DATA ANALYTICS & MACHINE LEARNING
No. 3

Q.5

→

Candidate	T_L	T_R
1	Occupation: Service	Occupation: Sales Management, Staff
2	Occupation: Management	Occupation: Service Sales, Staff
3	Occupation: Sales	Occupation: Service Management, Staff
4	Occupation: Staff	Occupation: Service Management, Sales
5	Gender: Male	Gender: Female
6	Age < 30	Age >= 30
7	Age < 40	Age >= 40

Split	P_L	P_R	$P(j \mid t_L)$	$P(j \mid t_R)$	$2P_L P_R$	$Q(s \mid t)$	$\phi(s \mid t)$
1	3/11 = 0.272	0.73	$L_1 = 0.33$	$L_1 = 0.125$	0.39	0.58	0.22
			$L_2 = 0.33$	$L_2 = 0.25$			
			$L_3 = 0.33$	$L_3 = 0.375$			
			$L_4 = 0$	$L_4 = 0.25$			
2	0.36	0.64	$L_1 = 0$	$L_1 = 0.285$	0.46	1.428	0.65
			$L_2 = 0$	$L_2 = 0.428$			
			$L_3 = 0.5$	$L_3 = 0.285$			
			$L_4 = 0.5$	$L_4 = 0$			
3	0.181	0.81	$L_1 = 0$	$L_1 = 0.222$	0.146	0.861	0.12
			$L_2 = 0.5$	$L_2 = 0.22$			
			$L_3 = 0.5$	$L_3 = 0.33$			
			$L_4 = 0$	$L_4 = 0.22$			

4

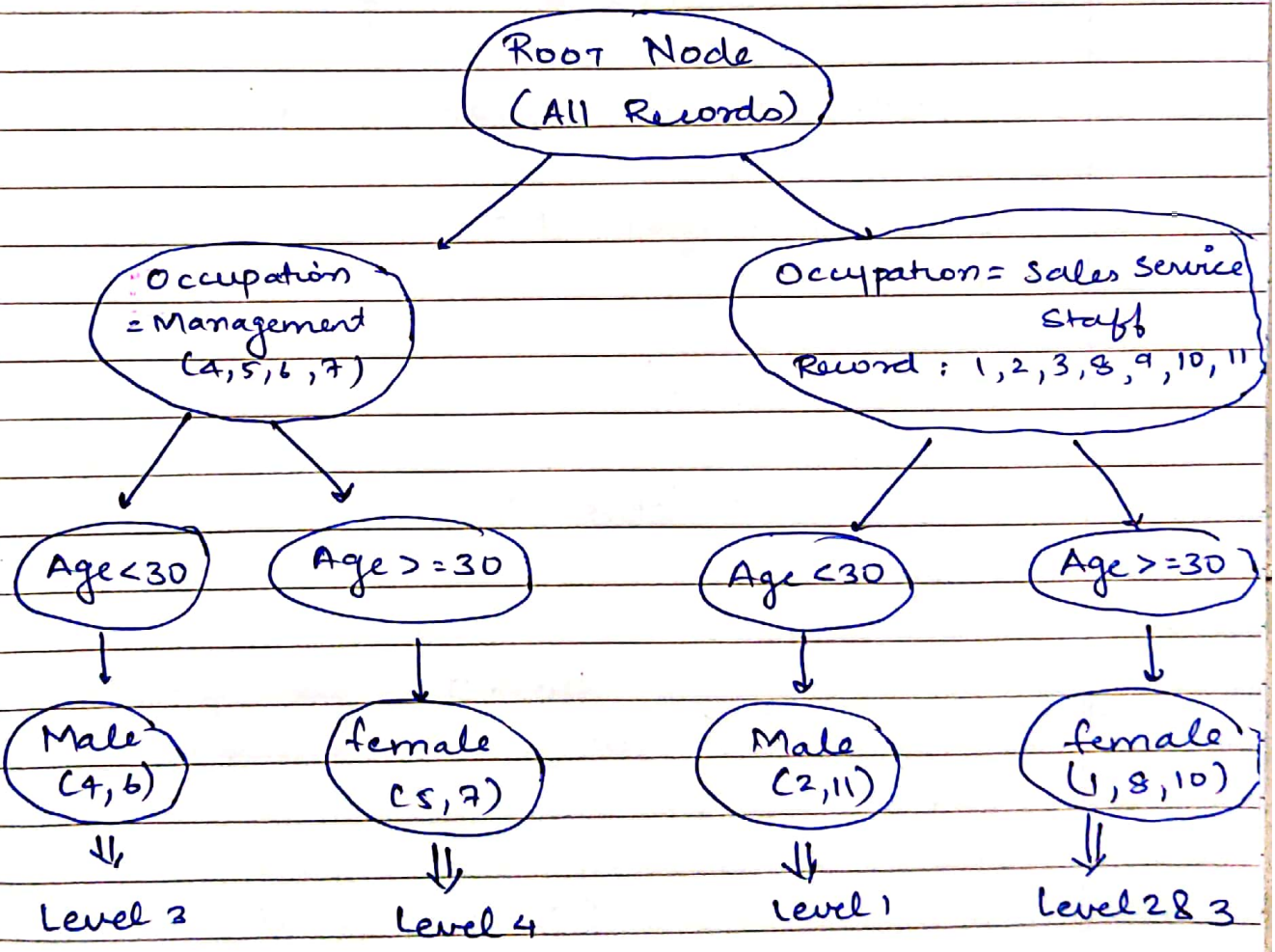
Split	P_L	P_R	$P(j \mid t_R)$	$P(j \mid t_R)$	$2P_L P_R$	$Q(s \mid t)$	$\phi(s \mid t)$
4	0.181	0.81	$L_1 = 0.5$	$L_1 = 0.111$	0.146	1.33	0.194
			$L_2 = 0.5$	$L_2 = 0.222$			
			$L_3 = 0$	$L_3 = 0.444$			
			$L_4 = 0$	$L_4 = 0.222$			
5	0.54	0.45	$L_1 = 0.33$	$L_1 = 0$	0.486	0.93	0.451
			$L_2 = 0.33$	$L_2 = 0.2$			
			$L_3 = 0.33$	$L_3 = 0.4$			
			$L_4 = 0$	$L_4 = 0.4$			
6	0.36	0.64	$L_1 = 0.5$	$L_1 = 0$	0.46	1.428	0.656
			$L_2 = 0$	$L_2 = 0.428$			
			$L_3 = 0.5$	$L_3 = 0.2857$			
			$L_4 = 0$	$L_4 = 0.2857$			
7	0.63	0.36	$L_1 = 0.285$	$L_1 = 0$	0.46	0.6429	0.2957
			$L_2 = 0.285$	$L_2 = 0.25$			
			$L_3 = 0.285$	$L_3 = 0.5$			
			$L_4 = 0.1428$	$L_4 = 0.25$			

After split for Node 2

Split	P_L	P_R	$P(J\|t_L)$	$P(J\|t_R)$	$2P_LP_R$	$Q(s\|T)$	$\Phi(s\|T)$
1	0.429	0.571	$L_1=0.33$ $L_2=0.33$ $L_3=0.33$ $L_4=0$	$L_1=0.25$ $L_2=0.5$ $L_3=0.25$ $L_4=0$	0.48	0.33	0.163
2	0.286	0.714	$L_1=0$ $L_2=0.5$ $L_3=0.5$ $L_4=0$	$L_1=0.4$ $L_2=0.4$ $L_3=0.2$ $L_4=0$	0.40	0.80	0.327
3	0.286	0.714	$L_1=0.5$ $L_2=0.5$ $L_3=0$ $L_4=0$	$L_1=0.2$ $L_2=0.4$ $L_3=0.4$ $L_4=0$	0.408	0.80	0.327
4	0.571	0.429	$L_1=0.5$ $L_2=0.5$ $L_3=0$ $L_4=0$	$L_1=0$ $L_2=0.33$ $L_3=0.66$ $L_4=0$	0.4898	1.33	0.653
5	0.478	0.571	$L_1=0$ $L_2=0.33$ $L_3=0.66$ $L_4=0$	$L_1=0.5$ $L_2=0.5$ $L_3=0$ $L_4=0$	0.488	1.33	0.651
6	0.286	0.714	$L_1=1$ $L_2=0$ $L_3=0$ $L_4=0$	$L_1=0$ $L_2=0.66$ $L_3=0.4$ $L_4=0$	0.4082	2.0	0.816

Split	P_L	P_R	$P(J/t_L)$	$P(J/t_R)$	$2P_LP_R$	$Q(s/t)$	$\Phi(s/t)$
7	0.571	0.429	$L_1 = 0.5$	$L_1 = 0$	0.48	1.33	0.653
			$L_2 = 0.5$	$L_2 = 0.33$			
			$L_3 = 0$	$L_3 = 0.66$			
			$L_4 = 0$	$L_4 = 0$			

CART DECISION TREE



Q. 6. C4.5

→ Entropy before splitting for entire dataset :

i) less than 35000 Level 1 P = 2/11

ii) between ≥ 35000 & ≤ 45000 Level 2 P = 3/11

iii) Between ≥ 45000 & ≤ 55000 Level 3 P = 4/11

iv) ≥ 55000 Level 4 P = 2/11

$$H(T) = -\sum_j P_j \log_2(P_j)$$

$$= -\frac{2}{11}\log_2(2/11) - \frac{3}{11}\log_2(3/11) - \frac{4}{11}\log(4/11)$$

$$= -2/11 \log(2/11)$$

$$= 1.927 \text{ bits}$$

Split on Occupation :

Entropy of 4 Branches :

$$H(\text{service}) = -\frac{1}{3}\log(1/3) - 1/3(\log(1/3)) - 1/3\log(1/3)$$

$$H(\text{service}) = 1.587$$

$$H(\text{management}) = 1$$

$$H(\text{sales}) = 1$$

$$H(\text{staff}) = 1$$

$$H_{\text{occupation}}(T) = \frac{3}{11} \times (1.587) + \frac{4}{11}(1) + \frac{2}{11}(1) + \frac{2}{11} \times (1)$$

$$= 1.158$$

$H(T) = -H_{occupation}(T) = 0.769$ bits

$H(male) = 1.583$

$H(female) = 1.51$

$$H_G(T) = \sum_{i}^{k} P_i H_G(T_i) = \frac{6}{11} \times (1.583) + \frac{5}{11} (1.51)$$

$= 1.54$ bits

$H(T) - H_G(T) = 1.927 - 1.54 = 0.377$ bits

Split on Age:

$H(Age \leq 25) = 0.9$

$H(Age > 25) = 1.5$

$$H_{Age}(T) = \frac{3}{11} \times 0.9 + \frac{8}{11} \times 1.5 = 1.3$$

$H(T) - H_{Age}(T) = 1.9 - 1.3 = 0.6$

Age Split: $H(Age \leq 35) = 1.9$

$H(Age > 35) = 1.5$

$$H_{Age}(AT) = \frac{7}{11}(1.9) + \frac{4}{11}(1.5) = 1.7 \text{ bits}$$

Split on Age

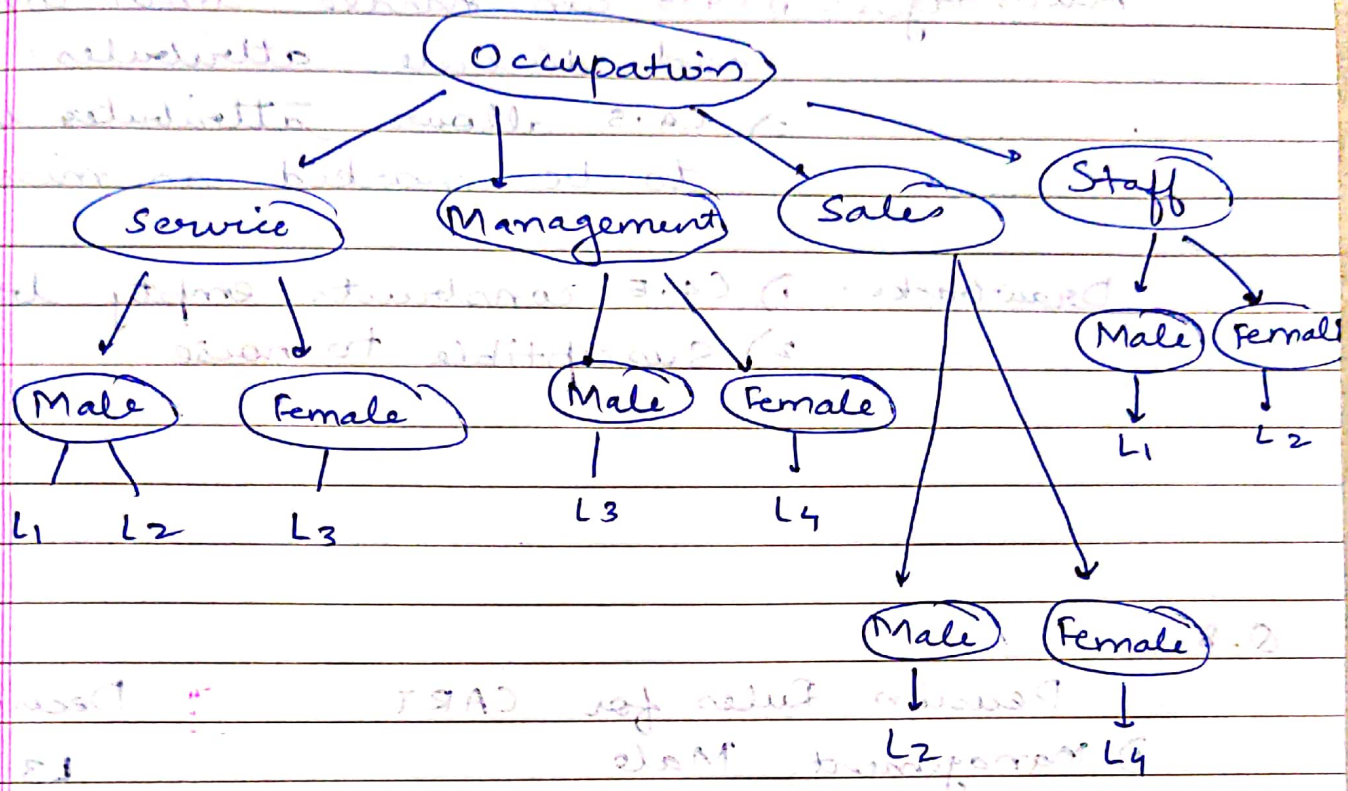
$P(\leq 45)$ $H(Age \leq 45) = 1.9$

$P(>45) = \frac{1}{11}$ $H(Age > 45) = 0.3$

$$H_{Age}(T) = \sum_{i=1}^{k} P_i H_{Age}(T_i) = \frac{10}{11}(1.9) + \frac{1}{11}(0.3)$$

$= 1.75$

H(T) - Hage(T) = 1.927 - 1.75 = 0.172 bits



Q.7.

CART:

Benefits : 1) CART can easily handle both numerical and categorical variables
2) CART can easily handle outliers

Drawbacks: 1) CART splits only by one variable
2) CART may have unstable decision tree. Insignificant modification of learning sample such as eliminating several observations and cause changes in decision Tree : increase or decrease of Tree complexity.

C4.5

Advantages : 1) C4.5 can handle both continuous and discrete attributes
2) C4.5 allows attributes values to be marked as missing

Drawbacks : 1) C4.5 constructs empty branches
2) Susceptible to noise

Q.8

Decision Rules for CART	Decision
1) Management, Male	L3
2) Management, Female	L4
3) High Age, Service, Male	L1
4) High Age, Service, Female	L3
5) High Age, Sales	L1
6) High Age, Staff	L2
7) Low Age, Service	L1
8) Low Age, Sales	L2
9) Low Age, Staff	L1

Q.9. Decision Rules for C4.5

Decision Rules for C4.5	Decision
1) Service, Male, Low Age	L_1
2) Service, Male, High Age	L_2
3) Service, Female	L_3
4) Management, Male	L_3
5) Management, Female	L_4
6) Sales, Male	L_2
7) Sales, Female	L_4
8) Staff, Male	L_1
9) Staff, Female	L_2

Q.10. The decision rule for C4.5 is more precise and descriptive as compared to CART.

C4.5 gives a more detailed picture in comparision to CART

C4.5 splitting criteria is based on gain ratio.