Homework 2

Data Mining

Problem 1 :-

1.1

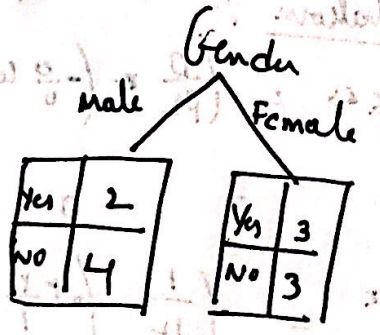
Sun	
Yes	5
No	7

$$E(P) = \left(\frac{-5}{12}\right)\log_2\left(\frac{5}{12}\right) - \left(\frac{7}{12}\right)\log_2\left(\frac{7}{12}\right)$$

$$= +.525 + .453 + .525$$

$$= .978$$

$$E(\text{Survived, Gen}) = \frac{6}{12}\left(\frac{-3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right) + \left(\frac{6}{12}\left(\frac{-2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}\right)\right)$$



$$= \frac{6}{12}\left(1 + .918\right)$$

$$= \frac{6}{12}\left(1.918\right) \quad 0.959$$

$$\approx .924$$

$$E(\text{Survived, Gen}) = .959$$

$$\text{Information gain} = E(\text{Survived}) - E(\text{Survived, Gen})$$

$$= .024$$

Scanned with CamScanner

E(survived, age)

		1		2																	
Cheatsheet		Yes		Yes		No		Yes		No		No		No		Yes		No		Yes	
Age		8		12		20		24		26		32		34		43		45		47	
		≤ 6 <		≤ 10 <		≤ 16 <		≤ 22 <		≤ 25 <		≤ 29 <		≤ 33 <		≤ 39 <		≤ 41 <		≤ 45 <	
Yes	0	5	1	4	2	3	2	3	3	2	3	2	3	2	3	2	3	1	4	1	4	1
No	0	7	0	7	0	7	1	6	1	6	2	5	3	4	4	3	5	2	5	2	6	1
Entropy		.970		.86		.73		.91		.92		.908		.95							
Gain		0		.11		.24		.062		.16		.072									

Cheatsheet		Yes		No		Yes		No		No		No		Yes		No		Yes		N	
Age		55		60																	
		≤ 49 <		≤ 70 <																	
Yes	5	0	5	0																	
No	6	1	7	0																	
Entropy																					
Gain																					

Calculations :-

$$E(\text{<}) = \left(\frac{0}{12} * \left(-\frac{0}{0}\log_2\left(\frac{0}{0}\right) - \frac{12}{12}\log_2\left(\frac{0}{0}\right)\right) + \frac{12}{12}\left(-\frac{5}{12}\log_2\left(\frac{5}{12}\right) - \frac{7}{12}\log_2\left(\frac{7}{12}\right)\right)$$

$$= 0$$

$$E(8<) = \left(\frac{1}{12}\left(-\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right)\right) + \frac{11}{12}\left(-\frac{4}{11}\log\left(\frac{4}{11}\right) - \frac{7}{12}\log_2\left(\frac{7}{11}\right)\right)$$

$$= .86$$

$$E(16<) = \left(\frac{2}{12}\left(-\frac{2}{2}\log\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right)\right) + \frac{10}{12}\left(-\frac{3}{10}\log\left(\frac{3}{11}\right) - \frac{7}{10}\log\left(\frac{7}{10}\right)\right)$$

$$= .73$$

$$E(20<) = \left(\frac{3}{12}\left(-\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right)\right)\right) + \frac{9}{12}\left(-\frac{3}{9}\log\left(\frac{3}{9}\right) - \frac{6}{9}\log\left(\frac{6}{9}\right)\right)$$

$$E(24<) = \left(\frac{4}{12}\left(-\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right)\right)\right) + \frac{8}{12}\left(-\frac{2}{8}\log\left(\frac{2}{8}\right) - \frac{6}{8}\log\left(\frac{6}{8}\right)\right)$$

$$= .91$$

Scanned with CamScanner

E(survived, age)

		1		2									
Cheat sheet	Yes	Yes	No	Yes	No	No	No	No	Yes	No	Yes	No	
Age	8	12	20	24	26	32	34	43	45	47	55	65	
	≤ 6 & <	≤ 10 & <	≤ 16 & <	≤ 22 & <	≤ 25 & <	≤ 28 & <	≤ 33 & <	≤ 39 & <	≤ 44 & <	≤ 46 & <	≤ 51 & <	≤ 60 & <	≤ 70 & <
Yes	0 5	1 4	2 3	2 3	3 2	3 2	3 2	3 2	3 2	4 1	4 1	5 0	5 0
NO	0 7	0 7	0 7	1 6	1 6	2 5	3 4	4 3	5 2	5 2	6 1	6 1	7 0
Entropy	.772	.86	.73	.91	.92	.908	.95						
Gain	0	.11	.24	.062	.16	.072							

Calculations :-

$$E(\text{Age} < 6) = \left(\frac{0}{12} * \left(-\frac{0}{0}\log_2\left(\frac{0}{0}\right) - \frac{12}{12}\log_2\left(\frac{0}{0}\right)\right) + \frac{12}{12}\left(-\frac{5}{12}\log_2\left(\frac{5}{12}\right) - \frac{7}{12}\log_2\left(\frac{7}{12}\right)\right)\right)$$

$$= 0$$

$$E(\text{Age} < 8) = \left(\frac{1}{12}\left(-\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right)\right) + \frac{11}{12}\left(-\frac{4}{11}\log\left(\frac{4}{11}\right) - \frac{7}{12}\log_2\left(\frac{7}{11}\right)\right)$$

$$= .86$$

$$E(\text{Age} < 16) = \left(\frac{2}{12}\left(-\frac{2}{2}\log\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right)\right) + \frac{10}{12}\left(-\frac{3}{10}\log\left(\frac{3}{11}\right) - \frac{7}{10}\log\left(\frac{7}{10}\right)\right)$$

$$= .73$$

$$E(\text{Age} < 20) = \left(\frac{3}{12}\left(-\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right)\right) + \frac{9}{12}\left(-\frac{3}{9}\log\left(\frac{3}{9}\right) - \frac{6}{9}\log\left(\frac{6}{9}\right)\right)\right)$$

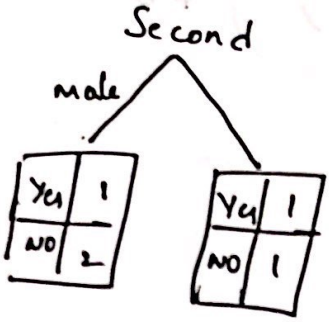
$$E(\text{Age} < 24) = \left(\frac{4}{12}\left(-\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right)\right)\right) + \frac{8}{12}\left(-\frac{2}{8}\log\left(\frac{2}{8}\right) - \frac{6}{8}\log\left(\frac{6}{8}\right)\right)$$

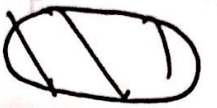
$$= .91$$

Scanned with CamScanner

Performing Split again
on second level of Main tree

$E(p)$





Reduction $= \frac{3}{5}\left(\frac{-1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)\right) + \frac{2}{5}\left(\frac{-1}{2}\log\left(\frac{1}{2}\right) + \frac{-1}{2}\log\left(\frac{1}{2}\right)\right)$

$= .970$

Scanned with CamScanner

E(survived, Gender) $= \frac{2}{4}\left(-\frac{1}{2}\log\left(\frac{1}{2}*\frac{1}{2}\right)\right) + 0$

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

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

$= .5$

Information gain $= .811 - .5$

$= .311$

Age

Cheatsheet		Yes		Yes		NO		Yes	
				12		20		24	
		≤ 60 <		≤ 16 <		≤ 22 <		≤ 26 <	
Yes	0	.5	1	.4	2	3	2	1	3	0
NO	0	2	0	2	0	7	1	0	1	0
Entropy										
Egain	0		.32		.3		.12		0	

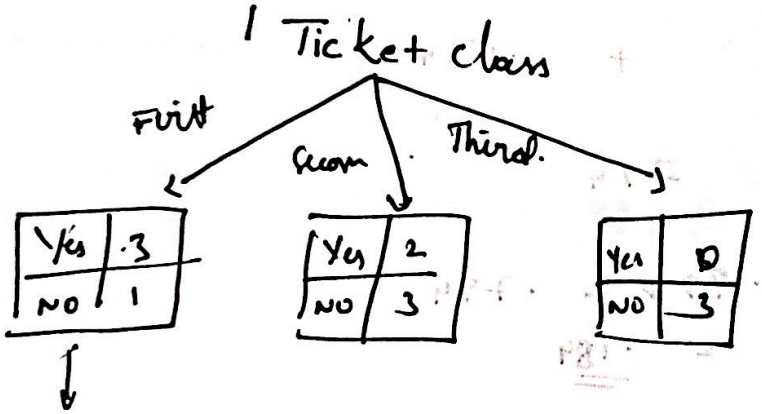
Scanned with CamScanner

Information gain = .978 − .67

= .309

~~we to consider option (B) as we are getting higher Information gain.~~

Now ~~consider~~ Ticket class is the Parent. as it has given highest information among the dependent variables

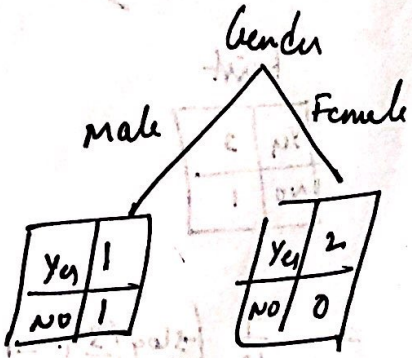


$$\text{Entropy(first)} = \left(-\frac{3}{4}\log(3/4) - \frac{1}{4}\log(1/4)\right)$$

.811

~~Entropy (Second)~~

Take gender First Part



$$E(\text{Survived, Gender}) = \frac{2}{4}\left(-\frac{1}{2}\log(1/2) - \frac{1}{2}\log(1/2)\right) + \frac{2}{4}\left(-\frac{2}{2}\log(2/2) - 0\right)$$

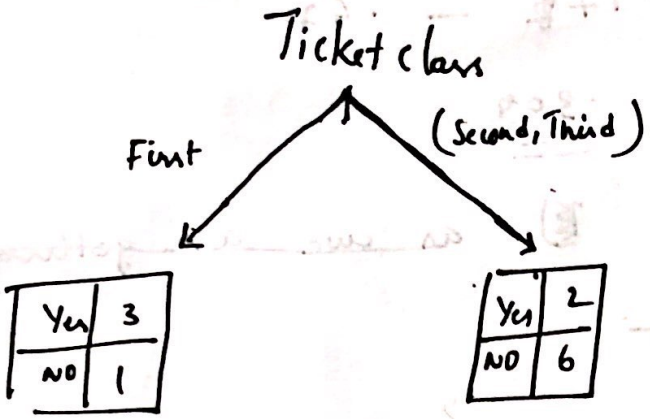
=

Scanned with CamScanner

E (survived, Ticket class)

option

(a)



$$E = \frac{4}{12}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) + \frac{8}{12}\left(-\frac{2}{8}\log_2\left(\frac{2}{8}\right) - \frac{6}{8}\log_2\left(\frac{6}{8}\right)\right)$$

$$= .2688 + \frac{1}{3}(+1.5) + (.3025)$$

$$= .26 + .534$$

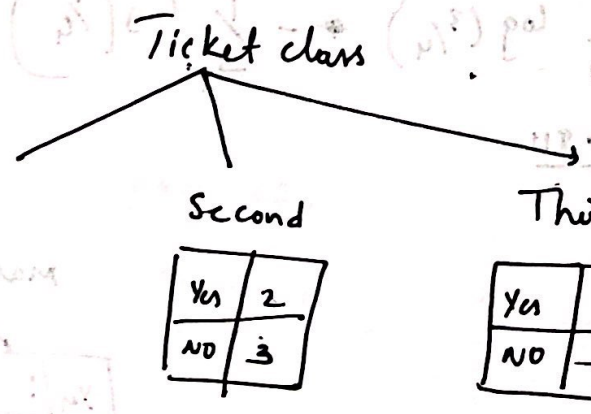
$$= .794$$

Information gain = $.978 - .794$

$= .184$

option

(b)



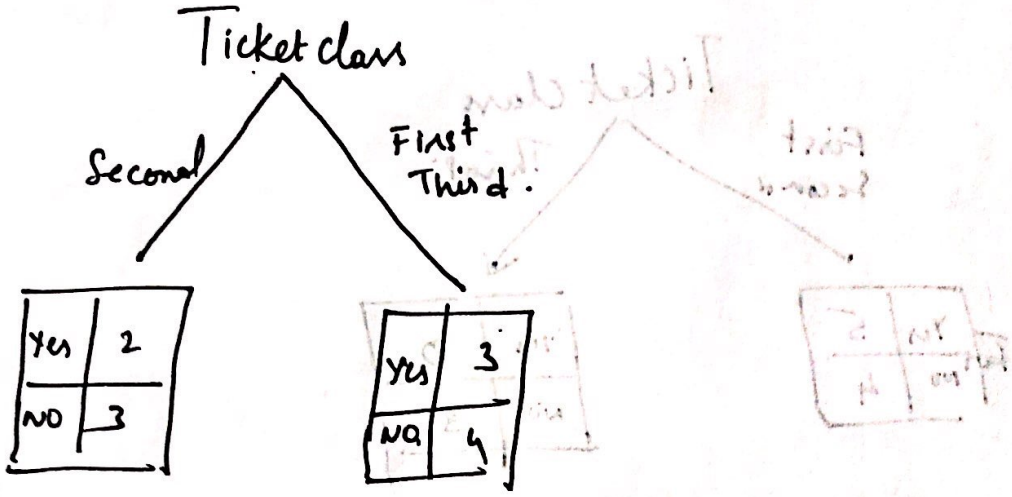
$$E = \frac{4}{12}\left(-\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right)\right) + \frac{5}{12}\left(-\frac{2}{5}\log\left(\frac{2}{5}\right) + -\frac{3}{5}\log\left(\frac{3}{5}\right)\right) + \left(0 * -\frac{3}{3}\log\frac{3}{3}\right)$$

$\searrow 0$

$= \frac{4}{12} * .62$

Scanned with CamScanner

E (survived, Ticket Class)



E (survived, Ticket Class) =

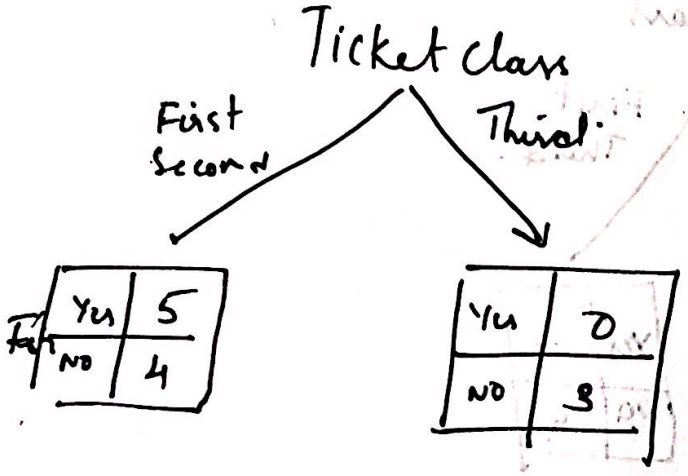
$$\frac{5}{12}\left(-\frac{2}{5}\log\left(\frac{2}{5}\right)-\frac{3}{5}\log\left(\frac{3}{5}\right)\right)+\frac{7}{12}\left(-\frac{3}{7}\log\left(\frac{3}{7}\right)+-\frac{4}{7}\log\left(\frac{4}{7}\right)\right)$$

$$= .973$$

Information gain = .978 − .973

= .005

E (Survived, Ticket Class)



E (survived, Ticketclass)

$$= \frac{9}{12}\left(-\frac{5}{9}\log_2\left(\frac{5}{9}\right) - \frac{4}{9}\log_2\left(\frac{4}{9}\right)\right) + \frac{3}{12}\left(0 - \frac{3}{3}\log_2\left(\frac{3}{3}\right)\right)$$

$$= \frac{9}{12}\left(.4711 + .52\right)$$

$$(1.04)\times\frac{9}{12}$$

$$= .78$$

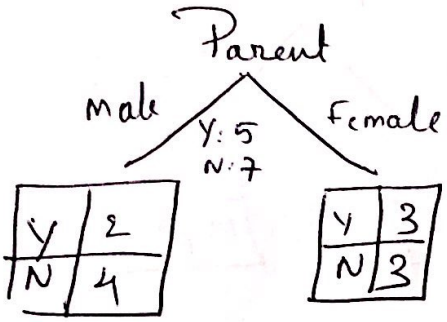
Information gain = $.98 - .78$

$$= \underline{\underline{.198}}$$

Scanned with CamScanner

1.2

Splitting Using GINI Index)



$$GINI(T) = 1 - \sum_j [P(j/T)]^2$$

$$G(\text{male}) = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = .44$$

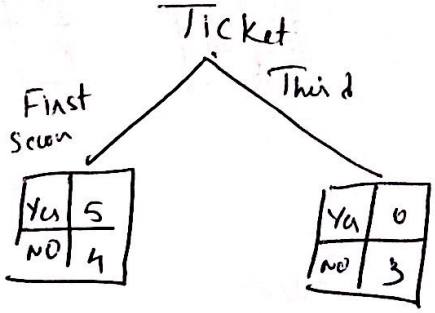
$$G(\text{female}) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = .5$$

$$GINI(\text{Parent}) = \frac{6}{12}(.44) + \frac{6}{12}(.5) = .47$$

$$1 - \left(\frac{5}{12}\right)^2 - \left(\frac{7}{12}\right)^2 = .486$$

Ticket class

option ①



$$G(\text{First, Second}) = 1 - \left(\frac{5}{9}\right)^2 - \left(\frac{4}{9}\right)^2 = .49$$

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

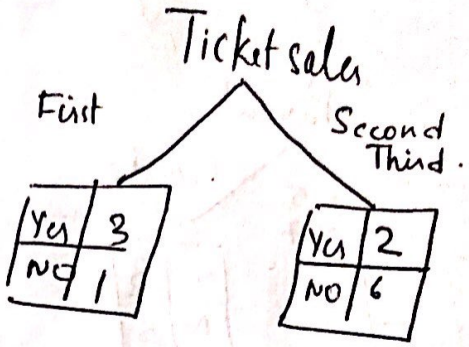
$$\frac{9}{12}(.49) + \frac{3}{12}(0) = .37$$

Scanned with CamScanner

option (2)

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

$= .375$



$G(\text{Second, Third}) = 1 - \left(\frac{2}{8}\right)^2 - \left(\frac{6}{8}\right)^2$

$= .375$

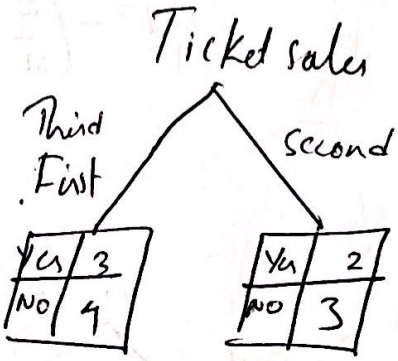
$G(\text{First})G + G(\text{Second, Third}) = \frac{4}{12}(.375) + \frac{8}{12}(.375)$

$= .375$

option (3) $G(\text{First, Third})$

$= 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2$

$= .48$



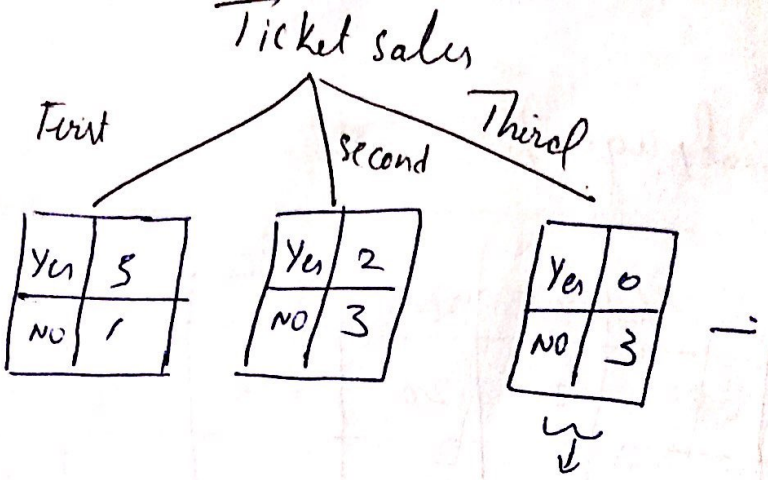
$G(\text{Second}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = .48$

$G(\text{First, Third}) = \frac{7}{12}(.48) + \frac{5}{12}(.48)$

$+ G(\text{Second})$

$= .40$

Scanned with CamScanner



Stopping condition

$$\text{Gini(Ticket Class)} = \frac{4}{12}\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right) + \frac{5}{12}\left(1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2\right)$$

$$+\frac{3}{12}\left(1-\left(\frac{0}{3}\right)^2-\left(\frac{3}{3}\right)^2\right)$$

$$= \frac{4}{12}(.375)+\frac{5}{12}(.48) = .325$$

Scanned with CamScanner

Cheat Ex

→ G(Survived, age)

Cheat sheet			Yes	Yes	NO	Yes	NO	NO	NO	NO	Yes	No	
A age			8	12	20	24	26	32	34	43	up 5	47	
	≤6<		≤10<	≤16<	≤22<	≤25<	≤28<	≤33<	≤38<	≤44<	≤46<	≤51<	≤
Yes	0	5	1	4	2	3	2	3	3	2	3	2	3	2	3	2	3	2	4	1	4	1	5
NO	0	7	0	7	0	7	1	6	1	6	2	5	3	4	4	3	5	2	5	2	6	1	6
Gini		.406		.42		.35		.44		.37	.43		.44		.48		.47		.481		.485		

↓ Least gini.

→ Least gini from all the attributes belongs to Ticket class

So we make Ticket class as the Parent node.

Next

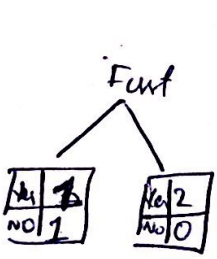
- Gender

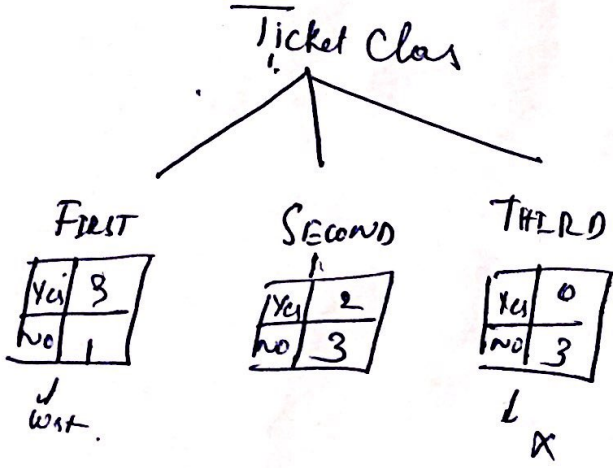
$Gini(male) = 1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2$

$= 0.5$

$Gini(female) = 1-[(1)^2+0^2] = 0$

Gain here $Gini(Gender) = \frac{2}{4}(.5) = .025$





Worst.

Scanned with CamScanner

G(Age)

(Cheat sheet) Survived	Y		Y		N		Y			
Age	8		12		x		24			
Split	≤4 <		≤10 <		≤16 <		≤22 <		≤25 <	
Yes	0	3	1	2	2	1	2	1	3	0
NO	0	1	0	1	0	1	1	0	1	0
Gini	.37		.33		.25		.33		.37	

↓

Least Gini

Since we got the gini index equal for both Age and gender we can Choose any of the following and perform split by using age

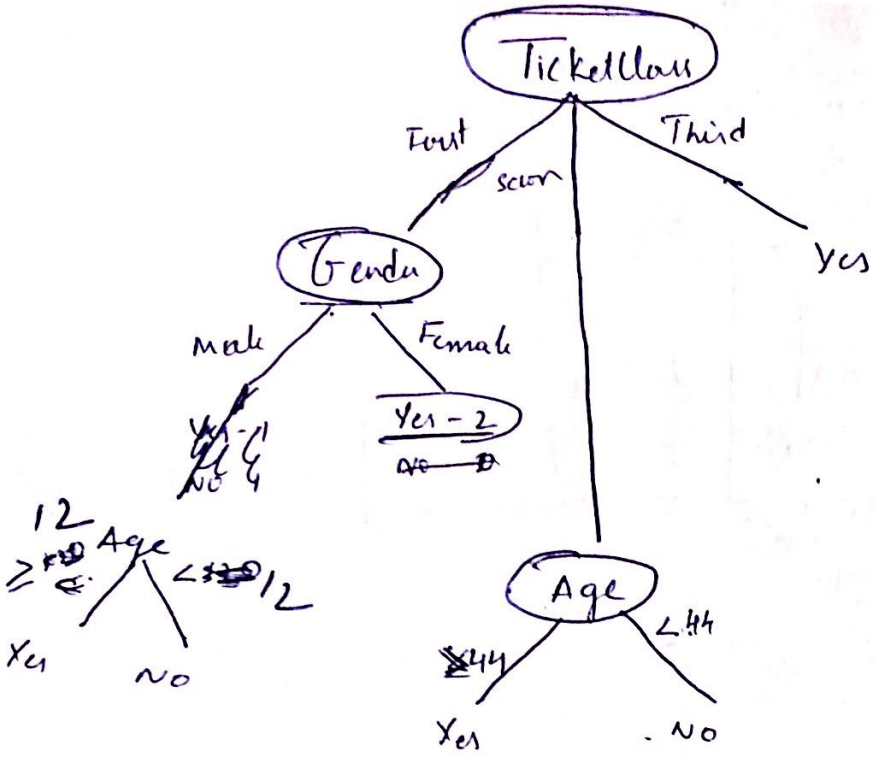
Cheat sheet	NO		NO		NO		Yes		Yes	
	32		34		43		45		55	
	≤30 <		≤33 <		≤38 <		≤44 <		≤50 <	≤60≤65
Yes	0	2	0	2	0	2	0	2	1	1	2	0
NO	0	3	1	2	2	1	3	0	3	0	3	0
Gini	.48		.4		.25		0		.3			

↓

least .

Scanned with CamScanner

Result



Scanned with CamScanner

1.3) Error given to each node = .5

$$e(T) = \sum_i \frac{[e(t_i) + \Omega(t_i)]}{\sum n(t_i)}.$$

$\sum_{i=1}^{n} t_i = T$

$$e'(T) = \frac{e(T) + \Omega(T)}{n(T)}$$

$e(T)$ = classification Error

n = total number of records

$\Omega(t) = 0.4$

$$= \frac{1 + (0.5)(5)}{12} = \frac{5.5}{12}$$

$$e'(T) = \underline{\underline{.29}}$$

Scanned with CamScanner