Problem #1

CART

Part (a)

Occupation	Gender	Age	Salary
Service	Female	45	L3
Service	Male	25	L1
Service	Male	33	L2
Management	Male	25	L3
Management	Female	35	L4
Management	Male	26	L3
Management	Female	45	L4
Management	Female	40	L3
Sales	Male	30	L2
Sales	Female	50	L2
Staff	Female	25	L1
Staff	Male	25	L1

Target variable = Salary.

$1-\sum_{i=1}^{c-1} P(i)^2$

Impurity Before Splitting

	Salary
L1	2
L2	3
L3	4
L4	2
	11

$$P = 1 - \sum_{i=0}^{c-1} P_i(t)^2$$

$$P = 1 - \left(\frac{2}{11}\right)^2 - \left(\frac{3}{11}\right)^2 - \left(\frac{4}{11}\right)^2 - \left(\frac{2}{11}\right)^2 = 0.7272$$

Gini Index for Occupation

(A)

Occupation	Service, Management	Sales, Staff
L1	1	1
L2	1	2
L3	3	1
L4	2	0
	7	4

$$Gini(N1) = 1 - \left(\frac{1}{7}\right)^2 - \left(\frac{1}{7}\right)^2 - \left(\frac{3}{7}\right)^2 - \left(\frac{2}{7}\right)^2 = 0.6938$$

$$Gini(N2) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{2}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0.625$$

$$\text{Weighted Gini} = \sum_{i=1}^{k} \frac{n_i}{n} Gini(i)$$

$$M = \frac{7}{11}(0.6938) + \frac{4}{11}(0.625) = 0.6687$$

$$Gain = P - M = 0.7272 - 0.6687 = \boxed{0.0585}$$

(B)

	service, sales	Management, staff
L1	1	1
L2	2	1
L3	2	2
L4	0	2
	5	6

$$Gini(N1) = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{2}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.64$$

$$Gini(N2) = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{1}{6}\right)^2 - \left(\frac{2}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.7222$$

$$M = \frac{5}{11}(0.64) + \frac{6}{11}(0.722) = 0.6848$$

$$Gain = P - M = 0.7272 - 0.6848 = \boxed{0.0424}$$

(C)

	service, staff	Management, sales
L1	2	0
L2	2	1
L3	1	3
L4	0	2
	5	6

$$Gini(N1) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0.64$$

$$Gini(N2) = 1 - \left(\frac{0}{6}\right)^2 - \left(\frac{1}{6}\right)^2 - \left(\frac{3}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.6111$$

$$M = \frac{5}{11}(0.64) + \frac{6}{11}(0.611) = 0.6242$$

$$Gain = P - M = 0.7272 - 0.6242 = \boxed{0.103}$$

(D)

occupation	N1 Sales	N2 Service, staff Management
L1	0	2
L2	1	2
L3	1	3
L4	0	2
	2	9

$Gini(N_1) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5$

$Gini(N_2) = 1 - (\frac{2}{9})^2 - (\frac{2}{9})^2 - (\frac{3}{9})^2 - (\frac{2}{9})^2 = 0.7407$

$M = \frac{2}{11}(0.5) + \frac{9}{11}(0.7407)$

$= 0.6969$

$Gain = P - M = 0.7272 - 0.6969$

$= \boxed{0.0303}$

(E)

occupation	N1 Staff	N2 Service, Sales Management
L1	1	1
L2	1	2
L3	0	4
L4	0	2
	2	9

$Gini(N1) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5$

$Gini(N2) = 1 - (\frac{1}{9})^2 - (\frac{2}{9})^2 - (\frac{4}{9})^2 - (\frac{2}{9})^2$

$= 0.6913$

$M = \frac{2}{11}(0.5) + \frac{9}{11}(0.6913)$

$= 0.6565$

$Gain = P - M = 0.7272 - 0.6565$

$= \boxed{0.0707}$

(F)

occupation	N1 Service	N2 sales, staff Management
L1	1	1
L2	1	2
L3	1	3
L4	0	2
	3	8

$Gini(N1) = 1 - (\frac{1}{3})^2 - (\frac{1}{3})^2 - (\frac{1}{3})^2$

$= 0.666$

$Gini(N2) = 1 - (\frac{2}{8})^2 - (\frac{1}{8})^2 - (\frac{3}{8})^2 - (\frac{2}{8})^2$

$= 0.7187$

$M = \frac{3}{11}(0.666) + \frac{8}{11}(0.7187)$

$= 0.7043$

$Gain = P - M = 0.7272 - 0.7043$

$= \boxed{0.0229}$

(G)

occupation	N1 Management	N2 sales, staff service
L1	0	2
L2	0	3
L3	2	2
L4	2	0
	4	7

$Gini(N1) = 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5$

$Gini(N2) = 1 - (\frac{2}{7})^2 - (\frac{3}{7})^2 - (\frac{2}{7})^2$

$= 0.6530$

$M = \frac{4}{11}(0.5) + \frac{7}{11}(0.6530)$

$= 0.5973$

$Gain = P - M$

$= 0.7272 - 0.5973$

$= \boxed{0.1299}$

Gini Index For Gender

	N1 Male	N2 Female
L1	2	0
L2	2	1
L3	2	2
L4	0	2
	6	5

$Gini(N1) = 1-\left(\frac{2}{6}\right)^2-\left(\frac{2}{6}\right)^2-\left(\frac{2}{6}\right)^2$
$= 0.666$

$Gini(N2) = 1-\left(\frac{1}{5}\right)^2-\left(\frac{2}{5}\right)^2-\left(\frac{2}{5}\right)^2$
$= 0.64$

$M = \frac{6}{11}(0.666)+\frac{5}{11}(0.64)$
$= 0.6541$

$Gain = P-M = 0.7272-0.6541$
$= \boxed{0.0731}$

Gini Index for Age

Salary levels	L1	L1	L3	L3	L2	L2	L4	L3	L3	L4	L2
Age →	25	25	25	26	30	33	35	40	45	45	50
	24	25	25	25	28	31	34	37	42	45	47	50

	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
L1	0	2	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0
L2	0	3	0	3	0	3	0	3	0	3	1	2	2	1	2	1	2	1	2	1	2	1	3	0
L3	0	4	1	3	1	3	1	3	2	2	2	2	2	2	2	2	3	1	4	0	4	0	4	0
L4	0	2	0	2	0	2	0	2	0	2	0	2	0	2	1	1	1	1	2	0	2	0	2	0
Gini	0.7272		0.5983		0.5983		0.5983		0.5973		0.6541		0.6541		0.7077		0.7043		0.6545		0.6545		0.7272	

(i) For 24

	N1 <=	N2 >
L1	0	2
L2	0	3
L3	0	4
L4	0	2
	0	11

$Gini(N1) = 1-0-0-0-0 = 1$

$Gini(N2) = 1-\left(\frac{2}{11}\right)^2-\left(\frac{3}{11}\right)^2-\left(\frac{4}{11}\right)^2-\left(\frac{2}{11}\right)^2 = 0.7272$

$weighted\ M = \frac{0}{11}(1)+\frac{11}{11}(0.7272) = 0.7272$

(ii) For 25

	<=	>
L1	2	0
L2	0	3
L3	1	3
L4	0	2
	3	8

$Gini(N1) = 1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2 = 0.444$

$Gini(N2) = 1-\left(\frac{3}{8}\right)^2-\left(\frac{3}{8}\right)^2-\left(\frac{2}{8}\right)^2$
$= 0.6562$

$weighted = \frac{3}{11}(0.444)+\frac{8}{11}(0.6562)$
$\simeq 0.5983$

For 28

	<=	>
L1	2	0
L2	0	3
L3	2	2
L4	0	2
	4	7

$Gini(N1) = 1-(\frac{2}{4})^2-(\frac{2}{4})^2 = 0.5$

$Gini(N2) = 1-(\frac{3}{7})^2-(\frac{2}{7})^2-(\frac{2}{7})^2 = 0.6530$

$M = \frac{4}{11}(0.5)+\frac{7}{11}(0.6530)$

$= 0.5973$

(iv) For 31

	<=	>
L1	2	0
L2	1	2
L3	2	2
L4	0	2
	5	6

$Gini(N1) = 1-(\frac{2}{5})^2-(\frac{1}{5})^2-(\frac{2}{5})^2$

$= 0.64$

$Gini(N2) = 1-(\frac{2}{6})^2-(\frac{2}{6})^2-(\frac{2}{6})^2$

$= 0.666$

$M = \frac{5}{11}(0.64)+\frac{6}{11}(0.666)$

$= 0.6541$

(v) For 34

	<=	>
L1	2	0
L2	2	1
L3	2	2
L4	0	1
	6	5

$Gini(N1) = 1-(\frac{2}{6})^2-(\frac{2}{6})^2-(\frac{2}{6})^2 = 0.666$

$Gini(N2) = 1-(\frac{1}{5})^2-(\frac{2}{5})^2-(\frac{2}{5})^2 = 0.64$

$M = \frac{6}{11}(0.666)+\frac{5}{11}(0.64)$

$= 0.6541$

(vi) For 37

	<=	>
L1	2	0
L2	2	1
L3	2	2
L4	1	1
	7	4

$Gini(N1) = 1-(\frac{2}{7})^2-(\frac{2}{7})^2-(\frac{2}{7})^2-(\frac{1}{7})^2$

$= 0.7551$

$Gini(N2) = 1-(\frac{1}{4})^2-(\frac{2}{4})^2-(\frac{1}{4})^2$

$= 0.625$

$M = \frac{7}{11}(0.755)+\frac{4}{11}(0.625)$

$= 0.7077$

(vii) For 42

	<=	>
L1	2	0
L2	2	1
L3	3	1
L4	1	1
	8	3

$Gini(N1) = 1-(\frac{2}{8})^2-(\frac{2}{8})^2-(\frac{3}{8})^2-(\frac{1}{8})^2 = 0.7187$

$Gini(N2) = 1-(\frac{1}{3})^2-(\frac{1}{3})^2-(\frac{1}{3})^2 = 0.666$

$M = \frac{8}{11}(0.7187)+\frac{3}{11}(0.666)$

$= 0.7043$

(viii) For 45

	<=	>
L1	2	0
L2	2	1
L3	4	0
L4	2	0
	10	1

$Gini(N1) = 1-(\frac{2}{10})^2-(\frac{2}{10})^2-(\frac{4}{10})^2-(\frac{2}{10})^2$

$= 0.72$

$Gini(N2) = 1-(\frac{1}{1})^2 = 0$

$M = \frac{10}{11}(0.72)+\frac{1}{11}(0)$

$= 0.6545$

(ix) For 47

	<=	>
L1	2	0
L2	2	1
L3	4	0
L4	2	0
	10	1

$Gini(N1) = 1 - \left(\frac{2}{10}\right)^2 - \left(\frac{2}{10}\right)^2 - \left(\frac{4}{10}\right)^2 - \left(\frac{2}{10}\right)^2 = 0.72$

$Gini(N2) = 1 - \left(\frac{1}{1}\right)^2 = 0$

$M = \frac{10}{11}(0.72) + \frac{1}{11}(0)$

$= 0.6545$

(x) For 50

	<=	>
L1	2	0
L2	3	0
L3	4	0
L4	2	0
	11	0

$Gini(N1) = 1 - \left(\frac{2}{11}\right)^2 - \left(\frac{3}{11}\right)^2 - \left(\frac{4}{11}\right)^2 - \left(\frac{2}{11}\right)^2$

$= 0.7272$

$Gini(N2) = 1 - 0 = 1$

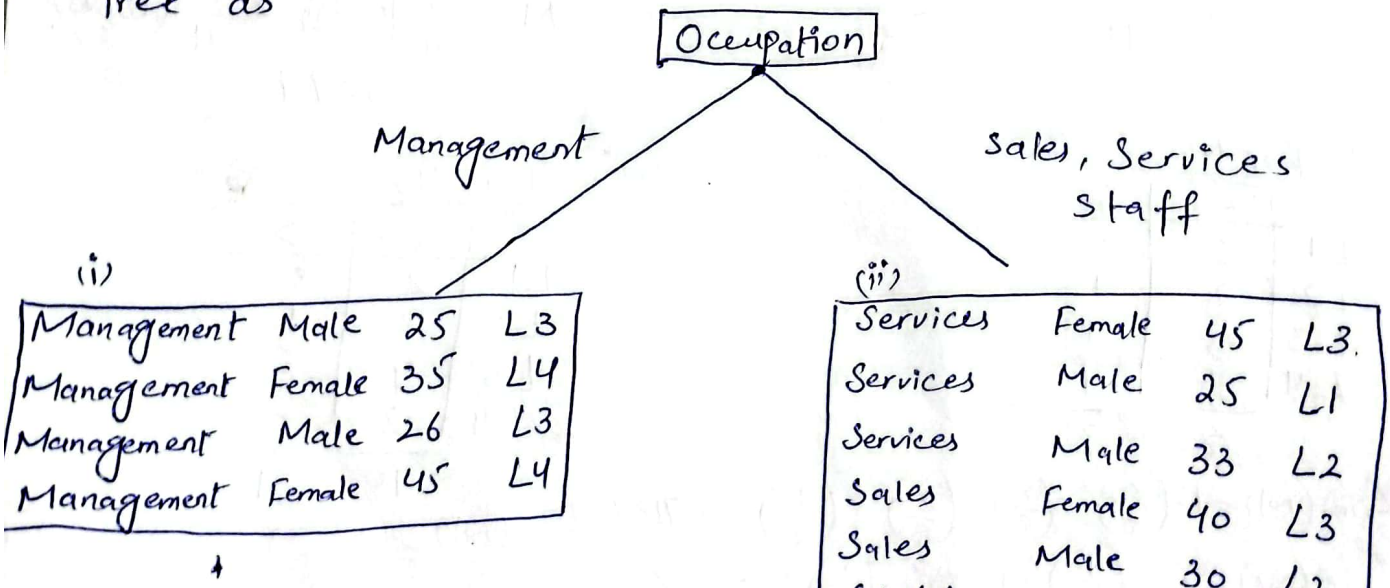
$M = \frac{11}{11}(0.7272) + \frac{0}{11}(1)$

$= 0.7272$

As Occupation with split (binary) of Management / sales, services, staff And Age with threshold of 28 has same Smallest gini index So we can pick anyone of these.

So, we selected Occupation as Root node.

Tree as



(i)

Management	Male	25	L3
Management	Female	35	L4
Management	Male	26	L3
Management	Female	45	L4

(ii)

Services	Female	45	L3
Services	Male	25	L1
Services	Male	33	L2
Sales	Female	40	L3
Sales	Male	30	L2
Staff	Female	50	L2
Staff	Male	25	L1

om Table (ii)

	Parent/salary
L1	2
L2	3
L3	2
L4	0
	7

$$Gini = 1 - \left(\frac{2}{7}\right)^2 - \left(\frac{3}{7}\right)^2 - \left(\frac{2}{7}\right)^2$$

$$P = 0.6530$$

Gini Index for Gender

	Female	Male
L1	0	2
L2	1	2
L3	2	0
L4	0	0
	3	4

$$Gini(N1) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2$$

$$= 1 - 0.111 - 0.444$$

$$= 0.444$$

$$Gini(N2) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2$$

$$= 1 - 0.25 - 0.25$$

$$= 0.5$$

weighted Gini

$$M = \frac{3}{7}(0.444) + \frac{4}{7}(0.5)$$

$$= 0.476$$

Gini Index for Age

L1	L1	L2	L2	L3	L3	L2
Age						
25	25	30	33	40	45	50

	22		25		27		31		36		42		47		50	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
1	0	2	2	0	2	0	2	0	2	0	2	0	2	0	2	0
2	0	3	0	3	0	3	1	2	2	1	2	1	2	1	3	0
3	0	2	0	2	0	2	0	2	0	2	1	1	2	0	2	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gini	0.6530		0.3428		0.3428		0.476		0.476		0.6		0.5708		0.6530	

For 22

$Gini(N1) = 1 - 0 - 0 - 0$

$= 1$

$Gini(N2) = 1 - (\frac{2}{7})^2 - (\frac{3}{7})^2 - (\frac{2}{7})^2$

$= 0.6530$

$M = \frac{0}{7}(1) + \frac{7}{7}(0.65)$

$= 0.6530$

For 25

$Gini(N1) = 1 - (\frac{2}{2})^2 = 0$

$Gini(N2) = 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48$

$M = \frac{2}{7}(0) + \frac{5}{7}(0.48)$

$= 0.3428$

For 27

$Gini(N1) = 1 - (\frac{2}{2})^2 = 0$

$Gini(N2) = 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48$

$M = 0.3428$

For 31

$Gini(N1) = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2$

$= 0.44$

$Gini(N2) = 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2$

$= 0.5$

$M = \frac{3}{7}(0.44) + \frac{4}{7}(0.5)$

$= 0.476$

For 36

$M = 0.476$

For 42

$Gini(N1) = 0.64$

$Gini(N2) = 0.5$

$M = 0.6$

For 47

$Gini(N1) = 0.66$

$Gini(N2) = 0$

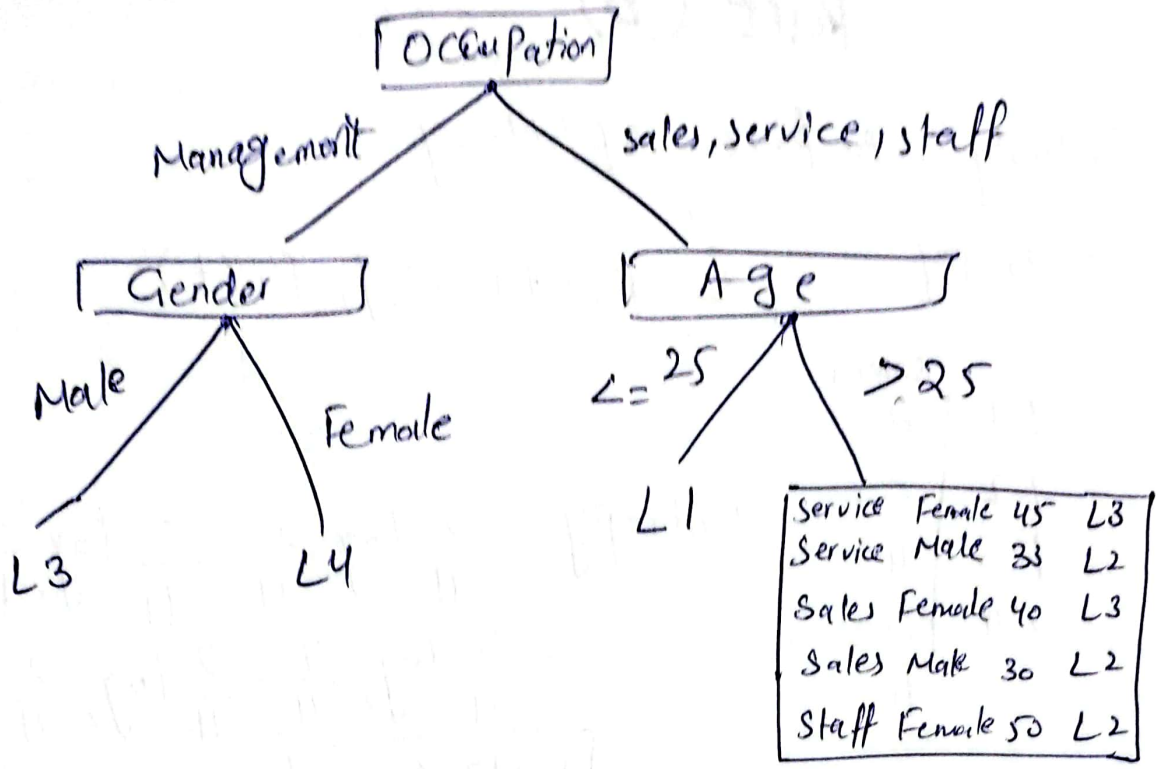
$M = 0.5708$

For 50

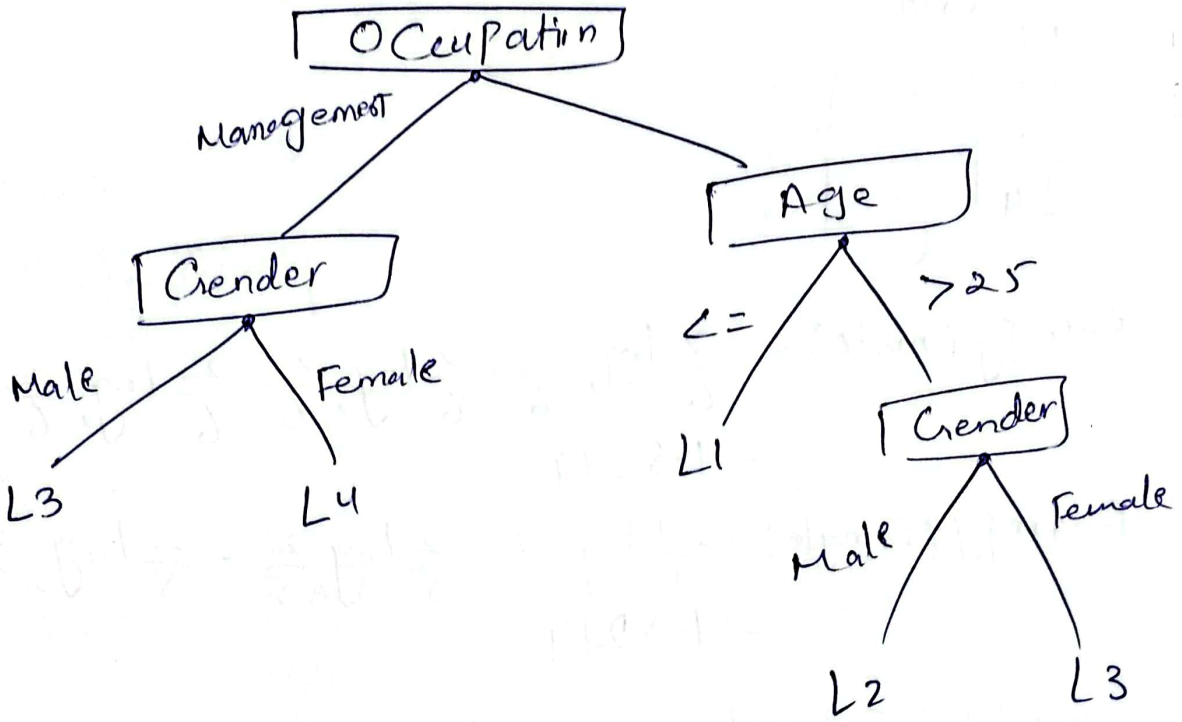
$Gini(N1) = 0.6530$

$Gini(N2) = \phi$

$M = 0.6530$



As occupation, Age already been covered so we can split on gender only - further.



One Record is misclassified with this tree.

Part (b) ID3

Entropy before splitting.

$$\text{Entropy} = -\sum_{i=1}^{c-1} p_i(t) \log_2 p_i(t).$$

	Salary
L1	2
L2	3
L3	4
L4	2
	11

$$\text{Entropy} = -\frac{2}{11}\log_2\frac{2}{11} - \frac{3}{11}\log_2\frac{3}{11} - \frac{4}{11}\log_2\frac{4}{11} - \frac{2}{11}\log_2\frac{2}{11}$$

$$= \boxed{1.9362}$$

Entropy for Gender

	Male	Female
L1	2	0
L2	2	1
L3	2	2
L4	0	2
	6	5

$$\text{Entropy (male)} = -\frac{2}{6}\log_2\frac{2}{6} - \frac{2}{6}\log_2\frac{2}{6} - \frac{2}{6}\log_2\frac{2}{6}$$

$$= 1.5849$$

$$\text{Entropy (Female)} = -\frac{1}{5}\log_2\frac{1}{5} - \frac{2}{5}\log_2\frac{2}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= 1.5219$$

Weighted Entropy

$$= \frac{6}{11}(1.5849) + \frac{5}{11}(1.5219)$$

$$= \boxed{1.5562}$$

28 Copies

Q#

Entropy for Age

age <=25 Total=3

Age >25 Total=8

	<=25	>25
L1	2	0
L2	0	3
L3	1	3
L4	0	2
	3	8

$$\text{Entropy}(N1) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.9182$$

$$\text{Entropy}(N2) = -\frac{3}{8}\log_2\frac{3}{8} - \frac{3}{8}\log_2\frac{3}{8} - \frac{2}{8}\log_2\frac{2}{8}$$

$$= 1.5612$$

$$\text{weighted Entropy} = \frac{3}{11}(0.9182) + \frac{8}{11}(1.5612)$$

$$= \boxed{1.3858}$$

Entropy for Occupation

(a)

	service, staff management	sales
L1	2	0
L2	2	1
L3	3	1
L4	2	0
	9	2

$$\text{Entropy}(N1) = -\frac{2}{9}\log_2\frac{2}{9} - \frac{2}{9}\log_2\frac{2}{9} - \frac{3}{9}\log_2\frac{3}{9} - \frac{2}{9}\log_2\frac{2}{9}$$

$$= 1.9749$$

$$\text{Entropy}(N2) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$\text{weighteed Entropy} = \frac{9}{11}(1.9749) + \frac{2}{11}(1)$$

$$= \boxed{1.7976}$$

(b)

	staff	sales, service Management
L1	1	1
L2	1	2
L3	0	4
L4	0	2
	2	9

$Entropy(N1) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$Entropy(N2) = -\frac{1}{9}\log_2\frac{1}{9} - \frac{2}{9}\log_2\frac{2}{9} - \frac{4}{9}\log_2\frac{4}{9} - \frac{2}{9}$
$= 1.8365$

weighted Entropy $= \frac{2}{11}(1) + \frac{9}{11}(1.8365) = \boxed{1.6844}$

(c)

	service	sales, staff Management
L1	1	1
L2	1	2
L3	1	3
L4	0	2
	3	8

$Entropy(N1) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3}$
$= 1.5849$

$Entropy(N2) = -\frac{1}{8}\log_2\frac{1}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{3}{8}\log_2\frac{3}{8} - \frac{2}{8}\log_2\frac{2}{8} = 1.9056$

weighted Entropy $= \frac{3}{11}(1.5849) + \frac{8}{11}(1.9056)$
$= \boxed{1.8181}$

(d)

	Management	sales, staff services
L1	0	2
L2	0	3
L3	2	2
L4	2	0
	4	7

$Entropy(N1) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$

$Entropy(N2) = -\frac{2}{7}\log_2\frac{2}{7} - \frac{3}{7}\log_2\frac{3}{7} - \frac{2}{7}\log_2\frac{2}{7}$
$= 1.5566$

weighted $= \frac{4}{11}(1) + \frac{7}{11}(1.5566) = \boxed{1.3542}$

(e)

	service, management	sales, staff
L1	1	1
L2	1	2
L3	3	1
L4	2	0
	7	4

$Entropy(N1) = 1.8423$

$Entropy(N2) = 1.5$

weighted $= \boxed{1.7178}$

(f)

	service, sales	staff, Management
L1	1	1
L2	2	1
L3	2	2
L4	0	2
	5	6

$Entropy(N1) = 1.5219$

$Entropy(N2) = 1.9182$

weighted $= \boxed{1.7380}$

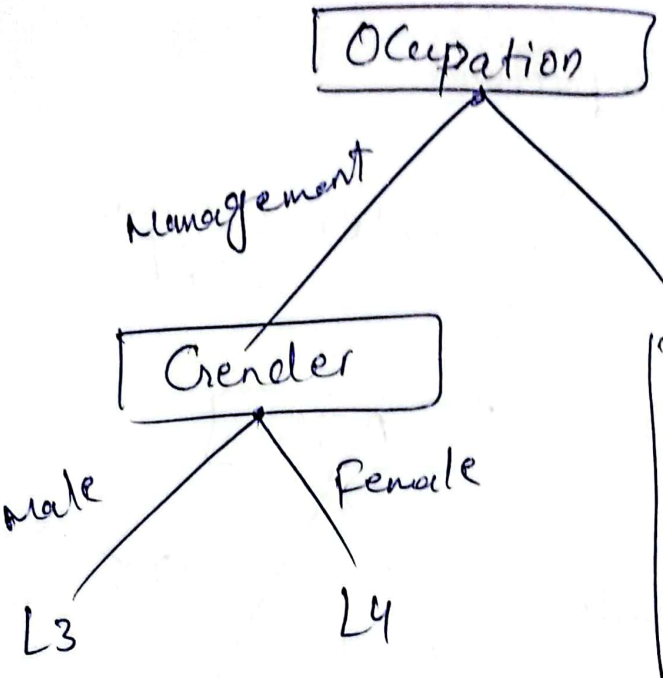
(g)

	staff, service	Management sales
L1	2	0
L2	2	1
L3	1	3
L4	0	2
	5	6

$Entropy(N1) = 1.5219$

$Entropy(N2) = 1.4591$

weighted $= \boxed{1.4876}$



Service	F	45	L3
Service	M	25	L1
Service	M	33	L2
sales	F	40	L3
sales	M	30	L2
staff	F	50	L2
staff	M	25	L1

Now for table.

Entropy for Gender

	Male	Female
L1	2	0
L2	2	1
L3	0	2
L4	0	0
	4	3

Entropy (male) = 1

Entropy (Female) = 0.918

weighted Entropy = $\frac{4}{7}(1) + \frac{3}{7}(0.918)$

$= 0.5714 + 0.393$

$= \boxed{0.9648}$

Entropy for Age

Age <=25

Age >25

Total = 2

Total = 5

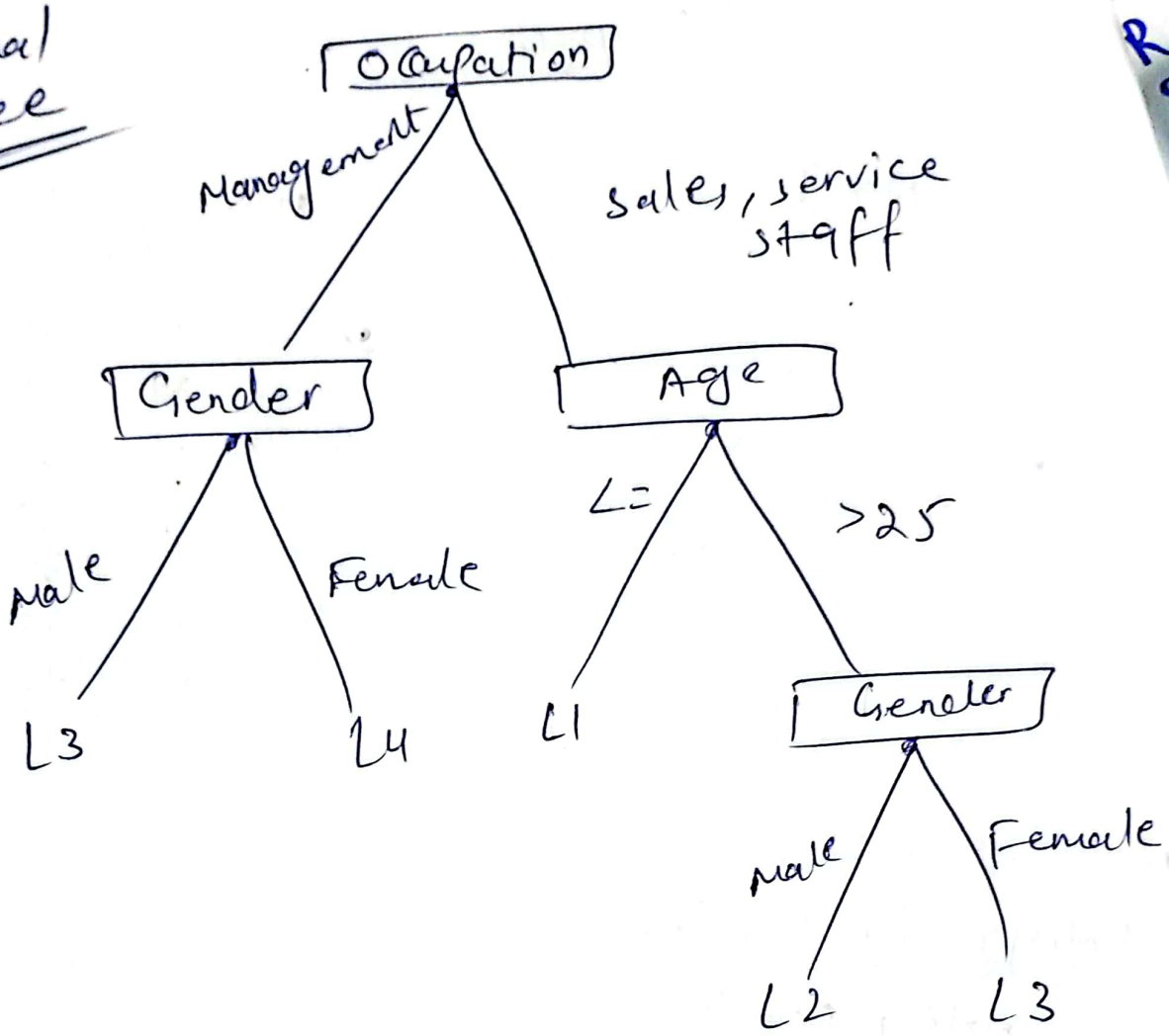
	<=25	>25
L1	1	0
L2	0	3
L3	0	2
L4	0	0
	1	5

Entropy (N1) = 0

Entropy (N2) = 0.9709

weighted = $\boxed{0.6935}$

Final Tree



One record is misclassified in this tree as well.

(c) Both trees are same.

(d) For Part(a) and (b)

① R1 (Occupation = Management) ∧ (Gender = Male) → L3

② R2 (Occupation = Management) ∧ (Gender = Female) → L4

③ R3 (Occupation = sales/service/staff) ∧ (Age <= 25) → L1

④ R4 (Occupation = sales/service/staff) ∧ (Gender = Male) ∧ (Age > 25) → L2

⑤ R5 (Occupation = sales/service/staff) ∧ (Age > 25) ∧ (Gender = Female) → L3.

Problem #2

Solution

(a) No, the rules are not mutually exclusive as more than one rules are covering its diff combinations.

(b) Yes, the rule set is exhaustive as it is covering the every combination of attributes.

(c) Rule (3,4) can be put at top to get maximum coverage by putting toughest requirement first as it involves many tests.

(d) No, we don't need any default Class as rule set is exhaustive.

oblem #3

Solution :-

(i) R1 ⟶ + (covers 12 positive and 4 negative examples)

R1 Rule's Accuracy $= \frac{12}{12+4} = \frac{12}{16} = 75\%$

R2 Rule's Accuracy $= \frac{45}{45+15} = \frac{45}{60} = 75\%$

R3 Rule's Accuracy $= \frac{115}{115+100} = \frac{115}{215} = 53\%$

R1 & R2 are the best rules according to Accuracy and R3 is the worst rule.

$P0 = 150$ $N0 = 320$

R1 Foil's Information gain.

$$= P1 * \left(\log_2 \frac{P1}{P1+N1} - \log_2 \frac{P0}{P0+N0}\right)$$

$P1 = 12$ $N1 = 4$

$$= 12 * \left(\log_2 \frac{12}{12+4} - \log_2 \frac{150}{150+320}\right)$$

$$= 14\%$$

R2 $P1 = 45$ $N1 = 15$

$$= 45 * \left(\log_2 \frac{45}{45+15} - \log_2 \frac{150}{150+320}\right)$$

$$= 55\%$$

R3 $P1 = 115$ $N1 = 100$

$$= 115 * \left(\log_2 \frac{115}{115+100} - \log_2 \frac{150}{150+320}\right) = 85\%$$

R1 is worst and R3 is best rule as per Foil info gain

Problem #4

Solution:- Nearest neighbor unweighted Majority voting

X1 = 4.0

X	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5	label
Y	-	-	+	+	+	-	-	+	-	-	
X1=4.0	3.5	1.0	0.5	0.6	0.9	1.2	1.3	1.5	3	5.5	
1			+								-
3			+	+	+						+
5		-	+	+	+	-					+
9	-	-	+	+	+	-	-	+	-		-

X2 = 7.5

X	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5	Label
Y	-	-	+	+	+	-	-	+	-	-	
X2=7.5	7	4.5	3	2.9	2.6	2.3	2.2	2	0.5	2	
1									-		-
3								+	-	-	-
5						-	-	+	-	-	-
9		-	+	+	+	-	-	+	-	-	-

stance Weighted Majority Voting

X1=4.0

Distance weight $= \frac{1}{d^2}$

X	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5	Label
y	−	−	+	+	+	−	−	+	−	−	
X1=4	3.5	1	0.5	0.6	0.9	1.2	1.3	1.5	3	5.5	
Distance weight	0.08	1	4	2.77	1.23	0.69	0.59	0.44	0.11	0.33	
1			+								+
3			+	+	+						+
5		−	+	+	+	−					+
9	−	−	+	+	+	−	−	+	−		+

For 5 $[(0.5)(4) + (0.6)(2.77) + (0.9)(1.23)$ OR $(1.2)(0.69) + (1)(1)]$

$[4.769$ OR $1.828]$ As $4.76 > 1.828$

$+ > -$

So "+" would be the label.

For 9 $[(0.5)(4) + (0.6)(2.77) + (0.9)(1.23) + (1.5)(0.44)$ OR

$[(3.5)(0.08) + (1)(1) + (1.2)(0.69) + (1.3)(0.59) + (3)(0.11)] = [5.429$ OR $3.205]$

As $[5.429 > 3.205]$ so "+" is label

X2=7.5

X	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5	Label
y	−	−	+	+	+	−	−	+	−	−	
X2=7.5	7	4.5	3	2.9	2.6	2.3	2.2	2	0.5	2	
Distance weight	0.020	0.049	0.11	0.118	0.14	0.18	0.20	0.25	4	0.25	
1									−		−
3								+	−	−	−
5						−	−	+	−	−	−
9		−	+	+	+	−	−	+	−	−	−

For 3

$$\left[(2)(0.25) \quad OR \quad (0.5)(1)+(2)(0.25)\right] = \left[\underset{(+)}{0.5} \quad OR \quad \underset{(-)}{2.5}\right]$$

As $2.5 > 0.5$ $(- > +)$ so "−" is the label.

For 5

$$\left[(2)(0.25) \quad OR \quad (0.5)(1)+(2)(0.25)+(2.2)(0.20)+(2.3)(0.18)\right]$$

$$\left[0.5 \quad OR \quad 3.354\right]$$ As 3.354 is > 0.5 so "−" is label.

For 9

$$\left[(2)(0.25)+(2.6)(0.14)+(2.9)(0.11)+(3)(0.11) \quad OR \quad (0.5)(1)+(2)(0.25)+(2.2)(0.20)+(2.3)(0.18)+(4.5)(0.04)\right]$$

$$= \left[\underset{(+)}{1.513} \quad OR \quad \underset{(-)}{3.534}\right]$$ As $3.534 > 1.513$ $(-) > (+)$ so "−" is the label.

Problem # 5

Solution:-

(a) R1 Accuracy $= \frac{12}{12+3} = \frac{12}{15} = 0.8 = 80\%$

R3 Accuracy $= \frac{8}{8+4} = \frac{8}{12} = 0.66 = 66\%$

R2 Accuracy $= \frac{7}{7+3} = \frac{7}{10} = 0.7 = 70\%$

R1 Best and R3 Worst.

R1 Accuracy $= \frac{12}{12+3} = \frac{12}{15} = 80\%$

R2 Accuracy $= \frac{7}{7+3} = \frac{7}{10} = 70\%$

R3 Accuracy $= \frac{6}{6+4} = \frac{6}{10} = 60\%$

R1 is best and R3 is Worst.

(c) R1 Accuracy $= \frac{12}{12+3} = \frac{12}{15} = 80\%$

R3 Accuracy $= \frac{6}{6+2} = \frac{6}{8} = 75\%$

R2 Accuracy $= \frac{7}{7+3} = \frac{7}{10} = 70\%$

R1 is best and R2 is Worst.

Problem #6

Solution:-

(i) For x-y dimensions.

(a) sorted for x

10	30
50	90
60	70
65	50
70	60
75	25
80	85
90	65
95	75

Median = 70

(b) sorted for y

10	30
65	50
60	70
50	90

Median = 60

75	25
90	65
95	75
80	85

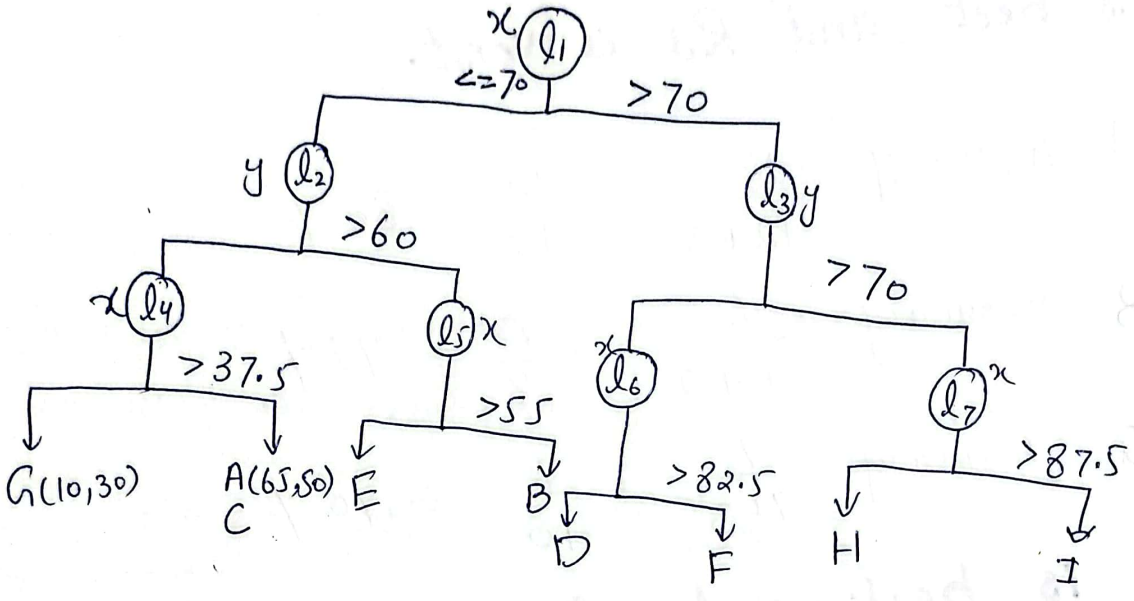
Median = 70

(c) Sorted for x

10	30	50	90
65	50	60	70
Median = 37.5		Median = 55	

75	25	80	85
90	65	95	75
Median = 82.5		Median = 87.5	

KD-Tree.



Z=(55,60) is close to point A(65,50) traversing tree as $l_1 \rightarrow l_2 \rightarrow l_4 \rightarrow A$.

For y - x dimensions

(a) Sorted for y

75	25
10	30
65	50
70	60
90	65
60	70
95	75
80	85
80	90

Median = 65

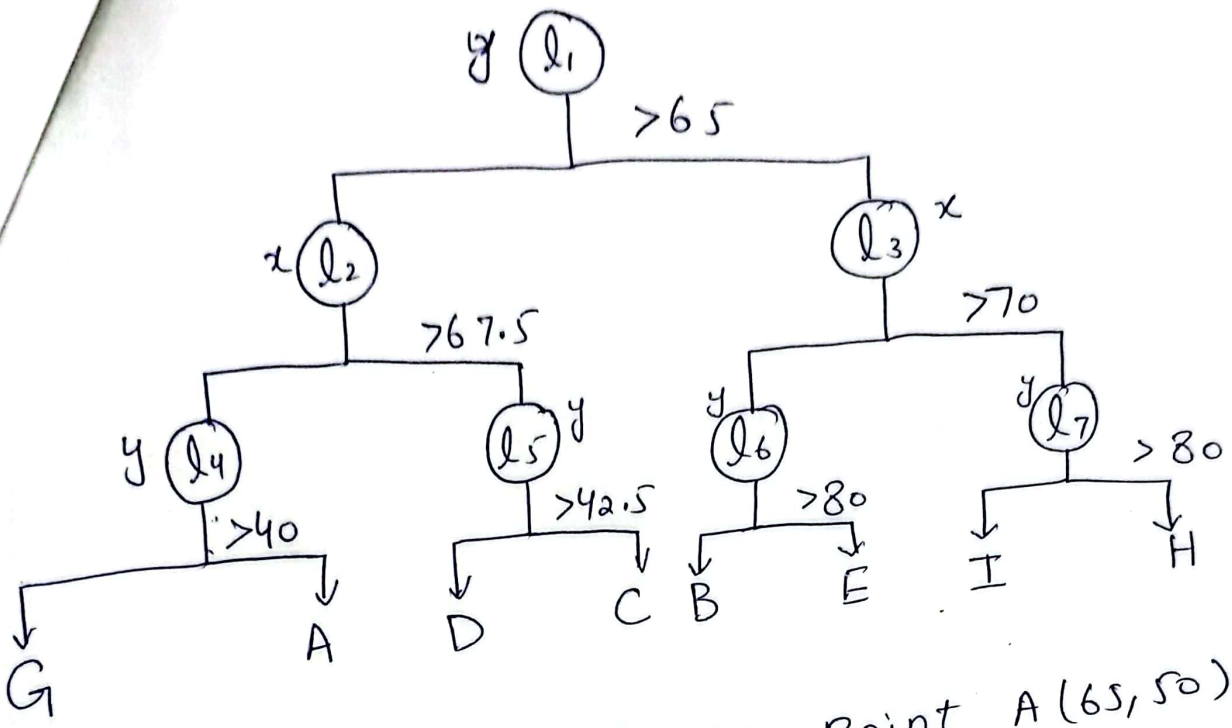
(b) Sorted for x

10	30	50	90
65	50	60	70
70	60	80	85
75	25	95	75
Median = 67.5		Median = 70	

(c) Sorted for y

10	30	75	25
65	50	70	60
Median = 40		Median = 42.5	
60	70	95	75
50	90	80	85
Median = 80		Median = 80	

KD-Tree



Z=(55,60) is closest to point A(65,50) traversing tree as $l_1 \rightarrow l_2 \rightarrow l_4 \rightarrow$ A.

There is no difference in the perform of both trees. Both Trees classifi new point in same manner and still find the same nearest neighbor for th point.