4+1}\right)^2 - \left(\frac{1}{4+1}\right)^2$$

$$= 1 - \frac{16}{25} - \frac{1}{25}$$

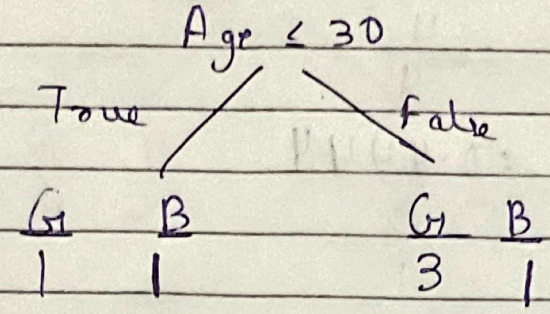
$$= \frac{8}{25}$$

$$\text{Gini}(\text{Age} < 23) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2)$$

$$= \frac{1}{6}(0) + \frac{5}{6}\left(\frac{8}{25}\right)$$

$$= \frac{4}{15} = 0.2667$$

(b) Age ≤ 30:



Gini(D_1) $= 1 - \left(\frac{1}{1+1}\right)^2 - \left(\frac{1}{1+1}\right)^2$

$= 0.5$

Gini(D_2) $= 1 - \left(\frac{3}{1+3}\right)^2 - \left(\frac{1}{1+3}\right)^2$

$= 1 - \frac{9}{16} - \frac{1}{16}$

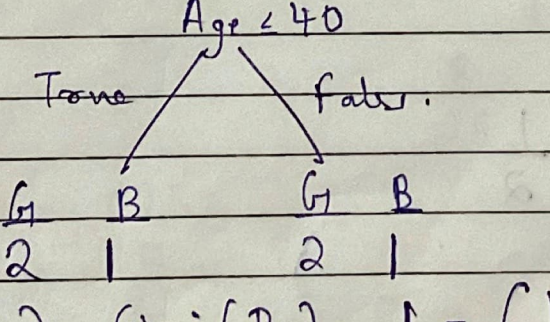
$= \frac{6}{16} = \frac{3}{8}$

Gini(Age ≤ 30) $= \frac{2}{6}(0.5) + \frac{4}{6} \times \frac{3}{8}$

$= \frac{1}{6} + \frac{1}{4}$

$= \frac{5}{12} = 0.4167$

(c) Age ≤ 40:



Gini(D_1) = Gini(D_2) $= 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2$

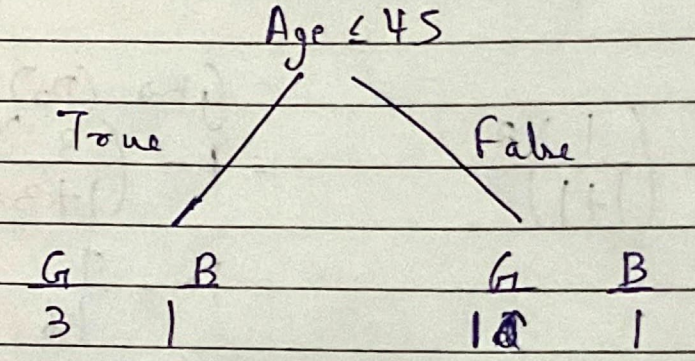
$= \frac{4}{9}$

$$Gini(Age < 40) = \frac{3}{6} \times \frac{4}{9} + \frac{3}{6} \times \frac{4}{9}$$

$$= \frac{4}{9}$$

$$= 0.4444$$

(d) Age < 45:



~~Gini (Age < 45) =~~

Gini (D_1)

$$= 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$$

$$= \frac{3}{8}$$

Gini (D_2)

$$= 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2$$

$$= \frac{4}{9}$$

Gini (D_2)

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

$$= 0.5$$

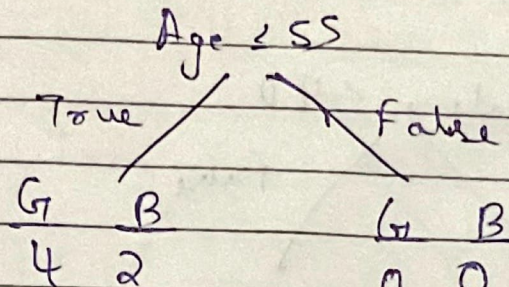
Gini (Age < 45)

$$= \frac{4}{6} \times \frac{3}{8} + \frac{2}{6} \times \frac{1}{2}$$

$$= \frac{1}{4} + \frac{1}{6}$$

$$= \frac{5}{12} = 0.4167$$

(c) Age < 55:



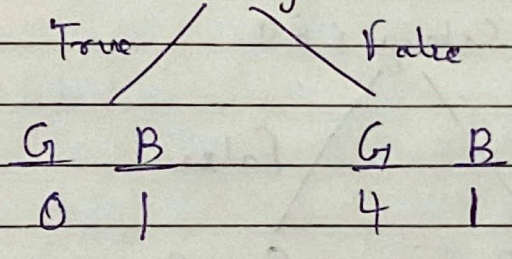
$$\text{Gini}(\text{Age} < 55) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2$$

$$= \frac{4}{9}$$

$$= 0.4444$$

(B) Salary as Attribute

(a) Salary < 15:



$$\text{Gini}(D_1) = 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2$$

$$= 0$$

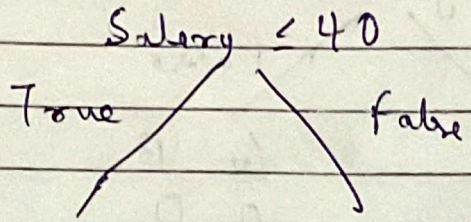
$$\text{Gini}(D_2) = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2$$

$$= \frac{8}{25}$$

$$\text{Gini}(\text{Salary} < 15) = \frac{1}{6} \times (0) + \frac{5}{6} \times \frac{8}{25}$$

$$= \frac{4}{15} = 0.26667$$

(b) Salary ≤ 40:



G	B	G	B
0	2	4	0

$Gini(D_1) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2$

$= 0$

$Gini(D_2) = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2$

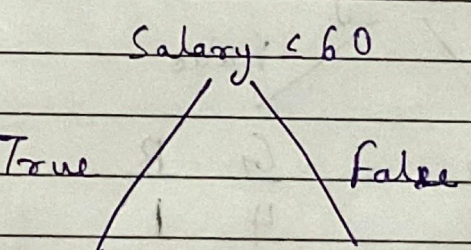
$= 1 - 1$

$= 0$

$Gini(Salary \leq 40) = \frac{2}{6} \times 0 + \frac{4}{6} \times 0$

$= 0.$

(c) Salary ≤ 60:



G	B	G	B
1	2	3	0

$Gini(D_1)$

$= 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2$

$= 4/9$

$Gini(D_2)$

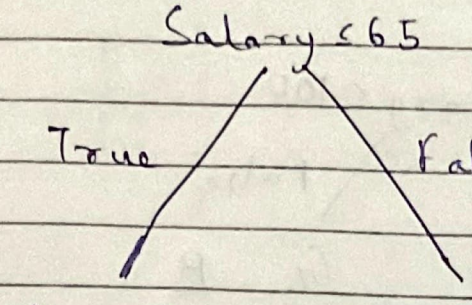
$= 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2$

$= 0$

$Gini(Salary \leq 60) = \frac{3}{6} \times \frac{4}{9} + \frac{3}{6} \times 0$

$= \frac{2}{9} = 0.2222$

(d) Salary < 65 :



G	B
2	2

G	B
2	0

Gini (D_1)

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

$= 0.5$

Gini (D_2)

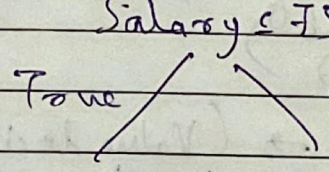
$= 1 - \left(\frac{2}{2}\right)^2$

$= 0$

Gini (Salary < 65) $= \frac{4}{6} \times \frac{1}{2} + \frac{2}{6} \times 0$

$= \frac{1}{3} = 0.3333$

(e) Salary < 75 :



G	B
3	2

G	B
1	0

Gini (D_1)

$= 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{3}{5}\right)^2$

$= \frac{12}{25}$

Gini (D_2)

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

$= 0$

Gini (Salary < 75) $= \frac{5}{6} \times \frac{12}{25} + \frac{1}{6} \times 0 = 0.4$

(f) Salary < 100 :

Salary < 100

True / False

G	B	G	B
4	2	0	0

$$Gini(D_1) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2$$

$$= 4/9$$

$$Gini(Salary < 100) = 4/9 = 0.4444$$

Best Split is given by the smallest Gini Impurity value.

Smallest Gini Impurity is for the Split: Salary < 40.

The same split can be given by the strict cond^n: $Salary < \left(\frac{40+60}{2}\right)$

$\Rightarrow$ Salary < 50 → (Value decided based on the pre-sorting order)

Gini(Salary < 50) = 0

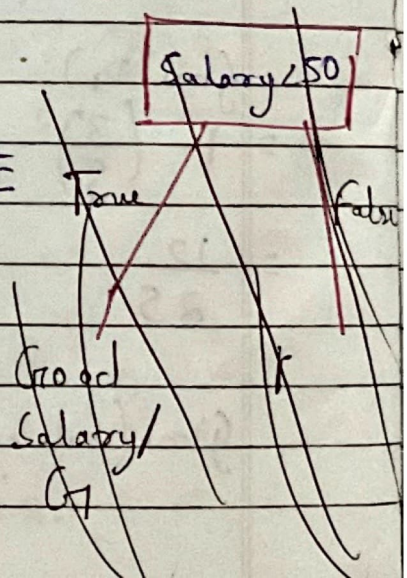
Salary < 50

True / False

G	B	G	B
0	2	4	0

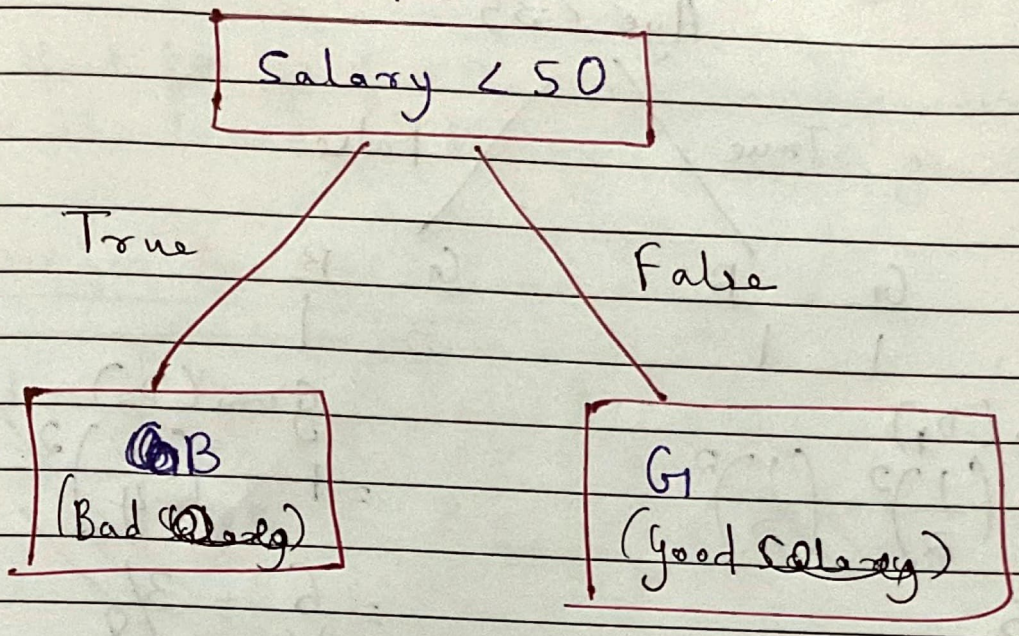
(We don't need to classify further as these nodes are pure. Hence, these are leaf nodes)

≡

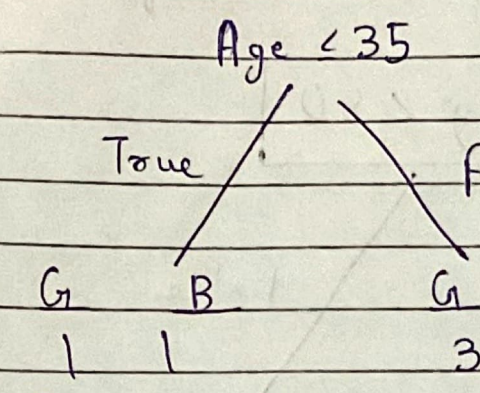


Date
Page

Based on the best split the classification tree is given as:



(ii) Root Partition Keeping (Age < 35)



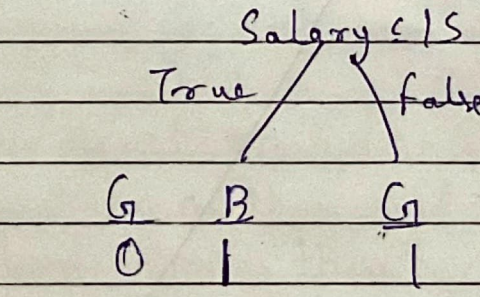
Gini(D_1)
$= 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2$
$= 0.5$

Gini(D_2)
$= 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$
$= \frac{6}{16} = \frac{3}{8}$
$= 0.375$

We see that the nodes D_1 & D_2 are both impure. We can try classifying the tree further by calculating the gini impurity value using Salary as an attribute. Note that we'll proceed only if the subsequent Gini values are smaller than Gini(D_1) & Gini(D_2) i.e., the any further classification should lead to purer classes than the current ones.

Classes for Age < 35 : True are:
(Pre-Sorted based on Salary)

Index	Age	Salary	Class
2	23	15	B
1	30	65	G



Gini(D_1) = $1 - (1)^2 = 0$
Gini(D_2) = $1 - (1)^2 = 0$

Gini (Salary < 15) = $0 \times \frac{1}{2} + 0 \times \frac{1}{2} = 0$.

∴ Best Split for the node corresponding to (Age < 35 : True) is Salary < 15 with gini impurity of 0.

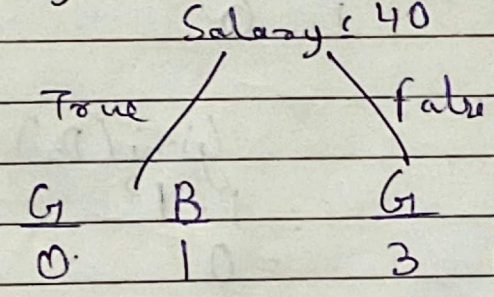
Classes for Age < 35 : False are

Index	Age	Salary	Class
	40	75	
	45	60	
	55		
	55		

(Pre-sorted based on salary)

Index	Age	Salary	Class
4	55	40	B
6	45	60	G
3	40	75	G
5	55	100	G

(a) Salary < 40 :

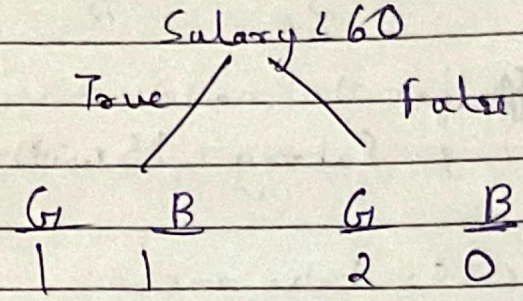


Gini (D_1) $= 1 - 1^2 = 0$

Gini (D_2) $= 1 - \left(\frac{3}{3}\right)^2 = 1 - 1 = 0$.

Gini (Salary < 40) $= \frac{1}{4} \times 0 + \frac{3}{4} \times 0 = 0$.

(b) Salary < 60 :

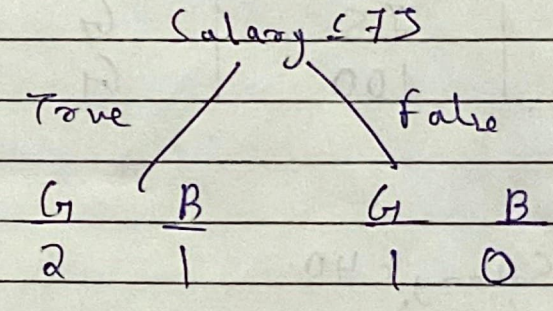


Gini (D_1)
$= 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2$
$= 0.5$

Gini (D_2)
$= 1 - \left(\frac{2}{2}\right)^2$
$= 0$

Gini (Salary < 60)
$= \frac{2}{4} \times 0.5 + \frac{2}{4} \times 0$
$= 0.25$

(c) Salary < 75 :

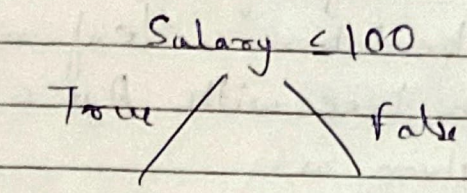


Gini (D_1)
$= 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2$
$= 4/9$

Gini (D_2)
$= 1 - 1^2$
$= 0$

Gini (Salary < 75)
$= \frac{3}{4} \times 4/9 = 1/3$
$= 0.3333$

(d) Salary < 100



True		False	
G	B	G	B
3	1	0	0

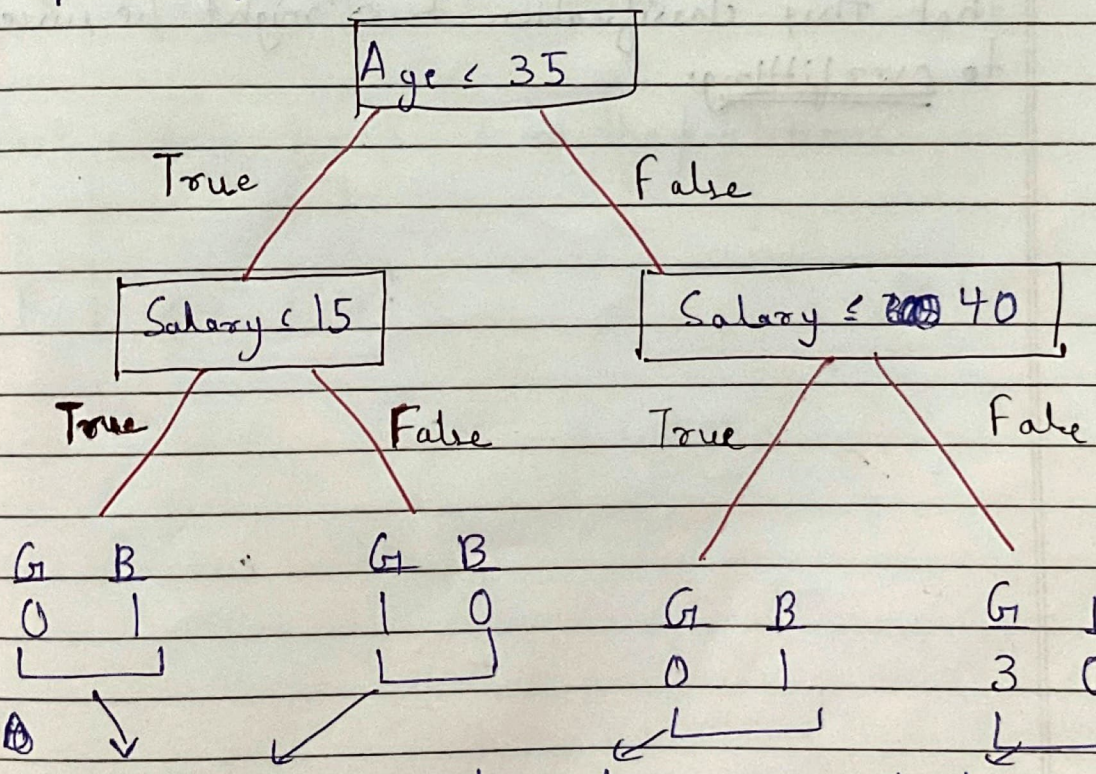
$$\text{Gini}(P_1) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$$

$$= \frac{3}{8}$$

$$\text{Gini}(\text{Salary} \leq 100) = \frac{4}{4} \times \frac{3}{8}$$

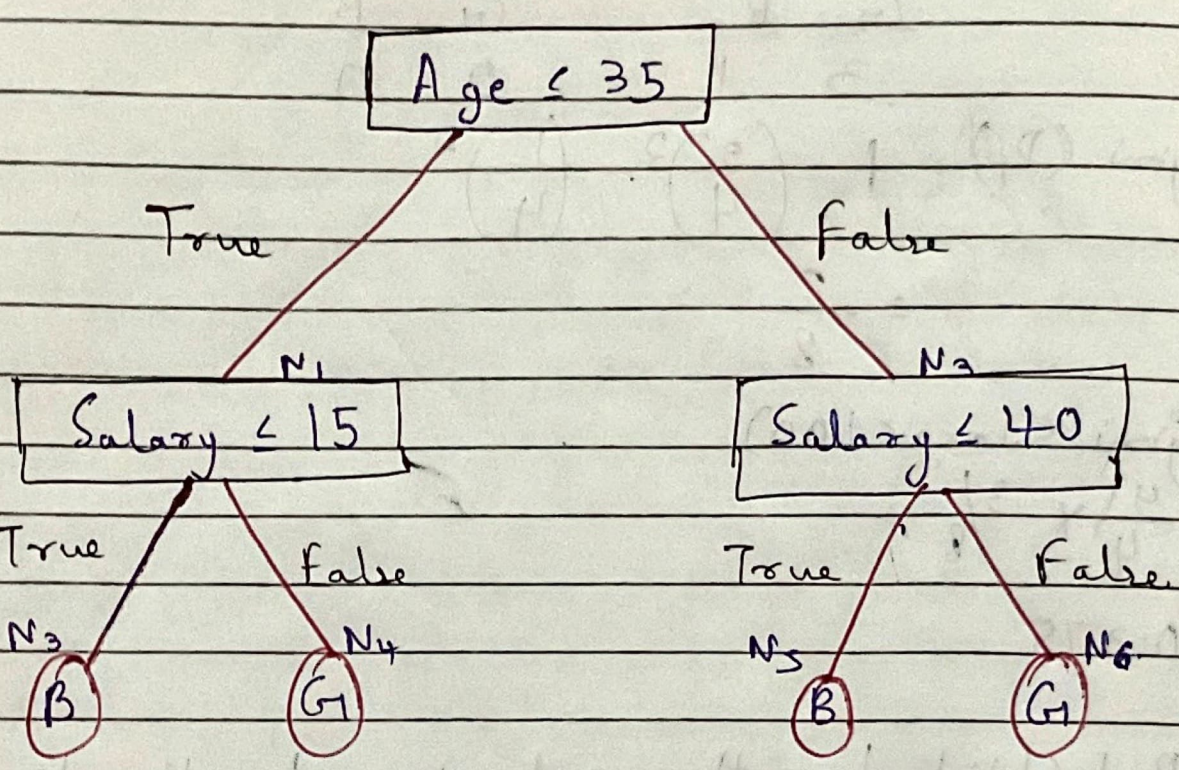
$$= 0.375$$

∴ Best Split for the node corresponding to (Age < 35 : False) is Salary ≤ 40 with Gini impurity of 0.



All these nodes are pure & give an exact classification

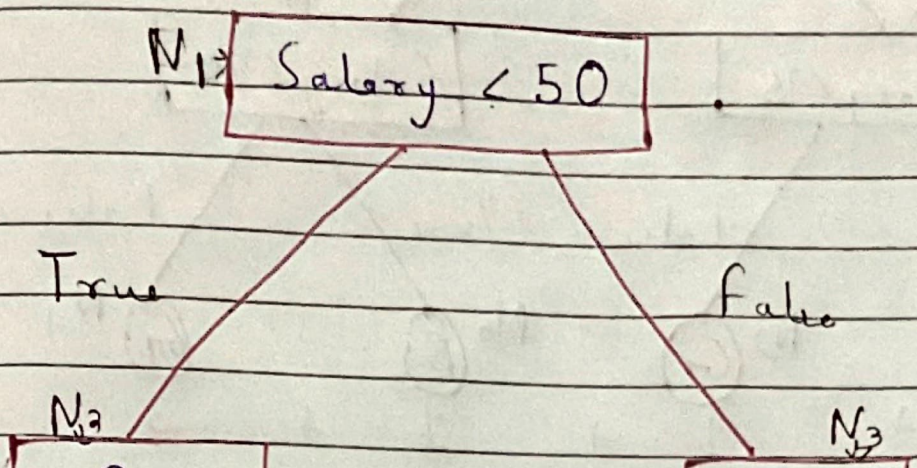
So there's no need for further classification. These 4 can be taken as leaf nodes. So the decision tree with Age ≤ 35 as root partition is given as:



Note: ~~The nodes~~ Only one training example, each make it to the nodes: N_3, N_4, N_5. So it is important to note that this classification tree might be susceptible to overfitting.

(2) Here, we have to compute the accuracy of the classifier from Q2 at every node using class histogram.

Ideal Classification tree (Refer (1) (i))



2 + 4 = 6

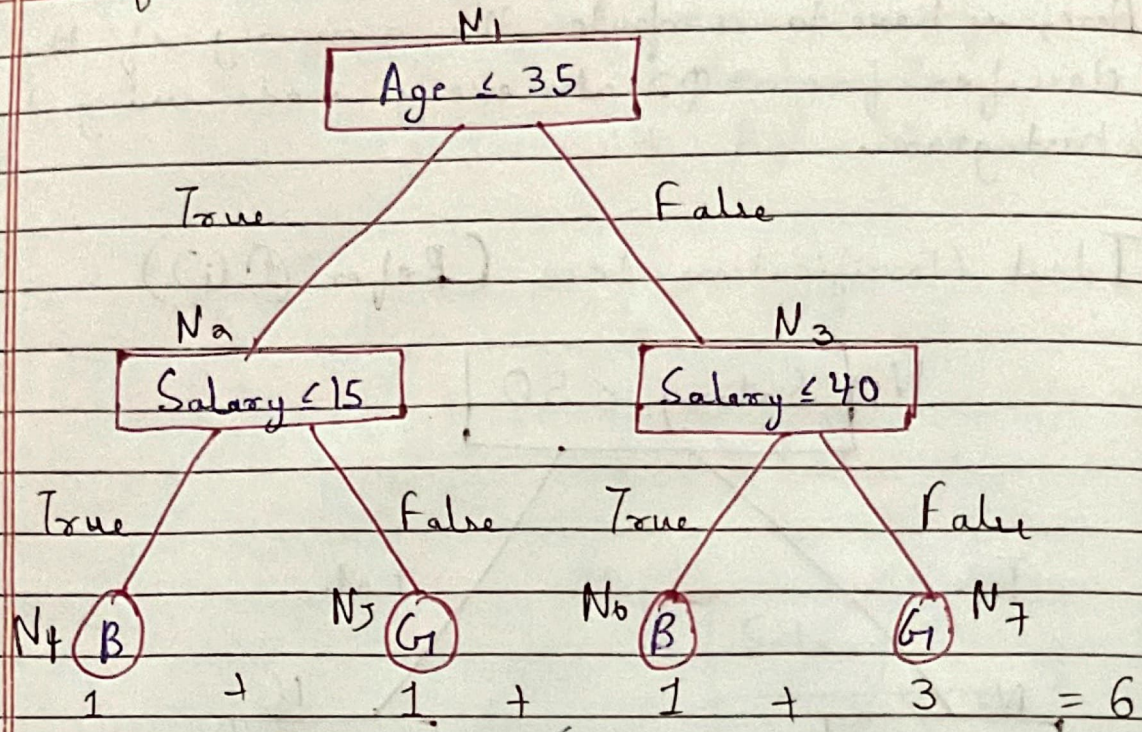
Intial Histogram for N_1:

	B	G
L	0	0
R	2	4

Post classification based on the partition:

	B	G
L	2	0
R	0	4

Classification Tree using root node Age < 35



$= 6$

Histogram post classification for N_1:

	B	G
(N_2) L	1	1
(N_3) R	1	3

Accuracy for left Branch / to get N_2 node

$= \frac{\text{No. of B's in } N_2}{\text{Expected No. of B's in } N_2} \times 100$

$= \frac{1}{2} \times 100$

$= 50.1.$

Accuracy for Right Branch / to get N_3 node

$= \frac{\text{No. of G's in } N_3}{\text{Expected no. of G's in } N_3} \times 100$

$= \frac{3}{4} \times 100$

$= 75.1.$

Accuracy of getting nodes N_4, N_5, N_6 & N_7 is 100% because these are leaf nodes, i.e., they offer a perfect split

When we the root partition as (Age < 35), we do not get the perfect split as given by the SLIQ model. Hence, we calculate the accuracy for both the left and right branches (how accurately does it identify the labels 'B' in the left branch & the label 'G' in the right branch)