Building a Decision Tree (Example)

Dataset

Training examples: 9 yes / 5 no

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

The dataset features the weather conditions for the past two weeks and whether or not John played tennis.

Task: Predict if John will play tennis given that:

New data:

D15	Rain	High	Weak	?

by building a decision tree.

Recall: ID3 Algorithm

Suppose feature A is the **best attribute** to split on.

- Split entire training set on attribute A
- For each subset/ child node:
  - If subset is pure: stop
  - Else: split subset

Recall: Entropy

$$E(S) = -p_{yes}\log_2 p_{yes} - p_{no}\log_2 p_{no}$$

S – subset of training examples

p_{yes} – proportion of positive (yes) examples

p_{no} – proportion of negative (no) examples

Recall: Information Gain

Information Gain – expected reduction in entropy after a split on an attribute

$$Gain(S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v)$$

A – Attribute

$Values(A)$ – possible values of A

S – subset of training examples

S_v – subset of S for which attribute A have value v

1. Calculate the entropy of the entire training set

S: (9 Yes / 5 No)

$$E(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.940$$

Training examples: 9 yes / 5 no

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Step 1:

Calculate the Information Gain of each feature (Outlook, Humidity, Wind) for the entire dataset

Humidity

Values(Humidity) = {High, Normal}

S: (9 Yes / 5 No)
|S| = 14
E(S) = 0.940

$$
\begin{tikzcd}
& \boxed{\text{Humidity}} \arrow[dl, "\text{High}"'] \arrow[dr, "\text{Normal}"] & \\
\; & & \;
\end{tikzcd}
$$

S_{High}: (3 Yes / 4 No)

$|S_{High}| = 7$

$$E(\,S_{High}) = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}$$
$$= 0.985$$

S_{Normal}: (6 Yes / 1 No)

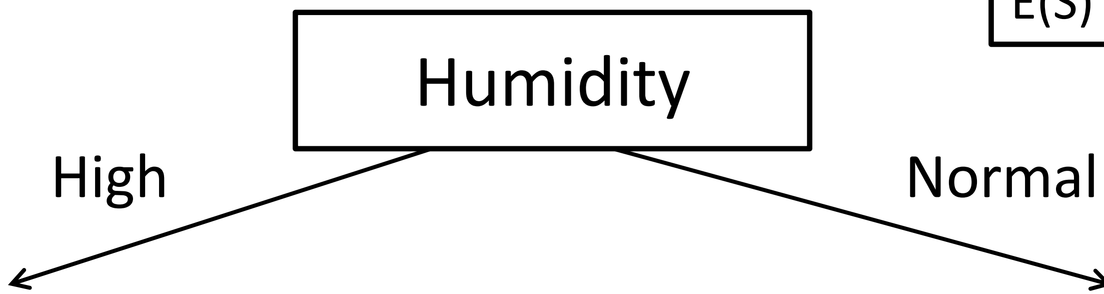
$|S_{Normal}| = 7$

$$E(\,S_{Normal}) = -\frac{6}{7}\log\frac{6}{7} - \frac{1}{7}\log\frac{1}{7}$$
$$= 0.592$$

Humidity

Values(Humidity) = {High, Normal}

S: (9 Yes / 5 No)
|S| = 14
E(S) = 0.940



S_{High}: (3 Yes / 4 No)

$|S_{High}| = 7$

$$E(S_{High}) = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}$$
$$= 0.985$$

S_{Normal}: (6 Yes / 1 No)

$|S_{Normal}| = 7$

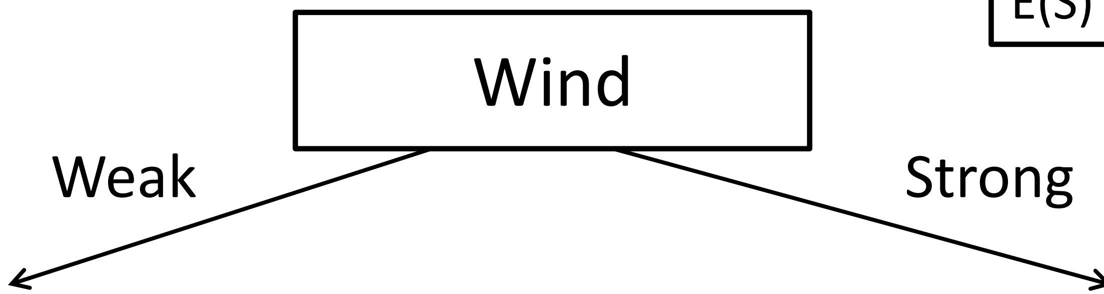
$$E(S_{Normal}) = -\frac{6}{7}\log\frac{6}{7} - \frac{1}{7}\log\frac{1}{7}$$
$$= 0.592$$

$$Gain(S, Humidity) = E(S) - \frac{|S_{High}|}{|S|}E(S_{High}) - \frac{|S_{Normal}|}{|S|}E(S_{Normal})$$
$$= 0.940 - \frac{7}{14}0.985 - \frac{7}{14}0.592 = \boxed{\mathbf{0.151}}$$

Wind

Values(Wind) = {Weak, Strong}

S: (9 Yes / 5 No)
|S| = 14
E(S) = 0.940



S_{Weak}: (6 Yes / 2 No)

$|S_{Weak}| = 8$

$$E(\, S_{Weak}) = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}$$
$$= 0.811$$

S_{Strong}: (3 Yes / 3 No)

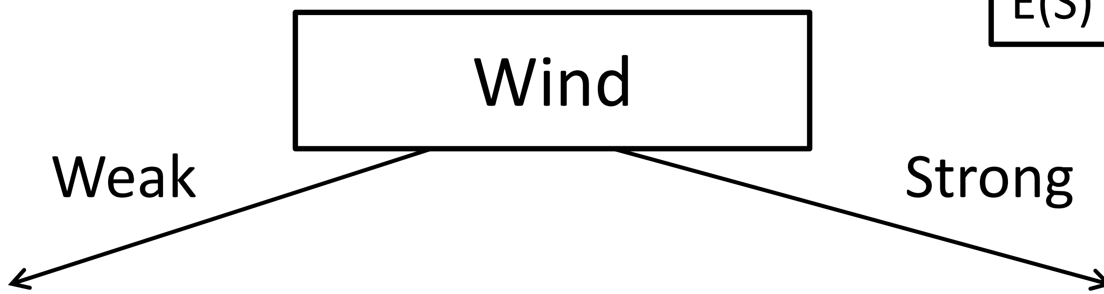
$|S_{Strong}| = 6$

$$E\left(\, S_{Strong}\right) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6}$$
$$= 1.00$$

Wind

Values(Wind) = {Weak, Strong}

S: (9 Yes / 5 No)
|S| = 14
E(S) = 0.940



S_{Weak}: (6 Yes / 2 No)

$|S_{Weak}| = 8$

$$E(\,S_{Weak}) = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}$$
$$= 0.811$$

S_{Strong}: (3 Yes / 3 No)

$\left|S_{Strong}\right| = 6$

$$E\left(\,S_{Strong}\right) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6}$$
$$= 1.00$$

$$Gain(S, Humidity) = E(S) - \frac{|S_{Weak}|}{|S|}E(S_{Weak}) - \frac{\left|S_{Strong}\right|}{|S|}E\left(S_{Strong}\right)$$

$$= \ 0.940\ - \frac{8}{14}0.811 - \frac{6}{14}1.00 = \boxed{\mathbf{0.048}}$$

Outlook

Values(Outlook) = {Sunny, Overcast, Rain}

S: (9 Yes / 5 No)
|S| = 14
E(S) = 0.940

Outlook

Sunny

Overcast

Rain

S_{Sunny}: (2 Yes / 3 No)

$|S_{Sunny}| = 5$

$$E(\,S_{Sunny}) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}$$
$$= 0.971$$

S_{Rain}: (3 Yes / 2 No)

$|S_{Rain}| = 5$

$$E(\,S_{Rain}) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}$$
$$= 0.971$$

$S_{Overcast}$: (4 Yes / 0 No)

$|S_{Overcast}| = 4$

$$E(\,S_{Overcast}) = -\frac{4}{4}\log\frac{4}{4} - \frac{0}{4}\log\frac{0}{4}$$
$$= 0$$

Outlook

Values(Outlook) = {Sunny, Overcast, Rain}

S: (9 Yes / 5 No)
|S| = 14
E(S) = 0.940

$$Gain(S, Outlook)$$
$$= E(S) - \frac{|S_{Sunny}|}{|S|} E(S_{Sunny}) - \frac{|S_{Overcast}|}{|S|} E(S_{Overcast}) - \frac{|S_{Rain}|}{|S|} E(S_{Rain})$$

$$Gain(S, Outlook) = 0.940 - \frac{5}{14}(0.971) - \frac{4}{14}(0) - \frac{5}{14}(0.971)$$

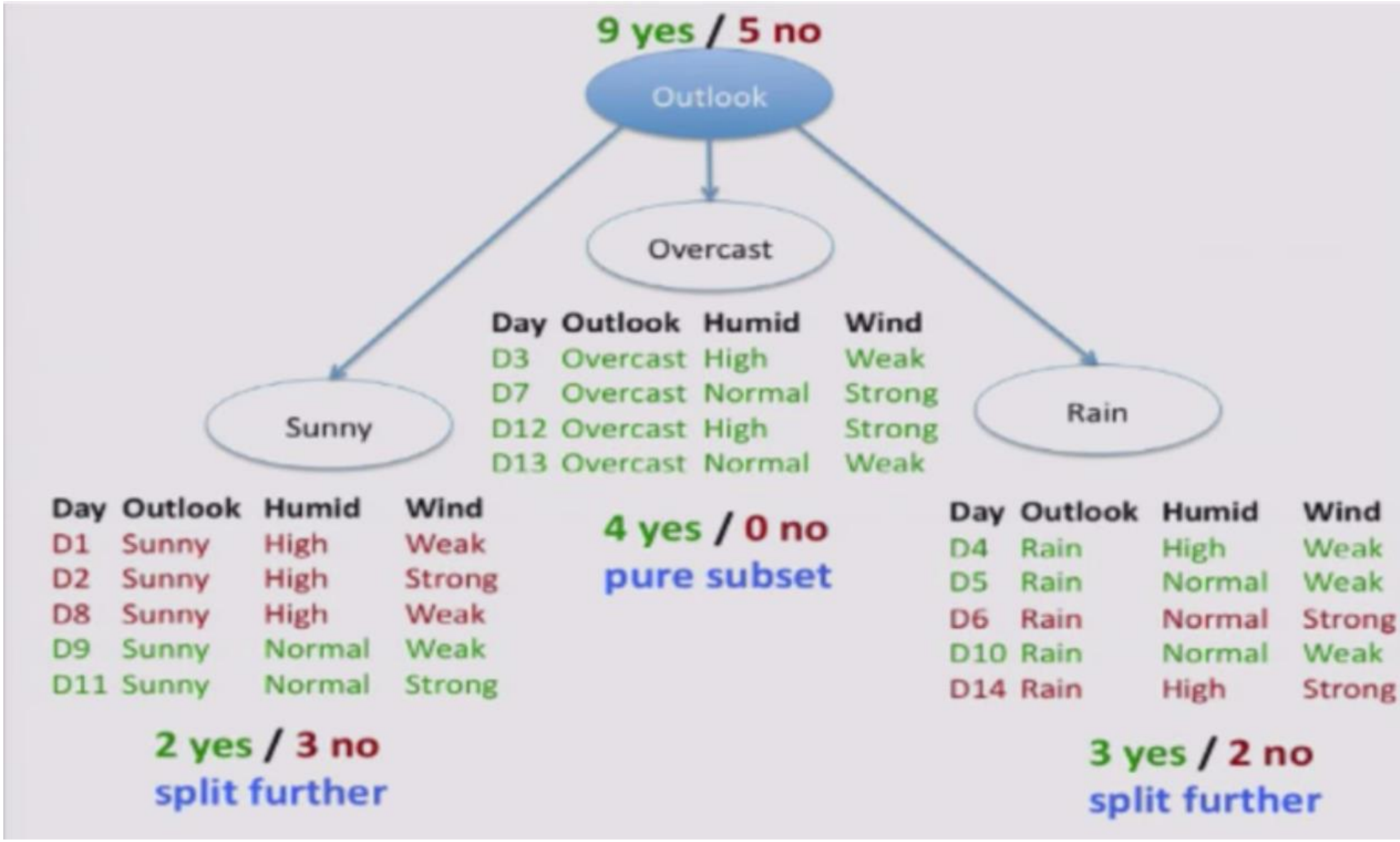
$$= \boxed{\mathbf{0.246}}$$

- Gain(S, Humidity) = 0.151
- Gain(S, Wind) = 0.048
- Gain(S, Outlook) = 0.246

Outlook has the great Information Gain!

We therefore split the dataset, according to Outlook, into three subsets:

- S_{Sunny} : (2 Yes / 3 No)
- $S_{Overcast}$: (4 Yes / 0 No) → Pure Subset!
- S_{Rain}: (3 Yes / 2 No)



Since $S_{Overcast}$ is pure, no need to split further.

S_{Sunny} and S_{Rain} are not pure and requires further splitting.

Step 2:

Calculate the Information Gain of each of the remaining features (Humidity, Wind) for S_{Sunny} and S_{Rain}.

Let's start with S_{Sunny}.

Recall that:

S_{Sunny}: (2 Yes / 3 No)

$|S_{Sunny}| = 5$

$E(\, S_{Sunny}) = 0.971$

Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

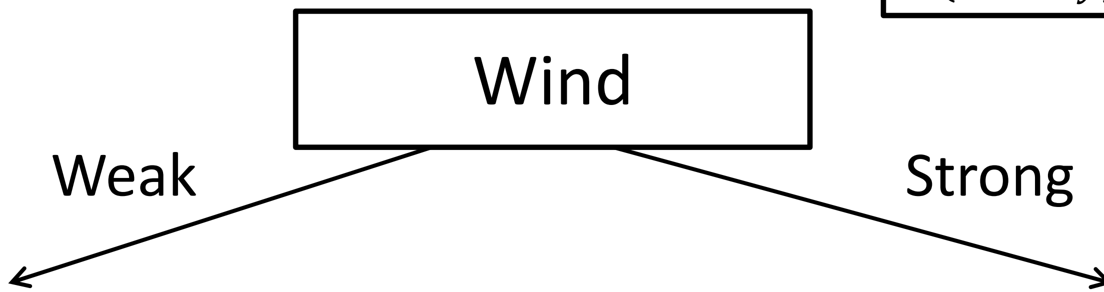
2 yes / 3 no

Let's start with S_{Sunny}.

Wind

Values(Wind) = {Weak, Strong}

S_{Sunny}: (2 Yes / 3 No)
$|S_{Sunny}| = 5$
$E(S_{Sunny}) = 0.971$



S_{Weak}: (1 Yes / 2 No)
$|S_{Weak}| = 3$
$E(S_{Weak}) = 0.918$

S_{Strong}: (1 Yes / 1 No)
$|S_{Strong}| = 2$
$E(S_{Strong}) = 1.00$

$$Gain(S_{Sunny}, Humidity) = E(S_{Sunny}) - \frac{|S_{Weak}|}{|S|} E(S_{Weak}) - \frac{|S_{Strong}|}{|S|} E(S_{Strong})$$

$$= 0.971 - \frac{3}{5} 0.918 - \frac{2}{5} 1.00 = \boxed{\mathbf{0.020}}$$

Let's start with S_{Sunny}.

Humidity

Values(Humidity) = {High, Normal}

S_{Sunny}: (2 Yes / 3 No)
$|S_{Sunny}| = 5$
$E(S_{Sunny}) = 0.971$

Humidity

High

Normal

S_{High}: (0 Yes / 3 No)
$|S_{High}| = 3$
$E(S_{High}) = 0$

S_{Normal}: (2 Yes / 0 No)
$|S_{Normal}| = 2$
$E(S_{Normal}) = 0$

$$Gain(S_{Sunny}, Humidity) = E(S_{Sunny}) - \frac{|S_{High}|}{|S|} E(S_{High}) - \frac{|S_{Normal}|}{|S|} E(S_{Normal})$$

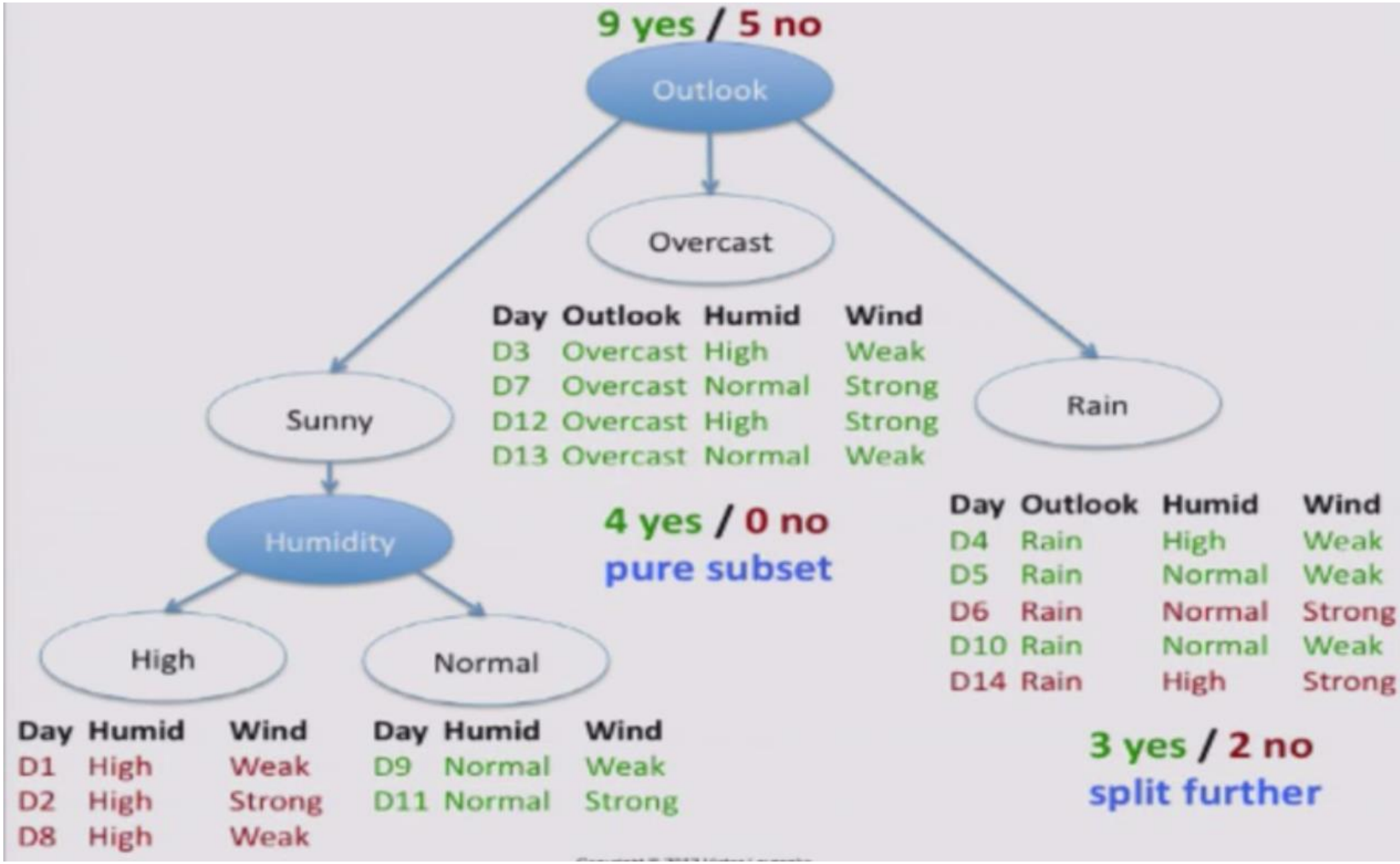
$$= 0.971 - \frac{3}{5}(0) - \frac{2}{5}(0) = \boxed{\mathbf{0.971}}$$

- Gain(S_{Sunny}, Wind) = 0.020
- Gain(S_{Sunny}, Humidity) = 0.971

Humidity has the greater Information Gain!
We therefore split S_{Sunny} into 2 subsets:

- S_{High} : (0 Yes / 3 No) → Pure Subset!
- S_{Normal} : (2 Yes / 0 No) → Pure Subset!

S_{High} and S_{Normal} are both pure; no need to split further.


9 yes / 5 no
Outlook
Overcast
Sunny
Rain
Day Outlook Humid Wind
D3 Overcast High Weak
D7 Overcast Normal Strong
D12 Overcast High Strong
D13 Overcast Normal Weak
4 yes / 0 no
pure subset
Humidity
High
Normal
Day Humid Wind
D1 High Weak
D2 High Strong
D8 High Weak
Day Humid Wind
D9 Normal Weak
D11 Normal Strong
Day Outlook Humid Wind
D4 Rain High Weak
D5 Rain Normal Weak
D6 Rain Normal Strong
D10 Rain Normal Weak
D14 Rain High Strong
3 yes / 2 no
split further

Going now to S_{Rain}.

Recall that:

S_{Rain}: (3 Yes / 2 No)
$|S_{Rain}| = 5$
$E(\, S_{Rain}) = 0.971$

Day	Outlook	Humid	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

3 yes / 2 no

Going now to S_{Rain}.

Humidity

Values(Humidity) = {High, Normal}

S_{Rain}: (3 Yes / 2 No)
$|S_{Rain}| = 5$
$E(S_{Rain}) = 0.971$

Humidity

High | Normal

S_{High}: (1 Yes / 1 No)
$|S_{High}| = 2$
$E(S_{High}) = 1$

S_{Normal}: (2 Yes / 1 No)
$|S_{Normal}| = 3$
$E(S_{Normal}) = 0.918$

$$Gain(S_{Rain}, Humidity) = E(S_{Rain}) - \frac{|S_{High}|}{|S|}E(S_{High}) - \frac{|S_{Normal}|}{|S|}E(S_{Normal})$$

$$= 0.971 - \frac{2}{5}(1) - \frac{3}{5}(0.918) = \boxed{\mathbf{0.020}}$$

Going now to S_{Rain}.

Wind

Values(Wind) = {Weak, Strong}

S_{Rain}: (3 Yes / 2 No)
$|S_{Rain}| = 5$
$E(\, S_{Rain}) = 0.971$

Wind

Weak

Strong

S_{Weak}: (3 Yes / 0 No)
$|S_{Weak}| = 3$
$E(\, S_{Weak}) = 0$

S_{Strong}: (0 Yes / 2 No)
$|S_{Strong}| = 3$
$E(\, S_{Strong}) = 0$

$$Gain(S_{Rain}, Humidity) = E(S_{Rain}) - \frac{|S_{Weak}|}{|S|} E(S_{Weak}) - \frac{|S_{Strong}|}{|S|} E(S_{Strong})$$

$$= 0.971 - \frac{3}{5}(0) - \frac{2}{5}(0) = \boxed{\mathbf{0.971}}$$

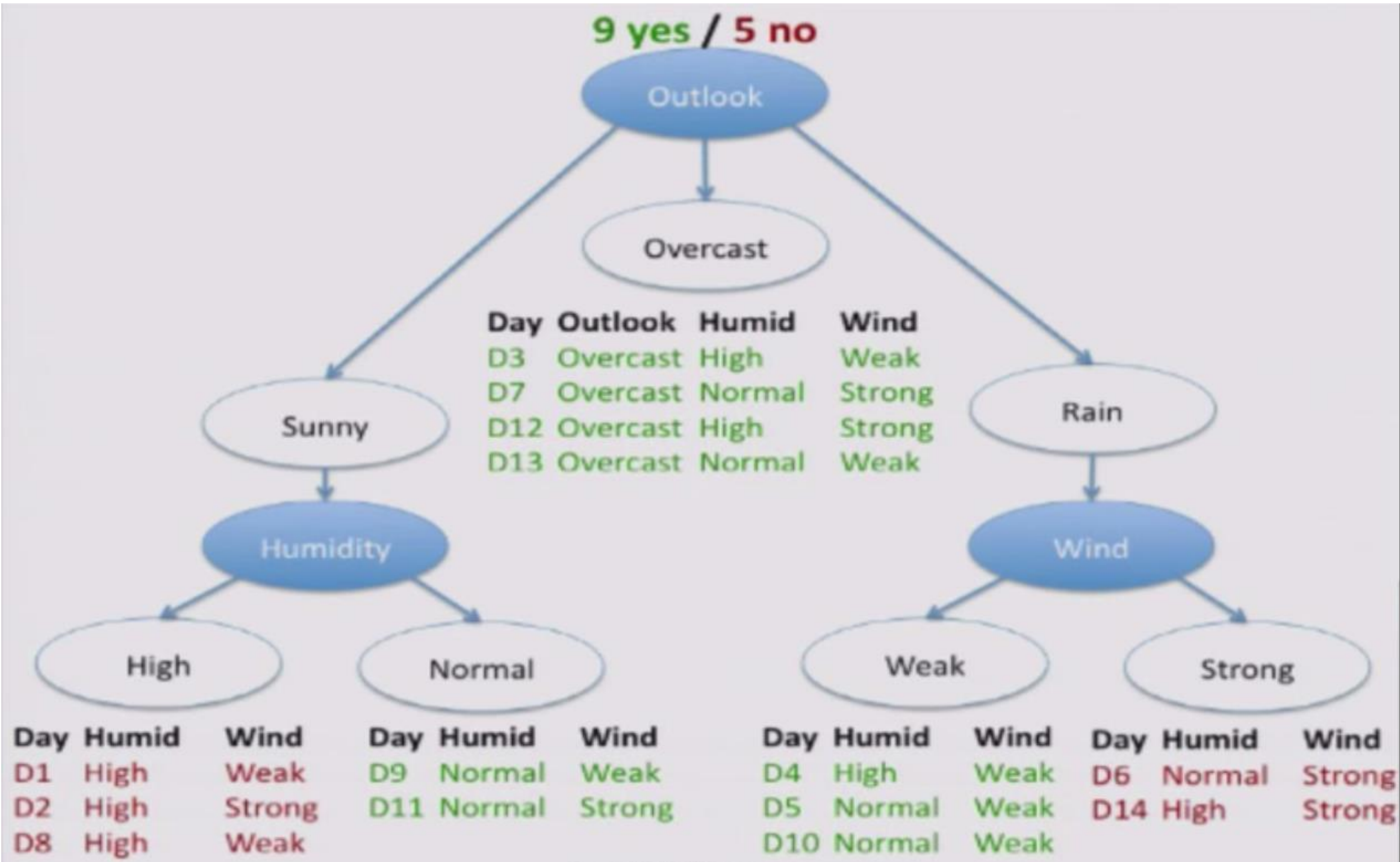
- Gain(S_{Rain}, Humidity) = 0.020
- Gain(S_{Rain}, Wind) = 0.971

Wind has the greater Information Gain!

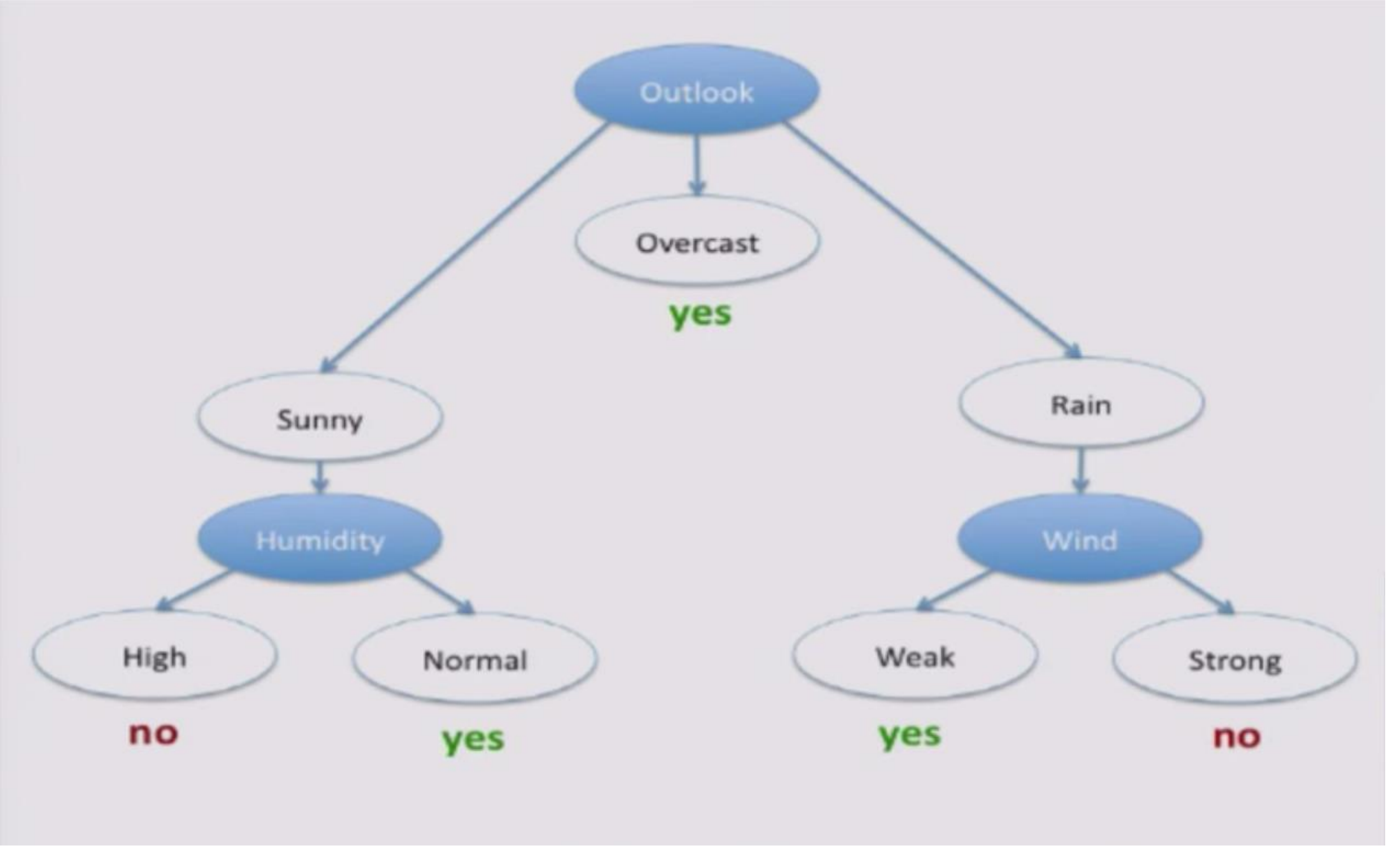
We therefore split S_{Rain} into 2 subsets:

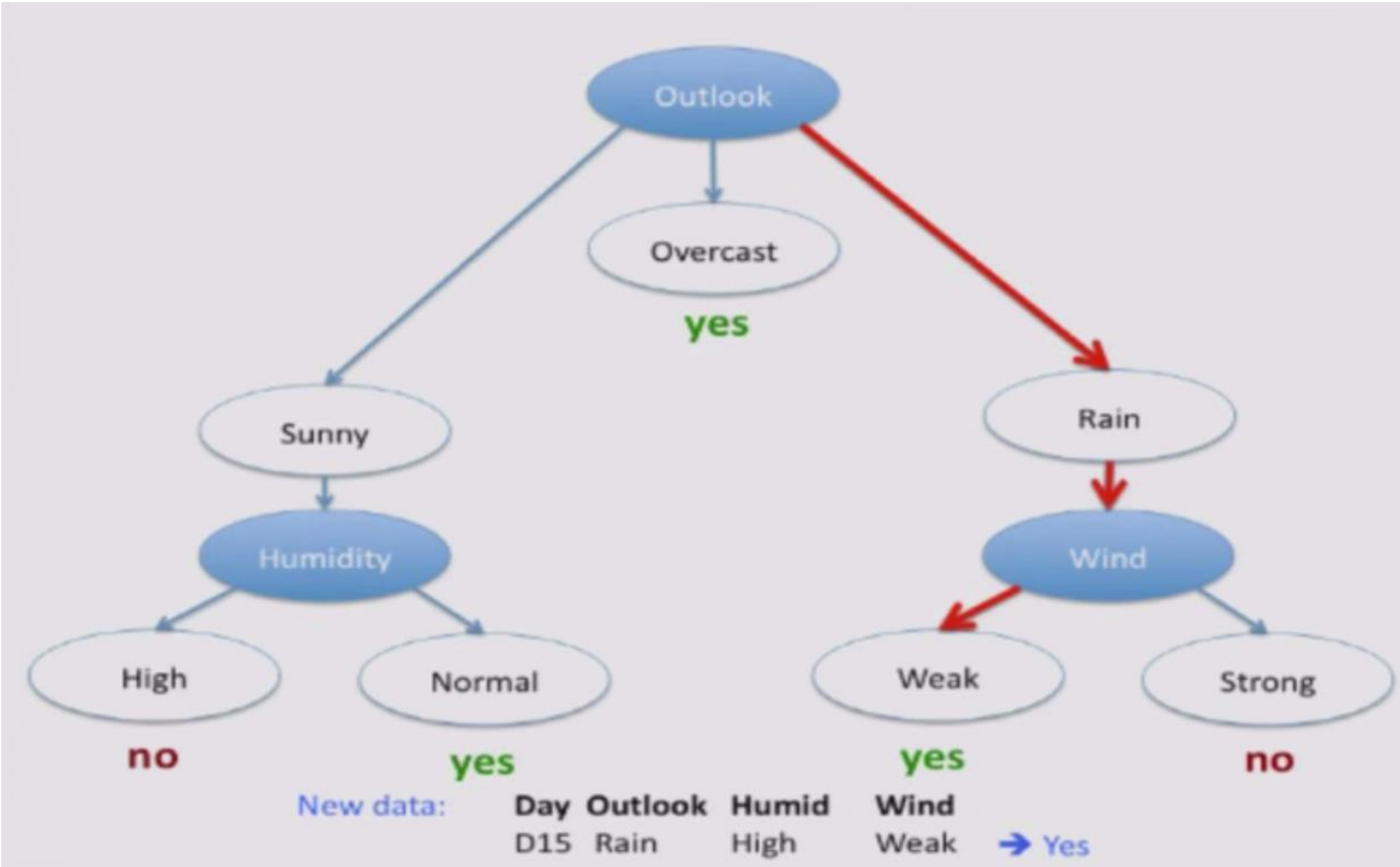
- S_{Weak} : (3 Yes / 0 No) → Pure Subset!
- S_{Strong} : (0 Yes / 2 No) → Pure Subset!

S_{Weak} and S_{Strong} are both pure; no need to split further.

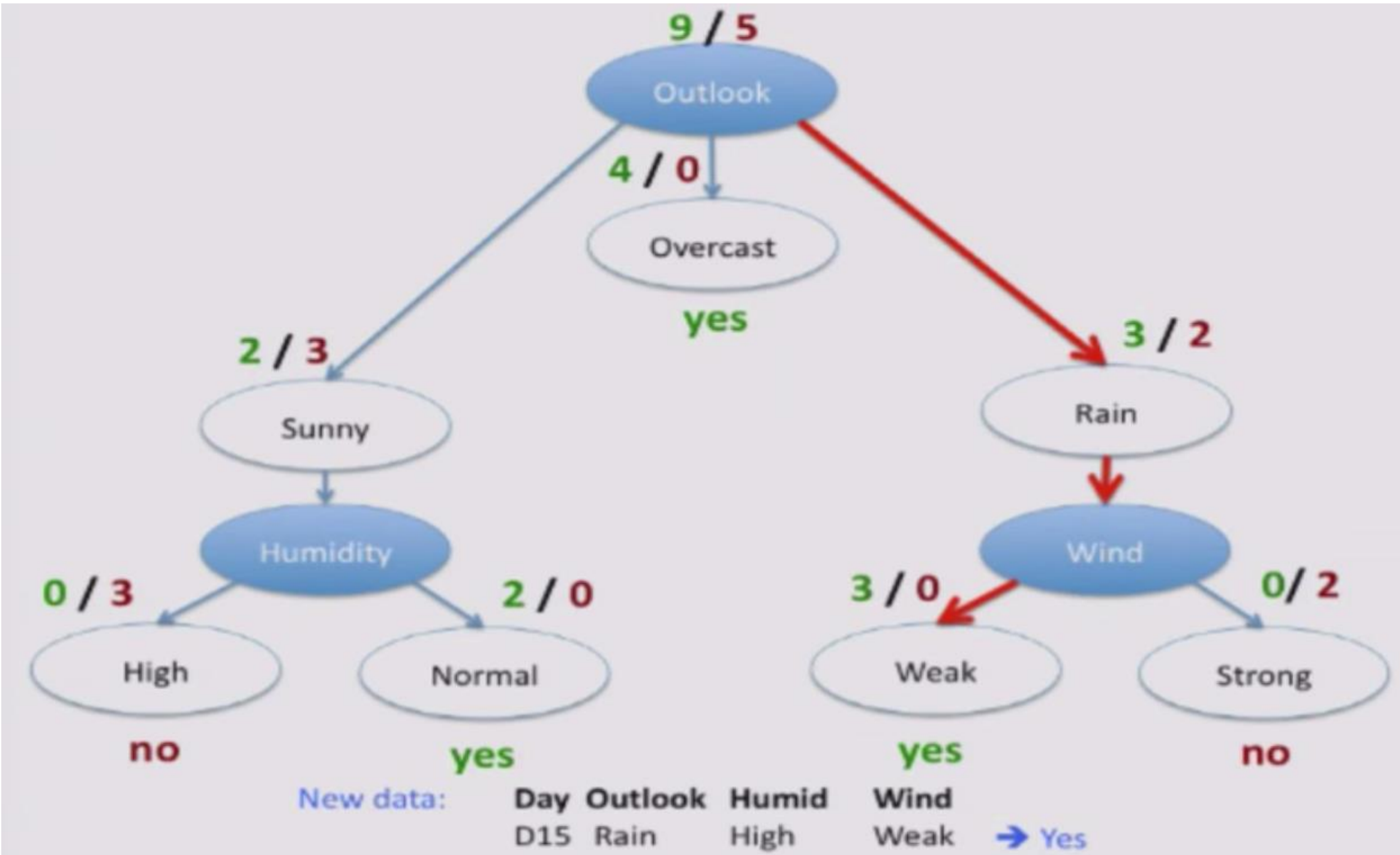

9 yes / 5 no
Outlook
Overcast
Day Outlook Humid Wind
D3 Overcast High Weak
D7 Overcast Normal Strong
D12 Overcast High Strong
D13 Overcast Normal Weak
Sunny
Rain
Humidity
Wind
High
Normal
Weak
Strong
Day Humid Wind
D1 High Weak
D2 High Strong
D8 High Weak
Day Humid Wind
D9 Normal Weak
D11 Normal Strong
Day Humid Wind
D4 High Weak
D5 Normal Weak
D10 Normal Weak
Day Humid Wind
D6 Normal Strong
D14 High Strong

Final Decision Tree





Outlook
Overcast
yes
Sunny
Rain
Humidity
Wind
High
Normal
Weak
Strong
no
yes
yes
no
New data:
Day Outlook Humid Wind
D15 Rain High Weak → Yes



9 / 5
Outlook
4 / 0
Overcast
yes
2 / 3
Sunny
3 / 2
Rain
Humidity
Wind
0 / 3
2 / 0
3 / 0
0/ 2
High
Normal
Weak
Strong
no
yes
yes
no
New data:
Day Outlook Humid Wind
D15 Rain High Weak
Yes