Institut Teknologi Sepuluh Nopember

Decision Trees

Examples

Predict if John will play tennis

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

Training examples: 9 yes / 5 no

New data:

D15	Rain	High	Strong	?

Testing examples

Predict if John will play tennis

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

New data:

D15	Rain	High	Strong	?

- **Hard to guess**
- **Try to understand when John plays**
- **Divide & Conquer:**
  - Split into subsets
  - Are they pure
  - If yes: stop
  - If no: repeat

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

9 yes / 5 no

Outlook

Sunny

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
Split further

Overcast

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

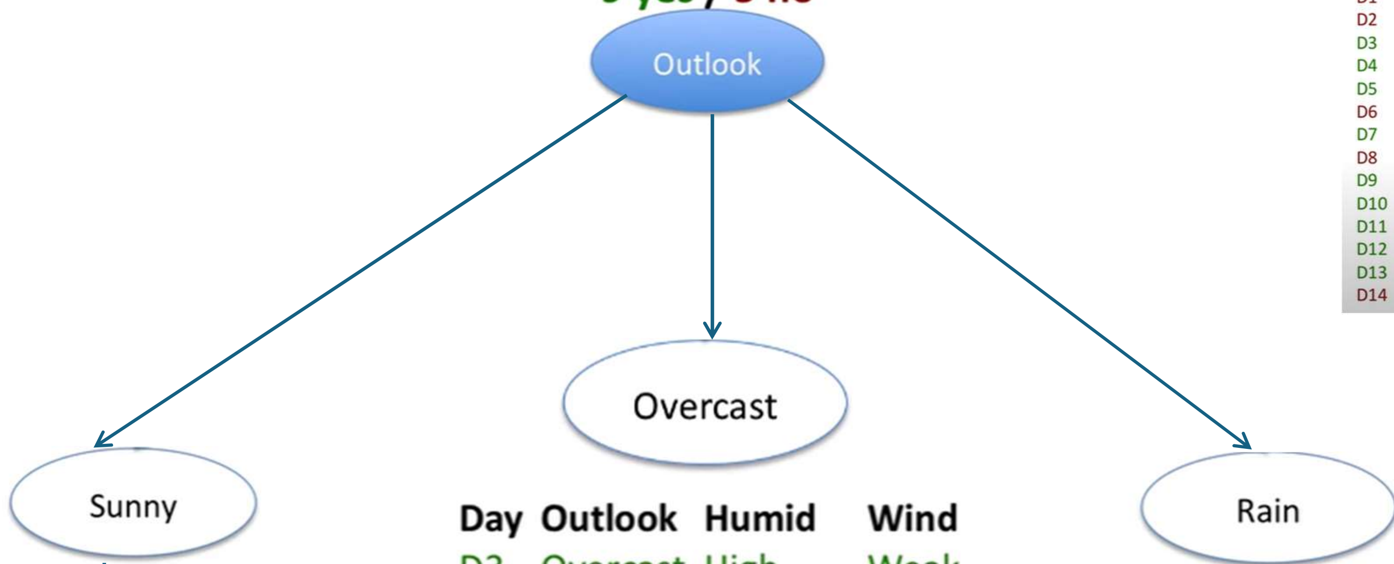
Rain

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

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

9 yes / 5 no

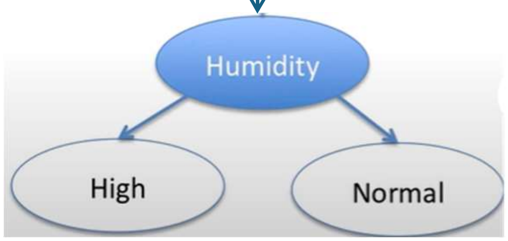


Day	Outlook	Humid	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

4 yes / 0 no
Pure subset

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

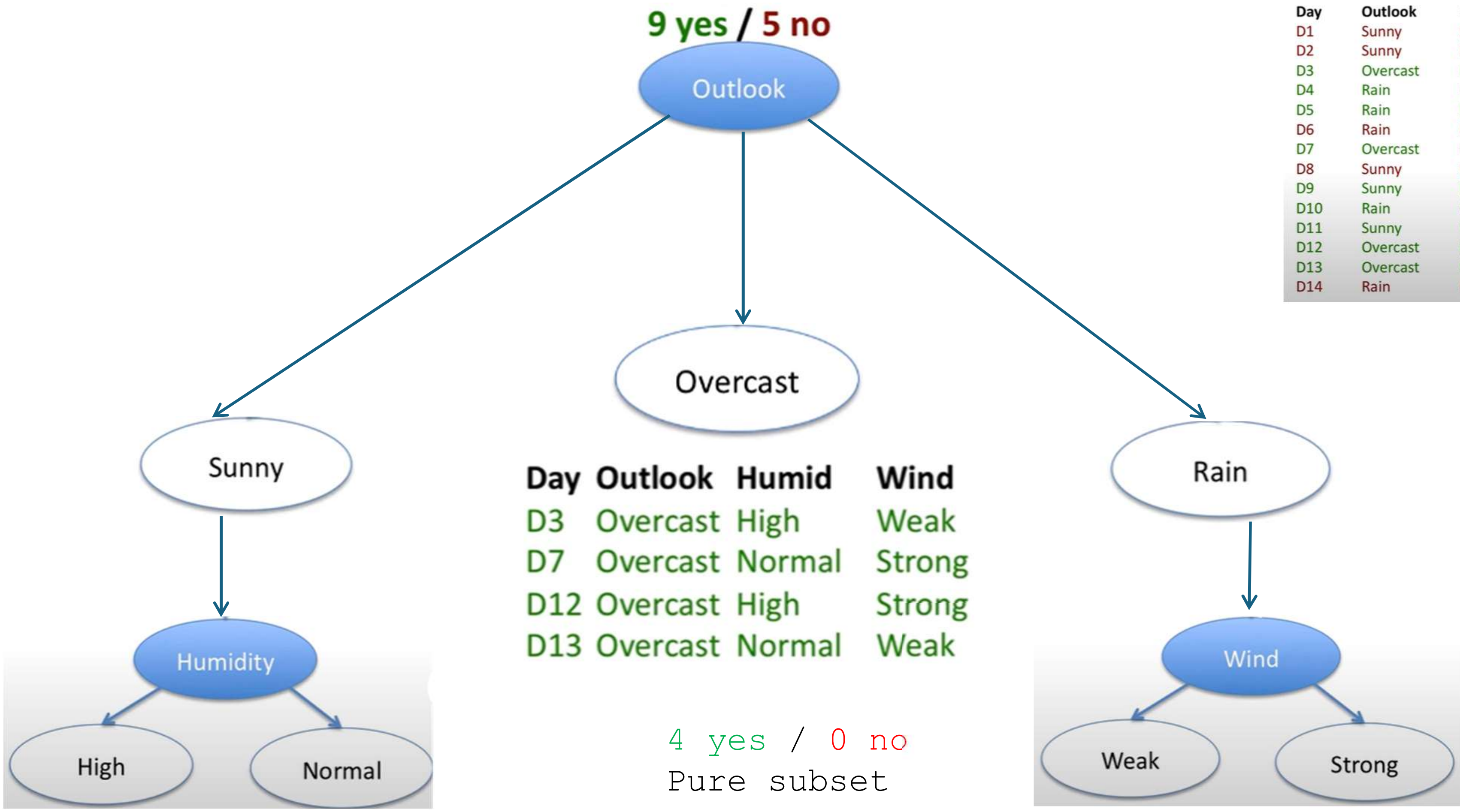
3 yes / 2 no
Split further



Day	Humid	Wind
D1	High	Weak
D2	High	Strong
D8	High	Weak

Day	Humid	Wind
D9	Normal	Weak
D11	Normal	Strong

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



Day	Outlook	Humid	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

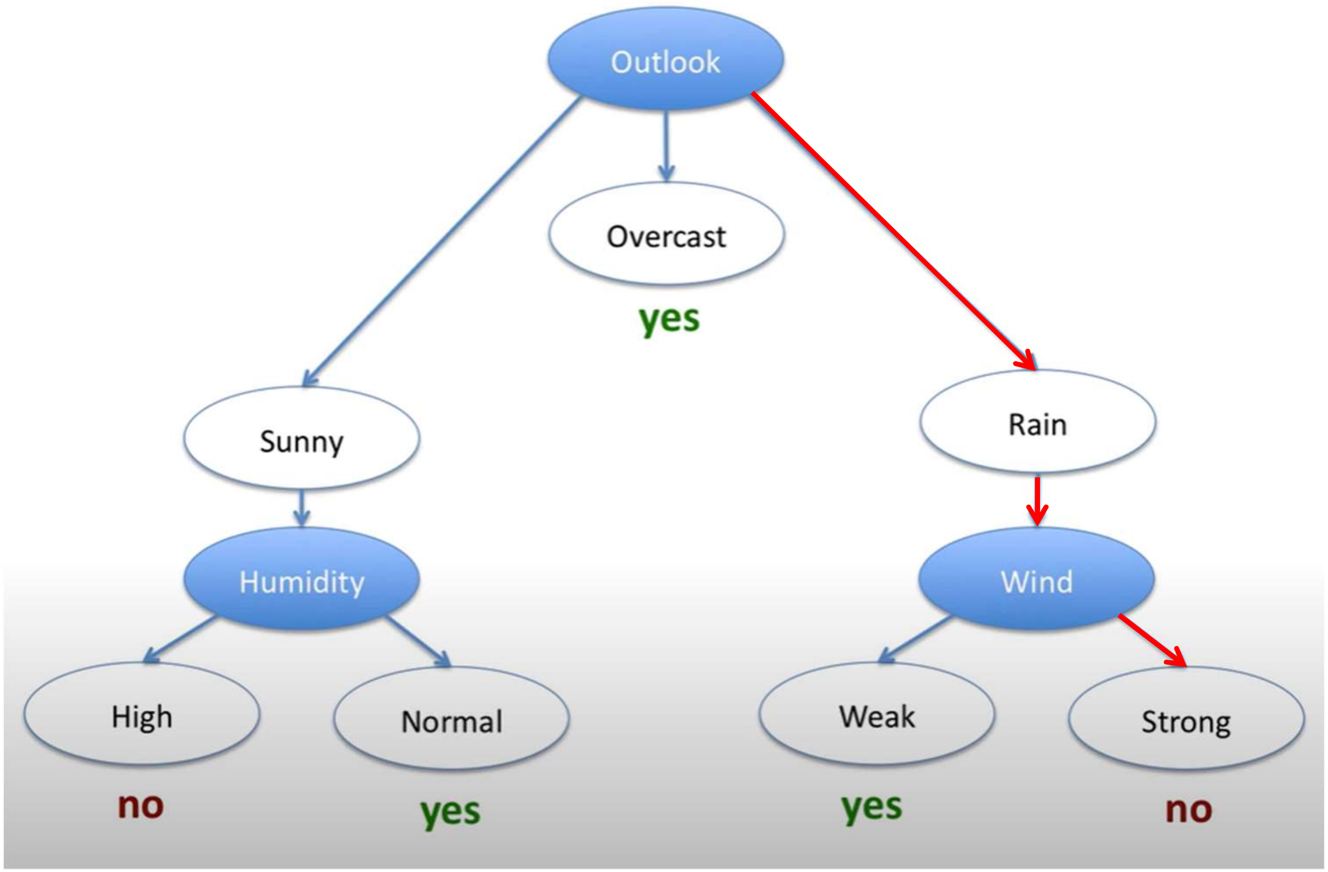
4 yes / 0 no
Pure subset

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



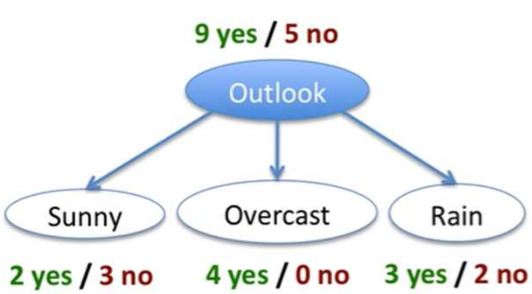
New data:

D15 Rain High Strong ? → No

ID3 Algorithm

- Split (node, {examples}):
  1. A ← the $\boxed{\text{best attribute}}$ for splitting the {examples}
  2. Decision attribute for this node ← A
  3. For each value of A, create new child node
  4. Split training {examples} to child nodes
  5. For each child node / subset:
     - If subset is pure: STOP
     - Else: Split (child_node, {subset})

Which attribute to split on first?



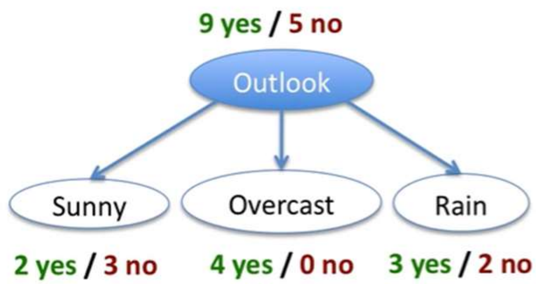
Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes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

- Want to measure “purity” of the split
  - ✓ More certain about yes/no after the split
    - Pure set (4 yes / 0 no) => completely certain (100%)
    - Impure set (3 yes / 3 no) => completely uncertain (50%)

Entropy

How pure the se

$$H(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$



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

$$H(S_{\text{outlook}}) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$
$$= 0.94$$

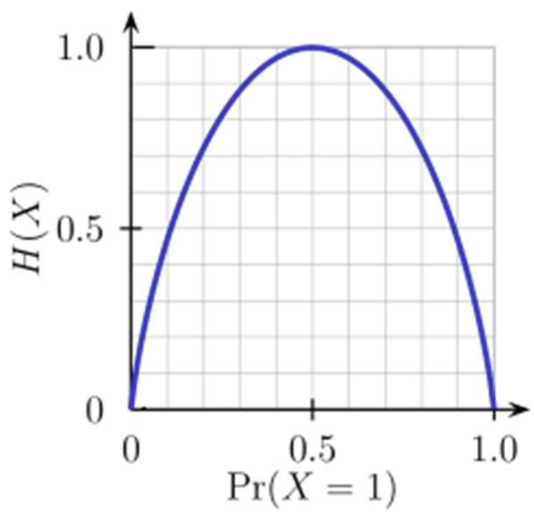
$$H(S_{\text{sunny}}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$
$$= 0.97$$

$$H(S_{\text{overcast}}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}$$
$$= 0$$

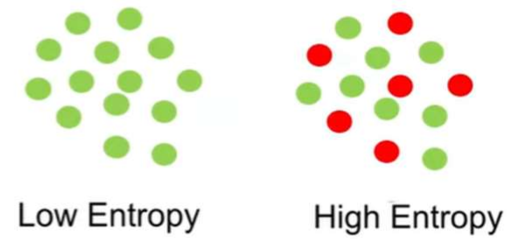
Entropy

- Entropy is a measure of the randomness in the information being processed
- The higher the entropy, the harder it is to draw any conclusions from that information

$$H(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

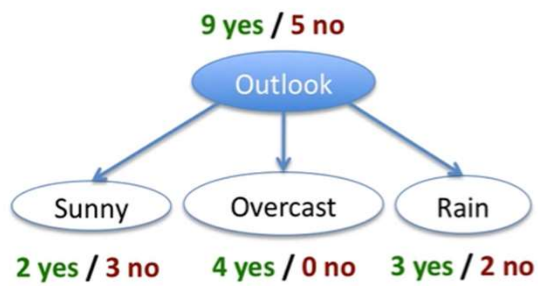


Entropy



How pure the sets?

$$H(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$



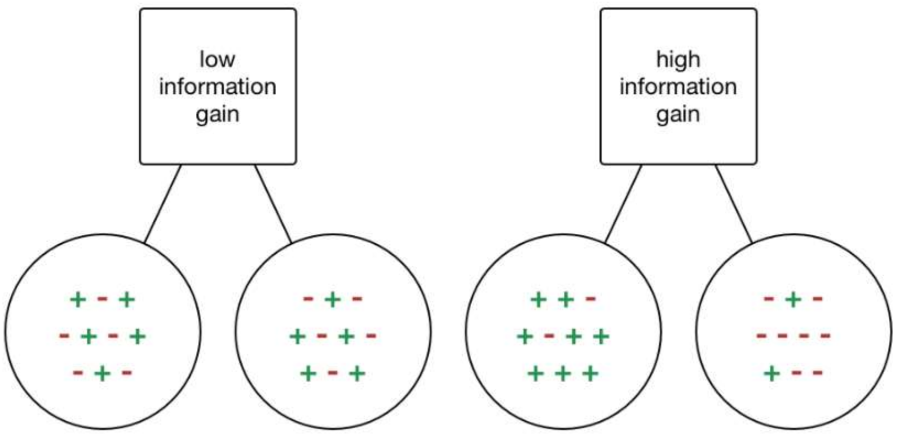
$$H(S_{\text{outlook}}) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$
$$= 0.94$$

$$H(S_{\text{sunny}}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$
$$= 0.97$$

$$H(S_{\text{overcast}}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}$$
$$= 0$$

Information Gain

- Want many items in pure sets
- Expected drop in entropy after split



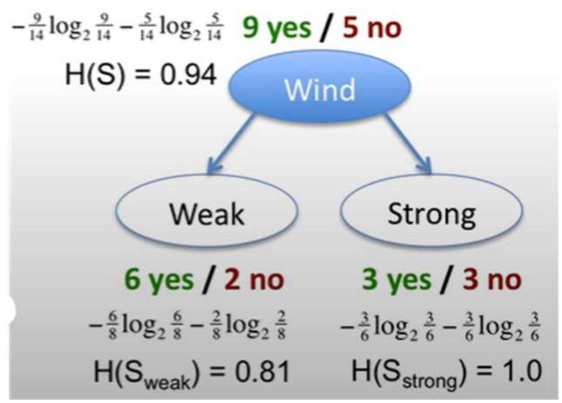
Information Gain

- Want many items in pure sets
- Expected drop in entropy after split:

$$Gain(S,A) = H(S) - \sum_{V \in Values(A)} \frac{|S_V|}{|S|} H(S_V)$$

V … possible values of A
S … set of examples {X}
S_V … subset where $X_A = V$

- Our goal is to maximize the *Information Gain*



$$\begin{aligned} Gain(S, Wind) &= H(S) - \frac{8}{14} H(S_{Weak}) - \frac{6}{14} H(S_{Strong}) \\ &= 0.94 - \frac{8}{14} 0.81 - \frac{6}{14} 1.0 \\ &= 0.049 \end{aligned}$$

ID3 Algorithm: example

Split (node, {examples}):

1. A ← the best attribute for splitting the {examples}
2. Decision attribute for this node ← A
3. For each value of A, create new child node
4. Split training {examples} to child nodes
5. For each child node / subset:
   - If subset is pure: STOP
   - Else: Split (child_node, {subset})

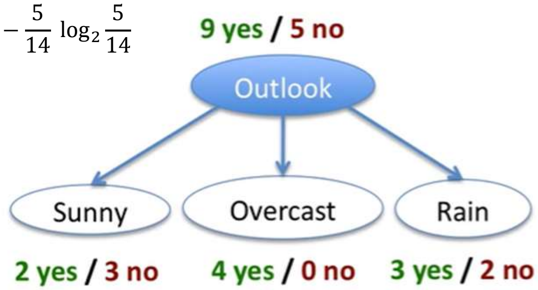
1. Choose the best attribute

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 best attribute based on information gain!

$$-\frac{9}{14}\log_2\frac{9}{14}-\frac{5}{14}\log_2\frac{5}{14}$$

$H(S) = 0.94$



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

$$-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}$$

$H(S_{sunny}) = 0.97$

$$-\frac{4}{4}\log_2\frac{4}{4}-\frac{0}{4}\log_2\frac{0}{4}$$

$H(S_{overcast}) = 0$

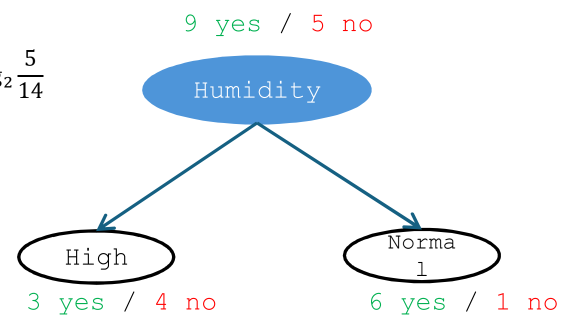
$$-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}$$

$H(S_{rain}) = 0.97$

$$\begin{aligned} Gain(S, Outlook) &= H(S) - \frac{5}{14}H(S_{sunny}) - \frac{4}{14}H(S_{overcast}) - \frac{5}{14}H(S_{rain}) \\ &= 0.94 - \frac{5}{14}\,0.97 - \frac{4}{14}\,0 - \frac{5}{14}\,0.97 \\ &= 0.25 \end{aligned}$$

$$-\frac{9}{14}\log_2\frac{9}{14}-\frac{5}{14}\log_2\frac{5}{14}$$

$H(S) = 0.94$



$$-\frac{3}{7}\log_2\frac{3}{7}-\frac{4}{7}\log_2\frac{4}{7}$$

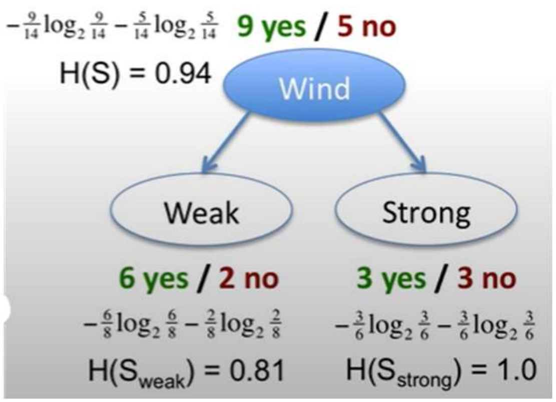
$H(S_{high}) = 0.98$

$$-\frac{6}{7}\log_2\frac{6}{7}-\frac{1}{7}\log_2\frac{1}{7}$$

$H(S_{normal}) = 0.59$

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

$$\begin{aligned} Gain(S, Humidity) &= H(S) - \frac{7}{14}H(S_{high}) - \frac{7}{14}H(S_{normal}) \\ &= 0.94 - \frac{7}{14}0.98 - \frac{7}{14}0.59 \\ &= 0.15 \end{aligned}$$



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

$$
\begin{aligned}
Gain(S, Wind) &= H(S) - \frac{8}{14}H(S_{Weak}) - \frac{6}{14}H(S_{Strong}) \\
&= 0.94 - \frac{8}{14}\,0.81 - \frac{6}{14}\,1.0 \\
&= 0.049
\end{aligned}
$$

Split (node, {examples}):

1. A ← the best attribute for splitting the {examples}
2. Decision attribute for this node ← A
3. For each value of A, create new child node
4. Split training {examples} to child nodes
5. For each child node / subset:
   - If subset is pure: STOP
   - Else: Split (child_node, {subset})

$$Gain(S, Outlook) = 0.25$$

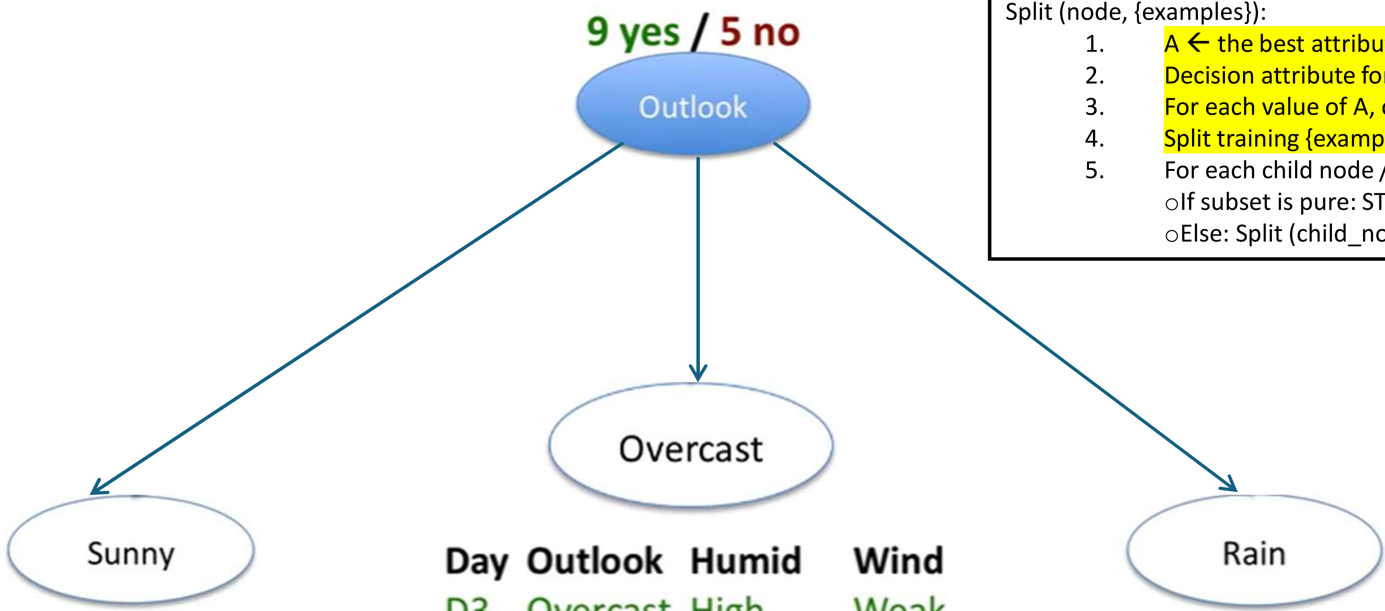
The best attribute based on information gain!

$$Gain(S, Humidity) = 0.15$$

$$Gain(S, Wind) = 0.049$$

Split (node, {examples}):

1. A ← the best attribute for splitting the {examples}
2. Decision attribute for this node ← A
3. For each value of A, create new child node
4. Split training {examples} to child nodes
5. For each child node / subset:
   - If subset is pure: STOP
   - Else: Split (child_node, {subset})



Sunny

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
Split further

Overcast

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

Rain

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

2 yes / 3 no

Sunny

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

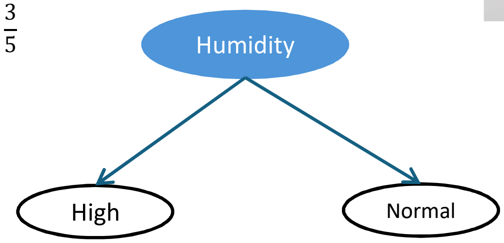
Split (node, {examples}):

1. A ← the best attribute for splitting the {examples}
2. Decision attribute for this node ← A
3. For each value of A, create new child node
4. Split training {examples} to child nodes
5. For each child node / subset:
   - If subset is pure: STOP
   - Else: Split (child_node, {subset})

2 yes / 3 no

$-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}$

$H(S) = 0.97$



0 yes / 3 no

$-\frac{0}{3}\log_2\frac{0}{3}-\frac{3}{3}\log_2\frac{3}{3}$

$H(S_{high}) = 0$

2 yes / 0 no

$-\frac{2}{2}\log_2\frac{2}{2}-\frac{0}{2}\log_2\frac{0}{2}$

$H(S_{normal}) = 0$

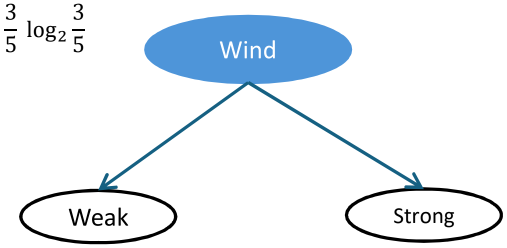
$$\begin{aligned} Gain(S, Humidity) &= H(S) - \frac{3}{5}H(S_{High}) - \frac{2}{5}H(S_{Normal}) \\ &= 0.97 - \frac{3}{5}\,0 - \frac{2}{5}\,0 \\ &= 0.97 \end{aligned}$$

Humidity is the best attribute based on information gain!

2 yes / 3 no

$-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}$

$H(S) = 0.97$



1 yes / 2 no

$-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}$

$H(S_{high}) = 0.92$

1 yes / 1 no

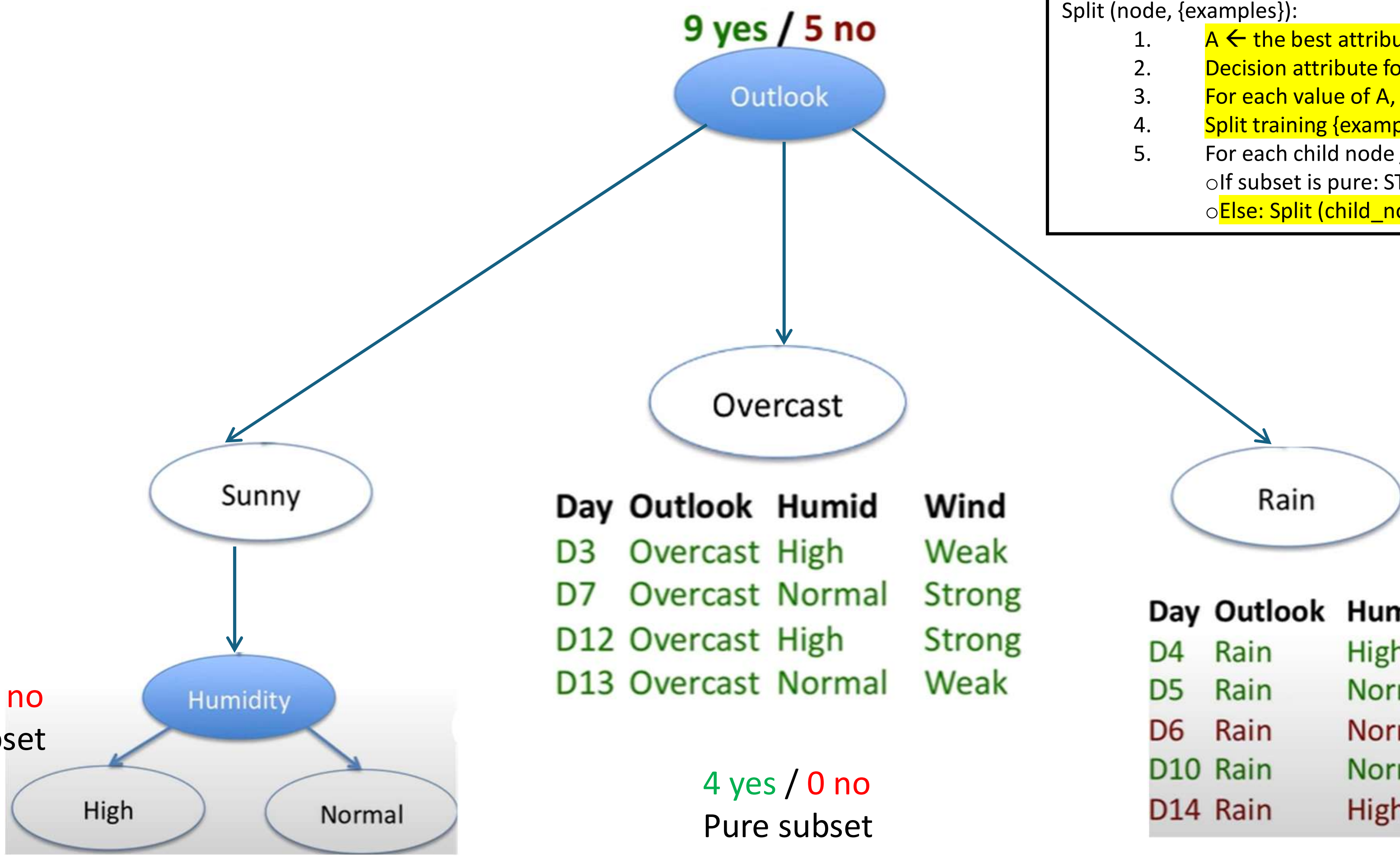
$-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}$

$H(S_{normal}) = 1$

$$\begin{aligned} Gain(S, Humidity) &= H(S) - \frac{3}{5}H(S_{Weak}) - \frac{2}{5}H(S_{Strong}) \\ &= 0.97 - \frac{3}{5}\,0.92 - \frac{2}{5}\,1 \\ &= 0.018 \end{aligned}$$

Split (node, {examples}):

1. A ← the best attribute for splitting the {examples}
2. Decision attribute for this node ← A
3. For each value of A, create new child node
4. Split training {examples} to child nodes
5. For each child node / subset:
   - If subset is pure: STOP
   - Else: Split (child_node, {subset})



Sunny → Humidity

3 yes / 2 no

Pure subset

Day	Humid	Wind
D1	High	Weak
D2	High	Strong
D8	High	Weak

Day	Humid	Wind
D9	Normal	Weak
D11	Normal	Strong

Overcast

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

Rain

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

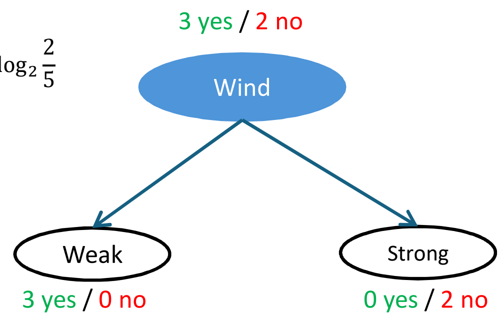
Split further

Split (node, {examples}):

1. A ← the best attribute for splitting the {examples}
2. Decision attribute for this node ← A
3. For each value of A, create new child node
4. Split training {examples} to child nodes
5. For each child node / subset:
   - If subset is pure: STOP
   - Else: Split (child_node, {subset})

$$-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$H(S) = 0.97$$



$$-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$$

$$H(S_{high}) = 0$$

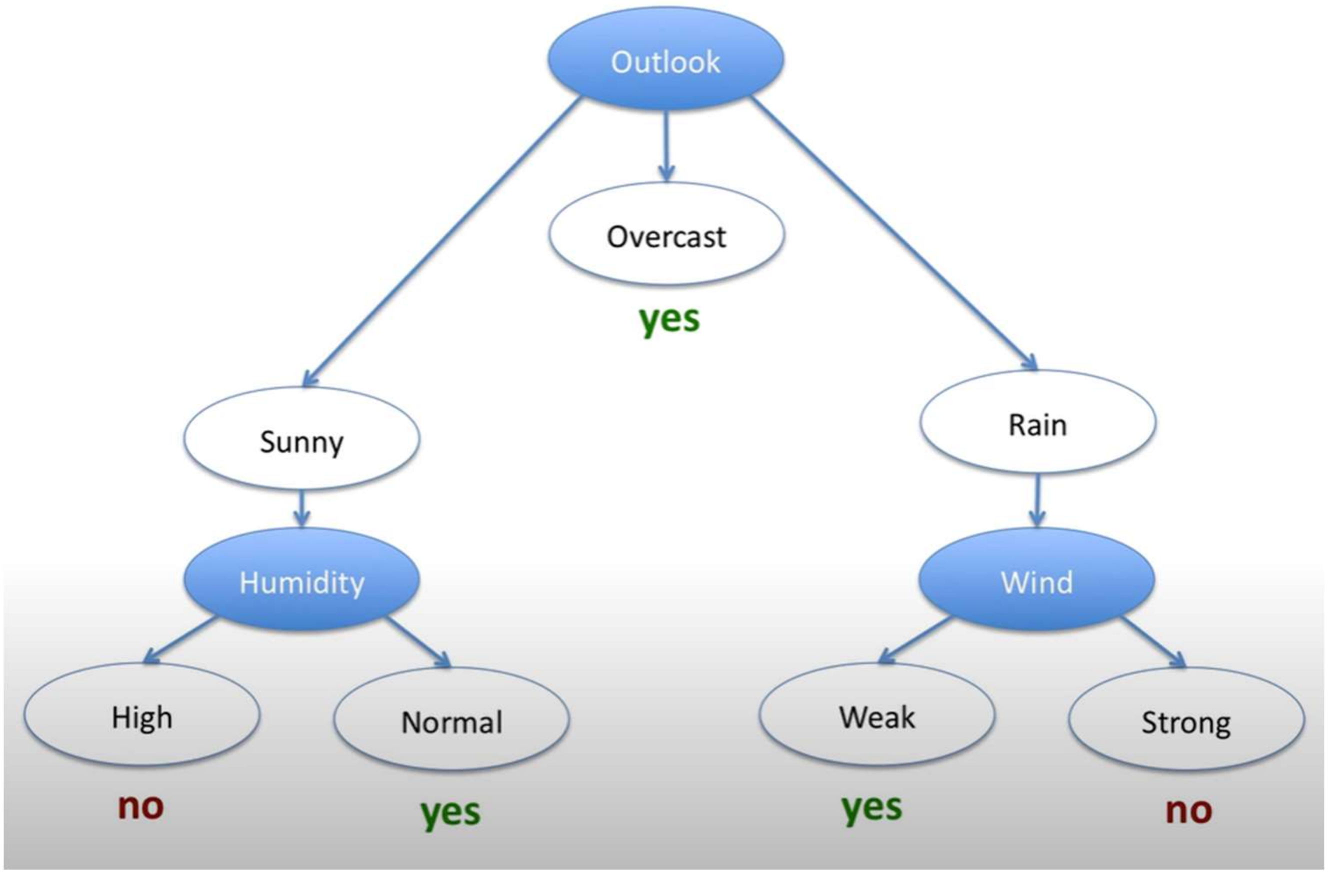
$$-\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}$$

$$H(S_{normal}) = 0$$

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

$$\begin{aligned} Gain(S, Humidity) &= H(S) - \frac{3}{5}H(S_{Weak}) - \frac{2}{5}H(S_{Strong}) \\ &= 0.97 - \frac{3}{5}0 - \frac{2}{5}0 \\ &= 0.97 \end{aligned}$$

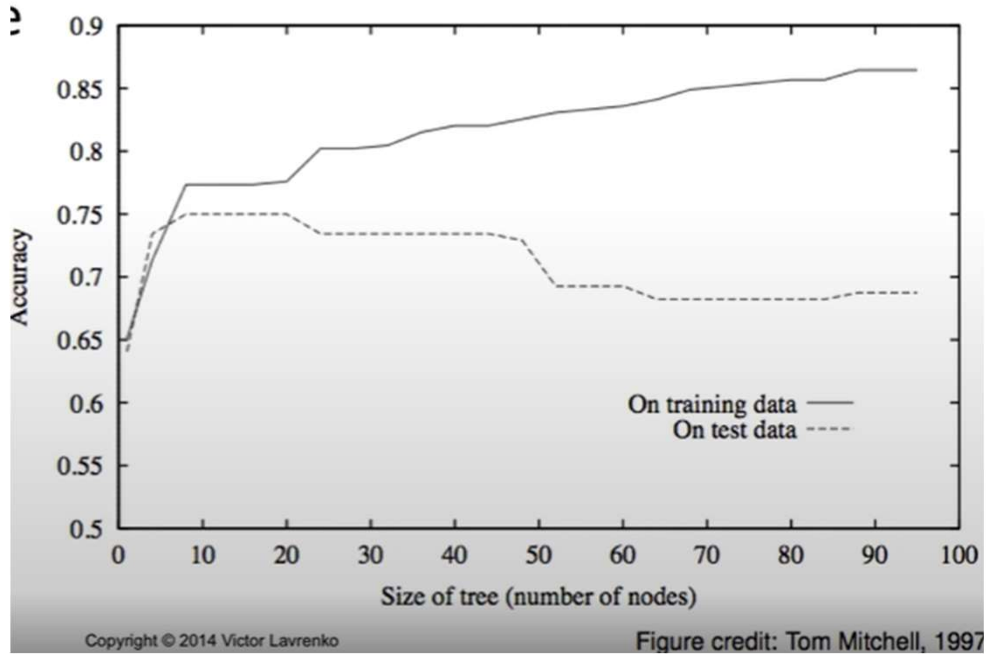
Final tree



Outlook
Overcast
yes
Sunny
Rain
Humidity
Wind
High
Normal
Weak
Strong
no
yes
yes
no

Overfitting in Decision Trees

- Can always classify training examples perfectly
- Keep splitting until each node contains 1 example
- Singleton = pure
- Doesn't work on new data



Task 1

ID	Gender	Car type	Cost	Buy?
1	F	Sport	Cheap	No
2	F	Sport	Expensive	Yes
3	F	Family	Cheap	Yes
4	F	Family	Expensive	No
5	F	Sport	Cheap	Yes
6	F	Sport	Expensive	Yes
7	F	Family	Cheap	Yes
8	F	Family	Expensive	No
9	M	Sport	Cheap	No
10	M	Family	Cheap	No
11	M	Sport	Expensive	No
12	M	Family	Expensive	No

- Generate decision trees using the ID3 algorithm, and calculate entropy and information gain for each node and leaf

Other Splitting Method

- Information Gain
- Gini Index
- Information Gain Ratio
- Others

Reference

- https://www.youtube.com/playlist?list=PLBv09BD7ez_4_UoYeGrzvqveIR_USBEKD