

Assignment-1

Question-1

$$0.2093 = (0.37)3$$

Training	Fever	Vomiting	Diarrhea	Shivering	classification
D ₁	No	No	No	No	healthy (H)
D ₂	avg	No	No	No	influenza (I)
D ₃	high	No	No	Yes	influenza (I)
D ₄	high	Yes	Yes	No	S P (S)
D ₅	avg	No	Yes	No	S P (S)
D ₆	No	Yes	Yes	No	B I (B)
D ₇	avg	Yes	Yes	No	B I (B)

1.1 Calculate entropy of the dataset

$$E(S) = \sum_{i=1}^C -P_i \log_2 P_i$$

Total no. of classification - 7

for Healthy - 1 (H)

for Influenza - 2 (I)

for Salmonella poisoning - 2 (S)

for Bowel Inflammation - 2 (B)

Now By applying the formula of Entropy

$$E(D) = \sum_{i=1}^C -P_i \log_2 P_i, \text{ we get}$$

$$E(D) = - \left(\left(\frac{1}{7}\right) \log_2 \left(\frac{1}{7}\right) + \left(\frac{2}{7}\right) \log_2 \left(\frac{2}{7}\right) + \left(\frac{2}{7}\right) \log_2 \left(\frac{2}{7}\right) + \left(\frac{2}{7}\right) \log_2 \left(\frac{2}{7}\right) \right)$$

$$= 1.950212$$

Entropy of Dataset

1.2

calculate the information gain (IG) for each of the attribute

Total no. of attributes - 4

Fever, Vomiting, Diarrhea, Shivering

• first we need to calculate entropy of each attribute
and we'll also calculate weighted entropy with

$$\rightarrow \text{Entropy}(x, y) = \sum_{c \in y} P(c) E(c)$$

1 $E(\text{classification}, \text{Fever}) \rightarrow \text{no, avg, high}$

2 $E(\text{classification}, \text{Vomiting}) \rightarrow \text{no, yes}$

3 $E(\text{classification}, \text{diarrhea}) \rightarrow \text{no, yes}$

4 $E(\text{classification}, \text{Shivering}) \rightarrow \text{no, yes}$

① For Fever $E(\text{fever}) = P(\text{no}) E(\text{no}) + P(\text{avg}) E(\text{avg}) + P(\text{high}) E(\text{high})$

$$\text{for } E(\text{no}) = - \left(\left(\frac{1}{2} \right) \log_2 \left(\frac{1}{2} \right) + \left(\frac{1}{2} \right) \log_2 \left(\frac{1}{2} \right) \right)$$

$$= 1$$

$$\text{for } E(\text{avg}) = - \left(\left(\frac{1}{3} \right) \log_2 \left(\frac{1}{3} \right) + \left(\frac{1}{3} \right) \log_2 \left(\frac{1}{3} \right) + \left(\frac{1}{3} \right) \log_2 \left(\frac{1}{3} \right) \right)$$

$$= 1.585$$

$$\text{for } E(\text{high}) = - \left(\left(\frac{1}{2} \right) \log_2 \left(\frac{1}{2} \right) + \left(\frac{1}{2} \right) \log_2 \left(\frac{1}{2} \right) \right)$$

$$= 1$$

	H	I	class	S	B
fever	No	1	0	0	1
	avg	0	1	1	1
	high	0	1	1	0

For fever

$$\text{Entropy(class, Fever)} = \left(\frac{2}{7} \times 1 \right) + \left(\frac{3}{7} \times 1.585 \right) + \left(\frac{2}{7} \times 1 \right) = 1.2507$$

→ for E(vomiting) = P(no) E(no) + P(yes) E(yes)

$$E(\text{no}) = - \left(\frac{1}{4} \log_2 \left(\frac{1}{4} \right) + \frac{2}{4} \log_2 \left(\frac{2}{4} \right) + \frac{1}{4} \log_2 \left(\frac{1}{4} \right) \right) = 1.5$$

$$E(\text{yes}) = - \left(\frac{1}{3} \log_2 \left(\frac{1}{3} \right) + \frac{2}{3} \log_2 \left(\frac{2}{3} \right) \right) = 0.9183$$

	No	1	2	1	0	-4
Vomiting	Yes	0	0	1	2	-3

$$\text{Entropy(class, Vomit)} = \left(\frac{3}{7} \times 0.9183 \right) + \left(\frac{4}{7} \times 1.5 \right) = 1.2507$$

→ for E(Diarrhea) = P(no) E(no) + P(yes) E(yes)

$$E(\text{no}) = - \left(\frac{1}{3} \log_2 \left(\frac{1}{3} \right) + \frac{2}{3} \log_2 \left(\frac{2}{3} \right) \right) = 0.9183$$

$$E(\text{yes}) = - \left(\frac{2}{4} \log_2 \left(\frac{2}{4} \right) + \frac{2}{4} \log_2 \left(\frac{2}{4} \right) \right) = 1$$

Diarrhea	Yes	0	0.9183	-4
	No	1	2	0.9183 - 3

$$E(\text{class}, \text{Diarrhea}) = \left(\left(\frac{3}{7} \times 0.9183 \right) + \left(\frac{4}{7} \times 1 \right) \right)$$

$$= 0.9649$$

→ for $E(\text{Shivering}) = P(\text{no}) E(\text{no}) + P(\text{yes}) E(\text{yes})$

$$E(\text{no}) = - \left(\left(\frac{1}{6} \log_2 \left(\frac{1}{6} \right) \right) + \left(\frac{1}{6} \log_2 \left(\frac{1}{6} \right) \right) + \left(\frac{2}{6} \log_2 \left(\frac{2}{6} \right) \right) + \left(\frac{2}{6} \log_2 \left(\frac{2}{6} \right) \right) \right)$$

$$= 1.9183$$

$$E(\text{yes}) = - \left(\left(\frac{1}{1} \log_2 \left(\frac{1}{1} \right) \right) \cdot 1 \right) = 0$$

	H	I	S	B
Shivering	Yes	0	1	0
	No	1	1	2

$$\text{Entropy}(\text{class}, \text{Shivering}) = \left(\left(\frac{6}{7} \times 1.9183 \right) + \left(\frac{1}{7} \times 0 \right) \right)$$

$$= 1.6442$$

→ calculating Information Gain ($I(G)$) for each Attribute

$$\text{Gain}(x, y) = \text{Entropy}(x) - \text{Entropy}(x|y)$$

$$\rightarrow \text{Gain(class, fever)} = E(\text{class}) - E(\text{class, fever})$$

$$= 1.9502 - 1.2507 \quad \text{ask per burd} \\ = 0.6995$$

$$\rightarrow \text{Gain(class, vomiting)} = E(\text{class}) - E(\text{class, vomiting})$$

$$= 1.9502 - 1.2507 \\ = 0.6995$$

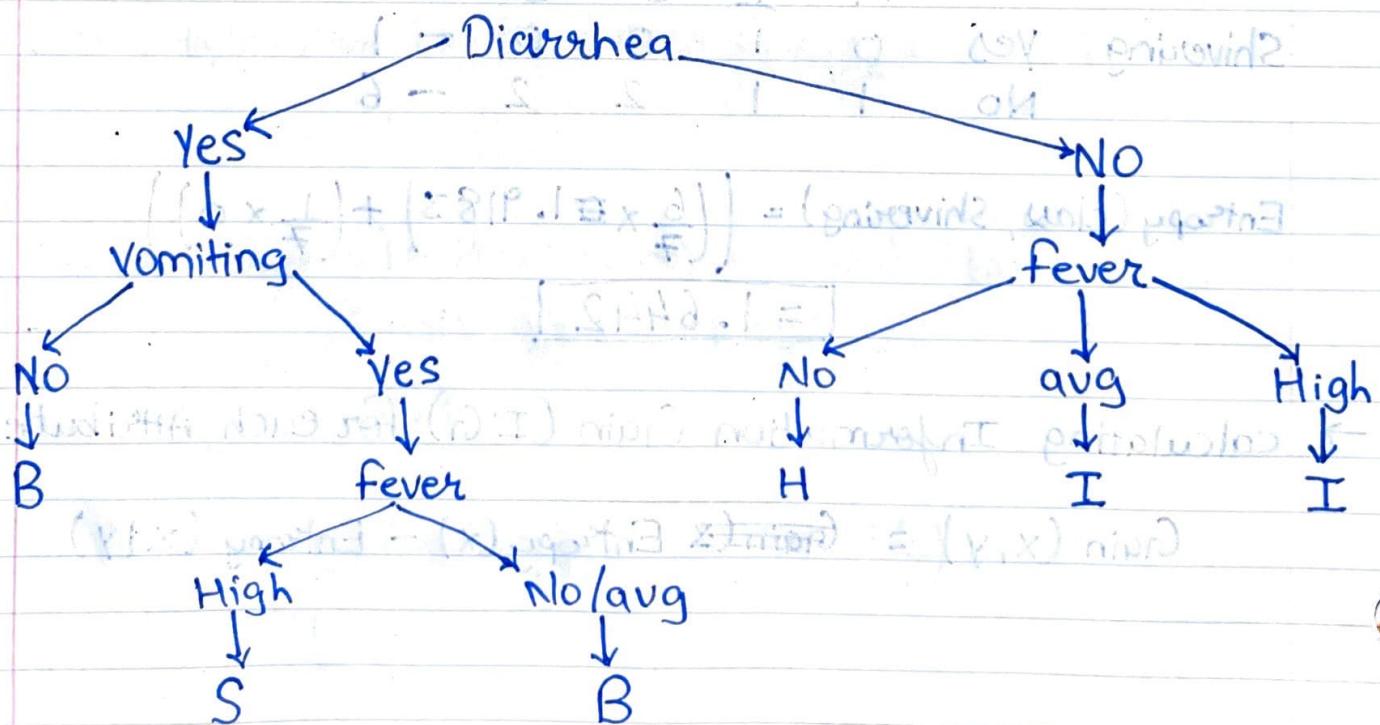
$$\rightarrow \text{Gain(class, Diarrhea)} = E(\text{class}) - E(\text{class, Diarrhea})$$

$$= 1.9502 - 0.9649 \\ = 0.9853$$

$$\rightarrow \text{Gain(class, Shivering)} = E(\text{class}) - E(\text{class, Shivering})$$

$$= 1.9502 - 1.6442 \\ = 0.3060$$

1.3 → Highest(I₆) is of Diarrhea, so that Diarrhea will be root node



Question 2

(a) $Z = \{x, y\}$ ~~given~~

Instance	a_1	a_2	a_3	Target class
1	T	T	71.0	-
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

2.1 → calculate the entropy of dataset

$$E(S) = - \sum_{i=1}^9 p_i \log_2 p_i$$

Total no. of target class / Instance = 9

for + = 4
for - = 5

By applying the formula, we get.

$$E(D) = - \left(\left(\frac{4}{9}\right) \log_2 \left(\frac{4}{9}\right) + \left(\frac{5}{9}\right) \log_2 \left(\frac{5}{9}\right) \right)$$

$$= 0.9910$$

Entropy of Dataset

2.2 calculate Information Gain (IG) of a_1 & a_2

$a_1 = T \ F$

$a_2 = T \ F$

We'll also calculate weighted entropy

$$\text{Entropy}(x, y) = \sum_{c \in y} P(c) E(c)$$

S. notes

$$1. E(\text{Target}, a_1) = T.F$$

$$2. F(\text{Target}, a_2) = T.F$$

$$E(a_1) = P(T) E(T) + P(F) E(F)$$

$$E(T) = -\left(\left(\frac{3}{4}\right) \log_2 \left(\frac{3}{4}\right) + \left(\frac{1}{4}\right) \log_2 \left(\frac{1}{4}\right)\right)$$

$$= 0.8112$$

$$E(F) = -\left(\left(\frac{4}{5}\right) \log_2 \left(\frac{4}{5}\right) + \left(\frac{1}{5}\right) \log_2 \left(\frac{1}{5}\right)\right)$$

$$= 0.7219$$

$$\text{Entropy}(\text{Target}, a_1) = \left(\left(\frac{4}{9} \times 0.8112\right) + \left(\frac{5}{9} \times 0.7219\right) \right)$$

$$= 0.7615$$

$$\text{For } E(a_2) = P(T) E(T) + P(F) E(F)$$

$$E(T) = -\left(\left(\frac{2}{5}\right) \log_2 \left(\frac{2}{5}\right) + \left(\frac{3}{5}\right) \log_2 \left(\frac{3}{5}\right)\right)$$

$$= 0.9709$$

$$E(F) = -\left(\left(\frac{2}{4}\right) \log_2 \left(\frac{2}{4}\right) + \left(\frac{2}{4}\right) \log_2 \left(\frac{2}{4}\right)\right)$$

$$= 1$$

$$\begin{array}{ccccc} & + & - & & \\ q_2 T & 2 & 3 & - 5 \\ \text{Ent} F & 2 & 2 & - 4 \end{array}$$

$$\text{Entropy}(\text{Target}, a_2) = \left(\frac{5}{9} \times 0.9709 \right) + \left(\frac{4}{9} \times 1 \right)$$

$= 0.9838$

Calculating Information Gain (IG)

$$\begin{aligned} \underline{2.2} \quad \text{Entropy}(\text{Target}, a_1) &= E(\text{Target}) - E(\text{Target}, a_1) \\ \text{Gain} &= 0.9910 - 0.7615 \end{aligned}$$

$\boxed{IG = 0.2295}$

$$\begin{aligned} \text{Entropy}(\text{Target}, a_2) &= E(\text{Target}) - E(\text{Target}, a_2) \\ \text{Gain} &= 0.9910 - 0.9838 \end{aligned}$$

$\boxed{IG = 0.0072}$

2.3 Based on IG (higher) the best split will be with a_1

Question 3.

ID	time	gender	area	Risk
1	1-2	M	urban	low
2	2-7	M	Rural	high
3	7	F	Rural	low
4	1-2	F	Rural	high
5	7	M	Rural	high
6	1-2	(M) rur	Rural	high
7	2-7	F	Urban	low
8	(2-7)	M	Urban	low

3.1 Create a decision tree

- first we need entropy of the dataset
- Second we need Information gain to create the best split for the decision tree

→ calculating entropy of the dataset

$$E(S) = - \sum_{i=1}^C p_i \log_2 p_i$$

Total no. of risk = 8

for High = 4

for low = 4

By applying the formula, we get

$$E(R) = - \left(\frac{4}{8} \log_2 \left(\frac{4}{8} \right) + \frac{4}{8} \log_2 \left(\frac{4}{8} \right) \right)$$

= 1 Entropy of dataset.

Now, we need to find the entropy of the attributes.

$$E(\text{Risk, area}) - \text{Urban, Rural}$$

$$E(\text{Risk, gender}) - \text{Male, Female}$$

$$E(\text{Risk, time}) - 1-2, 2-7, >7$$

$$E(\text{area}) = P(\text{urban}) E(\text{urban}) + P(\text{rural}) E(\text{rural})$$

$$E(\text{urban}) = - \left(\left(\frac{3}{5} \right) \log_2 \left(\frac{3}{5} \right) \right)$$

$$= 0$$

$$E(\text{rural}) = - \left(\left(\frac{1}{5} \right) \log_2 \left(\frac{1}{5} \right) + \left(\frac{4}{5} \right) \log_2 \left(\frac{4}{5} \right) \right)$$

$$= 0.7219$$

		high	low	
area	urban	0	3	- 3
rural	4	1	- 5	

$$\text{Entropy}(\text{Risk, area}) = \left(\left(\frac{3}{8} \times 0 \right) + \left(\frac{5}{8} \times 0.7219 \right) \right)$$

$$= 0.4511$$

$$E(\text{gender}) = P(M) E(M) + P(F) E(F)$$

$$E(M) = - \left(\left(\frac{2}{5} \right) \log_2 \left(\frac{2}{5} \right) + \left(\frac{3}{5} \right) \log_2 \left(\frac{3}{5} \right) \right)$$

$$= 0.9709$$

$$E(F) = - \left(\left(\frac{2}{3} \right) \log_2 \left(\frac{2}{3} \right) + \left(\frac{1}{3} \right) \log_2 \left(\frac{1}{3} \right) \right)$$

$$= 0.9182$$

	Risk	High	low
gender	M	3	2 - 5
	F	1	2 - 3

$$E(\text{area Risk, gender}) = \left(\left(\frac{5}{8} \times 0.9709 \right) + \left(\frac{3}{8} \times 0.9182 \right) \right) = 0.9511$$

$$E(\text{time}) = P(1-2) E(1-2) + P(2-7) E(2-7) + P(>7) E(>7)$$

$$E(1-2) = - \left(\left(\frac{2}{3} \right) \log_2 \left(\frac{2}{3} \right) + \left(\frac{1}{3} \right) \log_2 \left(\frac{1}{3} \right) \right) = 0.9182$$

$$E(2-7) = - \left(\left(\frac{1}{3} \right) \log_2 \left(\frac{1}{3} \right) + \left(\frac{2}{3} \right) \log_2 \left(\frac{2}{3} \right) \right) = 0.9182$$

$$E(>7) = - \left(\left(\frac{1}{2} \right) \log_2 \left(\frac{1}{2} \right) + \left(\frac{1}{2} \right) \log_2 \left(\frac{1}{2} \right) \right) = 1$$

	high	low
Time	1-2	$\frac{3}{8} \times 0.9182$
	2-7	$\frac{3}{8} \times 0.9182$
	>7	$\frac{1}{8} \times 1$

$$\text{Entropy (risk, time)} = \left(\left(\frac{3}{8} \times 0.9182 \right) + \left(\frac{3}{8} \times 0.9182 \right) + \left(\frac{1}{8} \times 1 \right) \right)$$

$$= 0.9386$$

→ Calculating Information gain (IG)

$$\text{Gain (Risk, area)} = E(\text{Risk.}) - E(\text{Risk, area})$$

$$= 1 - 0.4511$$

$$= 0.5489$$

$$\text{Gain}(\text{risk, gender}) = E(\text{risk}) - E(\text{risk, gender})$$

$$= 1 - 0.9511$$

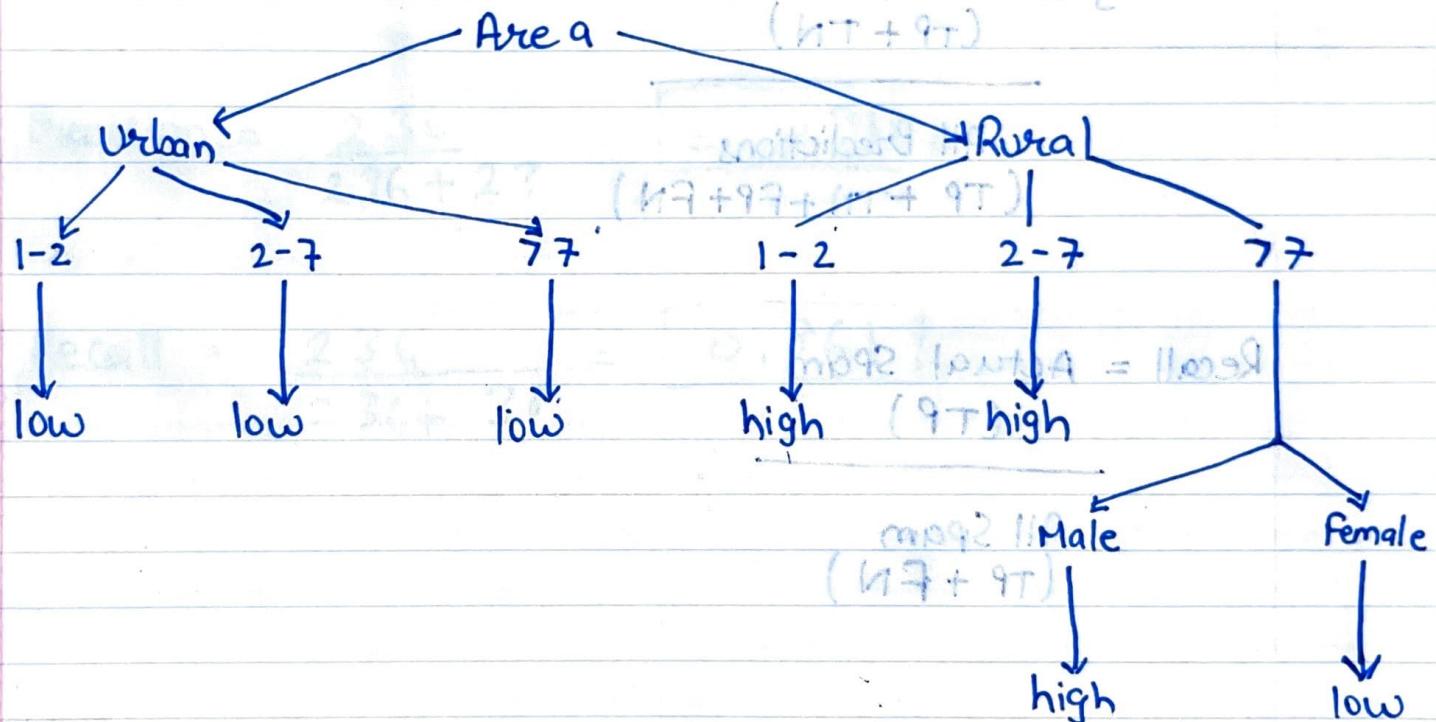
$$= 0.0489$$

$$\text{Gain}(\text{risk, time}) = E(\text{risk}) - E(\text{risk, time})$$

$$= 1 - 0.9386$$

$$= 0.0614$$

3.1 Now, we can make the decision tree with highest IG as root node:



3.2 Predict result :

ID	time	Gender	area	Risk (Predicted)
A	1-2	F	(4T + 9T) Rural	High
B	2-7	M	(9T) Urban	Low
C	1-2	F	Urban	Low

Question 4 Predicted = (Actual, plain) word!

4.3



True = (Actual, predicted)

Positive P. & negative

$$[TP, FN] =$$

Actual { Positive

Negative

False

Positive wise

true

negative

Accuracy = true predictions
 $\frac{(TP + TN)}{(TP + TN + FP + FN)}$

All Predictions

$$(TP + TN + FP + FN)$$

Recall = Actual Spam

$$\frac{TP}{(TP + FN)}$$

All Spam

$$(TP + FN)$$

Precision = Actual spam
 $\frac{TP}{(TP + FP)}$

Predicted Spam
 $(TP + FP)$

		Predicted	
		a	b
actual		236	38
		23	303

$$\text{Accuracy} = \frac{236 + 303}{236 + 38 + 23 + 303}$$

$$= 0.8983$$

$$= 89.833\%$$

$$\text{Precision} = \frac{236}{236 + 23} = 0.911$$

$$\text{Recall} = \frac{236}{236 + 38} = 0.861$$