Question :- 1 Decision - tree for Wether forecasting.

Total :- 14 Sunny : 8 Rainy 6

⇒ Before spliting :-

Probability $P(S) = 8/14$

$P(R) = 6/14$

Entropy :- $H(x) = -\sum_{K=1}^{K} P(X = x_K) \log_2 P(X = x_K)$

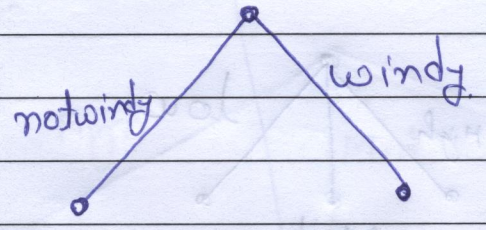
$$= -\left[\frac{8}{14}\log_2\left(\frac{8}{14}\right) + \frac{6}{14}\log_2\left(\frac{6}{14}\right)\right]$$

$$= -\left[\frac{8}{14}(-0.2430) + \frac{6}{14}(-036798)\right]$$

$$= 0.2965 \text{ bit}$$

⇒ After spliting

⇒ Windy :



sunny = 6	sunny = 2
Rainy = 2	Rainy = 4

$$\text{windy} = H(x) = -\left[\frac{2}{6}\log_2\left(\frac{2}{6}\right) + \left(\frac{4}{6}\right)\log_2\left(\frac{4}{6}\right)\right]$$

$$= 0.2764 \text{ bit}$$

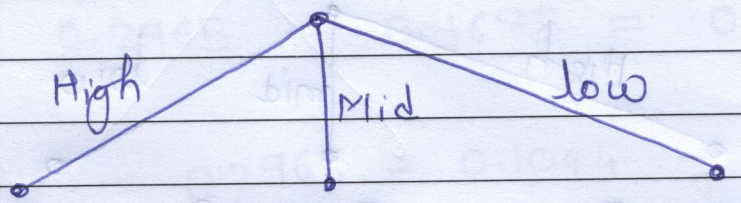
$$\text{Notwindy} = P(x) = -\left[\frac{6}{8}\log_2\left(\frac{6}{8}\right) + \left(\frac{2}{8}\right)\log_2\left(\frac{2}{8}\right)\right]$$

$$= 0.0843 \text{ bit}$$

$$\text{Weighted Average} = E(H(x)) = \left[\frac{6}{14}(0.2764) + \frac{8}{14}(0.0843)\right]$$

$$= 0.1666 \text{ bit}$$

⇒ Humidity :-



Sunny = 1	Sunny = 4	Sunny = 3
Rainy = 4	Rainy = 2	Rainy = 0

$$\text{High: } H(x) = -\left[\left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right) + \left(\frac{4}{5}\right)\log_2\left(\frac{4}{5}\right)\right]$$

$$= 0.2335 \text{ bit}$$

$$\text{Mid: } H(x) = -\left[\left(\frac{4}{6}\right)\log_2\left(\frac{4}{6}\right) + \left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right)\right]$$

$$= 0.2764 \text{ bit}$$

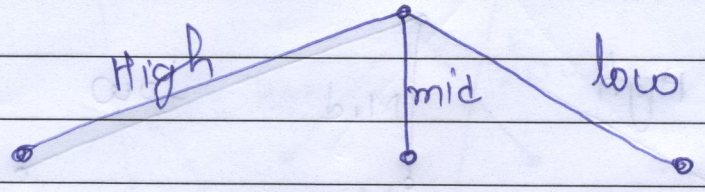
$$\text{low: } H(x) = -\left[\left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right) + \left(\frac{0}{3}\right)\log_2\left(\frac{0}{3}\right)\right]$$

$$= 0 \text{ bit}$$

$$\text{Weighted Average} = \left[\left(\frac{3}{14}\right)(0.2335) + \left(\frac{6}{14}\right)(0.2764) + \left(\frac{5}{14}\right)(0)\right]$$

$$= 0.1685 \text{ bit}$$

⇒ temperature :-



Sunny :- 5	Sunny : 3	Sunny : 0
Rainy :- 0	Rainy : 2	Rainy : 4

$$\text{High} : H(x) = -\left[\frac{5}{5}\log_2\left(\frac{5}{5}\right) + \left(\frac{0}{5}\right)\log_2\left(\frac{0}{5}\right)\right]$$

$$= 0$$

$$\text{mid} : H(x) = -\left[\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)\right]$$

$$= 0.2923$$

$$\text{low} : H(x) = -\left[\frac{0}{4}\log_2\left(\frac{0}{4}\right) + \left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right)\right]$$

$$= 0$$

$$\text{Weighted Average} = \left[\frac{5}{14}(0) + \frac{5}{14}(0.2923) + \frac{4}{14}(0)\right]$$

$$= 0.1044 \text{ bit.}$$

windy = 0.2965 - 0.1666 = 0.1299 bit.

Humidity = 0.2965 - 0.1685 = 0.128 bit

Temperature = 0.2965 - 0.1044 = **0.1921 bit**

⇒ In this example, spliting the data samples based on the temperature will minimize the entropy (unpredictability) compare to other features

⇒ In other word, splitting based on temperature provide the maximum amount of information gain at this level

⇒ so that best feature to split the data samples at the top of the decision tree is temperature