Weather Dataset - Decision Tree Creation

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

No. of Yes	**9**
No. of No	**5**

	Entropy for Yes	$-\frac{9}{14} * \log_2 \frac{9}{14} = 0.41$
	Entropy for No	$-\frac{5}{14} * \log_2 \frac{5}{14} = 0.53$
Total Entropy of Weather dataset	**H(S)**	E(Yes) + E(No) = **0.94**

Attribute 1: Outlook

Outlook has 3 options - sunny, overcast, rainy

Option 1: sunny

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Entropy for Sunny	5 instances out of 5, 2 YES and 3 NO	$-\frac{2}{5} * \log_2\frac{2}{5} - \frac{3}{5} * \log_2\frac{3}{5} = 0.5288 + 0.4422 = 0.971$

Option 2: overcast

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Entropy for Overcast	4 instances out of 4, 4 YES and 0 NO	$-\frac{4}{4} * \log_2\frac{4}{4} - \frac{0}{4} * \log_2\frac{0}{4} = 0$

Option 3: rainy

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Entropy for Rainy	5 instances out of 5, 3 YES and 2 NO	$-\frac{3}{5} * \log_2\frac{3}{5} - \frac{2}{5} * \log_2\frac{2}{5} = 0.4422 + 0.5288 = 0.971$

Now, we have to calculate the weighted average - out of the 4 instances, 5 instances for sunny, 4 for overcast, and 5 for rainy.

Weighted Average Entropy for Outlook	$\frac{5}{14} * 0.971 + \frac{4}{14} * 0 + \frac{5}{14} * 0.971 = 0.6936$

Information Gain of Outlook attribute

Information Gain of Outlook	Total Entropy - Weighted Average Entropy of Outlook = 0.94 - 0.6936 = 0.2464

Repeat this for all remaining attributes - Windy, Temperature, and Humidity.

Attribute 2 - Windy

Windy has 2 options - True & False

Entropy for True	6 instances Out of 6, 3 YES and 3 NO	$-\frac{3}{6} * \log_2\frac{3}{6} - \frac{3}{6} * \log_2\frac{3}{6} = 0.5+0.5=1$
Entropy for False	8 instances Out of 8, 6 YES and 2 NO	$-\frac{6}{8} * \log_2\frac{6}{8} - \frac{2}{8} * \log_2\frac{2}{8} = 0.3113 + 0.5 = 0.8113$
	Weighted Average Entropy for Windy	$\frac{6}{14} * 1 + \frac{8}{14} * 0.8113 = 0.4636 + 0.4286 = 0.8922$
	Information Gain of Windy	Total Entropy - Average Entropy of Windy = 0.94 - 0.8922 = 0.0478

Attribute 3 - Temperature

Temperature has 3 options - Cool, Hot & Mild

Entropy for Cool	4 instances out of 4, 3 YES and 1 NO	$-\frac{3}{4} * \log_2\frac{3}{4} - \frac{1}{4} * \log_2\frac{1}{4} = 0.8113$
Entropy for Hot	4 instances out of 4, 2 YES and 2 NO	$-\frac{2}{4} * \log_2\frac{2}{4} - \frac{2}{4} * \log_2\frac{2}{4} = 1$
Entropy for Mild	6 instances out of 5, 4 YES and 2 NO	$-\frac{4}{6} * \log_2\frac{4}{6} - \frac{2}{6} * \log_2\frac{2}{6} = 0.9183$
	Weighted Average Entropy for Temperature	$\frac{4}{14} * 0.8113 + \frac{4}{14} * 1 + \frac{6}{14} * 0.9183 = 0.9111$
	Information Gain of Temperature	Total Entropy - Average Entropy of Windy = 0.94 - 0.9111 = 0.0289

Attribute 4 - Humidity

Humidity has 2 options - High & Normal

Entropy for High	7 instances out of 7, 3 YES and 4 NO	$-\frac{3}{7} * \log_2\frac{3}{7} - \frac{4}{7} * \log_2\frac{4}{7} = 0.9852$
Entropy for Normal	7 instances out of 7, 6 YES and 1 NO	$-\frac{6}{7} * \log_2\frac{6}{7} - \frac{1}{7} * \log_2\frac{1}{7} = 0.5918$

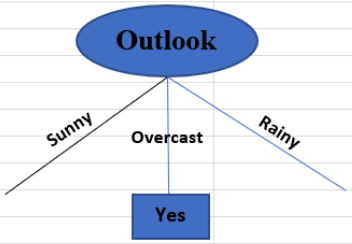
Weighted Average Entropy for Humidity	$\frac{7}{14} * 0.9852 + \frac{7}{14} * 0.5918 = 0.7885$

Information Gain of Temperature	Total Entropy - Average Entropy of Windy = 0.94 - 0.7885 = 0.1515

Summary

Attribute	Average Entropy	Information Gain
Outlook	0.6936	0.2464
Temperature	0.9111	0.0289
Humidity	0.7885	0.1515
Windy	0.8922	0.0478

Highest Information Gain



Level 2

Subset 1

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Subset 2

Outlook	Temperat	Humidity	Windy	Play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes

Subset 3

Outlook	Temperati	Humidity	Windy	Play
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no

Continue the same process for subsets to create subtrees for sunny and rainy options. As all cases of overcast are yes, we can draw the leaf there.