11/04/23

? For a Sunburn dataset given below, find the first splitting attribute for the decision tree by using the ID3 algorithm.

Name	Hair	Height	Weight	Lotion	Class
Sarah	Blonde	Average	Light	No	Sunburn
Dana	Blonde	Tall	Average	Yes	None
Alex	Brown	Tall	Average	Yes	None
Annie	Blonde	Short	Average	No	Sunburn
Emily	Red	Average	Heavy	No	Sunburn
Pete	Brown	Tall	Heavy	No	None
john	Brown	Average	Heavy	No	None
katie	Blonde	short	light	Yes	None

Step 1 :- Calculate the entropy of data set

T = data set

$$Info(T) = Entropy(p) = P_{yes} \log_2 P_{yes}$$
$$- P_{sunburn} \log_2 P_{sunburn}$$
$$- P_{none} \log_2 P_{none}$$

$$\text{Entropy}\left[\frac{3}{8}, \frac{5}{8}\right] = -\sum_{i=1}^{n} P_i \log_2 P_i$$

$$= -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8}$$

$$= \underline{\underline{0.954}}$$

Step 2 :- for each attribute · Hair, Height, weight, Lotion find entropy for all categorical values and also then find the informⁿ gain for the features.

First attribute :- Hair

Categorical values are :. Blonde, Brown, Red

$$\text{info}(\text{Hair}, T) = \sum_{i=1}^{3} \frac{|T_i|}{|T|}\,\text{info}(T_i)$$

$$= \frac{4}{8}\,\text{info}(\text{Blonde}) + \frac{3}{8}\,\text{info}(\text{Brown}) + \frac{1}{8}\,\text{info}(\text{Red})$$

$$= \frac{4}{8}\left[\frac{2}{5}\log_2 \frac{2}{5} - \right.$$

$$= \frac{4}{8}\left[-\frac{2}{4}\log_2 \frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right] + \frac{3}{8}\left[-\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3}\right]$$

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

$= 0.5 + 0 + 0$

$= \underline{\underline{0.5}}$

$\text{Gain}(\text{Hair}, T) = \text{info}(T) - \text{info}(\text{Hair})$

$= 0.954 - 0.5$

$= \underline{\underline{0.454}}$

Second attribute = Height

Categorical values are = Average, tall, short

$\text{info}(\text{Height}, T) = \frac{3}{8}\,\text{info}(\text{Average}) + \frac{3}{8}\,\text{info}(\text{Tall}) + \frac{2}{8}\,\text{info}(\text{short})$

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

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

$$\frac{2}{8}\left[-\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2}\right]$$

$= 0.344 + 0 + 0.25$

$= \underline{0.594}$

Gain (Height, T) = info (T) - info (Height)

$= 0.954 - 0.594$

$= \underline{0.36}$

Third attribute = weight

Categorical values are = Light, Average, Heavy

Info (weight, T) $= \frac{2}{8}$ info (Light) $+ \frac{3}{8}$ info (Average) $+ \frac{3}{8}$ info (Heavy)

$$= \frac{2}{8}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] +$$

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

$$\frac{3}{8}\left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right]$$

$= 0.25 + 0.344 + 0.344$

$= \underline{0.938}$

Gain (weight, T) = info (T) - info (weight)

$= 0.954 - 0.938$

$= 0.016$ //

Fourth attribute = Lotion

Categorical values are = No, Yes

$$\text{info}(\text{Lotion}, T) = \frac{5}{8}\,\text{info}(\text{No}) + \frac{3}{8}\,\text{info}(\text{Yes})$$

$$= \frac{5}{8}\left[-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2 \frac{2}{5}\right]$$

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

$$= \cancel{0.606}\ 0.607 + 0$$

$$= \underline{\underline{0.607}}$$

$$\text{Gain}(\text{Lotion}, T) = \text{info}(T) - \text{info}(\text{Lotion})$$

$$= 0.954 - 0.607$$

$$= \underline{\underline{0.347}}$$

step 3 :- Here the attribute with maximum informⁿ gain is Hair. So Hair is the root of decision tree.

attribute	inform^n gain
Hair	0.454
Height	0.36
weight	0.016
Lotion	0.347