Name - Prayag Nikul Purani

SJSU ID - 017416737

(1)

Home-work - 2

Q1 Ch4, Ex - 3

a) Entropy of data set.

$$H(\text{Annual Income}, D) = -\sum_{n} P(x) \log_2(P_x)$$

Annual Income ← 5 (25k - 50k)
2 (< 25k)
1 (> 50k)

$$H(AI, D) = -\left(\frac{5}{8}\log_2\left(\frac{5}{8}\right)\right) + -\left(\frac{2}{8}\log_2\left(\frac{2}{8}\right)\right) + -\left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right)$$

$$H(AI, D) = -(-0.5 - 0.45 - 0.375)$$

$$= 1.299 \text{ bits}$$

b) Gini Index (Annual Index) =

$$1 - \sum P(x)^2$$

$$1 - \left[\left(\frac{5}{8}\right)^2 + \left(\frac{2}{8}\right)^2 + \left(\frac{1}{8}\right)^2 \right]$$

$$1 - (0.0625 + 0.3906 + 0.015625)$$

$$\simeq 0.5312$$

Gini Index = 0.5312

c) So, first we need to sort the data set

ID	Age	AI	
3	18	≤25k	
6	24	≤25k	} 26
4	28	25k-50k	
5	37	25k-50k	
1	39	25k-50k	} 39.5
8	40	≥50k	} 45
2	50	25k-50k	
7	52	25k-50k	

Boundary b/6 ID 446 is $\frac{24+28}{2} = 26$

Boundary b/6 ID 148 is $\frac{39+40}{2} = 39.5$

Boundary b/6 ID 842 = $\frac{40+50}{2} = 45$

H for data set = 1.299 bits

Age ← 26, 39.5, 45

→ Age ⟹ 26

$$rem(d, D) = \sum \underbrace{\frac{|Dd=i|}{|D|}}_{\text{weighting}} \times \underbrace{H(t, Dd=i)}_{\text{Entropy of position}}$$

$$= \frac{2}{8}(H(<26) + \frac{6}{8}(H(>26))$$

$$= 0 + \frac{3}{4}\left[\frac{5}{6}\log_2\left(\frac{5}{6}\right) + -\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right]$$

$$= 0 + \frac{3}{4}[0.64]$$

$$= 0.48$$

$$IG = H(t, D) - rem(a, D)$$

$$= 1.3 - 0.48$$

$$= 0.82$$

Now, for Age $\Rightarrow$ 39.5

$$rem(a, D) = 0.94$$

$$IG = H(t, D) - rem(a, D)$$

$$= 1.3 - 0.94$$

$$= 0.35$$

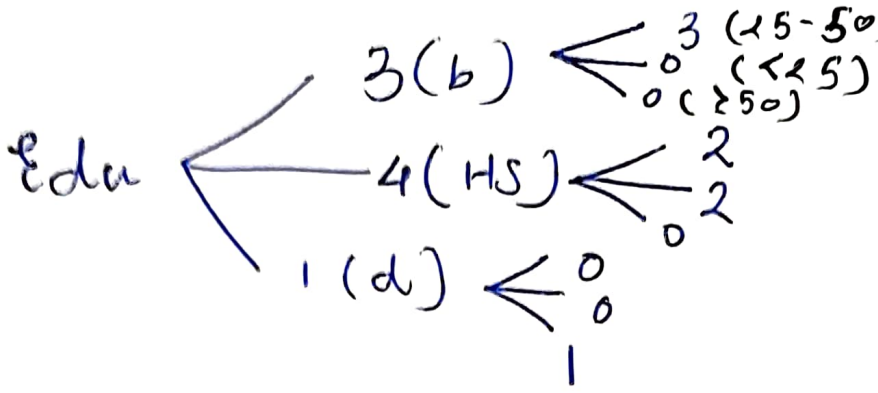
Now, for Age $\Rightarrow$ 45

$$rem(a, D) = 1.095$$

$$IG = 1.3 - 1.095$$

$$= 0.215$$

d)



$$IG(t, D) = H(t, d) - rem(d, D)$$

$$rem(Ed, D) =$$

$$\frac{3}{8}\left[-\left(\frac{3}{3}\log_2\frac{3}{3}\right)-\left(\frac{0}{3}\log_2\frac{0}{3}\right)-\left(\frac{0}{3}\log_2\frac{0}{3}\right)\right]$$

$$+\frac{4}{8}\left[-\left(\frac{2}{4}\log_2\frac{2}{4}\right)-\left(\frac{2}{4}\log_2\frac{2}{4}\right)-\left(\frac{0}{4}\log_2\frac{0}{4}\right)\right]$$

$$+\frac{1}{8}\left[-\left(\frac{1}{1}\log_2\frac{1}{1}\right)-\left(\frac{0}{1}\log_2\frac{0}{1}\right)-\left(\frac{0}{1}\log_2\frac{0}{1}\right)\right]$$

$$= 0.5$$

$$IG = 1.3 - 0.5 = 0.8$$

Information gain for Education = 0.8

from (Occupation)

$$\Rightarrow \frac{2}{8}\left[-\left(\frac{2}{2}\log_2\frac{2}{2}\right)+0+0\right]+$$

$$\Rightarrow \frac{3}{8}\left[-\left(\frac{2}{3}\log_2\frac{2}{3}\right)+\left(\frac{1}{3}\log_2\frac{1}{3}\right)+0\right]+$$

$$\Rightarrow \frac{2}{8}\left[-\left(\frac{1}{2}\log_2\frac{1}{2}\right)-\left(\frac{1}{2}\log_2\frac{1}{2}\right)+0\right]+$$

$$\frac{1}{8}\left[-\left(\frac{1}{1}\log_2\frac{1}{1}\right)-\left(\frac{0}{1}\log_2\frac{0}{1}\right)-\left(\frac{0}{1}\log_2\frac{0}{1}\right)\right]$$

$$= 0.5944$$

$$IG = 1.3 - 0.5944$$

$$= \underline{0.7056}$$

IG for Occupation is 0.7056

$$\text{Rem(Marital Status)} =$$

$$\frac{3}{8}\left[-\left(\frac{1}{3}\log_2\frac{1}{3}\right)-\left(\frac{2}{3}\log_2\frac{2}{3}\right)-0\right]+$$

$$\frac{4}{8}\left[-\left(\frac{3}{4}\log_2\frac{3}{4}\right)-\left(\frac{1}{4}\log_2\frac{1}{4}\right)-0\right]+$$

$$\frac{1}{8}\left[-\left(\frac{1}{1}\log_2\frac{1}{1}\right)-0-0\right]$$

$$= 0.75$$

$$IG = 1.3 - 0.75$$

$$= \underline{0.55}$$

IG for Marital Status = 0.55

e) $IG(Ed, D) = 0.8$

$IG(MS, D) = 0.55$

$IG(Occ, D) = 0.7$

entropy Ed = 1.4 bits

entropy MS = 1.4 bits

entropy Occ = 1.9 bits

$$IGR = \frac{IG}{H}$$

$$IGR(Ed, D) = \frac{0.8}{1.4} = 0.567$$

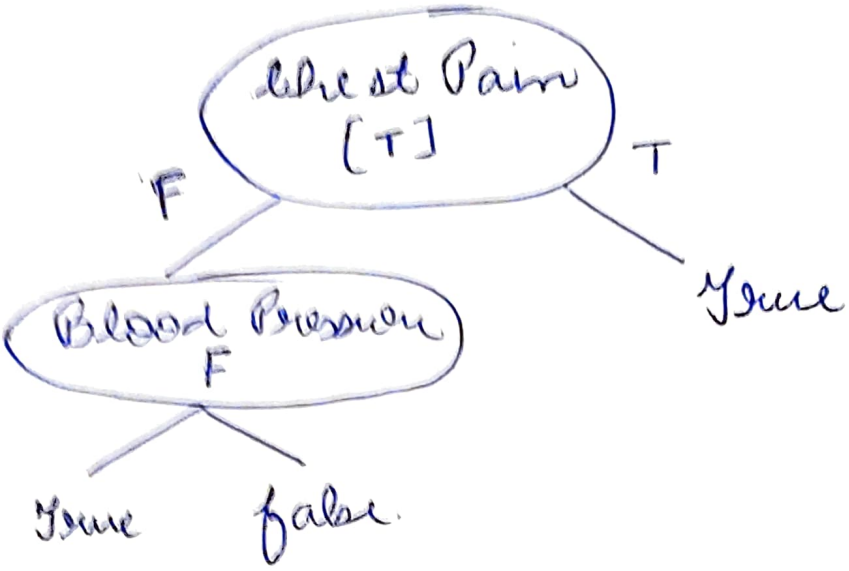
$$IGR(MS, D) = \frac{0.55}{1.4} = 0.39$$

$$IGR(Occ, D) = \frac{0.705}{1.9} = 0.36$$

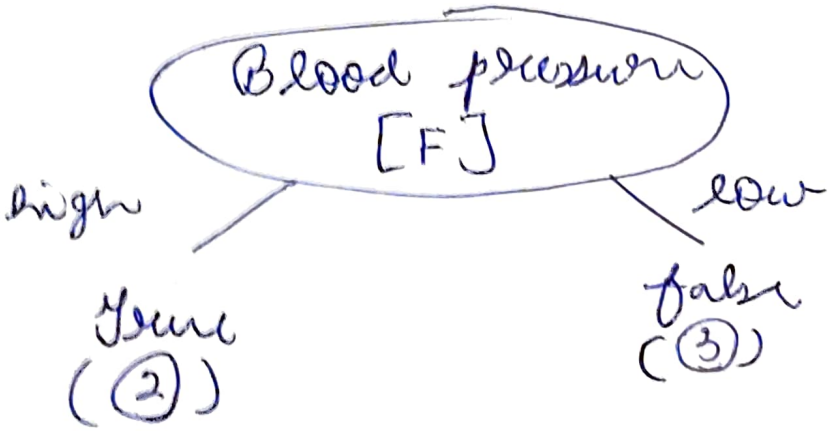
f)

feature	Level	Partition GI	Bem	IG
Education	HS	0.5	0.25	0.2513
	bachelors	0		
	doctorate	0		
Matrial Status	NA	0.45	0.3542	0.1771
	M	0.375		
	D	0		
Occupation	T	0	0.2917	0.2396
	P	0.45		
	A	0.5		
	AF	0		

Ch-4, Exercise-4



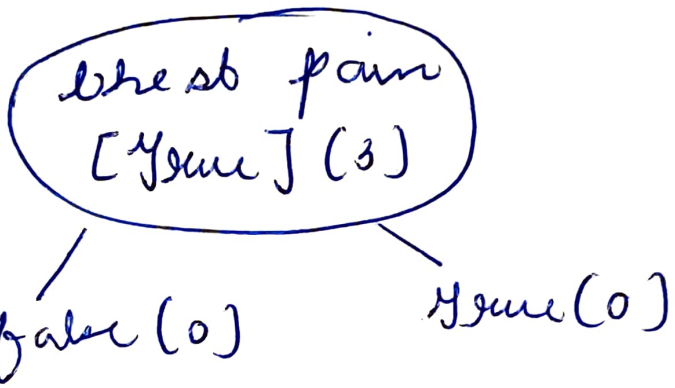
Given, decision tree to predict heart disease, blood pressure & chest pain are descriptive feature & heart disease is the target feature.



- The number in round brackets represent no. of error in pruning set of that node.

- The number in square bracket represent the majority target level from that node.

- The number of error coming from leaf node is greater than that of the number of error coming from the root node.



2nd iteration for pruning root node predicts Yes which causes 3 mistake our pruning set the algorithm will stop here.

Thus the final decision tree is produced.

Q3 Ch-4, Ex-5

a) $H(risk) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$

$= 0.72$ bits

Sample A

$Rem(Ex) = \frac{1}{5}\left[-\left(\frac{1}{1}\log_2\frac{1}{1}\right)\right] + \frac{2}{5}\left[-\frac{2}{2}\log_2\frac{2}{2}\right] + \frac{2}{5}\left[-\frac{2}{2}\log\frac{2}{2} + 0\right]$

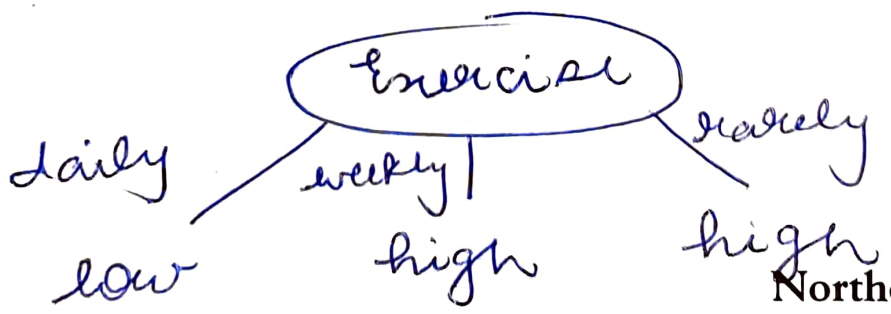
$= 0$

$Rem(family) = \frac{3}{5}\left[-\left(\frac{1}{3}\log_2\frac{1}{3}\right) - \left(\frac{2}{3}\log_2\frac{2}{3}\right)\right] + \frac{2}{5}\left[-\left(\frac{2}{2}\log_2\frac{2}{2}\right) + 0\right]$

$= 0.55$

$IG(Ex) = 0.72 - 0 = 0.72$ ✓

$IG(family) = 0.72 - 0.55 = 0.17$.



- as IG of Exercise is highest so we will choose that

for sample B

Entropy = 0.72 bits

$$\text{rem(smoker)} = \frac{1}{5}\left[\left(-\frac{1}{1}\log_2\frac{1}{1}\right)+0\right] + \frac{4}{4}\left[\left(-\frac{4}{4}\log_2\frac{4}{4}\right)\right]$$

$$= 0$$

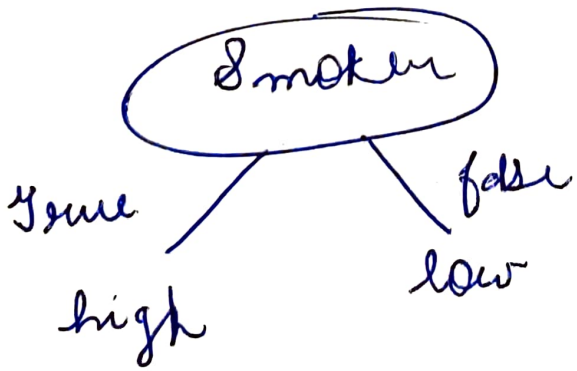
$$\text{rem(obese)} = \frac{3}{5}\left[\left(-\frac{1}{3}\log_2\frac{1}{3}\right)+\left(-\frac{2}{3}\log_2\frac{2}{3}\right)\right] + \frac{2}{3}\left[\left(-\frac{2}{2}\log_2\frac{2}{2}\right)+\left(-\frac{0}{2}\log_2\frac{0}{2}\right)\right]$$

$$= 0.55$$

IG(smoker) = 0.721 − 0 = 0.721 ✓

IG(obese) = 0.721 − 0.55 = 0.17

Smoker has higher IG



for sample C

$$\text{Entropy} = \left(-\frac{2}{5}\log_2\frac{2}{5}\right) - \left(\frac{3}{5}\log_2\frac{3}{5}\right)$$

$$= 0.971 \text{ bits}$$

$$\text{Rem (Class)} = \frac{3}{5}\left(\frac{-2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)$$

$$+ \frac{2}{5}\left(\frac{-1}{2}\log_2\frac{1}{2} - \frac{0}{2}\log_2\frac{0}{2}\right)$$

$$= 0.55$$

$$\text{Rem (family)} = \frac{4}{5}\left[\left(\frac{-2}{4}\log_2\frac{2}{4}\right) + \left(\frac{2}{4}\log_2\frac{2}{4}\right)\right]$$

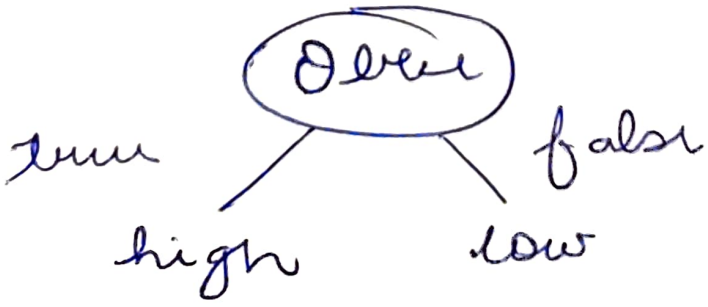
$$+ \frac{1}{5}\left[\left(\frac{-1}{1}\log_2\frac{1}{1}\right)\right]$$

$$= 0.8$$

$$IG(\text{Class}) = 0.971 - 0.55 = 0.42$$

$$IG(\text{family}) = 0.971 - 0.8 = 0.17$$

Hence, Class has higher IG



b) Exercise = rarely
smoker = false
obese = true
family = yes

Tree 1

Exercise → rarely → risk → high

Tree 2

Smoker → false → risk → low

Tree 3

Obses → true → risk → high

So, majority value for risk is high so it will predict high as the output.

Q4 Ch-4 Ex-6

a) $H(\text{risk}) = \left(-\frac{3}{6}\log_2\frac{3}{6}\right) + \left(-\frac{3}{6}\log_2\frac{3}{6}\right)$

$= 1$ bit.

$\text{rem}(\text{chest}) = 2 \times \frac{3}{6} \times \left[\left(-\frac{1}{3}\log_2\frac{1}{3}\right) + \left(-\frac{2}{3}\log_2\frac{2}{3}\right)\right]$

$= 0.9183$

$\text{rem}(\text{smoker}) = \frac{2}{6}\left[\left(-\frac{2}{2}\log_2\frac{2}{2}\right) + 0\right] + \frac{4}{6}\left[\left(-\frac{3}{4}\log_2\frac{3}{4}\right) + \left(-\frac{1}{4}\log_2\frac{1}{4}\right)\right]$

$= 0.54$

$\text{rem}(\text{Drinks alcohol}) = \frac{5}{6}\left[\left(-\frac{2}{5}\log_2\frac{2}{5}\right) + \left(-\frac{3}{5}\log_2\frac{3}{5}\right)\right] + \frac{1}{6}\left[\left(-\frac{1}{1}\log_2\frac{1}{1}\right)\right]$

$= 0.809$

IG(Obese) = 1 − 0.9183 = 0.0817

IG(Smoker) = 1 − 0.54 = 0.46 ✓

IG(Drink alcohol) = 1 − 0.809 = 0.191

Smoker has the highest IG so it will be the root node according ID3 algorithm.

b) Other feature which can be more descriptive feature such as workout, skincare & medication, can also be added to the dataset that can possibly indicate the low target level.

85 Ch-4, En-7

a) $H(Buys, D) = \left[-\frac{9}{14}\log_2\frac{9}{14}\right] + \left[-\frac{5}{14}\log_2\frac{5}{14}\right]$

$= 0.9403$

$Rem(Student) = \frac{7}{14}\left[-\left(\frac{4}{7}\log_2\frac{4}{7}\right) - \left(\frac{3}{7}\log_2\frac{3}{7}\right)\right]$

$= 0.788$

$IG(student) = 0.9397 - 0.788$

$= 0.1517$

$IG(student) < IG(age)$,

so is not a better option.

b) for ID

all the values in ID column is unique & different

so $rem(ID) = 0$

hence, it is not a good idea to use ID as a root as it will results in overfitting of the model