Anand Karna
2K20/CO/065

Assignment

Tuple#	Age	Salary	Job	Performance	Select
1	Y	H	P	A	N
2	Y	H	P	E	N
3	M	H	P	A	Y
4	O	M	P	A	Y
5	O	L	G	A	Y
6	O	L	G	E	N
7	M	L	G	E	Y
8	Y	M	P	A	N
9	Y	L	G	A	Y
10	O	M	G	A	Y
11	Y	M	G	E	Y
12	M	M	P	E	Y
13	M	H	G	A	Y
14	O	M	P	E	N

1. ID3 Algorithm

→ Computing the entropy of the data set:

$$E(S) = -P(Y)\log_2 P(Y) - P(N)\cdot\log_2 P(N)$$

$$\text{or, } E(S) = -(9/14)\times\log_2(9/14) - (5/14)\times\log_2(5/14)$$

$$\therefore E(S) = 0.94$$

For Age; Values(Age) = {Y, M, O}

$S(Y) = (-2/5\log 2/5 - 3/5\log_2 3/5) = 0.971$

$S(M) = -1\log_2 1 - 0\log_2 0 = 0$

$S(O) = -3/5\log_2 3/5 - 2/5\log_2 2/5 = 0.971$

$I(age) = 5/14\times 0.971 + 4/14\times 0 + 5/14\cdot 0971 = 0.693$

$\text{Gain(age)} = 0.94 - 0.693 = 0.247$

For performance, values(performance) = {A, E}

$S(A) = 1$

$S(E) = -P(Y)\cdot\log_2(Y) - P(N)\cdot\log_2(N)$

$= -P(6/8)\cdot\log_2(6/8) - P(2/8)\cdot\log_2(2/8)$

$= 0.811$

Information from performance

$= (8/14\times 0.811 + 6/14\times 1) = 0.892$

Gain (performance) = 0.94 - 0.892 = 0.048

For salary: Entropies of each feature: Values (salary) = {H, M, L}

$S(H) = -P(N) \cdot \log_2(N) - P(Y) \cdot \log_2(Y)$

$= -(2/4) \cdot \log_2(2/4) - P(2/4) \cdot \log_2(2/4)$

$= 1$

$S(M) = -(4/6) \cdot \log_2(4/6) - (2/6 \cdot \log_2 2/6)$

$= 0.39 + 0.53 = 0.92$

$S(L) = -(3/4) \log_2(3/4) - (1/4 \cdot \log_2 1/4)$

$= 0.3112 + 0.5 = 0.8112$

Information from salary,

$= 4/14 \times 1 + 6/14 \times 0.92 + 4/14 \times 0.8112$

$= 0.28 + 0.39 + 0.237$

$= 0.911$

gain(salary) = $0.94 - 0.911 = 0.029$

Again,

$\therefore$ gain (job) = $0.94 - 0.788 = 0.152$

Now,

selecting the attribute with maximum information gain,

$\therefore$ Age will be the root note.

Again,

calculating information gain of tuple = {1, 2, 8, 9, 11}

Values (salary) = {H, M, L}

So,

Tuple	Salary	Job	Performance	select
1	H	P	A	N
2	H	P	E	N
8	M	P	A	N
9	L	G	A	Y
11	M	G	E	Y

$S(Y) = 0.971$

$S(H) = 0 - 2/2 \cdot \log_2(2) = 0$

$S(M) = 1$

$S(L) = 0$

$\therefore$ gain (Y, salary) = $0.971 - \frac{2}{5} 0 - \frac{2}{5} \times 1 - \frac{1}{5} \times 0 = 0.570$

similarly, values(Job) = {P, G}

$S(Y) = 0.971$

$S(P) = 0$ $S(G) = 0$

$\therefore$ gain(Y, job) = 0.971

→ values (Performance) = {A, E}

$S(Y) = 0.971$

$S(E) = 1$

$S(A) = -1/3 \log_2 1/3 - 2/3 \log_2 2/3 = 0.9183$

$\therefore$ gain (Y, Performance) = 0.0192

similarly,

gain(O, salary) = 0.0192

gain(O, Job) = 0.0192

gain(O, performance) = 0.971

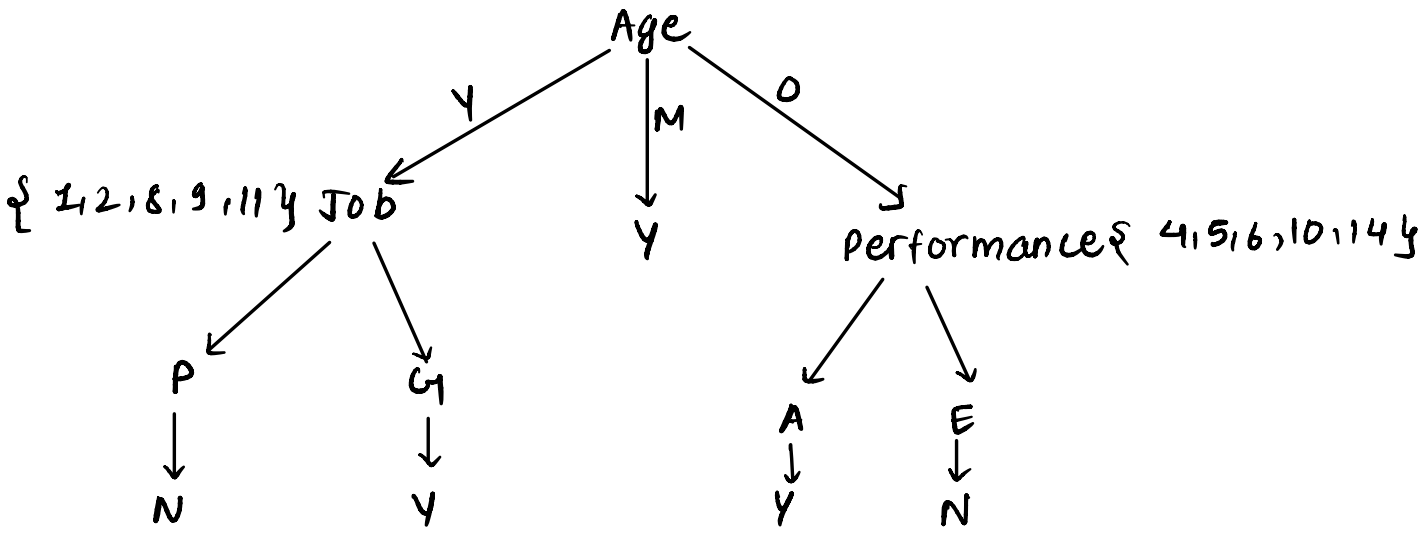


Fig: Decision Tree Using ID3 Algorithm

2. CART Algorithm

Here, attribute with minimum error will be considered as root note.

We distinguish error on the basis of correct classification's number.

→ Age

Values	Freq.		
Y	5	Y N	2 3
M	4	Y N	4 0
O	5	Y N	3 2

Attribute	Rules	Error	Total Error
Age	Y → N M → Y O → Y	2/5 0/4 2/5	4/14

→ Salary

Values	Frequency		
H	4	Y	2
		N	2
M	6	Y	4
		N	2
L	4	Y	3
		N	1

Attribute	Rules	Error	Total Error
salary	H → Y	2/4	5/14
	M → Y	2/6	
	L → Y	1/4	

→ Job

Values	Frequency		
P	7	Y	3
		N	4
G	7	Y	6
		N	1

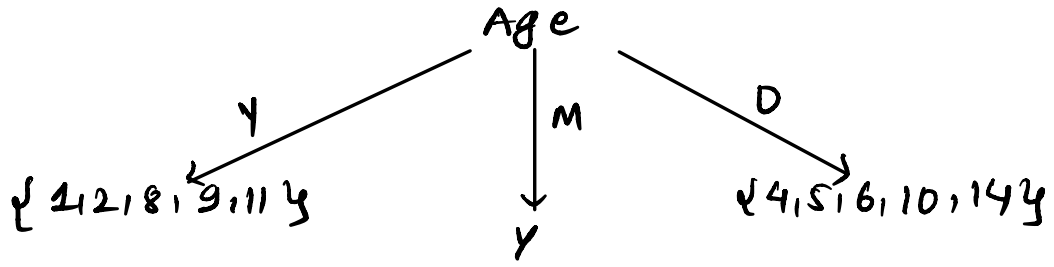
Attribute	Rules	Error	Total Error
Job	P → N	3/7	4/14
	G → Y	1/7	

→ Performance

Values	Frequency		
A	8	Y	6
		N	2
E	6	Y	3
		N	3

Attribute	Rules	Error	Total Error
performance	A → Y	2/8	5/14
	E → N	3/6	

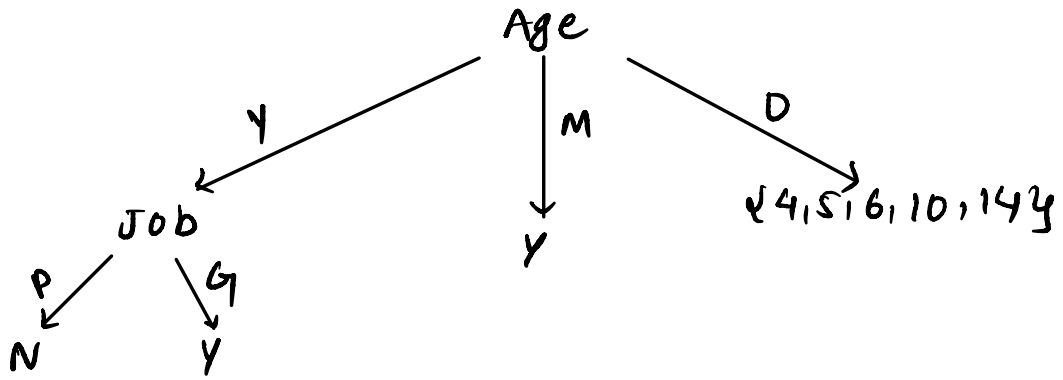
Here, total Error (Age) == total error (Job)
So, we check individual errors & find the rule with zero errors.
∴ 'M' from Age:
error = 0/4
∴ Age is selected as root node.



→ For 'Y'

Attribute	Rules	Error	Total Error
Salary	H → N	0/2	1/5
	M → N	1/2	
	L → Y	0/1	
Job	P → N	0/3	0/5
	G → Y	0/2	
Performance	A → N	1/3	2/5
	E → Y	1/2	

∴ Job is selected



→ For 'O'

Attribute	values	Frequency	Rules	Error	Total Error
salary	M L	3 2	Y Y	1/3 1/2	2/5
Job	P G	2 3	Y Y	1/2 1/3	1/5
Performance	A E	3 2	Y N	0/3 0/2	0/5

∴ Performance is selected.

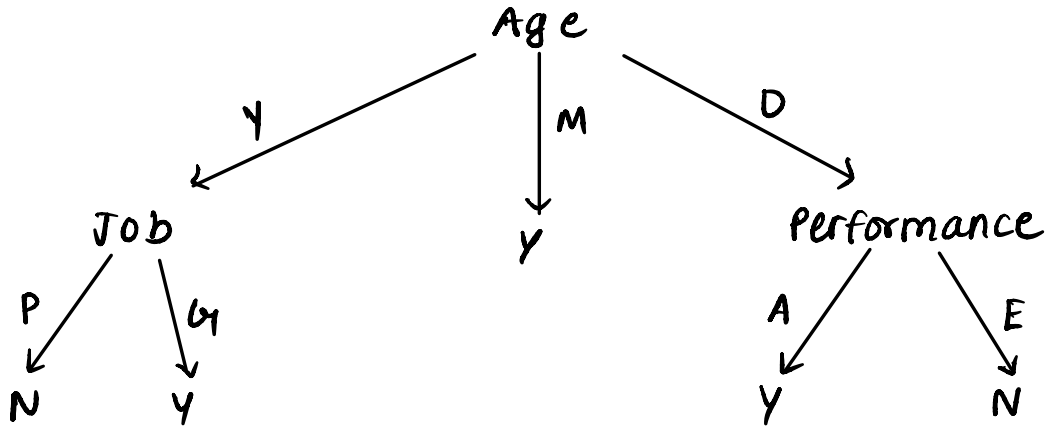


Fig: Decision Tree Using CART Algorithm

3. C4.5 Algorithm

First,

Entropy (dataset) = 0.94

For Age:

values (Age) = {Y, M, O}

gain(Y) = 0.971
gain(M) = 0.0
gain(O) = 0.971
∴ gain (Age) = 0.247

For Salary:

values (salary) = {H, M, L}

gain(H) = 1 gain(M) = 0.8112 gain(L) = 0.911

∴ gain(salary) = 0.029

similarly,

gain(job) = 0.152
gain(performance) = 0.048

→ Split Info & Gain Ratio

$$D_{salary} = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right)$$

$$= 1.557$$

$$\therefore \text{gain ratio(salary)} = \frac{0.029}{1.557} = 0.019$$

Again,

$$D_{Age} = -\frac{5}{14} \log_2\left(\frac{5}{14}\right) - \frac{4}{14} \log_2\left(\frac{4}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right)$$

$$= 0.53 + 0.516 + 0.53$$
$$= 1.576$$

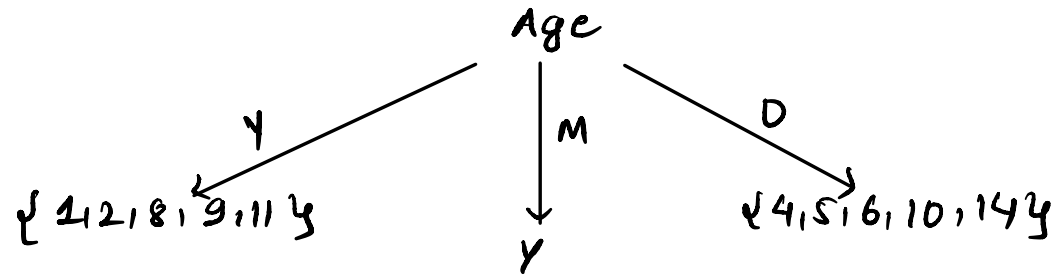
$$\text{gain ratio(Age)} = \frac{0.247}{1.576} = 0.1567$$

similarly,

$D_{job} = 1$ gain ratio (job) = 0.152

$D_{performance} = 0.9852$ gain ratio (performance) = 0.049

The attribute with maximum gain ratio is selected as splitting attribute. We select 'Age'.



Again,

For tuples: {1,2,8,9,11}

Tuple	Salary	Job	Performance	select
1	H	P	A	N
2	H	P	E	N
8	M	P	A	N
9	L	G	A	Y
11	M	G	E	Y

∴ Entropy (Age) = 0.971

$S(H) = 0 - 2/2 \cdot \log_2(2) = 0$

$S(M) = 1$

$S(L) = 0$

→ Salary

$\text{gain}(Y, \text{salary}) = 0.971 - \frac{2}{5}0 - \frac{2}{5} \times 1 - \frac{1}{5} \times 0 = 0.570$

$D(Y, \text{salary}) = 1.5219$

∴ $\text{gain ratio}(Y, \text{salary}) = \frac{0.570}{1.522} = 0.3751$

→ Job

$\text{gain}(Y, \text{job}) = 0.971$

$D(Y, \text{job}) = 0.9709$

∴ $\text{gain ratio} = \frac{0.971}{0.9709} = 1$

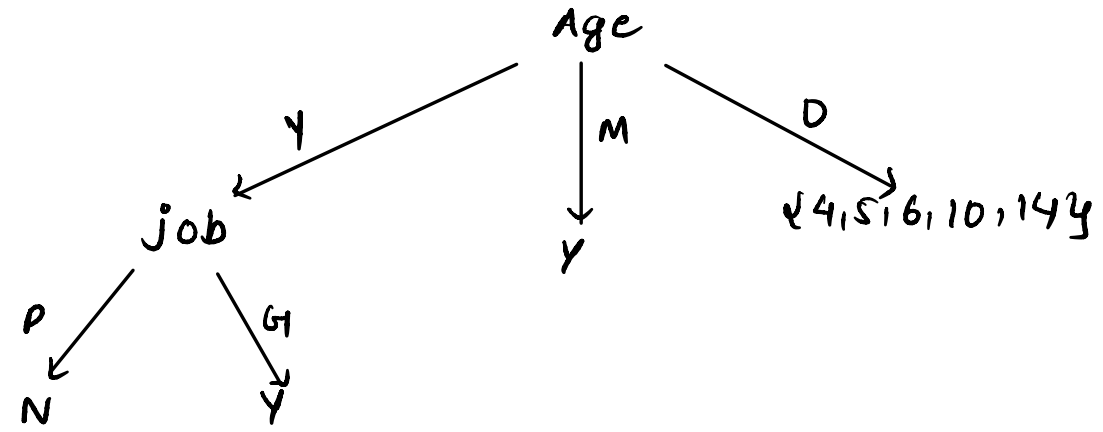
→ Performance

$\text{gain}(Y, \text{performance}) = 0.971 - 0.744 = 0.227$

$D(Y, \text{performance}) = 0.971$

∴ $\text{gain ratio}(Y, \text{performance}) = 0.234$

∴ Gain ratio of job is maximum.



Again, → Salary)

$E(O) = 0.971$

$gain(O, salary) = 0.0192$ $\quad \therefore gain\ ratio(O, salary) = 0.0205$

$D(O, salary) = 0.971$

→ Job

$gain(O, job) = 0.0192$ $\quad D(O, job) = 0.9709$

$\therefore gain\ ratio(O, job) = 0.0205$

→ Performance

$gain(O, performance) = 0.971$

$D(O, performance) = 0.9709$

$\therefore gain\ ratio(O, performance) = 1$ [Highest]

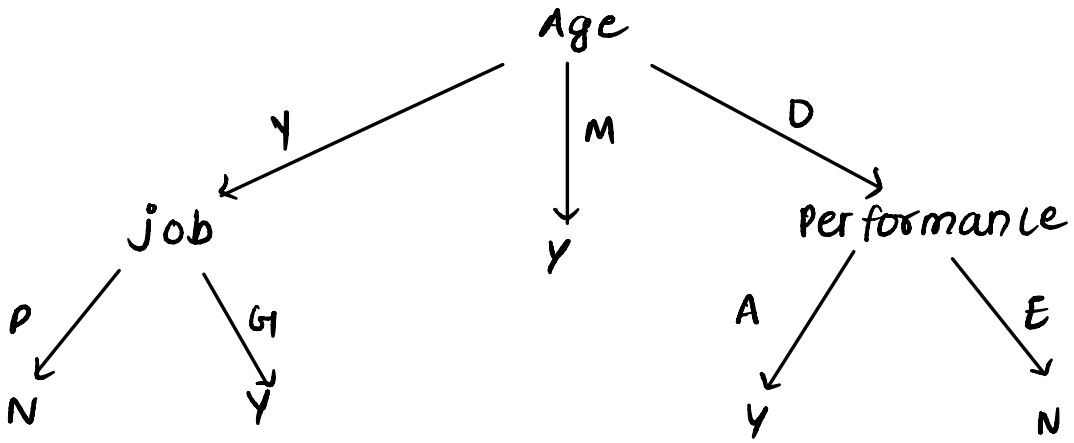


Fig: Decision Tree Using C4.5 Algorithm