Machine Learning Decision Tree – Solved Problem (ID3 algorithm)

Competition Description

Your goal is to find out when people will play outside through next week's weather forecast. You find out that the reason people decide whether to play or not depends on the weather. The following table is the decision table for whether it is suitable for playing outside.

Data Description

This can be taken as S.No. or Day. So there are 14 days... D1, D2,D3,.....D14

id	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Course Design

Choose your own way and programming language to implement the decision tree algorithm **(with code comments or notes)**. Divide the data in **Data Description** into training sets and test sets the get your answer.

Solution: I have followed ID 3 (Iterative Dichotomiser 3) Algorithm

We need to construct the Decision tree to predict whether people will play outside or not?

The following Dataset is given in the form of table

id	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Step 1: Compute Entropy (H) for entire Dataset

Number of samples = 14

Number of Attributes = 4 → Outlook, Temp, Humidity, Wind

Output variable = Play

Number of distinct outputs = 2 → Yes, No

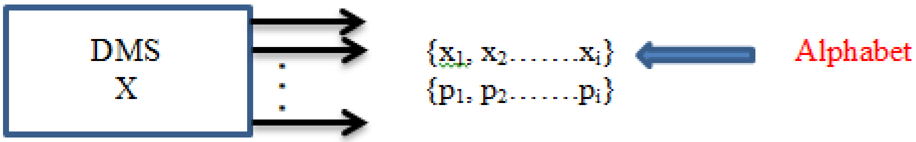
- Out of 14 samples, 9 samples belong to "Yes" category
- Out of 14 samples, 5 samples belong to "No" category

So, Number of "Yes" = 9
Number of "No" = 5

Now Total Entropy of given dataset $H = \sum_{i=1}^{L} p(x_i) \log_2 p(x_i)$

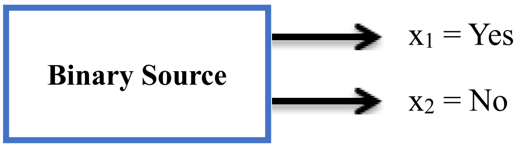
Here L = Number of symbols at the output of the DMS source. DMS is the Discrete Memoryless Source. Note that Decision Tree is a Binary tree.

A discrete information source is a source that has only a finite set of symbols as possible outputs. A discrete information source consists of a discrete (countable) set of letters or symbols.



Let X having alphabets $\{x_1, x_2, \ldots .. x_m\}$. Note that set of source symbols is called source alphabet.

A Binary source is described by the list of 2 symbols, probability assignment to these symbols a.



Total Entropy $H = \sum_{i=1}^{L} p(x_i) \log_2 \frac{1}{p(x_i)} = -\sum_{i=1}^{L} p(x_i) \log_2 p(x_i)$

$$p(x_1) = \frac{No.of\ favourables\ to\ Yes}{Total\ samples} = \frac{9}{14}$$

$$p(x_2) = \frac{No.of\ favourables\ to\ No}{Total\ samples} = \frac{5}{14}$$

$$\therefore H = -\{p(x_1) \log_2 p(x_1) + p(x_2) \log_2 p(x_2)\}$$

$$= -\left\{\frac{9}{14} \log_2 \frac{9}{14} + \frac{5}{14} \log_2 \frac{5}{14}\right\}$$

$$= -\{0.642857 \ \log_2 0.642857 + 0.357142857 \ \log_2 0.357142857\}$$

$$= -\{0.642857\ x\ (-0.63742992) + 0.357142857\ x\ (-1.4854268\ \}$$

$$= -\{-0.40977637 - 0.53050957\}$$
$$= 0.94028$$

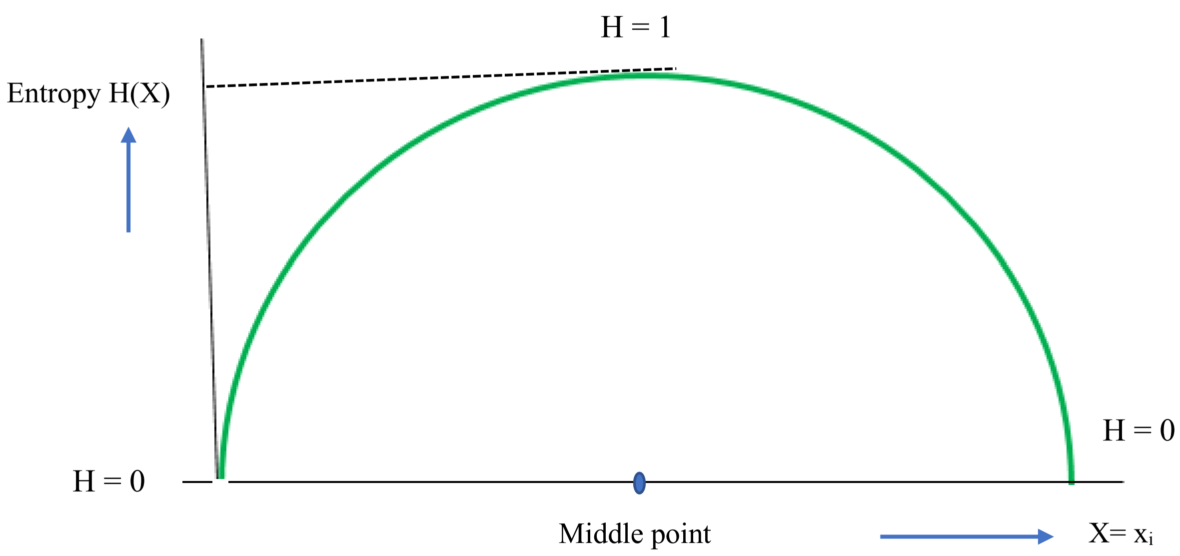
$$\log_2 0.642857 = \frac{\log_{10} 0.642857}{\log_{10} 2} = \frac{-0.19188526}{0.3010299} = -0.63742992$$

$$\log_2 0.357142857 = \frac{\log_{10} 0.357142857}{\log_{10} 2} = \frac{-0.447158}{0.3010299} = -1.4854268$$

Entropy concept

- Measures the uncertainty present in the data
- Entropy measures randomness in the data
- It is used to decide how a decision tree can split the data
- Entropy is the measure of the disorder of a system
- Entropy tends to be maximum in the middle with value 1 and minimum 0(zero) at the ends.
- The higher the entropy more the information content.
- Entropy is the average information contained in a message

Entropy $H(X) = \sum_{i=1}^{L} p(x_i)\log_2 \frac{1}{p(x_i)} = -\sum_{i=1}^{L} p(x_i)\log_2 p(x_i)$, where X is a source and L = number of symbols or messages generated by source. Binary source generates 2 symbols (example: Yes and No).



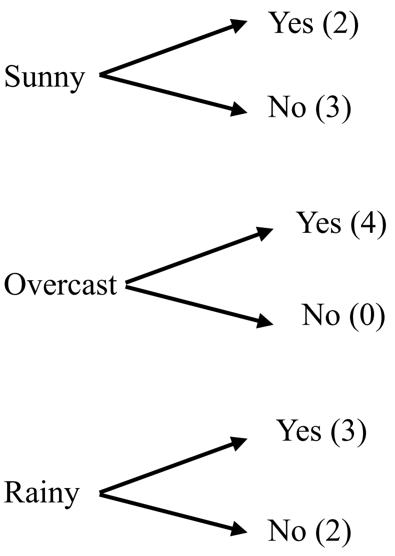
Entropy tends to be maximum in the middle with value 1 and minimum 0(zero) at the ends. x_i are the events or symbols or messages

Step 2: Calculations for every Attribute

Calculate Entropy and Information Gain for these different Attributes
In the given dataset, there 4 Attributes: Outlook, Temp, Humidity, Wind

(i) For Outlook attribute

Outlook has 3 different parameters: Sunny, Overcast, Rainy



id	outlook	play
1	sunny	no
2	sunny	no
3	overcast	yes
4	rainy	yes
5	rainy	yes
6	rainy	no
7	overcast	yes
8	sunny	no
9	sunny	yes
10	rainy	yes
11	sunny	yes
12	overcast	yes
13	overcast	yes
14	rainy	no

Entropy for Sunny: H(Outlook = Sunny) = $-\sum_{i=1}^{L} p(x_i)\log_2 p(x_i)$

$= -\{p(x_1)\log_2 p(x_1) + p(x_2)\log_2 p(x_2)\}$ Where x_1 = Yes and x_2 = No

From above data,

$$p(x_1) = \frac{Number\ of\ favourables\ for\ Yes}{Total\ samples} = \frac{2}{5}$$

$$p(x_2) = \frac{Number\ of\ favourables\ for\ No}{Total\ samples} = \frac{3}{5}$$

$$\therefore \text{H(Outlook = Sunny)} = -\left\{\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right\}$$
$$= -\{0.4\log_2 0.4 + 0.6\log_2 0.6\}$$
$$= 0.4\ x\ 1.321928 + 0.6\ x\ 0.736065$$
$$= 0.9709$$

$$\log_2 0.4 = \frac{\log_{10} 0.4}{\log_{10} 2} = \frac{-0.39794}{0.3010299} = -1.321928$$

$$\log_2 0.6 = \frac{\log_{10} 0.6}{\log_{10} 2} = \frac{-0.22184875}{0.3010299} = -0.736965$$

Entropy for Overcast: H(Outlook = Overcast) = $-\sum_{i=1}^{L} p(x_i) \log_2 p(x_i)$

$= -\{p(x_1) \log_2 p(x_1) + p(x_2) \log_2 p(x_2)\}$ Where x_1 = Yes and x_2 = No

From above data,

$$p(x_1) = \frac{Number\ of\ favourables\ for\ Yes}{Total\ samples} = \frac{4}{4} = 1$$

$$p(x_2) = \frac{Number\ of\ favourables\ for\ No}{Total\ samples} = \frac{0}{4} = 0$$

$$\therefore \text{H(Outlook = Overcast)} = -\{1 \log_2 1 + 0 \log_2 0\} = 0$$

Entropy for Rainy: H(Outlook = Rainy) = $-\sum_{i=1}^{L} p(x_i) \log_2 p(x_i)$

$= -\{p(x_1) \log_2 p(x_1) + p(x_2) \log_2 p(x_2)\}$ Where x_1 = Yes and x_2 = No

From above data,

$$p(x_1) = \frac{Number\ of\ favourables\ for\ Yes}{Total\ samples} = \frac{3}{5}$$

$$p(x_2) = \frac{Number\ of\ favourables\ for\ No}{Total\ samples} = \frac{2}{5}$$

$$\therefore \text{H(Outlook = Sunny)} = -\left\{\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5}\right\}$$

$$= -\{0.6 \log_2 0.6 + 0.4 \log_2 0.4\}$$

$$= 0.6\ x\ 0.736065 + 0.4\ x\ 1.321928$$

$$= 0.9709$$

Now we have to find Information Gain for attribute: Outlook

Information Gain = Entropy of Total Dataset – Information (Outlook)

Information of Outlook attribute is the weighted average and is given as:

$$I(Outlook) = \sum_{v \in (Sunny, Overcast, rainy)} \frac{|H_v|}{H}\ Entropy\ (H_v)$$

$$= \frac{5}{14} x\ 0.9709 + \frac{4}{14}\ x\ 0 + \frac{5}{14}\ x\ 0.9709$$

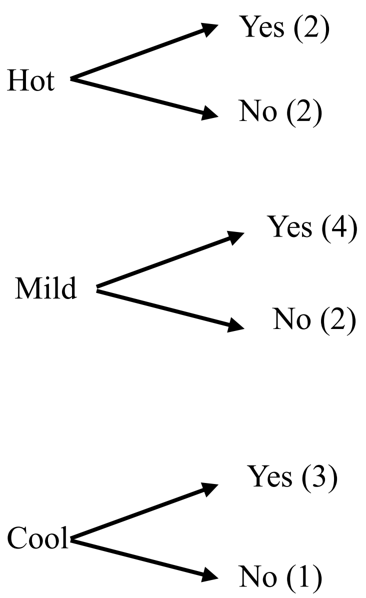
$$= 0.34675 + 0.34675 = 0.6935$$

$$\therefore Information\ Gain\ (OUtlook) = Total\ Entropy - I(Outlook)$$

$$= 0.94028 - 0.6935 = 0.24678$$

(ii) For Temperature Attribute

Temp has 3 different parameters: Hot, Mild, Cool



id	Temp	play
1	hot	no
2	hot	no
3	hot	yes
4	mild	yes
5	cool	yes
6	cool	no
7	cool	yes
8	mild	no
9	cool	yes
10	mild	yes
11	mild	yes
12	mild	yes
13	hot	yes
14	mild	no

Entropy for Hot: H(Temp = Hot) = $-\sum_{i=1}^{L} p(x_i)\log_2 p(x_i)$

$= -\{p(x_1)\log_2 p(x_1) + p(x_2)\log_2 p(x_2)\}$ Where x_1 = Yes and x_2 = No

From above data,

$$p(x_1) = \frac{Number\ of\ favourables\ for\ Yes}{Total\ samples} = \frac{2}{4}$$

$$p(x_2) = \frac{Number\ of\ favourables\ for\ No}{Total\ samples} = \frac{2}{4}$$

$$\therefore\ \text{H(Temp} = \text{Hot)} = -\left\{\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right\}$$

$$= -\{0.5\log_2 0.5 + 0.5\log_2 0.5\}$$

$$= -\{\log_2 0.5\}$$

$$= -\left\{\frac{\log_{10} 0.5}{\log_{10} 2}\right\}$$

$$= -\left\{\frac{-0.30102995}{0.30102995}\right\} = 1$$

Entropy for Mild: H(Temp = Mild) = $-\sum_{i=1}^{L} p(x_i)\log_2 p(x_i)$

$= -\{p(x_1)\log_2 p(x_1) + p(x_2)\log_2 p(x_2)\}$ Where x_1 = Yes and x_2 = No

From above data,

$$p(x_1) = \frac{Number\ of\ favourables\ for\ Yes}{Total\ samples} = \frac{4}{6}$$

$$p(x_2) = \frac{Number\ of\ favourables\ for\ No}{Total\ samples} = \frac{2}{6}$$

$$\therefore \text{ H(Temp = Mild)} = -\left\{\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right\}$$

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

$$= -\{0.6\ \log_2 0.6 + 0.4\log_2 0.4\}$$

$$= \frac{2}{3}\ x\ 0.5849626 + \frac{1}{3}\ x\ 1.5849627$$

$$= 0.389975 + 0.5283209 = 0.9183$$

$$\log_2\frac{2}{3} = \log_2 0.66667 = \frac{\log_{10} 0.66667}{\log_{10} 2} = \frac{-0.17609126}{0.3010299} = -0.5849626$$

$$\log_2\frac{1}{3} = \log_2 0.333333 = \frac{\log_{10} 0.333333}{\log_{10} 2} = \frac{-0.47712125}{0.3010299} = -1.5849627$$

Entropy for Cool: H(Temp = Cool) = $-\sum_{i=1}^{L} p(x_i)\log_2 p(x_i)$

$= -\{p(x_1)\log_2 p(x_1) + p(x_2)\log_2 p(x_2)\}$ Where x_1 = Yes and x_2 = No

From above data,

$$p(x_1) = \frac{Number\ of\ favourables\ for\ Yes}{Total\ samples} = \frac{3}{4}$$

$$p(x_2) = \frac{Number\ of\ favourables\ for\ No}{Total\ samples} = \frac{1}{4}$$

$$\therefore \text{ H(Temp = Mild)} = -\left\{\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right\}$$

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

$$= \frac{3}{4}\ x\ 0.0.41503752 + \frac{1}{4}\ x\ 2$$

$$= 0.311278 + 0.5 = 0.81127814$$

$$\log_2 \frac{3}{4} = \log_2 0.75 = \frac{\log_{10} 0.75}{\log_{10} 2} = \frac{-0.1249387}{0.3010299} = -0.41503752$$

$$\log_2 \frac{1}{4} = \log_2 0.25 = \frac{\log_{10} 0.25}{\log_{10} 2} = \frac{-0.60205999}{0.3010299} = -2$$

$$I(Temp) = \sum_{v \in (Hot, Mild, Cool)} \frac{|H_v|}{H} \; Entropy\ (H_v)$$

$$= \frac{4}{14} x\ 1 + \frac{6}{14}\ x\ 0.9183 + \frac{4}{14}\ x\ 0.81127814$$

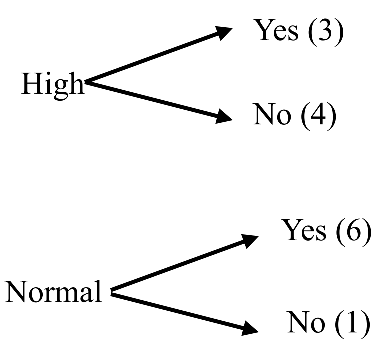
$$= 0.2857143 + 0.393557 + 0.23179 = 0.911065$$

$$\therefore Information\ Gain\ (Temp) = Total\ Entropy - I(Temp)$$

$$= 0.94028 - 0.911065 = 0.0292149$$

(iii) For Humidity Attribute

Temp has 2 different parameters: High, Normal



id	Temp	play
1	high	no
2	high	no
3	high	yes
4	high	yes
5	normal	yes
6	normal	no
7	normal	yes
8	high	no
9	normal	yes
10	normal	yes
11	normal	yes
12	high	yes
13	normal	yes
14	high	no

Entropy for High: H(Humidity = High) = $-\sum_{i=1}^{L} p(x_i) \log_2 p(x_i)$

$= -\{p(x_1) \log_2 p(x_1) + p(x_2) \log_2 p(x_2)\}$ Where x_1 = Yes and x_2 = No

From above data,

$$p(x_1) = \frac{Number\ of\ favourables\ for\ Yes}{Total\ samples} = \frac{3}{7}$$

$$p(x_2) = \frac{Number\ of\ favourables\ for\ No}{Total\ samples} = \frac{4}{7}$$

$$\therefore\ \text{H(Humidity} = \text{High)} = -\left\{\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right\}$$

$$= \frac{3}{7}\ x\ 1.2223926 + \frac{4}{7}\ x\ 0.807355$$

$$= 0.5238825 + 0.46134574$$

$$= -\{\log_2 0.5\}$$

$$= 0.98523$$

$$\log_2\frac{3}{7} = \log_2 0.42857 = \frac{\log_{10} 0.42857}{\log_{10} 2} = \frac{-0.36797678}{0.3010299} = -1.222392$$

$$\log_2\frac{4}{7} = \log_2 0.57143 = \frac{\log_{10} 0.57143}{\log_{10} 2} = \frac{-0.243038}{0.3010299} = -0.807355$$

Entropy for Normal: H(Humidity = Normal) = $-\sum_{i=1}^{L} p(x_i)\log_2 p(x_i)$

$= -\{p(x_1)\log_2 p(x_1) + p(x_2)\log_2 p(x_2)\}$ Where x_1 = Yes and x_2 = No

From above data,

$$p(x_1) = \frac{Number\ of\ favourables\ for\ Yes}{Total\ samples} = \frac{6}{7}$$

$$p(x_2) = \frac{Number\ of\ favourables\ for\ No}{Total\ samples} = \frac{1}{7}$$

$$\therefore\ \text{H(Humidity} = \text{High)} = -\left\{\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right\}$$

$$= \frac{6}{7}\ x\ 0.22239245 + \frac{1}{7}\ x\ 2.807355$$

$$= 0.190622 + 0.40105076$$

$$= 0.59167286$$

$$\log_2\frac{6}{7} = \log_2 0.857143 = \frac{\log_{10} 0.857143}{\log_{10} 2} = \frac{-0.0669467}{0.3010299} = -0.22239245$$

$$\log_2\frac{1}{7} = \log_2 0.142857 = \frac{\log_{10} 0.142857}{\log_{10} 2} = \frac{-0.845098}{0.3010299} = -2.807355$$

$$I(Humidity) = \sum_{v \in (High, Normal)} \frac{|H_v|}{H} \; Entropy\ (H_v)$$

$$= \frac{7}{14} x\ 0.98523 + \frac{7}{14}\ x\ 0.59167286$$

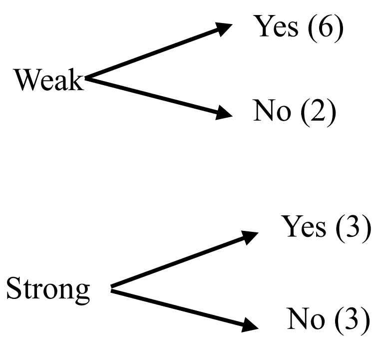
$$= 0.492615 + 0.29583643 = 0.78845143$$

$$\therefore Information\ Gain\ (Humidity) = Total\ Entropy - I(Humidity)$$

$$= 0.94028 - 0.78845143 = 0.15182857$$

(iv) For Wind Attribute

Temp has 2 different parameters: Weak, Strong



id	Wind	play
1	weak	no
2	strong	no
3	weak	yes
4	weak	yes
5	weak	yes
6	strong	no
7	strong	yes
8	weak	no
9	weak	yes
10	weak	yes
11	strong	yes
12	strong	yes
13	weak	yes
14	strong	no

Entropy for Weak: H(Wind = Weak) = $-\sum_{i=1}^{L} p(x_i) \log_2 p(x_i)$

$= -\{p(x_1) \log_2 p(x_1) + p(x_2) \log_2 p(x_2)\}$ Where x_1 = Yes and x_2 = No

From above data,

$$p(x_1) = \frac{Number\ of\ favourables\ for\ Yes}{Total\ samples} = \frac{6}{8}$$

$$p(x_2) = \frac{Number\ of\ favourables\ for\ No}{Total\ samples} = \frac{2}{8}$$

$\therefore$ H(Wind = Weak) $= -\left\{\frac{6}{8}\log_2\frac{6}{8} + \frac{2}{8}\log_2\frac{2}{8}\right\}$

$$= -\left\{\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right\}$$

$$= 0.311278 + 0.5$$

$$= 0.81278$$

Entropy for Strong: H(Wind = Strong) = $-\sum_{i=1}^{L} p(x_i)\log_2 p(x_i)$

$= -\{p(x_1)\log_2 p(x_1) + p(x_2)\log_2 p(x_2)\}$ Where x_1 = Yes and x_2 = No

From above data,

$$p(x_1) = \frac{Number\ of\ favourables\ for\ Yes}{Total\ samples} = \frac{3}{6}$$

$$p(x_2) = \frac{Number\ of\ favourables\ for\ No}{Total\ samples} = \frac{3}{6}$$

$\therefore$ H(Wind = Strong) $= -\left\{\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right\}$

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

$$I(Wind) = \sum_{v\in(High,Normal)} \frac{|H_v|}{H}\ Entropy\ (H_v)$$

$$= \frac{8}{14}x\ 0.81278 + \frac{6}{14}\ x\ 1$$

$$= 0.46444 + 0.4285714 = 0.893017$$

$\therefore Information\ Gain\ (Humidity) = Total\ Entropy - I(Humidity)$

$$= 0.94028 - 0.893017 = 0.04726288$$

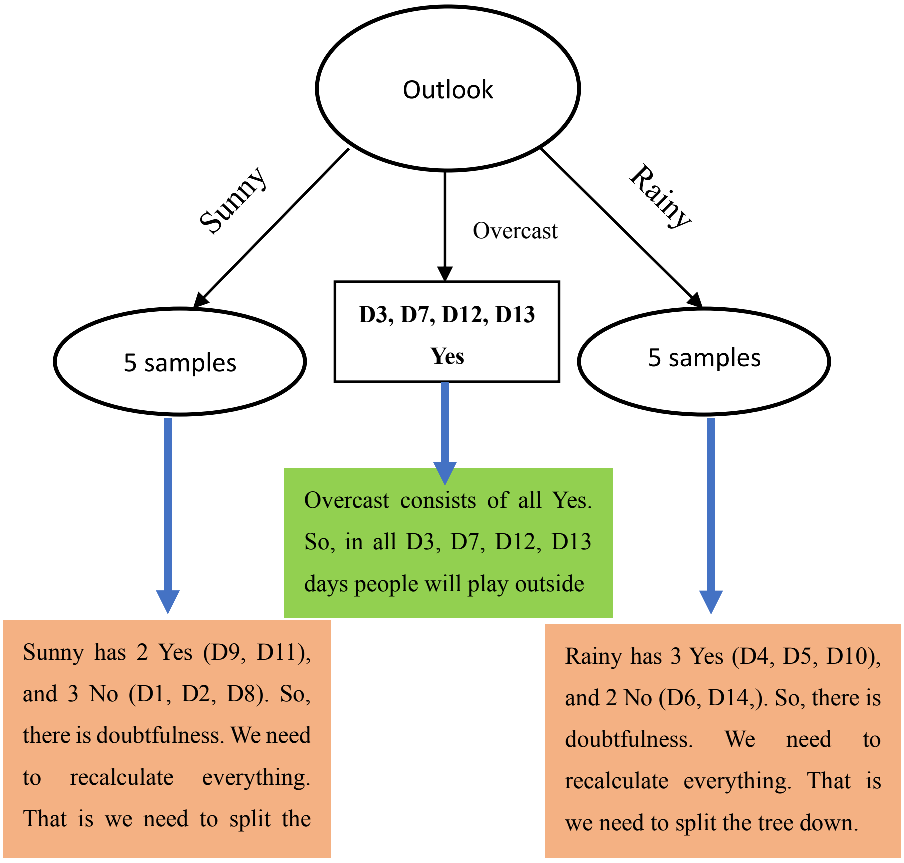
Information gains are reproduced below:

IG(Outlook) = 0.24678 → Highest Gain
IG(Temp) = 0.0292149
IG(Humidity) = 0.152
IG(Wind) = 0.04726288

The best attribute (predictor variable) is the one that separates dataset into different classes most effectively or it is the feature that best splits the dataset. Attribute with highest Information gain is taken as ROOT NODE. Here the Outlook attribute has highest information gain.

Drawing Decision tree

Select Outlook as Root node



Take Outlook = Sunny and proceed al the steps that we did for original dataset. Outlook already taken as Root node, so no need to write Outlook attribute in table. Take Sunny samples from original table and write down as shown below:

id	temperature	humidity	wind	play
1	hot	high	weak	no
2	hot	high	strong	no
8	mild	high	weak	no
9	cool	normal	weak	yes
11	mild	normal	strong	yes

For this table, we need to calculate EVERYTHING that we did for original table

Step 1: Compute Entropy of new dataset given in above table

Number of samples = 5

Number of Attributes = 4 → Temp, Humidity, Wind

Output variable = Play

Number of distinct outputs = 2 → Yes, No

- Out of 5 samples, 2 samples belong to "Yes" category
- Out of 5 samples, 3 samples belong to "No" category

So, Number of "Yes" = 2

Number of "No" = 3

Now Total Entropy of given dataset $H = \sum_{i=1}^{L} p(x_i) \log_2 p(x_i)$

Total Entropy $H = \sum_{i=1}^{L} p(x_i) \log_2 \frac{1}{p(x_i)} = -\sum_{i=1}^{L} p(x_i) \log_2 p(x_i)$

$$p(x_1) = \frac{No. of\ favourables\ to\ Yes}{Total\ samples} = \frac{2}{5}$$

$$p(x_2) = \frac{No. of\ favourables\ to\ No}{Total\ samples} = \frac{3}{5}$$

$$\therefore H = -\{p(x_1) \log_2 p(x_1) + p(x_2) \log_2 p(x_2)\}$$

$$= -\left\{\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5}\right\}$$

$$\log_2 \frac{2}{5} = \log_2 0.4 = \frac{\log_{10} 0.4}{\log_{10} 2} = \frac{-0.39794}{0.3010299} = -1.32193$$

$$\log_2 \frac{3}{5} = \log_2 0.6 = \frac{\log_{10} 0.6}{\log_{10} 2} = \frac{-0.2218487}{0.3010299} = -0.7369657$$

$$\therefore H = -\{0.4\ x\ 1.32193 + 0.6\ x\ 0.7369657\}$$

$$= 0.528772 + 0.4421794 = 0.9709514$$

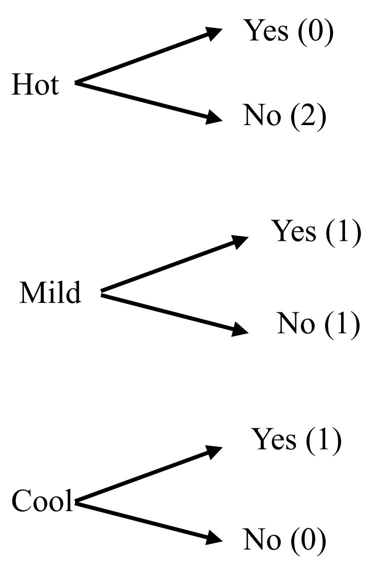
Step 2: Calculations for every attribute

Step 2: Calculations for every Attribute
Calculate Entropy and Information Gain for these different Attributes

In the given dataset, there 3 Attributes: Temp, Humidity, Wind

(i) For Temp attribute

Temp has 3 different parameters: Hot, Mild, Cool



id	temperature	play
1	hot	no
2	hot	no
8	mild	no
9	cool	yes
11	mild	yes

Directly we can place values of entropy by remembering properties of entropy. No mathematical calculations are required.

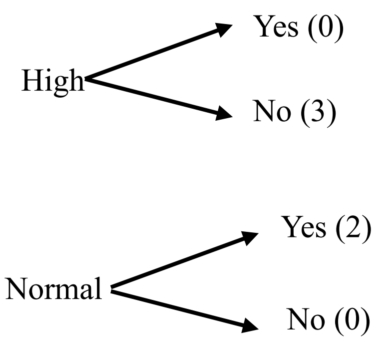
Entropy H(Temp = Hot) = 0 (because all No)
Entropy H(Temp = Mild) = 1 (because equal number of Yes and No)
Entropy H(Temp = Cool) = 0 (because all Yes)

$$I(Temp) = \frac{2}{5}\ x\ 0 + \frac{2}{5}\ x\ 1 + \frac{1}{5}\ x\ 0 = \frac{2}{5} = 0.4$$

$$IG(Temp) = 0.9709514 - 0.4 = 0.5709514$$

(ii) For Humidity attribute

Humidity has 2 different parameters: High, Normal



id	humidity	play
1	high	no
2	high	no
8	high	no
9	normal	yes
11	normal	yes

Directly we can place values of entropy by remembering properties of entropy. No mathematical calculations are required.

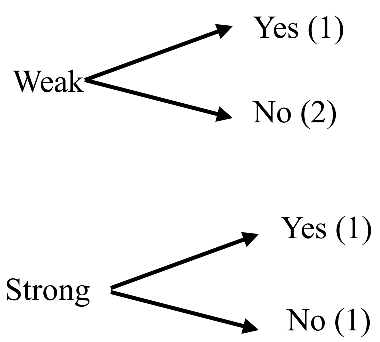
Entropy H(Humidity = High) = 0 (because all No)
Entropy H(Humidity = Normal) = 0 (because all Yes)

$$I(\text{Humidity}) = \frac{3}{5}\,x\,0 + \frac{2}{5}\,x\,0 = 0$$

$$IG(\text{Humidity}) = 0.9709514 - 0 = 0.9709514$$

(iii) For Wind attribute

Temp has 2 different parameters: Weak, Strong



id	wind	play
1	weak	no
2	strong	no
8	weak	no
9	weak	yes
11	strong	yes

Entropy H(Wind = Strong) = 1 (because equal number of Yes and No)

Entropy H(Wind = Weak) = $-\left\{\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right\} = 0.5283209 + 0.389975 = 0.9183$

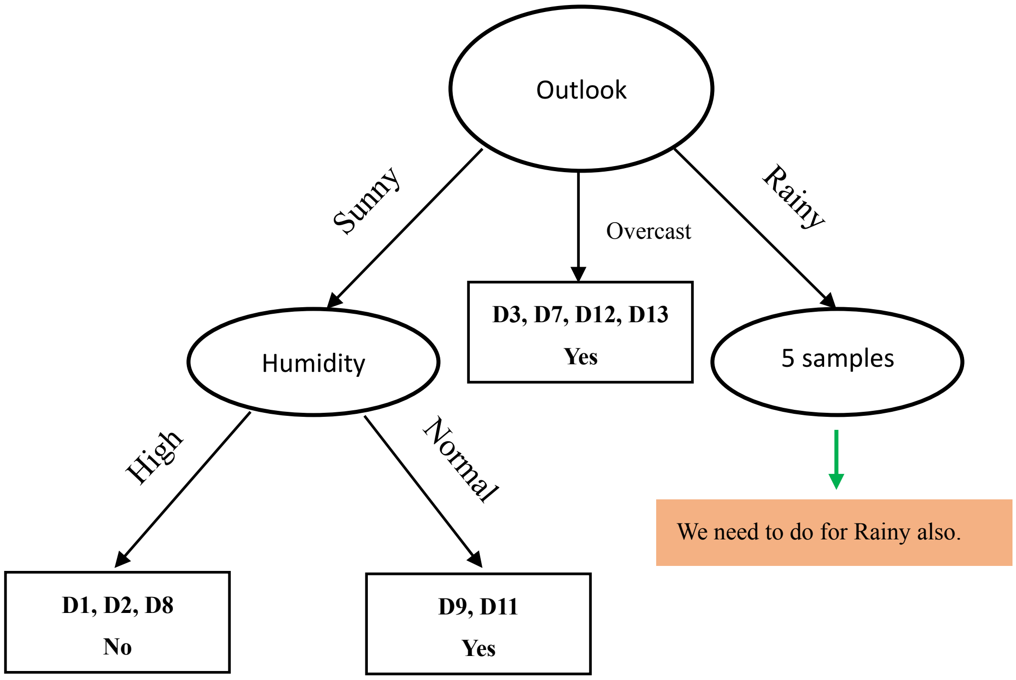
$$I(\text{Wind}) = \frac{3}{5}\,x\,0.9183 + \frac{2}{5}\,x\,1 = 0.95098$$

$$IG(\text{Wind}) = 0.9709514 - 0.95098 = 0.0199714$$

Information gains are reproduced below:

IG(Temp) = 0.5709514
IG(Humidity) = 0.9709514 → highest GAIN
IG(Wind) = 0.0199174

New Decision tree is shown below



NOW WE SHOULD WORK ON Outlook = Rainy condition

id	temperature	humidity	wind	play
4	mild	high	weak	yes
5	cool	normal	weak	yes
6	cool	normal	strong	no
10	mild	normal	weak	yes
14	mild	high	strong	no

Step 1: Compute Entropy of new dataset given in above table.

Number of samples = 5

Number of Attributes = 3 → Temp, Humidity, Wind

Output variable = Play

Number of distinct outputs = 2 → Yes, No

Out of 5 samples, 3 samples belong to “Yes” category
Out of 5 samples, 2 samples belong to “No” category

So, Number of “Yes” = 3
Number of “No” = 2

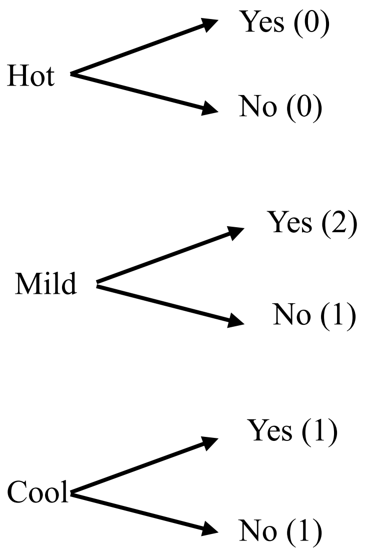
$$p(x_1) = \frac{No.of\ favourables\ to\ Yes}{Total\ samples} = \frac{3}{5}$$
$$p(x_2) = \frac{No.of\ favourables\ to\ No}{Total\ samples} = \frac{2}{5}$$

$$\therefore H = -\{p(x_1)\log_2 p(x_1) + p(x_2)\log_2 p(x_2)\}$$

$$= -\left\{\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5} +\right\} = 0.9709514$$

(iv) **For Temp attribute**

Temp has 3 different parameters: Hot, Mild, Cool



id	temperature	play
4	mild	no
5	cool	no
6	cool	no
10	mild	yes
14	mild	yes

Directly we can place values of entropy by remembering properties of entropy. No mathematical calculations are required.

Entropy H(Temp = Hot) = 0 (because all No)

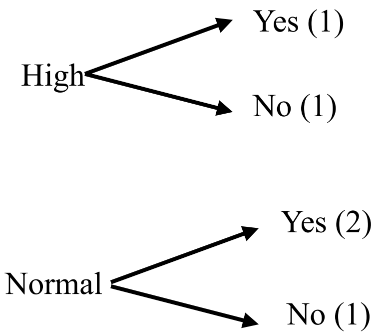
Entropy H(Temp = Mild) = $-\left\{\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3} +\right\} = 0.9183$

Entropy H(Temp = Cool) = 1 (because equal number of Yes and No)

$$I(Temp) = \frac{0}{0}\ x\ 0 + \frac{3}{5}\ x\ 0.9183 + \frac{2}{5}\ x\ 1 = 0.95098$$
$$IG(Temp) = 0.9709514 - 0.95098 = 0.0199714$$

(v) For Humidity attribute

Humidity has 2 different parameters: High, Normal



id	humidity	play
4	high	no
5	normal	no
6	normal	no
10	normal	yes
14	high	yes

Directly we can place values of entropy by remembering properties of entropy. No mathematical calculations are required.

Entropy H(Humidity = High) = 1 (because equal number of Yes and No)

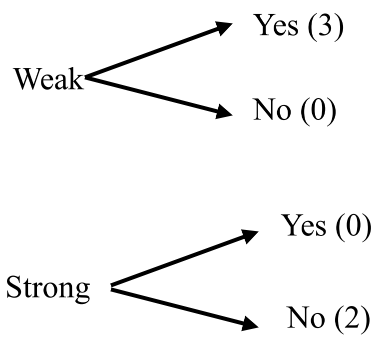
Entropy H(Humidity = Normal) = $-\left\{\frac{2}{3}\log_2\frac{2}{3}+\frac{1}{3}\log_2\frac{1}{3}+\right\} = 0.9183$

$I(\text{Humidity}) = \frac{2}{5}\ x\ 1 + \frac{3}{5}\ x\ 0.9183 = 0.95098$

$IG(\text{Humidity}) = 0.9709514 - 0.95098 = 0.0199714$

(vi) For Wind attribute

Humidity has 2 different parameters: Weak, Strong



id	wind	play
4	weak	no
5	weak	no
6	strong	no
10	weak	yes
14	strong	yes

Directly we can place values of entropy by remembering properties of entropy. No mathematical calculations are required.

Entropy H(Wind = Weak) = 0 (because all Yes)
Entropy H(Wind = Weak) = 0 (because all No)

$I(\text{wind}) = \frac{3}{5}\ x\ 0 + \frac{2}{5}\ x\ 0 = 0$

$IG(\text{Humidity}) = 0.9709514 - 0 = 0.0199714 = 0.9709514$

Information gains are reproduced below:

IG(Temp) = 0.0199714
IG(Humidity) = 0.0199714
IG(Wind) = 0.9709514 → highest GAIN
Complete Decision tree is shown below

Final Decision tree is shown below