Course Number: CSCE 5380 Name: Vishnu Vardhan Kaitepalli.
UNT ID: 11647518.

Homework No. 2.

Q1)

a)

$E = [5+, 5-]$

$\therefore$ Entropy $(E) = -\sum_{i=1}^{n} P_i \log_2 (P_i)$

$\therefore$ Entropy $(E) = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10}$

$\therefore$ Entropy $(E) = 1$

Attribute: Tobacco smoking.

values $\rightarrow$ [yes, No]

$S_{yes} \leftarrow [+4, -1]$ $\therefore$ Entropy $(S_{yes}) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2 1/5$

$= 0.7219$

$\therefore S_{No} \leftarrow [+1, -4]$ $\therefore$ Entropy $(S_{No}) = -\frac{1}{5}\log_2 1/5 - \frac{4}{5}\log_2\frac{4}{5}$

$= 0.7219$

$\therefore$ Gain (S) = Entropy $(E) - \sum_{v \in V} \frac{|S_v|}{|S|}$ Entropy (S_v).

$$Gain(S, \text{Tobacco smoking}) = Entropy(E) - \sum_{v \in (yes,No)} \frac{|S_v|}{S} Entropy(S_v).$$

$$= Entropy(E) - \frac{5}{10} Entropy(S_{yes}) - \frac{5}{10} Entropy(S_{No}).$$

$$= 1 - \frac{5}{10}(0.7219) - \frac{5}{10}(0.7219).$$

$$= 0.2781$$

$\therefore Gain(S, \text{Tobacco smoking}) = 0.2781$

Attribute : Random Exposure.

Values $\rightarrow$ [yes, No]

$\therefore S_{yes} \leftarrow [+2, -0]$ $\therefore Entropy(S_{yes}) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}$

$= 0.$

$\therefore S_{No} \leftarrow [+3, -5]$ $\therefore Entropy(S_{No}) = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8}$

$= 0.954$

$\therefore Gain(S, Exposure) = Entropy(E) - \sum_{v \in (yes,No)} \frac{|S_v|}{S} Entropy(S_v)$

$$= 1 - \frac{2}{10}\,\text{Entropy}(S_{yes}) - \frac{8}{10}\,\text{Entropy}(S_{No})$$

$$= 1 - \frac{2}{10}(0) - \frac{8}{10}(0.954)$$

$$= 0.2368.$$

$\therefore$ Gain $(S, \text{Exposure}) = 0.2368.$

Attribute: Chronic cough

values $\rightarrow$ [yes, No]

$S_{yes} \leftarrow [+4, -3]$ $\quad \therefore \text{Entropy}(S_{yes}) = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7}$

$= 0.9852$

$S_{No} \leftarrow [+1, -2]$ $\quad \therefore \text{Entropy}(S_{No}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

$= 0.9182$

$$\therefore \text{Gain}(S, \text{cough}) = \text{Entropy}(E) - \sum_{v \in (yes, No)} \frac{|S_v|}{S}\,\text{Entropy}(S_v)$$

$$= \text{Entropy}(E) - \frac{7}{10}\,\text{Entropy}(S_{yes}) - \frac{3}{10}\,\text{Entropy}(S_{No})$$

$$= 1 - \frac{7}{10}(0.9852) - \frac{3}{10}(0.9182).$$

$$= 0.0349$$

$\therefore$ Gain $(S, \text{cough}) = 0.0349.$

Attribute : Weight loss

values → [yes, NO]

$\therefore S_{yes} \leftarrow [+3, -2]$ $\quad \therefore Entropy(S_{yes}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.9709.$

$\therefore S_{NO} \leftarrow [+2, -3]$ $\quad \therefore Entropy(S_{NO}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.9709.$

$\therefore Gain(S, Weight\ loss) = Entropy(E) - \sum_{v \in (yes, NO)} \frac{|S_v|}{S} Entropy(S_v)$

$= Entropy(E) - \frac{5}{10} Entropy(S_{yes}) - \frac{5}{10} Entropy(S_{NO})$

$= 1 - \frac{5}{10}(0.9709) - \frac{5}{10}(0.9709)$

$= 0.0291$

$\therefore Gain(S, loss) = 0.0291$

Here by Gain.

$\therefore$ Gain (S, Tobacco smoking) = 0.2781

Gain (S, Radon Exposure) = 0.2368

Gain (S, Chronic cough) = 0.0349

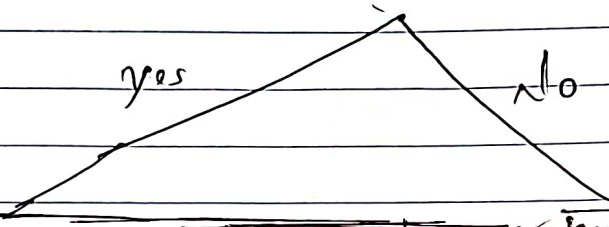
Gain (S, Weight loss) = 0.0291

∴ In the formation of decision Tree. We should Consider the Maximum Information Gain.

∴ So, here Tobacco smoking as a root.

→ As there is one 'No' for 'yes' in cancer. We have to repeat.

∴ Tobacco Smoking



Smoking	Exposure	cough	weight loss	cancer
yes	yes	yes	No	yes
yes	No	yes	No	yes
yes	No	yes	yes	yes
yes	No	yes	yes	yes
yes	No	No	No	No

Smoking	Exposure	cough	weight loss	cancer
No	Yes	No	yes	yes
No	No	yes	No	No
No	No	yes	yes	No
No	No	yes	No	No
No	No	No	yes	No

∴ For IN "YES"

∴ Entropy $(E) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.7219$

∴ Attribute: Exposure

$S_{yes} \leftarrow [+1, -0]$ ∴ Entropy $(S_{yes}) = -\frac{1}{1}\log_2\frac{1}{1} - 0$

$= 0$

$S_{No} \leftarrow [+3, -1]$ ∴ Entropy $(S_{No}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$

$= 0.8112$

$\therefore$ Gain(S, Exposure) $= 0.7219 - \frac{1}{5}(0) - \frac{4}{5}(0.8112)$

$\therefore$ Gain(S, Exposure) $= 0.07294$

Attribute: cough

$\therefore S_{yes} \leftarrow [+4, -0]$ $\therefore$ Entropy$(S_{yes}) = -\frac{4}{4}\log_2\frac{4}{4} - 0 = 0.$

$\therefore S_{NO} \leftarrow [+0, -1]$ $\therefore$ Entropy$(S_{NO}) = 0 - \frac{1}{1}\log_2\frac{1}{1} = 0.$

$\therefore$ Gain(S, cough) $= 0.7219 - 0 - 0$

$\therefore$ Gain(S, cough) $= 0.7219.$

Attribute: weight loss

$\therefore S_{yes} \leftarrow [+2, -0]$ $\therefore$ Entropy$(S_{yes}) = -\frac{2}{2}\log_2\frac{2}{2} - 0 = 0.$

$\therefore S_{NO} \leftarrow [+2, -1]$ $\therefore$ Entropy$(S_{NO}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$

$= 0.9182$

$\therefore$ Gain(S, weight loss) $= 0.7219 - \frac{2}{5}(0) - \frac{3}{5}(0.9182)$

$\therefore$ Gain(S, weight loss $= 0.17098.$

$\therefore$ Here

$\therefore$ Gain(S, Exposure) $= 0.07294$

$\therefore$ Gain(S, cough) $= 0.7219$

$\therefore$ Gain(S, weight loss) $= 0.17098.$ [$\therefore$ Maximum value of gain]

∴ IN "No"

∴ Entropy(E) $= -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.7219.$

Attribute: Exposure:

∴ $S_{yes} \leftarrow [+1, -0]$ ∴ Entropy$(S_{yes}) = 0.$

∴ $S_{No} \leftarrow [+0, -4]$ ∴ Entropy$(S_{No}) = 0.$

∴ Gain(S, Exposure) $= 0.7219.$

Attribute: cough

∴ $S_{yes} \leftarrow [+0, -3]$ ∴ Entropy$(S_{yes}) = 0.$

∴ $S_{No} \leftarrow [+1, -1]$ ∴ Entropy$(S_{No}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

∴ Gain(S, cough) $= 0.7219 - 0 - \frac{2}{5}(1) = 0.3219.$

Attribute: weight loss:

∴ $S_{yes} \leftarrow [+1, -2]$ ∴ Entropy$(S_{yes}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

$= 0.9182$

∴ $S_{No} \leftarrow [+0, -2]$ ∴ Entropy$(S_{No}) = 0.$

∴ Gain(S, loss) $= 0.7219 - \frac{3}{5}(0.9182) - 0.$

∴ Gain(S, weight loss) $= 0.1709$

∴ Here

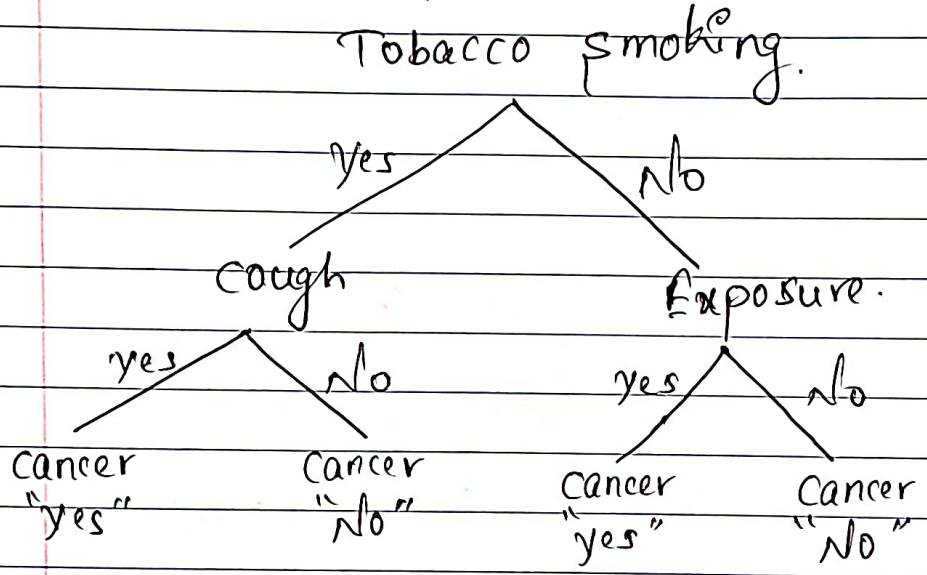
∴ Gain (S, Exposure) = 0.7219

∴ Gain (S, cough) = 0.3219

∴ Gain (S, weight loss) = 0.1709.

{ ∵ Maximum value of Gain. }

∴ The tree is.



Q2

c)

We first estimate prior probabilities for "status" class labels.

$\therefore P(\text{senior}) = 5/11$

and

$\therefore P(\text{junior}) = 6/11$

Now conditional probabilities.

$\therefore P(\text{department} / \text{status})$

class	sales	systems	marketing	secretary
senior	1/5	2/5	1/5	1/5
junior	2/6 = 1/3	2/6 = 1/3	1/6	1/6

$P(\text{age} / \text{status})$

class	21...25	26....30	31....35	36....40	41....45	46....50
senior	0/5 = 0	0/5 = 0	2/5	1/5	1/5	1/5
junior	1/6	3/6 = 1/2	2/6 = 1/3	0/6 = 0	0/6 = 0	0/6 = 0

$P(\text{salary}/\text{status})$

class	26k-30k	31k-35k	36k-40k	41k-45k	46k-50k	66k-70k
Senior	$0/5=0$	$0/5=0$	$1/5$	$0/5=0$	$2/5$	$2/5$
Junior	$2/6=1/3$	$1/6=0$	$0/6=0$	$1/6$	$2/6=1/3$	$0/6$

∴ So for Test instance.

$$V_{NB} = \underset{v_j \in [yes,no]}{\text{argmax}} \; P(v_j) \prod_i P(a_i/v_j)$$

∴ (systems, "26....30", "46k-50k")

Two labels → Senior, Junior.

$$\therefore P(\text{Senior}/a) = P(\text{senior}) * P(\text{systems}/\text{senior}) * P(26....30/\text{senior}) * P(46k-50k/\text{Senior})$$

$$= \frac{5}{11} * \frac{2}{5} * 0 * \frac{2}{5}$$

$$\therefore P(\text{Senior}/a) = 0$$

$$\therefore P(\text{Junior}/a) = P(\text{Junior}) * P(\text{systems}/\text{junior}) * P(26...30/\text{Junior}) * (P(46k-50k)/\text{Junior})$$

$$= \frac{6}{11} * \frac{1}{3} * \frac{1}{2} * \frac{1}{3}$$

$$\therefore P(\text{Junior}/a) = 0.030$$

$\therefore \boxed{P(\text{Junior}/a) = 0.030}$

∴ By comparing the two values, and by the maximum value was Junior.

∴ So.

Hence, the label for this instance (person) is "Junior"

∴ Since $P(\text{senior}/a) < P(\text{junior}/a)$.