MIS 637 ASSIGNMENT-3

POORVI RAUT - 20009560

Chapter 6, Page 127, Problems 5-10

Occupation	Gender	Age	Salary
Service	Female	45	$48,000
	Male	25	$25,000
	Male	33	$35,000
Management	Male	25	$45,000
	Female	35	$65,000
	Male	26	$45,000
	Female	45	$70,000
Sales	Female	40	$50,000
	Male	30	$40,000
Staff	Female	50	$40,000
	Male	25	$25,000

Consider the data in above Table. The target variable is salary. Start by discretizing salary as follows:

Less than $35,000 Level 1

$35,000 to less than $45,000 Level 2

$45,000 to less than $55,000 level 3

Above $55 000 Level 4.

Construct a classification & regression tree to classify salary based on the other variables. Do as much as you can by hand, before turning to the software.

Occupation	Gender	Age	Salary ($)	level
Service	female	45	$48,000	level 3
Service	male	25	$25,000	level 1
Service	male.	33	$35,000	level 2
Management	male	25	$45,000	level 3
Management	Female	35	$65,000	level 4
Management	male	26	$45,000	level 3
Management	Female	45	$70,000	level 4

CONTD page 2.

Occupation	Gender	Age	Salary ($)	Level
Sales	female	40	$50,000	level 3
Sales	male	30	$40,000	level 2
staff	female	50	$40,000	level 2
staff	male	25	$25,000	level 1

Candidate split for t = Root Node.

Candidate Split	Left Child Node, t_L	Right child node, t_R
1	Occupation : Service	Occupation : management, sales, 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

$\phi(s|t)$:- optimality measure of split (maximizes split over all possible splits at node t)

let $\phi(s|t)$ be measure of "goodness" of candidate split s at node t

where formula is

$$\phi(s|t) = 2 P_L P_R \sum_{j=1}^{\#classes} |P(j|t_L) - P(j|t_R)|$$

t_L = Left child node of node t

t_R = Right child node of node t

P_L = number of records at t_L / number of records in training set

P_R = number of records at t_R / number of records at training set

$P(j|t_L)$ = number of class j records at t_L / number of records at t

$P(j|t_R)$ = number of class j records at t_R / number of records at t

$$Q(s|t) = \sum_{j=1}^{\#\text{classes}} |P(j|t_L) - P(j|t_R)|$$

$Q(s|t)$ is large when distance between $P(j|t_L)$ & $P(j|t_R)$ is maximized across each value of target variable.

∴ The values for our candidate split for t = Root will be determined as for example t_L

1st candidate value for occupation = service

$f_L = \frac{3}{11} = 0.2727$ since service as occupation in dataset = 3, Total value = 11

$f_R = \frac{8}{11} = 0.7272$ // other occupation excluding service / Total

$P(j|t_L) =$ ~~of salary~~ Level 1 (service) / Total service in left child node $= \frac{1}{3} = 0.333$

Level 2 (service) / service in t_L $= \frac{1}{3} = 0.33$

Level 3 (service) / service in t_L $= \frac{1}{3} = 0.33$

Level 4 of salary in service / service in t_L $= \frac{0}{3} = 0$

$P(j|t_R) =$ Level 1 (other in t_R) / other occupation in t_R $= \frac{1}{8} = 0.125$

Level 2 (other t_R) / other occupation in t_R $= \frac{2}{8} = 0.25$

Level 3 (other t_R) / other ~~occupations~~ t_R $= \frac{3}{8} = 0.375$

Level 4 (other t_R) / Other occupations in t_R $= \frac{2}{8} = 0.25$

$2P_L \times P_R = 2 \times 0.2727 \times 0.7273 = 0.3966$

$$\Phi(S|T) = \sum_{j=1}^{\#class} |P(j|t_L) - P(j|t_R)|$$

$$= |\Sigma P(j|t_L) - \Sigma P(j|t_R)|$$

$$= |P(level\ 1|t_L) - P(level\ 1|t_R)| + |P(level\ 2|t_L - P(level\ 2|t_R$$
$$+ |P(level\ 3|t_L) - P(level\ 3|t_R)| + (P(level\ 4|t_L) - P(level\ 4|t_R)|$$

$$= |0.333 - 0.125| + |0.33 - 0.25| + |0.33 - 0.375| + |0 - 0.25|$$

$$= 0.208 + 0.08 + 0.045 + 0.25$$

$$\boxed{Q(S|t) = 0.583}$$

$$\phi(S|t) = Q(S|t) \times 2P_L P_R$$

$$= 0.583 \times 0.3966$$

$$= \boxed{0.2312}$$

Similarly we calculate values for other splits and determine value of $\phi(S|t)$ which needs to be maximized.

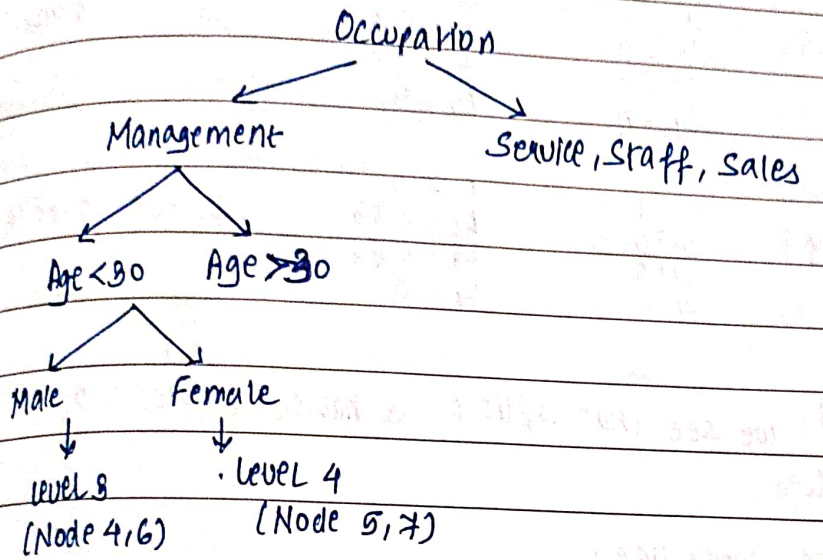
	P_L	P_R	$P(j\|t_L)$	$P(j\|t_R)$	$Q(S\|t)$	$2P_L P_R$	$\phi(S\|t)$
t_L Service 1.	0.2727	0.7273	L1 = 0.333 L2 = 0.33 L3 = 0.33 L4 = 0	L1 = 0.125 L2 = 0.25 L3 = 0.375 L4 = 0.25	0.583	0.3966	0.2312
t_L Mgmt 2	0.3636	0.6363	L1 = 0 L2 = 0 L3 = 0.5 L4 = 0.5	L1 = 0.285 L2 = 0.428 L3 = 0.285 L4 = 0	1.428	0.4626	0.66
t_L Sales 3.	0.1818	0.8181	L1 = 0 L2 = 0.5 L3 = 0.5 L4 = 0	L1 = 0.22 L2 = 0.22 L3 = 0.33 L4 = 0.22	0.889	0.2974	0.2643
t_L Staff 4	0.1818	0.8181	L1 = 0.5 L2 = 0.5 L3 = 0 L4 = 0	L1 = 0.11 L2 = 0.22 L3 = 0.44 L4 = 0.22	1.333	0.2974	0.3964
t_L male 5.	0.5454	0.4546	L1 = 0.33 L2 = 0.33 L3 = 0.33 L4 = 0	L1 = 0 L2 = 0.20 L3 = 0.40 L4 = 0.40	0.933	0.4958	0.4626
t_L Age<30 6.	0.3636	0.6363	L1 = 0.5 L2 = 0 L3 = 0.5 L4 = 0	L1 = 0 L2 = 0.428 L3 = 0.285 L4 = 0.28	1.4285	0.4627	0.66
≥30.							

P_L	P_R	$P(j\|t_L)$	$P(j\|t_R)$	$\Phi(s\|t)$	$2P_LP_R$	$\phi(s\|t)$
0.6363	0.3636	$L_1 = 0.285$ $L_2 = 0.285$ $L_3 = 0.165$ $L_4 = 0.142$	$L_1 = 0$ $L_2 = 0.25$ $L_3 = 0.5$ $L_4 = 0.25$	0.642	0.4627	0.297

from above table, row 2 & 6 has highest value of $\phi(s|t)$

∴ Random key : 2

∴ The decision tree looks like :-



Split	Left child node t_L	Right child node t_R
1	Occupation : Service	Occupation : Sales, staff
3	Occupation : sales	Occupation : service, staff
4	Occupation : staff	Occupation : service, sales
5	Gender : male	Gender : female
6	Age <30	Age >= 30
7	Age <40	Age >= 40

Calculate : $\phi(s|t) = 2P_LP_R \sum_{j=1}^{\#classes} |P(j|t_L) - P(j|t_R)|$

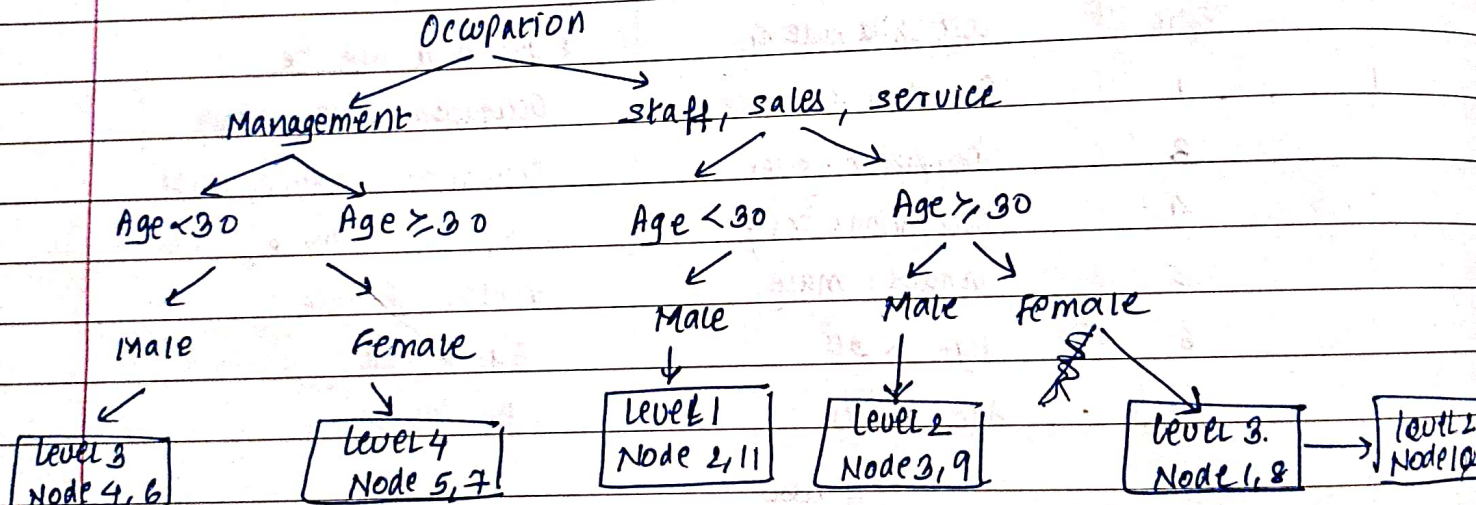
	P_L	P_R	$P(j\|t_L)$	$P(j\|t_R)$	$\Phi(s\|t)$	$2P_LP_R$	$\phi(s\|t)$
1.	3/7 = 0.428	4/7 = 0.5714	$L_1 = 1/3 = 0.33$ $L_2 = 0.33$ $L_3 = 0.33$ $L_4 = 0$	$L_1 = 1/4 = 0.25$ $L_2 = 2/4 = 0.5$ $L_3 = 1/4 = 0.25$ $L_4 = 0$	0.332	0.489	0.1625
3.	2/7 = 0.2857	5/7 = 0.7142	$L_1 = 0$ $L_2 = 1/2 = 0.5$ $L_3 = 1/2 = 0.5$ $L_4 = 0$	$L_1 = 2/5 = 0.4$ $L_2 = 2/5 = 0.4$ $L_3 = 1/5 = 0.2$ $L_4 = 0$	0.8	0.408	0.3264

Date:

	P_L	P_R	$P(j\|t_L)$	$P(j\|t_R)$	$\phi(s\|t)$	$2P_LP_R$	$\phi(s\|t)$
4 ✱	2/7 = 0.2857	5/7 = 0.7142	$L_1 = 1/2 = 0.5$ $L_2 = 1/2 = 0.5$ $L_3 = 0$ $L_4 = 0$	$L_1 = 1/5 = 0.2$ $L_2 = 2/5 = 0.4$ $L_3 = 2/5 = 0.4$ $L_4 = 0$	0.8	0.408	0.3264
5.	6/14 = 0.5714	0.4285	$L_1 = 0.5$ $L_2 = 0.5$ $L_3 = 0$ $L_4 = 0$	$L_1 = 0$ $L_2 = 0.33$ $L_3 = 0.66$ $L_4 = 0$	1.333 ~~0.4896~~	0.4896 ~~0.333~~	0.6526
→ 6	0.2857	0.7142	$L_1 = 1$ $L_2 = 0$ $L_3 = 0$ $L_4 = 0$	$L_1 = 0$ $L_2 = 0.6$ $L_3 = 0.4$ $L_4 = 0$	2	0.408	0.816
7	0.5714	0.4285	$L_1 = 0.5$ $L_2 = 0.5$ $L_3 = 0$ $L_4 = 0$	$L_1 = 0$ $L_2 = 0.33$ $L_3 = 0.66$ $L_4 = 0$	≈1.333	0.4896	0.652

Here in split 2nd: we see that split 6 is having $\phi(s|t) = 0.816$ highest among others values.

The decision tree looks like:



Decision Rules :-

If Occupation is = Management and Age<30 and gender = Male	then L_3 (Node 4, 6)
If Occupation = Management and Age≥30 and gender = Female	then L_4 (Node 5, 7)
If Occupation = staff, sales, service and Age<30 and gender = male	then L_1 (Node 2, 11)

If occupation = staff, sales, service and Age ≥ 30 and gender = male	then L2 (Node 3, 9)
If Occupation = staff, Sales, Service and Age ≥ 30 and gender = female	then L3 (Node 1, 8)
If Occupation = staff, Sales, Service and Age ≥ 30 and gender = Female	then L2 (Node 10)

6. Construct a C4.5 decision tree to classify salary based on the other variables, Do as much as you can by hand, before turning to the software

split	child nodes (left)	child node (right)
1	occupation : service, occupation : management	occupation : sales, staff
2	Gender = male	Gender = Female
3	Age ≤ 25	Age > 25
4	Age ≤ 35	Age > 35
5	Age ≤ 45	Age > 45

H(T) = Entropy

levels	salary (target)	Probability
level 1	less than 35,000$	2/11
level 2	Between ≥35000 and <45000	3/11
level 3	Between ≥45,000 and <55,000	4/11
level 4	>55000$	2/11

$$\text{Entropy } H(X) = -\sum_j p_j \log_2(p_j)$$

Variable X has K values with probabilities $p_1, p_2 \dots p_K$

for variables with several outcomes, use weighted sum of $-\log_2(p)$ to transmit result

$$= \frac{-2}{11}\log\left(\frac{2}{11}\right) \frac{-3}{11}\log\left(\frac{3}{11}\right) \frac{-4}{11}\log\left(\frac{4}{11}\right) \frac{-2}{11}\log\left(\frac{2}{11}\right)$$

$$H(T) = -p_j \log_2(p_j) = \boxed{1.927 \text{ bits}}$$

Split on Occupation :

P(service) = 3/11

P(Management) = 4/11

P(sales) = 2/11

P(staff) = 2/11

Occupation : service : $P(<55,000) = 1/3 = P(>55,000) = 0$

Occupation : Management : $P(<55,000) = 1/2 = P(>55,000) = 1/2$

Occupation : Sales : $P(<55,000) = 1/2 = P(>55,000) = 0$

Occupation : staff : $P(<55,000) = 0 = P(>55,000) = 0$

Entropy of 4 branches : service, Management, sales, staff

$H(\text{service}) = -1/3 \log(1/3) - 1/3 \log(1/3) - 1/3 \log(1/3) - 0 \log(0) = 1.58$

$H(\text{management}) = -0 \log(0) - 0 \log(0) - 1/2 \log(1/2) - 1/2 \log(1/2) = 1$

$H(\text{sales}) = -0 \log(0) - 1/2 \log(1/2) - 1/2 \log(1/2) - 0 \log(0) = 1$

$H(\text{staff}) = -1/2 \log(1/2) - 1/2 \log(1/2) - 0 \log(0) - 0 \log(0) = 1$

Combining Entropy of all 4 branches along with their proportion P_i

$$P_i = \frac{3}{11} * (1.58) + \frac{4}{11} * (1) + \frac{2}{11} * (1) + \frac{2}{11} * (1)$$

$$= \boxed{1.1572 \text{ bits}}$$

$$K H_o(T) = P_i H_o(T_i) \; i = 1$$

∴ From the split on Occupation we have the attribute.

$H(T) - T = 1.927 - 1.1572 = \boxed{0.7698 H_o \text{ bits}}$

Now split for Gender :

P(Male) = 6/11 P(Female) = 5/11

Gender ⇒ Male

$P(<35,000) = 1/3$

$P(\geq 35,000 \;\&\; <45,000) = 1/3$

$P(\geq 45,000 \;\&\; <55,000) = 1/3$

$P(>55,000) = 0$

Gender → Female

$P(<35,000) = 0$

$P(\geq 35000 \;\&\; 45,000) = 1/5$

$P(\geq 45,000 \;\&\; <55,000) = 2/5$

$P(>55,000) = 2/5$

Date: youvA

From above the Entropy

H(male) = -1/3 log(1/3) -1/3 log(1/3) -1/3 log(1/3) -0 log(0) = 1.58

H(female) = -0 log(0) -1/5 log(1/5) -2/5 log(2/5) = 1.52

Combining Entropy of 2 branches along with their proportion P_i

$= \frac{6}{11} * (1.58) + \frac{5}{11} (1.52) =$ 1.55 bits

$H_G(T) = \sum_{i=1} P_i H_G(T_i)$

∴ H(Gain) H(T) - T = 1.927 - 1.55 = 0.377 bits

H(G)

Split on Age :-

P(<=35) = 7/11 ⇒ P(>35) = 4/11

Age <=35 ⟹ P(<35,000) = 2/7

P(>=35,000 & <45,000) = 2/7

P(>=45,000 & <55,000) = 2/7

(1) Age >35

P(≥35,000) = 0

P(>=35,000 & <45,000) = 1/4

P(>=45,000 & <55,000) = 2/4

P(>55,000) = 1/4

Entropy of from the above 2 branches Age <=35 & Age >35

H(Age <= 35) = $-\frac{2}{7}\log(\frac{2}{7}) - \frac{2}{7}\log(\frac{2}{7}) - \frac{2}{7}\log(\frac{2}{7}) -$

1/7 log(1/7) = 1.93

H(Age >35) = -1/4 log(1/4) - 1/4 log(1/4) - 2/4 log(2/4) = 1.5

Combining entropy of above 2 branches, P_i

= 7/11 * (1.93) + 4/11(1.5) = 1.77 bits

$H_{(Age)}(T) = \sum_{i=1}^{k} P_i H_{age}(T_i)$ ⇒

∴ from the info gained by split on Gender attribute is

$H(T) - H_{Gage}(T) = 1.927 - 1.77$

$= \boxed{0.157 \text{ bits}}$

Split on Age: $P(<=45)$

$= 10/11 \Rightarrow P(>45) = 1/11$

Age: $<=45 = P(<35,000) = 2/10$

$P(>=35,000 \text{ and } <46,000) = 2/10$

$P(>=45,000 \text{ and } <55,000) = 4/10$

$P(>55,000) = 2/10$

Age: >45

$P(<35,000) = 0$

$P(>=35,000 \text{ and } <45,000) = 9/11$

$P(>=45,000 \text{ and } <55,000) = 0$

$P(>55,000) = 0$

Entropy for above two: Age <=45 and Age >45

$H(\text{Age} <=45) = -2/10 \log(2/10) - 2/10 \log(2/10) - 4/10 \log(4/10) - 2/10 \log(2/10) = 1.92$

$H(\text{Age} >45) = -1/11 \log(1/11) = 0.3136$

Combining the Entropy of 2 branches, along with their proportion Pi

$= 10/11 * 1.92 + 1/11 (0.3136) = 1.7735 \text{ bits}$

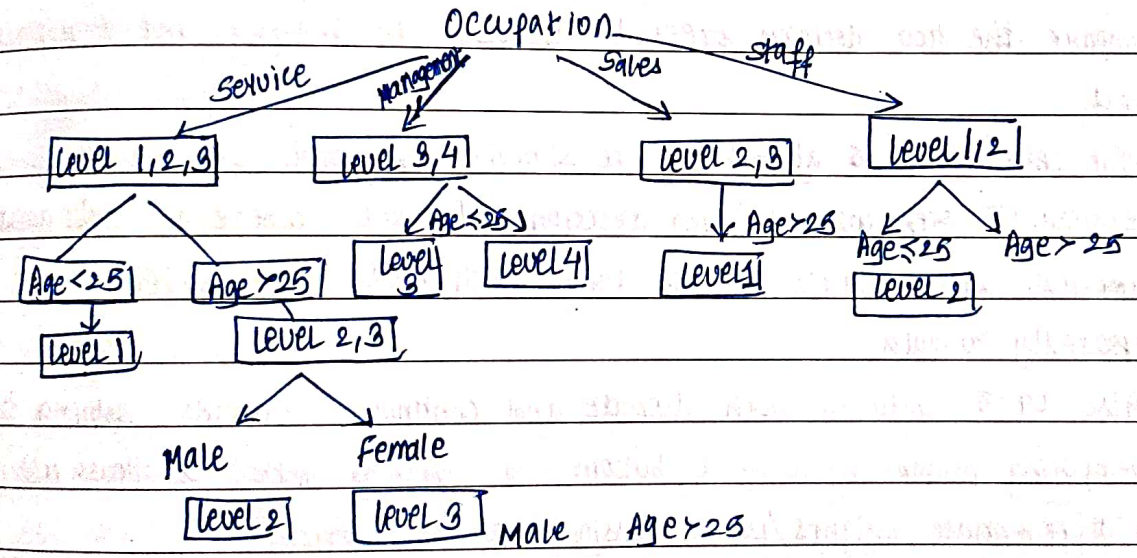
∴ from above values gained by split on gender $H(T) - H_{age}(T)$

$= 1.927 - 1.7735 = \boxed{0.1465 \text{ bits}}$

Page No.:	
Date:	YOUVA

Condition split	child node	Information Gain
1	Occupation : Service, Management, Sales, Staff	0.7689 bits
2	Gender = Male, Female	0.377 bits
3	Age $\leq$ 25 & Age $>$ 25	0.547 bits
4	Age $\leq$ 35 and Age $>$ 35	0.187 bits
5	Age $\leq$ 45 and Age $>$ 45	0.1465 bits

C4.5 Decision Tree :-



Decision Rules :-

	Antecedent	Consequent	Support	Confidence
1	if occupation = service & Age <= 25	Then L1	3/11	1
2	if occupation = service and Age > 25 and male	Then L2	1/3	1
3	if occupation = service and Age > 25 and Female	Then L3	1/3	1

	Antecedent	Consequent	Support	Confidence
4.	if Occupation = Management and Age <= 25	Then L3	3/11	1
5	if Occupation = Management and Age >25	Then L4	8/11	1
6	if Occupation = Sales and Age >25 and Male	Then L2	1/2	1
7	if Occupation = Sales and Age >25 and Female	Then L3	1/2	1
8	if Occupation = Staff and Age <= 25	Then L1	3/11	1
9.	if Occupation = Staff and Age >25	Then L2	8/11	1

7. Compare the two decision trees & discuss the benefits and drawbacks of each.

→ The CART and C4.5 algorithms are similar, however C4.5 uses an intermediate step when building decision rules set, whereas CART used numerical split criteria to split data which needs to be applied repeatedly to data.

→ Also C4.5 includes both discrete and continuous features, solving the overfitting problem by using a bottom-up pruning method & allows users to differentiate weights/values assigned to each attribute.

8. Generate the full-set of decision rules for the CART decision tree

1.	If Occupation = Management and Age <30 and Male	then L3 (Node 4, 6)
2.	If Occupation = Management and Age >=30 and Female	then L4 (Node 5, 7)

3	if Occupation = Staff, Sales, Service & Age	then L1 (Node 2, 11)
4	if occupation = staff, Sales, Service and Age >=30 and Male	then L2 (Node 3, 9)
5.	if occupation = staff, sales, service and Age >=30 & Female	then L3 (Node 1, 8)
6.	if occupation = staff, sales, service and Age >= 30 and female and staff	then L2 (Node 10)

9. Generate the full set of decision rules for the C4.5 decision tree.

	Antecedent	Consequent	Support	Confidence
1.	if occupation = Service and Age <= 25	Then L1	3/11	1
2	if Occupation = Service and Age > 25 & Male	Then L2	1/3	1
3	if Occupation = Service and Age > 25 & Female	Then L3	1/3	1
4.	if Occupation = Mgmt and Age > 25	Then L3	3/11	1
5	if occupation = Mgmt & Age > 25	Then L4	8/11	1
6	if occupation = Sales & Age > 25 & Male	Then L2	1/2	1
7	if occupation = sales Age > 25 & Female	Then L3	1/2	1
8	if Occupation = staff Age <= 25	Then L1	3/11	1
9	if occupation = staff & Age > 25	Then L2	8/11	1

10. Evaluate the advantages & disadvantages of the two sets of decision rules trees

CART advantages :-

(i) It builds binary trees, in ~~which~~ which each internal node has exactly two outgoing branches

(ii) Using the two criteria, splits are chosen & the resulting tree is pruned using cost complexity method

(iii) CART is able to handle variables that are both numerical & categorical, & it does so with outliers (works fine).

(iv) It has more general decision making rules

CART disadvantages :-

(i) It can only split based on one variable.

(ii) The decision tree growth can be unstable.

C4.5 advantages :-

(i) The algorithm inherently employs single pass pruning process to mitigate Overfitting

(ii) It can work with both Discrete & continuous data.

(iii) C4.5 can handle the issue of incomplete data very well.

C4.5 disadvantages :-

1. The algorithm suffers from Overfitting
2. It has poor attribute split technique, inability to handle continuous valued & missing valued attributes with high learning cost.

—x-x— END of ASSIGNMENT —x— x——