Play Tennis: Decision Tree - Complete example

Machine Learning 2024-25 Course Activity

Furno Francesco - francesco.furno@studenti.unipd.it - 2139507

November 1st, 2024

Table of contents

Abstract

Given the "Play Tennis" dataset provided by the professor during the lecture on October 14th, 2024, Decision Tree - Part I, we aim to build a decision tree using the ID3 algorithm. During this process, to provide a complete example, we calculated all the possible attributes, even when it was clear which attribute was optimal to choose.

Dataset

The dataset provided is the following:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Let's start calculating the **Entropy of PlayTennis**. Over 14 days:

- 9 "Yes" $\rightarrow P(\text{Yes}) = \frac{9}{14}$
- 5 "No" $\rightarrow P(\text{No}) = \frac{5}{14}$

$$\begin{aligned} E(\text{PlayTennis}) &= -\left(\log\left(\frac{9}{14}\right) \times \frac{9}{14} + \log\left(\frac{5}{14}\right) \times \frac{5}{14}\right) \\ &\approx -(0.643 \times (-0.644) + 0.357 \times (-1.485)) \\ &\approx 0.414 + 0.530 \\ &\approx 0.940 \end{aligned}$$

Remember that the ID3 algorithm uses the concepts of **Entropy** and **Information Gain**: for each possible node of the tree, we are looking for the attribute wich maximizes the Information Gain.

Root node

"Let's begin by calculating the Information Gain for all possible attributes at the root node.

Outlook:
There are 3 possible values for the Outlook attribute: "Sunny", "Overcast" and "Rain".

Let's start with the "Sunny" value. There are 5 occurrences of it:

Day	PlayTennis
D1	No
D2	No
D8	No
D9	Yes
D11	Yes

In total:

- 2 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{2}{5}$
- 3 occurrences of $\text{PlayTennis}_N \Rightarrow \frac{3}{5}$

Then:

$$\begin{aligned} E(\text{Outlook}_S) &= -\left(\log\left(\frac{2}{5}\right) \times \frac{2}{5} + \log\left(\frac{3}{5}\right) \times \frac{3}{5}\right) \\ &\approx 0.970 \end{aligned}$$

Let's analyze the "Overcast" value. There are 4 occurrences of it:

Day	PlayTennis
D3	Yes
D7	Yes

Day	PlayTennis
D12	Yes
D13	Yes

In total:

- 4 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{4}{4} = 1$
- 0 occurrences of $\text{PlayTennis}_N \Rightarrow 0$

Then:

$$E(\text{Outlook}_O) = 0$$

This particular Entropy is easy to calculate because the values are not distributed.

Let's analyze the "Rain" value. There are 5 occurrences of it:

Day	PlayTennis
D4	Yes
D5	Yes
D6	No
D10	Yes
D14	No

In total:

- 3 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{3}{5}$
- 2 occurrences of $\text{PlayTennis}_N \Rightarrow \frac{2}{5}$

Then:

$$\begin{aligned} E(\text{Outlook}_R) &= -\left(\log\left(\frac{3}{5}\right) \times \frac{3}{5} + \log\left(\frac{2}{5}\right) \times \frac{2}{5}\right) \\ &\approx 0.970 \end{aligned}$$

So the **Information Gain** from the "Outlook" attribute is:

$$\begin{aligned} G(S, \text{Outlook}) &= 0.940 - \left(\frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97\right) \\ &\approx 0.248 \end{aligned}$$

Temperature:
There are 3 possible values for the Temperature attribute: "Hot", "Mild" and "Cool".

Let's start with the "Hot" value. There are 4 occurrences of it:

Day	PlayTennis
D1	No
D2	No
D3	Yes

Day	PlayTennis
D13	Yes

In total:

- 2 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{2}{4} = \frac{1}{2}$
- 2 occurrences of $\text{PlayTennis}_N \Rightarrow \frac{2}{4} = \frac{1}{2}$

$$E(\text{Temperature}_H) = -\left(\log\left(\frac{1}{2}\right) \times \frac{1}{2} + \log\left(\frac{1}{2}\right) \times \frac{1}{2}\right)$$
$$\approx 1$$

This particular Entropy is easy to calculate because the values are perfectly distributed.

Let's analyze the "Mild" value. There are 6 occurrences of it:

Day	PlayTennis
D4	Yes
D8	No
D10	Yes
D11	Yes
D12	Yes
D14	No

In total:

- 4 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{4}{6} = \frac{2}{3}$
- 2 occurrences of $\text{PlayTennis}_N \Rightarrow \frac{2}{6} = \frac{1}{3}$

$$E(\text{Temperature}_M) = -\left(\log\left(\frac{2}{3}\right) \times \frac{2}{3} + \log\left(\frac{1}{3}\right) \times \frac{1}{3}\right)$$
$$\approx 0.918$$

Let's analyze the "Cool" value. There are 4 occurrences of it:

Day	PlayTennis
D5	Yes
D6	No
D7	Yes
D9	Yes

In total:

- 3 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{3}{4} = \frac{3}{4}$
- 1 occurrence of $\text{PlayTennis}_N \Rightarrow \frac{1}{4} = \frac{1}{4}$

$$E(\text{Temperature}_C) = -\left(\log\left(\frac{3}{4}\right) \times \frac{3}{4} + \log\left(\frac{1}{4}\right) \times \frac{1}{4}\right)$$
$$\approx 0.811$$

So the **Information Gain** from the "Temperature" attribute is:

$$\begin{aligned} G(S, \text{Temperature}) &= 0.940 - \left(\frac{4}{14} \times 1 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.811\right) \\ &= 0.030 \end{aligned}$$

Humidity:
There are 2 possible values for the Humidity attribute: "High" and "Normal".

Let's start with the "High" value. There are 7 occurrences of it:

Day	PlayTennis
D1	No
D2	No
D3	Yes
D4	Yes
D8	No
D12	Yes
D14	No

In total:

- 3 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{3}{7}$
- 4 occurrences of $\text{PlayTennis}_N \Rightarrow \frac{4}{7}$

Then:

$$\begin{aligned} E(\text{Humidity}_H) &= -\left(\log\left(\frac{3}{7}\right) \times \frac{3}{7} + \log\left(\frac{4}{7}\right) \times \frac{4}{7}\right) \\ &\approx 0.985 \end{aligned}$$

Let's analyze the "Normal" value. There are 7 occurrences of it:

Day	PlayTennis
D5	Yes
D6	No
D7	Yes
D9	Yes
D10	Yes
D11	Yes
D13	Yes

In total:

- 6 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{6}{7}$
- 1 occurrence of $\text{PlayTennis}_N \Rightarrow \frac{1}{7}$

Then:

$$E(\text{Humidity}_N) = -\left(\log\left(\frac{6}{7}\right) \times \frac{6}{7} + \log\left(\frac{1}{7}\right) \times \frac{1}{7}\right)$$
$$\approx 0.591$$

So the **Information Gain** from the "Humidity" attribute is:

$$G(S, \text{Humidity}) = 0.940 - \left(\frac{7}{14} \times 0.985 + \frac{7}{14} \times 0.591\right)$$
$$= 0.152$$

Wind:
There are 2 possible values for the Wind attribute: "Weak" and "Strong"

Let's start with the "Weak" value. There are 8 occurrences of it:

Day	PlayTennis
D1	No
D3	Yes
D4	Yes
D5	Yes
D8	No
D9	Yes
D10	Yes
D13	Yes

In total:

- 6 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{6}{8} = \frac{3}{4}$
- 2 occurrences of $\text{PlayTennis}_N \Rightarrow \frac{2}{8} = \frac{1}{4}$

Then:

$$E(\text{Wind}_W) = -\left(\log\left(\frac{3}{4}\right) \times \frac{3}{4} + \log\left(\frac{1}{4}\right) \times \frac{1}{4}\right)$$
$$\approx 0.811$$

Let's analyze the "Strong" value. There are 6 occurrences of it:

Day	PlayTennis
D2	No
D6	No
D7	Yes
D11	Yes
D12	Yes
D14	No

In total:

- 3 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{3}{6} = \frac{1}{2}$

- 3 occurrences of $\text{PlayTennis}_N \Rightarrow \frac{3}{6} = \frac{1}{2}$

Then:

$$\begin{aligned} E(\text{Wind}_S) &= -\left(\log\left(\frac{1}{2}\right) \times \frac{1}{2} + \log\left(\frac{1}{2}\right) \times \frac{1}{2}\right) \\ &= 1 \end{aligned}$$

So the Information Gain from the attribute "Wind" is:

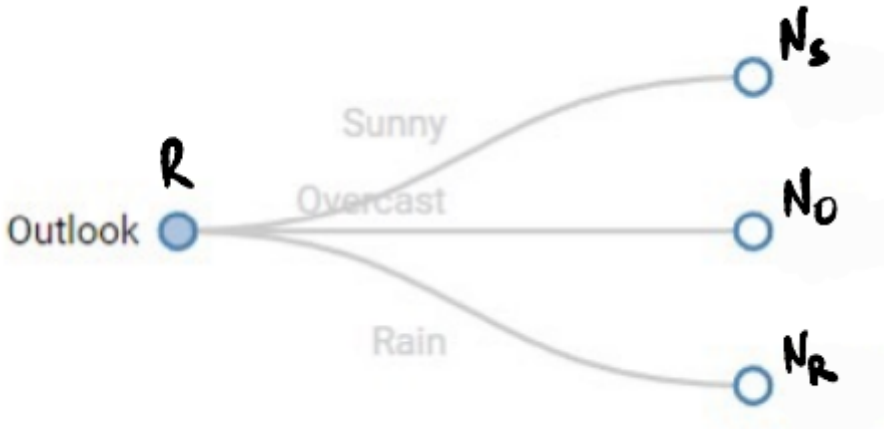
$$\begin{aligned} G(S, \text{Wind}) &= 0.940 - \left(\frac{8}{14} \times 0.811 + \frac{6}{14} \times 1\right) \\ &= 0.048 \end{aligned}$$

Conclusion:
Let's summarize all the I.G. for all the possible attributes of the root node:

Attribute	Information Gain
Outlook	0.248
Temperature	0.030
Humidity	0.152
Wind	0.048

The root node of the Decision Tree should have the "Outlook" attribute, which is the optimal attribute. The S set is partitioned according to the values that Outlook can take:

- N_S is the sub-tree where $S_{\text{Outlook = Sunny}}$
- N_O is the sub-tree where $S_{\text{Outlook = Overcast}}$
- N_R is the sub-tree where $S_{\text{Outlook = Rain}}$



N_S - Node "Sunny"

N_S is the sub-tree where $S_{\text{Outlook = Sunny}}$.

We have already calculated the Entropy: $E(\text{Outlook}_S) = 0.970$.

Temperature:
There are 3 possible values for the Temperature attribute: “Hot”, “Mild” and “Cool”.

Let’s start with the “Hot” value. There are 2 occurrences of it:

Day	PlayTennis
D1	No
D2	No

In total:

- 0 occurrences of $\text{PlayTennis}_Y \Rightarrow 0$
- 2 occurrences of $\text{PlayTennis}_N \Rightarrow \frac{2}{2} = 1$

$$E\Big(\text{Temperature}_{N_S-H}\Big) = 0$$

Let’s analyze the “Mild” value. There are 2 occurrences of it:

Day	PlayTennis
D8	No
D11	Yes

In total:

- 1 occurrence of $\text{PlayTennis}_Y \Rightarrow \frac{1}{2}$
- 1 occurrence of $\text{PlayTennis}_N \Rightarrow \frac{1}{2}$

$$E\Big(\text{Temperature}_{N_S-M}\Big) = 0.5$$

Let’s analyze the “Cool” value. There is 1 occurrence of it:

Day	PlayTennis
D9	Yes

In total:

- 1 occurrence of $\text{PlayTennis}_Y \Rightarrow 1$
- 0 occurrences of $\text{PlayTennis}_N \Rightarrow 0$

$$E\Big(\text{Temperature}_{N_S-C}\Big) = 0$$

So the **Information Gain** from the “Temperature” attribute is:

$$\begin{aligned} G(S_{O=S}, \text{Temperature}) &= 0.970 - \left(\frac{2}{5} \times 0 + \frac{2}{5} \times 0.5 + \frac{1}{5} \times 0\right) \\ &= 0.770 \end{aligned}$$

Humidity:
There are 2 possible values for the Humidity attribute: “High”, “Normal”.

Let’s start with the “High” value. There are 3 occurrences of it:

Day	PlayTennis
D1	No
D2	No
D8	No

In total:

- 0 occurrences of $\text{PlayTennis}_Y \Rightarrow 0$
- 3 occurrences of $\text{PlayTennis}_N \Rightarrow 1$

Then:

$$E\Big(\text{Humidity}_{N_S-H}\Big) = 0$$

Let's analyze the "Normal" value. There are 2 occurrences of it:

Day	PlayTennis
D9	Yes
D11	Yes

In total:

- 2 occurrences of $\text{PlayTennis}_Y \Rightarrow 1$
- 0 occurrences of $\text{PlayTennis}_N \Rightarrow 0$

Then:

$$E\Big(\text{Humidity}_{N_S-N}\Big) = 0$$

So the **Information Gain** from the "Humidity" attribute is:

$$\begin{aligned} G(S_{O=S}, \text{Humidity}) &= 0.970 - \left(\frac{3}{5} \times 0 + \frac{2}{5} \times 0\right) \\ &= 0.970 \end{aligned}$$

Wind:
There are 2 possible values for the attribute Wind: "Weak" and "Strong".

Let's start with the "Weak" value. There are 3 occurrences of it:

Day	PlayTennis
D1	No
D8	No
D9	Yes

In total:

- 1 occurrence of $\text{PlayTennis}_Y \Rightarrow \frac{1}{3}$
- 2 occurrences of $\text{PlayTennis}_N \Rightarrow \frac{2}{3}$

Then:

$$E\left(\text{Wind}_{N_S-W}\right) = -\left(\log\left(\frac{1}{3}\right) \times \frac{1}{3} + \log\left(\frac{2}{3}\right) \times \frac{2}{3}\right)$$
$$\approx 0.918$$

Let's analyze the "Strong" value. There are 2 occurrences of it:

Day	PlayTennis
D2	No
D11	Yes

In total:

- 1 occurrence of $\text{PlayTennis}_Y \Rightarrow \frac{1}{2}$
- 1 occurrence of $\text{PlayTennis}_N \Rightarrow \frac{1}{2}$

Then:

$$E\left(\text{Wind}_{N_S-S}\right) = 1$$

So the **Information Gain** from the attribute "Wind" is:

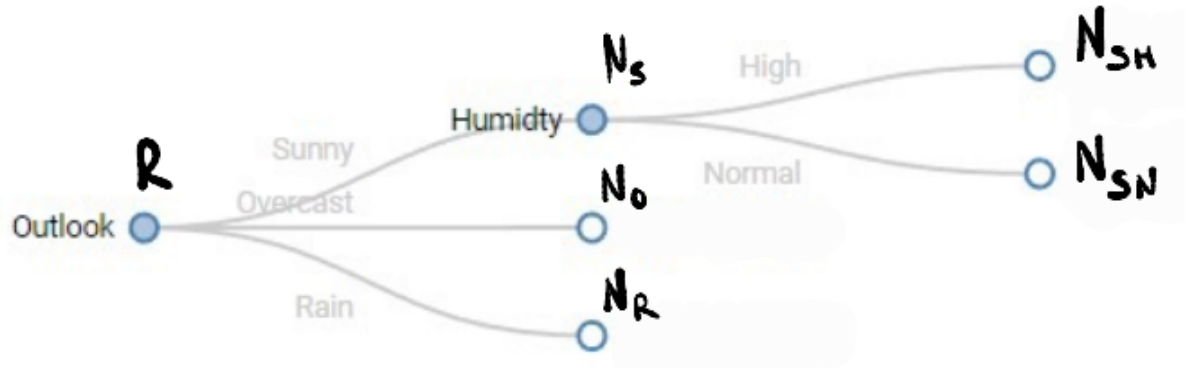
$$G(S_{O=S}, \text{Wind}) = 0.970 - \left(\frac{3}{5} \times 0.918 + \frac{2}{5} \times 1\right)$$
$$= 0.019$$

Conclusion:
Let's summarize all the I.G. for all the possible attributes of the Node N_S:

Attribute	Information Gain
Temperature	0.770
Humidity	0.970
Wind	0.019

The node N_S should have attribute "Humidity", which is the optimal attribute. The $S_{O=S}$ set is partitioned according to the values that Humidity can take:

- N_{SH} is the sub-tree where $S_{\text{Outlook = Sunny} \wedge \text{Humidity = High}}$
- N_{SN} is the sub-tree where $S_{\text{Outlook = Sunny} \wedge \text{Humidity = Normal}}$

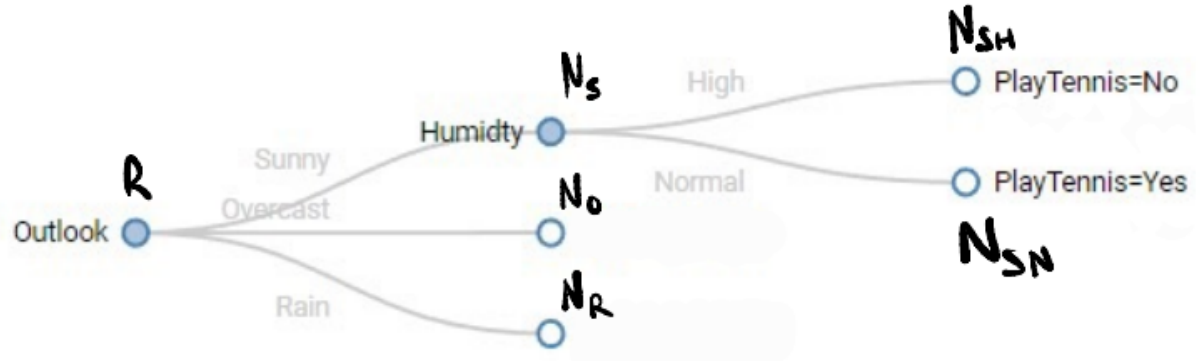


N_{SH} - Node "Sunny - High"

N_{SH} is the sub-tree where $S_{\text{Outlook = Sunny} \wedge \text{Humidity = High}}$.
We have already calculated the Entropy: $E(\text{Outlook}_{SH}) = 0$. $S_{O=S \wedge H=H}$ is pure: all the occurrences in this set belong to the class PlayTennis = No. The node N_{SH} is a leaf with the class **No** .

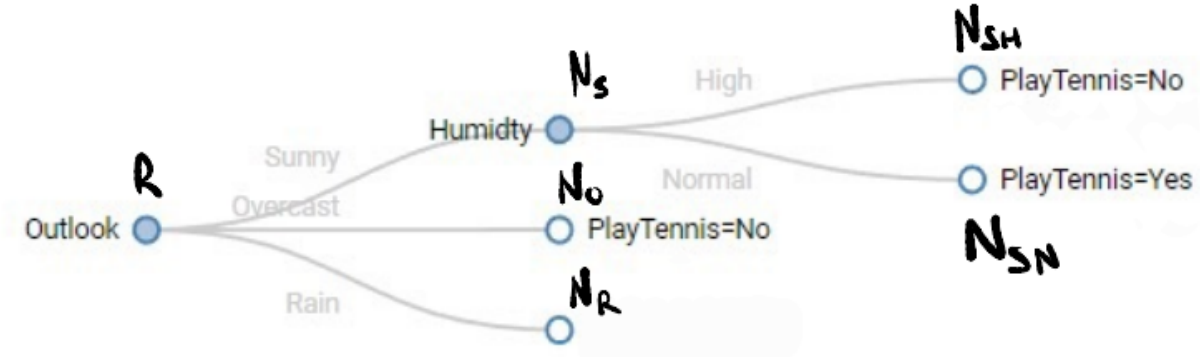
N_{SN} - Node "Sunny - Normal"

N_{SN} is the sub-tree where $S_{\text{Outlook = Sunny} \wedge \text{Humidity = Normal}}$.
We have already calculated the Entropy: $E(\text{Outlook}_{SN}) = 0$. $S_{O=S \wedge H=N}$ is pure: all the occurrences in this set belong to the class PlayTennis = Yes. The node N_{SN} is a leaf with the class **Yes** .



N_O - Node "Overcast"

N_O is the sub-tree where $S_{\text{Outlook = Overcast}}$.
We have already calculated the Entropy: $E(\text{Outlook}_O) = 0$. $S_{O=O}$ is pure: all the occurrences in this set belong to the class PlayTennis = Yes. The node N_{SH} is a leaf with the class **Yes** .



N_R - Node "Rain"

N_R is the sub-tree where $S_{\text{Outlook = Rain}}$.
We have already calculated the Entropy: $E(\text{Outlook}_R) = 0.970$.

Temperature:
There are 3 possible values for the Temperature attribute: "Hot", "Mild" and "Cool".

Let's start with the "Hot" value. There are no occurrences for it:

$$E\left(\text{Temperature}_{N_R - H}\right) = \emptyset$$

Let's analyze the "Mild" value. There are 3 occurrences of it:

Day	PlayTennis
D4	Yes
D10	Yes
D14	No

In total:

- 2 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{2}{3}$
- 1 occurrence of $\text{PlayTennis}_N \Rightarrow \frac{1}{3}$

$$E\Big(\text{Temperature}_{N_{R-M}}\Big) = -\left(\log\left(\frac{2}{3}\right) \times \frac{2}{3} + \log\left(\frac{1}{3}\right) \times \frac{1}{3}\right)$$
$$\approx 0.918$$

Let's analyze the "Cool" value. There are 2 occurrences of it:

Day	PlayTennis
D5	Yes
D6	No

In total:

- 1 occurrence of $\text{PlayTennis}_Y \Rightarrow \frac{1}{2}$
- 1 occurrence of $\text{PlayTennis}_N \Rightarrow \frac{1}{2}$

$$E\Big(\text{Temperature}_{N_{R-C}}\Big) = 0.5$$

So the **Information Gain** from the attribute "Temperature" is:

$$G(S_{O=R}, \text{Temperature}) = 0.970 - \left(\frac{3}{5} \times 0.918 + \frac{2}{5} \times 0.5\right)$$
$$= 0.219$$

Humidity:
There are 2 possible values for the Humidity attribute: "High" and "Normal".

Let's start with the "High" value. There are 2 occurrences of it:

Day	PlayTennis
D4	Yes
D14	No

In total:

- 1 occurrence of $\text{PlayTennis}_Y \Rightarrow \frac{1}{2}$
- 1 occurrence of $\text{PlayTennis}_N \Rightarrow \frac{1}{2}$

Then:

$$E\Big(\text{Humidity}_{N_R-H}\Big) = 0.5$$

Let's analyze the "Normal" value. There are 3 occurrences of it:

Day	PlayTennis
D5	Yes
D6	No
D10	Yes

In total:

- 2 occurrences of $\text{PlayTennis}_Y \Rightarrow \frac{2}{3}$
- 1 occurrence of $\text{PlayTennis}_N \Rightarrow \frac{1}{3}$

Then:

$$\begin{aligned} E\Big(\text{Humidity}_{N_R-N}\Big) &= -\left(\log\left(\frac{2}{3}\right) \times \frac{2}{3} + \log\left(\frac{1}{3}\right) \times \frac{1}{3}\right) \\ &\approx 0.918 \end{aligned}$$

So the **Information Gain** from the "Humidity" attribute is:

$$\begin{aligned} G(S_{O=R}, \text{Humidity}) &= 0.970 - \left(\frac{2}{5} \times 0.5 + \frac{3}{5} \times 0.918\right) \\ &= 0.219 \end{aligned}$$

Wind:
There are 2 possible values for the attribute Wind: "Weak" and "Strong".

Let's start with the "Weak" value. There are 3 occurrences of it:

Day	PlayTennis
D4	Yes
D5	Yes
D10	Yes

In total:

- 3 occurrences of $\text{PlayTennis}_Y \Rightarrow 1$
- 0 occurrences of $\text{PlayTennis}_N \Rightarrow 0$

Then:

$$E\Big(\text{Wind}_{N_R-W}\Big) = 0$$

Let's analyze "Strong" value. There is 2 occurrence of the value "Strong":

Day	PlayTennis
D6	No
D14	No

In total:

- 0 occurrence of $\text{PlayTennis}_Y \Rightarrow 0$
- 2 occurrences of $\text{PlayTennis}_N \Rightarrow 1$

Then:

$$E\left(\text{Wind}_{N_R - S}\right) = 0$$

So the **Information Gain** from the attribute "Wind" is:

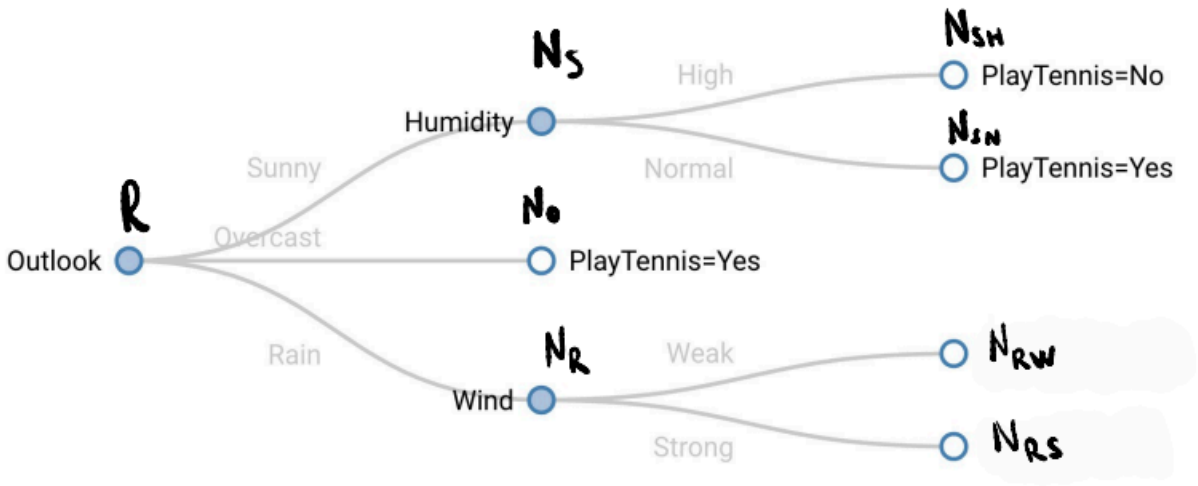
$$\begin{aligned} G(S_{O=R}, \text{Wind}) &= 0.970 - \left(\frac{3}{5} \times 0 + \frac{2}{5} \times 0\right) \\ &= 0.970 \end{aligned}$$

Conclusion:
Let's summarize all the I.G. for all the possible attributes of the Node N_R:

Attribute	Information Gain
Temperature	0.219
Humidity	0.219
Wind	0.970

The node N_R should have attribute "Wind", which is the optimal attribute. The $S_{O=R}$ set is partitioned according to the values that Humidity can take:

- N_{RW} is the sub-tree where $S_{\text{Outlook = Rain} \wedge \text{Wind = Weak}}$
- N_{RS} is the sub-tree where $S_{\text{Outlook = Rain} \wedge \text{Wind = Strong}}$



N_{RW} - Node "Rain - Weak"

N_{RW} is the sub-tree where $S_{\text{Outlook = Rain} \wedge \text{Wind = Weak}}$
We have already calculated the Entropy: $E(\text{Wind}_{RW}) = 0$. $S_{O=R \wedge W=W}$ is pure: all the occurrences in this set belong to the class PlayTennis = Yes. The node N_{RW} is a leaf with the class **Yes**

N_{RS} - Node "Rain - Strong"

N_{RS} is the sub-tree where $S_{\text{Outlook = Rain} \wedge \text{ Wind = Strong}}$
We have already calculated the Entropy: $E(\text{Wind}_{RS}) = 0$. $S_{O=R \wedge W=S}$ is pure: all the occurrences in this set belong to the class PlayTennis = No. The node N_{RS} is a leaf with the class **No**

Complete Decision Tree