MDL Assignment-4

Vikrant Dewangan, Roll No.- 2018111024

Contents

1   Problem Statement

Design a decision tree for the following Dataset, showing construction at each level.

Forecast	Temperature	Humidity	Wind	Go on a trip
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	**No**
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	**No**
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes

2   The Algorithm

3   Calculating Information Gain

At first, we shall calculate the entropy for the decision **Go on a trip**. To do this, we calculate $H\big(\textbf{Go on a trip}\big)$ -

Go on a trip	
Yes	**No**
6	6

$$
\begin{aligned}
H\big(\textbf{Go on a trip}\big) &= -\frac{p}{p+n} \cdot log\Big(\frac{p}{p+n}\Big) - \frac{n}{p+n} \cdot log\Big(\frac{n}{p+n}\Big) \\
&= -2 \cdot \frac{6}{12} \cdot log\Big(\frac{6}{12}\Big) \\
&= 1
\end{aligned}
$$

Now, we shall calculate the entropies for each of the 4 attributes **Forecast**, **Temperature**, **Humidity** and **Wind**.

1. $H\big(\textbf{Go on a trip}, \textbf{Temperature}\big)$

		Go on a trip		Total
		Yes	**No**	
	Hot	0	3	3
Temperature	Mild	4	1	5
	Cool	2	2	4

$$H\big(Hot\big) = 0$$

$$H\big(Mild\big) = -\frac{4}{5} \cdot log\left(\frac{4}{5}\right) - \frac{1}{5} \cdot log\left(\frac{1}{5}\right)$$

$$= log\big(5\big) + \frac{4}{5} \cdot log\big(4\big)$$

$$= 0.72$$

$$H\big(Cold\big) = -\frac{2}{4} \cdot log\left(\frac{2}{4}\right) - \frac{2}{4} \cdot log\left(\frac{2}{4}\right)$$

$$= 1$$

$$H\big(\textbf{Go on a trip}, \textbf{Temperature}\big) = P\big(Hot\big)H\big(Hot\big)+$$
$$P\big(Cold\big)H\big(Cold\big) + P\big(Mild\big)H\big(Mild\big)$$
$$= \frac{3}{12} \cdot H\big(0,3\big) + \frac{5}{12} \cdot H\big(4,1\big) + \frac{4}{12} \cdot H\big(2,2\big)$$
$$= 0 + \frac{5}{12} \cdot 0.72 + \frac{4}{12} \cdot 1$$
$$= 0.63$$

2. $H\big(\textbf{Go on a trip}, \textbf{Forecast}\big)$

		Go on a trip		Total
		Yes	**No**	
	Sunny	2	3	5
Forecast	Rain	3	1	4
	Overcast	1	2	3

$$H\left(Sunny\right) = -\frac{2}{5} \cdot log\left(\frac{2}{5}\right) - \frac{3}{5} \cdot log\left(\frac{3}{5}\right)$$
$$= 0.97$$
$$H\left(Rain\right) = -\frac{3}{4} \cdot log\left(\frac{3}{4}\right) - \frac{1}{4} \cdot log\left(\frac{1}{4}\right)$$
$$= 0.81$$
$$H\left(Overcast\right) = -\frac{1}{3} \cdot log\left(\frac{1}{3}\right) - \frac{2}{3} \cdot log\left(\frac{2}{3}\right)$$
$$= 0.92$$

$$H\left(\textbf{Go on a trip}, \textbf{Forecast}\right) = P\left(Sunny\right)H\left(Sunny\right) + P\left(Rain\right)H\left(Rain\right)$$
$$+ P\left(Overcast\right)H\left(Overcast\right)$$
$$= \frac{5}{12} \cdot H\left(2,3\right) + \frac{4}{12} \cdot H\left(3,1\right) + \frac{3}{12} \cdot H\left(1,2\right)$$
$$= \frac{5}{12} \cdot 0.97 + \frac{4}{12} \cdot 0.81 + \frac{3}{12} \cdot 0.92$$
$$= 0.90$$

3. $H\left(\textbf{Go on a trip}, \textbf{Humidity}\right)$

		Go on a trip		Total
		Yes	No	
Humidity	High	2	4	6
	Normal	4	2	6

$$H\left(High\right) = -\frac{2}{6} \cdot log\left(\frac{2}{6}\right) - \frac{4}{6} \cdot log\left(\frac{4}{6}\right)$$
$$= 0.92$$
$$H\left(Rain\right) = -\frac{4}{6} \cdot log\left(\frac{4}{6}\right) - \frac{2}{6} \cdot log\left(\frac{2}{6}\right)$$
$$= 0.92$$

$$H\left(\textbf{Go on a trip}, \textbf{Humidity}\right) = P\left(High\right)H\left(High\right) + P\left(Normal\right)H\left(Normal\right)$$
$$= \frac{6}{12} \cdot H\left(2,4\right) + \frac{6}{12} \cdot H\left(4,2\right)$$
$$= 0.92$$

		Go on a trip		Total
		Yes	No	
Wind	Weak	3	4	7
	Strong	3	2	5

4. $H(\textbf{Go on a trip}, \textbf{Wind})$

$$H(Weak) = -\frac{3}{7} \cdot log\left(\frac{3}{7}\right) - \frac{4}{7} \cdot log\left(\frac{4}{7}\right)$$
$$= 0.99$$
$$H(Strong) = -\frac{3}{5} \cdot log\left(\frac{3}{5}\right) - \frac{2}{5} \cdot log\left(\frac{2}{5}\right)$$
$$= 0.97$$

$$H(\textbf{Go on a trip}, \textbf{Wind}) = P(Strong)H(Strong) + P(Weak)H(Weak)$$
$$= \frac{5}{12} \cdot H(3,2) + \frac{7}{12} \cdot H(3,4)$$
$$= 0.98$$

Now we shall calculate information gain for each split -

$$Gain(\textbf{Go on a trip}, \textbf{Temperature}) = H(\textbf{Go on a trip}) - H(\textbf{Go on a trip}, \textbf{Temperature})$$
$$= 1 - 0.63$$
$$= 0.37$$
$$Gain(\textbf{Go on a trip}, \textbf{Forecast}) = H(\textbf{Go on a trip}) - H(\textbf{Go on a trip}, \textbf{Forecast})$$
$$= 1 - 0.90$$
$$= 0.10$$
$$Gain(\textbf{Go on a trip}, \textbf{Humidity}) = H(\textbf{Go on a trip}) - H(\textbf{Go on a trip}, \textbf{Humidity})$$
$$= 1 - 0.92$$
$$= 0.08$$
$$Gain(\textbf{Go on a trip}, \textbf{Wind}) = H(\textbf{Go on a trip}) - H(\textbf{Go on a trip}, \textbf{Wind})$$
$$= 1 - 0.98$$
$$= 0.02$$

4   Building further tree

Since we have the highest information gain at the Temperature split, our first **decision** shall be regarding checking the temperature.

Forecast	Temperature	Humidity	Wind	Go on a trip
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	**No**
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	**No**
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	High	Weak	No
Rain	Mild	High	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes

1. Now as we can clearly see, <u>Hot</u> temperature always results in **No** decision. In other words,

$$H\big(\textbf{Temp} = Hot\big) = 1$$

Thus one of our path to *leaf nodes* is decided.
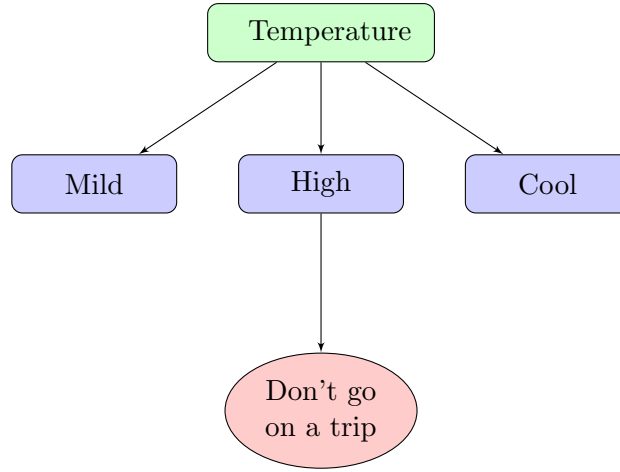
2. Now, for when the temperature is <u>Cool</u>, the following table results -

| | | Go on a trip | | Total | Entropy |
		Yes	No		
Wind	Weak	2	0	2	0
	Strong	0	2	2	0
Forecast	Rain	1	1	2	1
	Overcast	0	1	1	0
	Sunny	1	0	1	0
Humidity	Normal	2	2	4	1

$$H\big(\textbf{Go on a trip}, \textbf{Temperature} = Cool\big) = H\big(2, 2\big)$$
$$= 1$$

Now for $H\big(\textbf{Go on a trip}, \textbf{Temperature} = Cool\big)$, the best split would be that which can the highest information gain. Clearly as we can see,

$$H\big(\textbf{Go on a trip}, \textbf{Temperature} = Cool, \textbf{Wind}\big) = \frac{2}{4} \cdot 0 + \frac{2}{4} \cdot 0$$
$$= 0$$

$$H\big(\textbf{Go on a trip}, \textbf{Temperature} = Cool, \textbf{Forecast}\big) = \frac{2}{4} \cdot 1 + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 0$$
$$= 0.5$$

$$H\big(\textbf{Go on a trip}, \textbf{Temperature} = Cool, \textbf{Humidity}\big) = \frac{1}{1} \cdot 1$$
$$= 1$$

Thus $Gain\big(\mathbf{Go\ on\ a\ trip}, \mathbf{Temperature} = Cool, \mathbf{Wind}\big)$ is the highest. Here, we can distinctly make out that **Wind**=Weak results in Yes and **Wind**=Strong results in No.
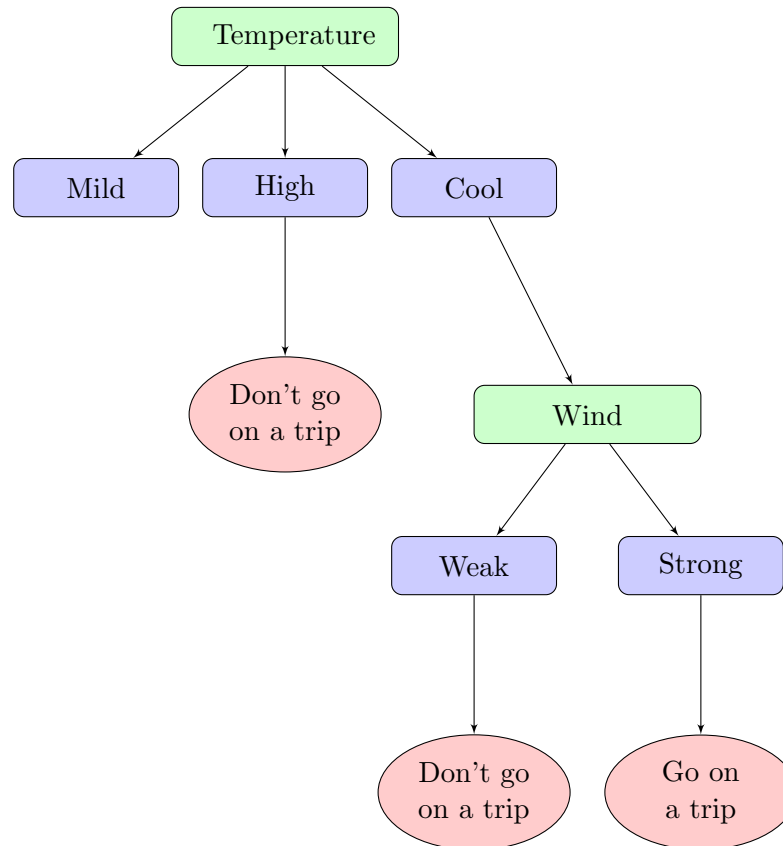
3. Now, for when the temperature is <u>Mild</u>, the following table results

		Go on a trip		Total	Entropy
		Yes	**No**		
Wind	Weak	2	1	3	0.92
	Strong	2	0	2	0
Forecast	Rain	2	0	2	0
	Overcast	1	0	1	0
	Sunny	1	1	2	1
Humidity	Normal	2	0	2	0
	High	2	1	3	0.92

$$H\big(\textbf{Go on a trip}, \textbf{Temperature} = Mild, \textbf{Wind}\big) = \frac{3}{5} \cdot 0.92 + \frac{2}{5} \cdot 0$$
$$= 0.55$$
$$H\big(\textbf{Go on a trip}, \textbf{Temperature} = Mild, \textbf{Forecast}\big) = \frac{2}{5} \cdot 0 + \frac{1}{5} \cdot 0 + \frac{2}{5} \cdot 1$$
$$= 0.4$$
$$H\big(\textbf{Go on a trip}, \textbf{Temperature} = Mild, \textbf{Humidity}\big) = \frac{2}{5} \cdot 0 + \frac{3}{5} \cdot 0.92$$
$$= 0.55$$

Clearly, as $H\big(\textbf{Go on a trip}, \textbf{Temperature} = Mild, \textbf{Forecast}\big)$ is smallest, we shall look into this split.

Forecast	Temperature	Humidity	Wind	Go on a trip
Sunny	Mild	High	Weak	No
Sunny	Mild	Normal	Strong	Yes
Rain	Mild	High	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Overcast	Mild	High	Strong	Yes

Clearly, when the **Forecast** is not Sunny the output is clearly **Yes**. When the **Forecast** is Sunny, we look at the **Humidity** or **Wind** parameters. Since both of them convey the same thing, we go for **Humidity** this time.
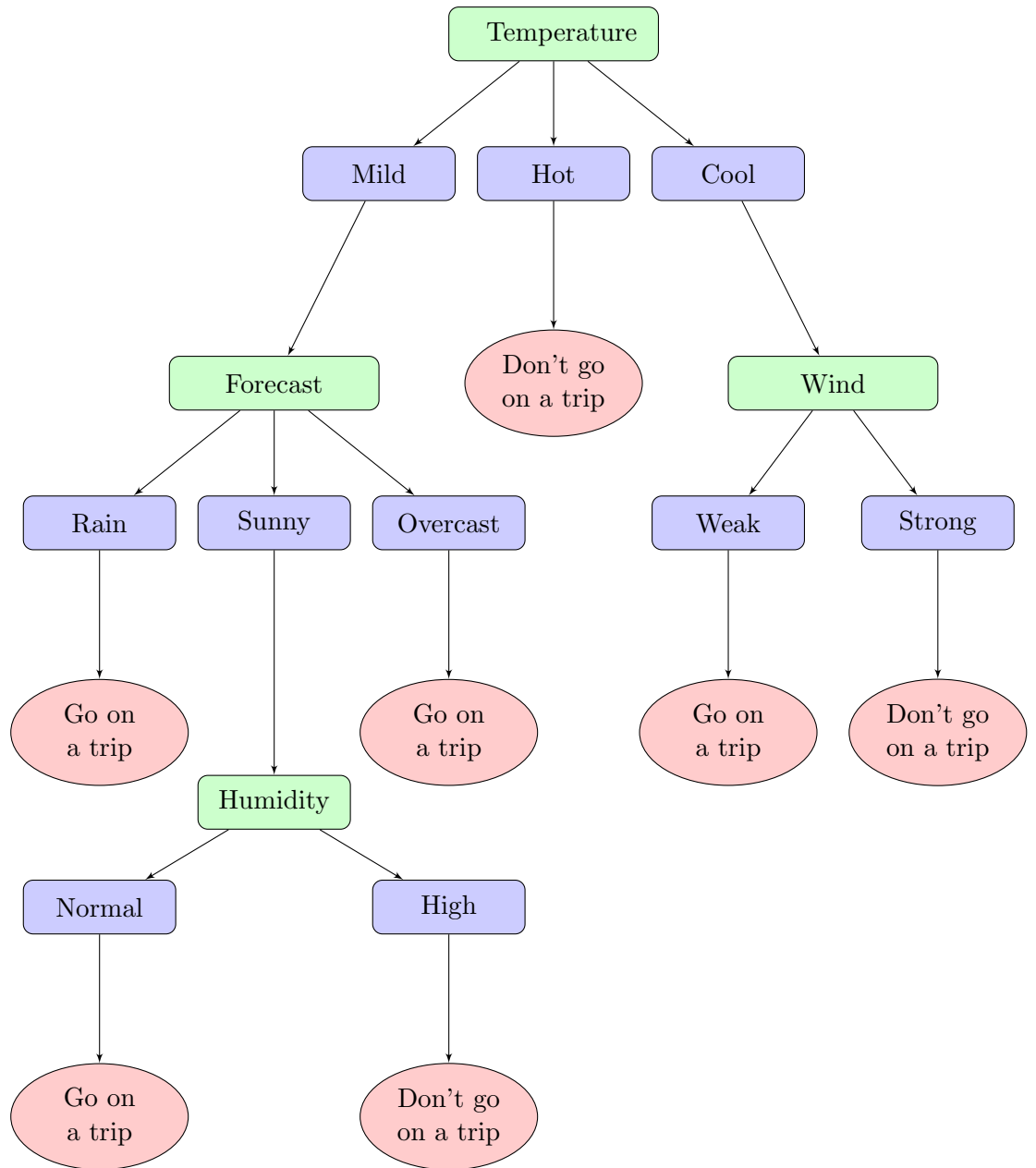
We shall use the following notation - Accordingly, the flowchart can be summarized as -

Function	Attribute
f(0)	Forecast
f(1)	Temperature
f(2)	Humidity
f(3)	Wind

f(1)=='Mild'?

Yes — No

f(0)=='Rain'?

f(1)=='Hot'?

Yes — No

True

f(0)=='Overcast'?

Yes — No

False

f(3)=='Weak'?

Yes — No

True

f(2)=='Normal'?

Yes — No

True — False

Yes — No

True — False