ML

ASSIGNMENT-2

NAME: SHRIYA DUTTA
ROLL NO.: 1805705
SECTION: G1

Q1) Construct a decision tree demonstrating step by step analysis.

Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
Yes	Yes	Yes	Yes	Yes	Strep Throat
No	No	No	Yes	Yes	Allergy
Yes	Yes	No	Yes	No	Cold
Yes	No	Yes	No	No	Strep Throat
No	Yes	No	Yes	No	Cold
No	No	No	Yes	No	Allergy
No	No	Yes	No	No	Strep Throat
Yes	No	No	Yes	Yes	Allergy
No	Yes	No	Yes	Yes	Cold
Yes	No	No	Yes	Yes	Cold

Solution:

$$\text{Info. Gain} = -\frac{p}{p+n+r}\log_2\frac{p}{(p+n+r)} - \frac{n}{p+n+r}\log_2\left(\frac{n}{p+n+r}\right) - \frac{r}{p+n+r}\log_2\left(\frac{r}{p+n+r}\right)$$

where, p = Strep Throat
n = Allergy
r = Cold

$$\therefore \text{Info Gain} = -\left[\frac{3}{10}\log_2\left(\frac{3}{10}\right) + \frac{3}{10}\log_2\left(\frac{3}{10}\right) + \frac{4}{10}\log_2\left(\frac{4}{10}\right)\right]$$

$$= -\left(0.6\log_2(0.3) + 0.4\log_2(0.4)\right)$$

$$= 1.57$$

For sore Throat :

	Strep Throat	Allergy	Cold
Yes	2	1	2
No	1	2	2

$$I(Yes) = -\left[\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$

$$= 1.52$$

$$I(No) = -\left[\frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$

$$= 1.52$$

$$\therefore E(\text{Sore Throat}) = \left[\frac{5}{10} \times I(Yes)\right] + \left[\frac{5}{10} \times I(No)\right]$$

$$= \left(\frac{5}{10} \times 1.52\right) + \left(\frac{5}{10} \times 1.52\right)$$

$$= 1.52$$

$$\text{Gain (Sore Throat)} = \text{Info. Gain} - E(\text{Sore Throat})$$

$$= 1.57 - 1.52$$

$$= 0.05$$

For fever :

	Strep Throat	Allergy	Cold
Yes	1	0	3
No	2	3	1

$$I(Yes) = -\left[\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right]$$

$$= 0.81$$

$$I(No) = -\left[\frac{2}{6}\log_2\left(\frac{2}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{1}{6}\log_2\left(\frac{1}{6}\right)\right]$$

$$= 1.09$$

$$\therefore E(Fever) = \left(\frac{4}{10} \times 0.81\right) + \left(\frac{6}{10} \times 1.09\right)$$

$$= 0.98$$

$$Gain(Fever) = Info.\ Gain - E(Fever)$$

$$= 1.57 - 0.98$$

$$= 0.59$$

For Swollen Glands:

	Strep Throat	Allergy	Cold
Yes	3	0	0
No	0	3	4

$$I(Yes) = -\frac{3}{3}\log_2\left(\frac{3}{3}\right)$$

$$= 0$$

$$I(No) = -\left[\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right]$$

$$= 0.985$$

$$\therefore E(\text{swollen glands}) = \left(0 \times \frac{3}{10}\right) + \left(\frac{7}{10} \times 0.985\right)$$

$$= 0.69$$

$$\therefore \text{Gain (swollen glands)} = 1.57 - 0.69$$

$$= 0.88$$

For congestion:

	Strep Throat	Allergy	Cold
Yes	1	3	4
No	2	0	0

$$I(\text{Yes}) = -\left[\frac{1}{8}\log_2\left(\frac{1}{8}\right) + \frac{3}{8}\log_2\left(\frac{3}{8}\right) + \frac{4}{8}\log_2\left(\frac{4}{8}\right)\right]$$

$$= 1.41$$

$$I(\text{No}) = -\frac{2}{2}\log_2\left(\frac{2}{2}\right)$$

$$= 0$$

$$\therefore E(\text{congestion}) = \left(1.41 \times \frac{8}{10}\right)$$

$$= 1.13$$

$$\therefore \text{Gain (congestion)} = \text{Info. Gain} - E(\text{congestion})$$

$$= 1.57 - 1.13$$

$$= 0.44$$

For Headache:

	Strep Throat	Allergy	Cold
Yes	1	2	2
No	2	1	2

$$I(Yes) = -\left[\frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$

$$= 1.52$$

$$I(No) = -\left[\frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$

$$= 1.52$$

$$\therefore E(\text{headache}) = \left(1.52 \times \frac{5}{10}\right) + \left(1.52 \times \frac{5}{10}\right)$$

$$= 1.52$$

$$\therefore \text{Gain(headache)} = \text{Info. Gain} - E(\text{headache})$$

$$= 1.57 - 1.52$$

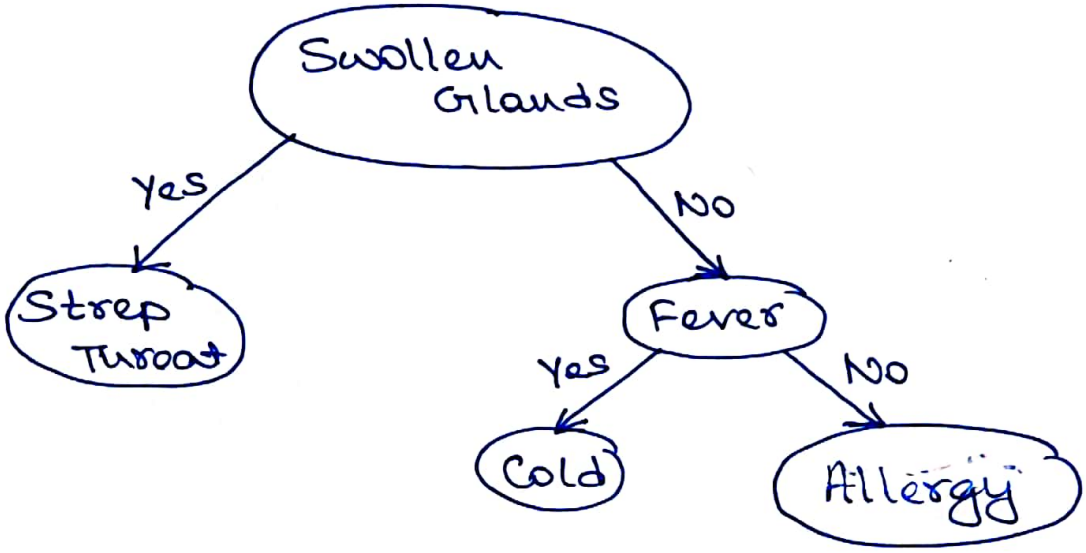
$$= 0.05$$

Gain for each attribute:

Attribute	Gain
Sore Throat	0.05
Fever	0.59
Swollen Glands	0.88
Congestion	0.44
Headache	0.05

Since the gain of swollen glands is the highest, i.e., 0.88, hence, it will be the splitting attribute.

Decision Tree:



Q2) Construct a decision tree demonstrating step by step analysis.

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Solution:

$$\text{Info. Gain} = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

where, p = Yes

n = No

$$\therefore \text{Info. gain} = -\left[\frac{9}{14}\log_2\left(\frac{9}{14}\right) + \frac{5}{14}\log_2\left(\frac{5}{14}\right)\right]$$

$$= 0.94$$

For outlook:

	Yes	No
Sunny	2	3
Overcast	4	0
Rain	3	2

$$I(\text{Sunny}) = -\left[\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]$$

$$= 0.97$$

$$I(\text{overcast}) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right)$$

$$= 0$$

$$I(\text{Rain}) = -\left[\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$

$$= 0.97$$

$$\therefore E(\text{outlook}) = \left[\frac{5}{14} \times I(\text{sunny})\right] + \left[\frac{4}{14} \times I(\text{overcast})\right] + \left[\frac{5}{14} \times I(\text{Rain})\right]$$

$$= \left(\frac{5}{14} \times 0.97\right) + \left(\frac{5}{14} \times 0.97\right)$$

$$= 0.35$$

$$\therefore \text{Gain}(\text{outlook}) = \text{Info. Gain} - E(\text{outlook})$$

$$= 0.94 - 0.35 = 0.59$$

for Temperature:

	Yes	NO
Hot	2	2
Mild	4	2
Cool	3	1

$$I(\text{Hot}) = -\left[\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right]$$

$$= 1$$

$$I(\text{Mild}) = -\left[\frac{4}{6}\log_2\left(\frac{4}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right]$$

$$= 0.92$$

$$I(\text{Cool}) = -\left[\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$$

$$= 0.81$$

$$\therefore E(\text{Temperature}) = \left(1 \times \frac{4}{14}\right) + \left(0.92 \times \frac{6}{14}\right) + \left(0.81 \times \frac{4}{14}\right)$$

$$= 0.91$$

$$\text{Gain(Temperature)} = \text{Info. Gain} - E(\text{temperature})$$

$$= 0.94 - 0.91$$

$$= 0.03$$

for Humidity:

	Yes	NO
High	3	4
Normal	6	1

$$I(\text{High}) = -\left[\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right]$$

$$= 0.99$$

$$I(\text{Normal}) = -\left[\frac{6}{7}\log_2\left(\frac{6}{7}\right) + \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right]$$

$$= 0.59$$

$$\therefore E(\text{Humidity}) = \left(0.99 \times \frac{7}{14}\right) + \left(0.59 \times \frac{7}{14}\right)$$

$$= 0.79$$

$$\text{Gain}(\text{Humidity}) = 0.94 - 0.79$$

$$= 0.15$$

For Wind:

	Yes	No
Weak	6	2
Strong	3	3

$$I(\text{weak}) = -\left[\frac{6}{8}\log_2\left(\frac{6}{8}\right) + \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right]$$

$$= 0.81$$

$$I(\text{Strong}) = -\left[\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right]$$

$$= 1$$

$$E(\text{wind}) = \left(0.81 \times \frac{8}{14}\right) + \left(1 \times \frac{6}{14}\right)$$

$$= 0.89$$

Gain (Wind) = Info. Gain − E(Wind)

= 0.94 − 0.89

= 0.05

Gain for each attribute:

Attribute	Gain
Outlook	0.59
Temperature	0.03
Humidity	0.15
Wind	0.05

Since the gain of Outlook is the highest, i.e., 0.59, hence, it will be the splitting attribute.

Decision Tree: