CS7.301 **18^{th} April, 2020**

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

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

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**	
Temperature	Hot	0	3	3
	Mild	4	1	5
	Cool	2	2	4

$$
\begin{aligned}
H\big(Hot\big) &= 0 \\
H\big(Mild\big) &= -\frac{4}{5} \cdot log\left(\frac{4}{5}\right) - \frac{1}{5} \cdot log\left(\frac{1}{5}\right) \\
&= log\big(5\big) + \frac{4}{5} \cdot log\big(4\big) \\
&= 0.72 \\
H\big(Cold\big) &= -\frac{2}{4} \cdot log\left(\frac{2}{4}\right) - \frac{2}{4} \cdot log\left(\frac{2}{4}\right) \\
&= 1
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
H\big(Sunny\big) &= -\frac{2}{5}\cdot log\bigg(\frac{2}{5}\bigg) - \frac{3}{5}\cdot log\bigg(\frac{3}{5}\bigg)\\
&= 0.97\\
H\big(Rain\big) &= -\frac{3}{4}\cdot log\bigg(\frac{3}{4}\bigg) - \frac{1}{4}\cdot log\bigg(\frac{1}{4}\bigg)\\
&= 0.81\\
H\big(Overcast\big) &= -\frac{1}{3}\cdot log\bigg(\frac{1}{3}\bigg) - \frac{2}{3}\cdot log\bigg(\frac{2}{3}\bigg)\\
&= 0.92
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
H\big(High\big) &= -\frac{2}{6}\cdot log\bigg(\frac{2}{6}\bigg) - \frac{4}{6}\cdot log\bigg(\frac{4}{6}\bigg)\\
&= 0.92\\
H\big(Rain\big) &= -\frac{4}{6}\cdot log\bigg(\frac{4}{6}\bigg) - \frac{2}{6}\cdot log\bigg(\frac{2}{6}\bigg)\\
&= 0.92
\end{aligned}
$$

$$
\begin{aligned}
H\big(\textbf{Go on a trip}, \textbf{Humidity}\big) &= P\big(High\big)H\big(High\big) + P\big(Normal\big)H\big(Normal\big)\\
&= \frac{6}{12}\cdot H\big(2,4\big) + \frac{6}{12}\cdot H\big(4,2\big)\\
&= 0.92
\end{aligned}
$$

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

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

$$\begin{aligned} H\big(String\big) &= -\frac{3}{7} \cdot log\left(\frac{3}{7}\right) - \frac{4}{7} \cdot log\left(\frac{4}{7}\right) \\ &= 0.99 \\ H\big(Weak\big) &= -\frac{4}{6} \cdot log\left(\frac{4}{6}\right) - \frac{2}{6} \cdot log\left(\frac{2}{6}\right) \\ &= 0.97 \end{aligned}$$

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

Now we shall calculate information gain for each split -

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

4 Building further tree

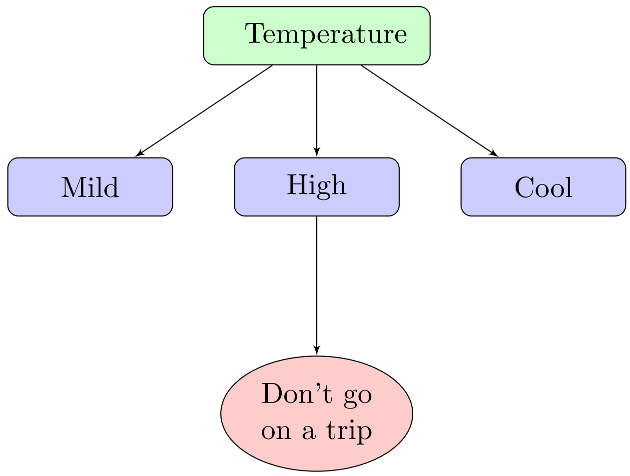
Since we have the highest information gain at the Temperature split, our first **decision** shall be regarding checking the temperature.

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, Hot temperature always results in **No** decision. In other words,

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

   Thus one of our path to *leaf nodes* is decided.



2. Now, for when the temperature is Cool, the following table results -

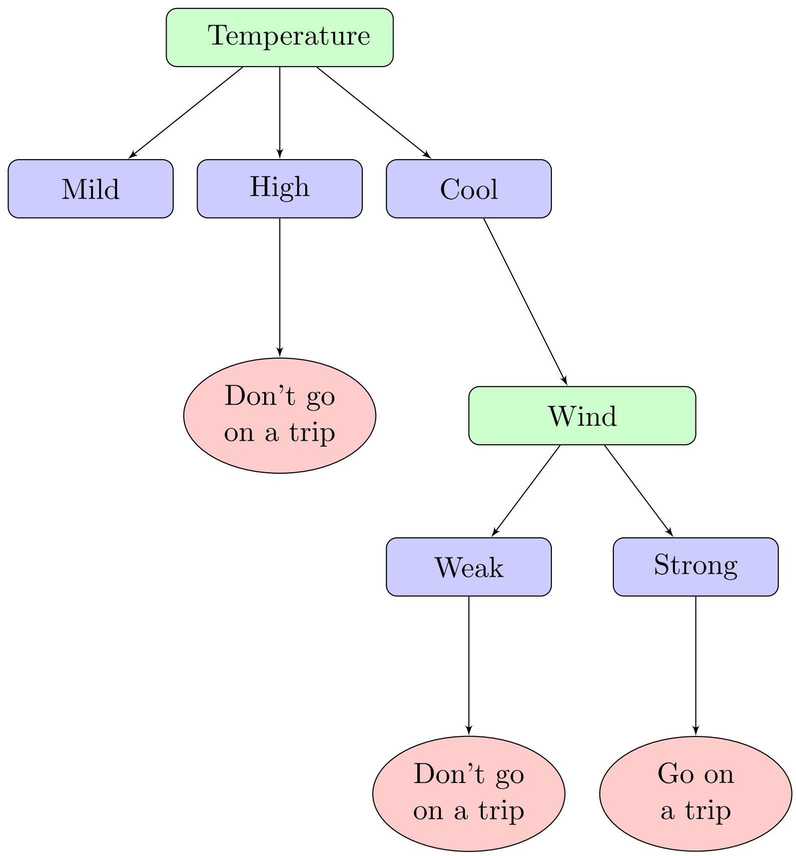
		Go on a trip		Total	Entropy
		Yes	**No**		
Wind	Weak	2	0	2	0
	Strong	0	2	2	0
Forecast	Rain	1	1	2	1
	Overcast	0	1	1	0
	Sunny	1	0	1	0
Humidity	Normal	2	2	4	1

$$\begin{aligned} H\big(\textbf{Go on a trip}, \textbf{Temperature} = \mathit{Cool}\big) &= H\big(2,2\big) \\ &= 1 \end{aligned}$$

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

$$\begin{aligned} H\big(\textbf{Go on a trip}, \textbf{Temperature} = \mathit{Cool}, \textbf{Wind}\big) &= \frac{2}{4}\cdot 0 + \frac{2}{4}\cdot 0 \\ &= 0 \\ H\big(\textbf{Go on a trip}, \textbf{Temperature} = \mathit{Cool}, \textbf{Forecast}\big) &= \frac{2}{4}\cdot 1 + \frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 0 \\ &= 0.5 \\ H\big(\textbf{Go on a trip}, \textbf{Temperature} = \mathit{Cool}, \textbf{Humidity}\big) &= \frac{1}{1}\cdot 1 \\ &= 1 \end{aligned}$$

Thus $Gain\big(\textbf{Go on a trip}, \textbf{Temperature} = Cool, \textbf{Wind}\big)$ 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 Mild, the following table results

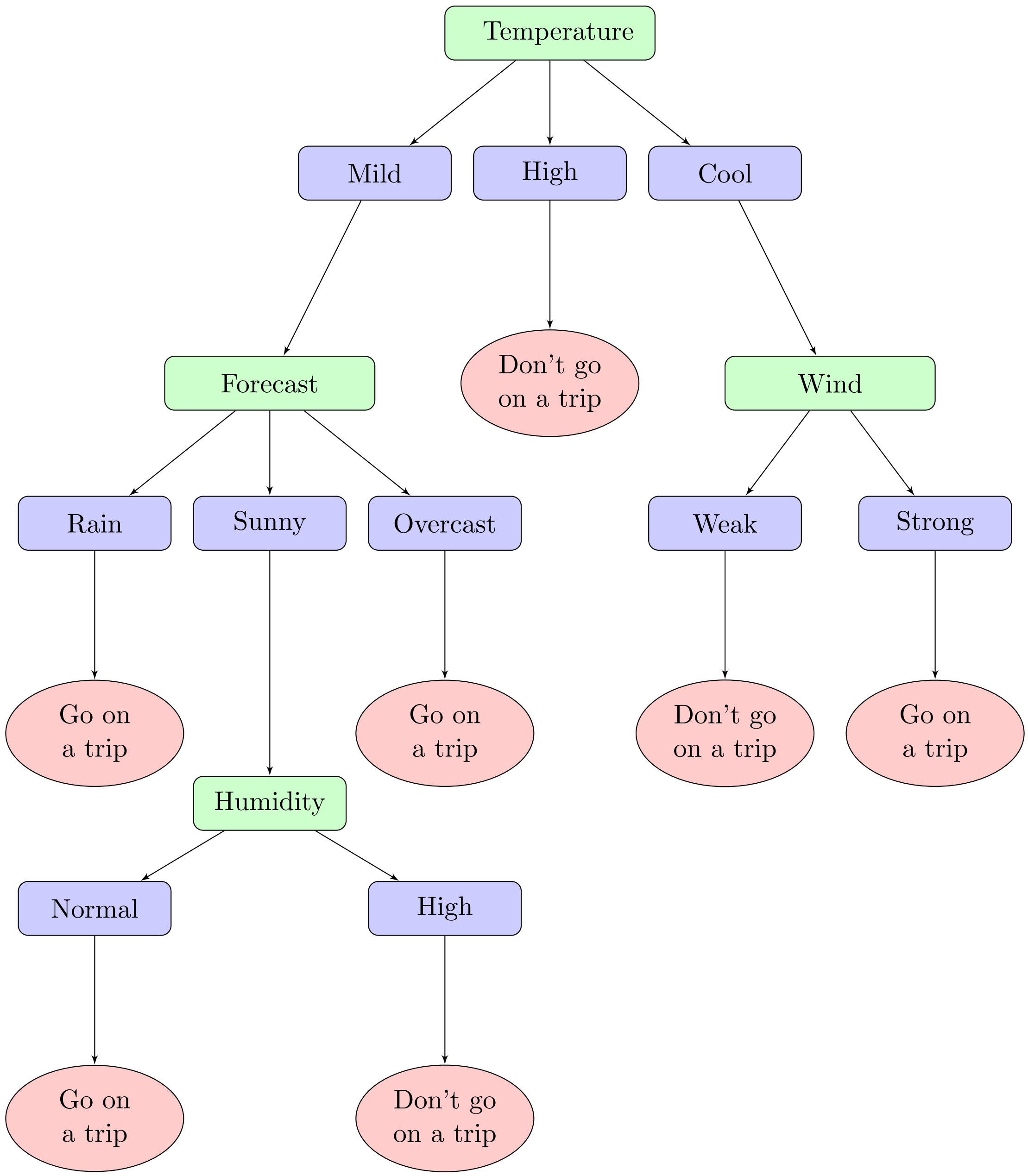
		Go on a trip		**Total**	**Entropy**
		Yes	**No**		
Wind	Weak	2	1	3	0.92
	Strong	2	0	2	0
Forecast	Rain	2	0	2	0
	Overcast	1	0	1	0
	Sunny	1	1	2	1
Humidity	Normal	2	0	2	0
	High	2	1	3	0.92

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

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.


Temperature
Mild
High
Cool
Forecast
Don't go on a trip
Wind
Rain
Sunny
Overcast
Weak
Strong
Go on a trip
Go on a trip
Don't go on a trip
Go on a trip
Humidity
Normal
High
Go on a trip
Don't go on a trip