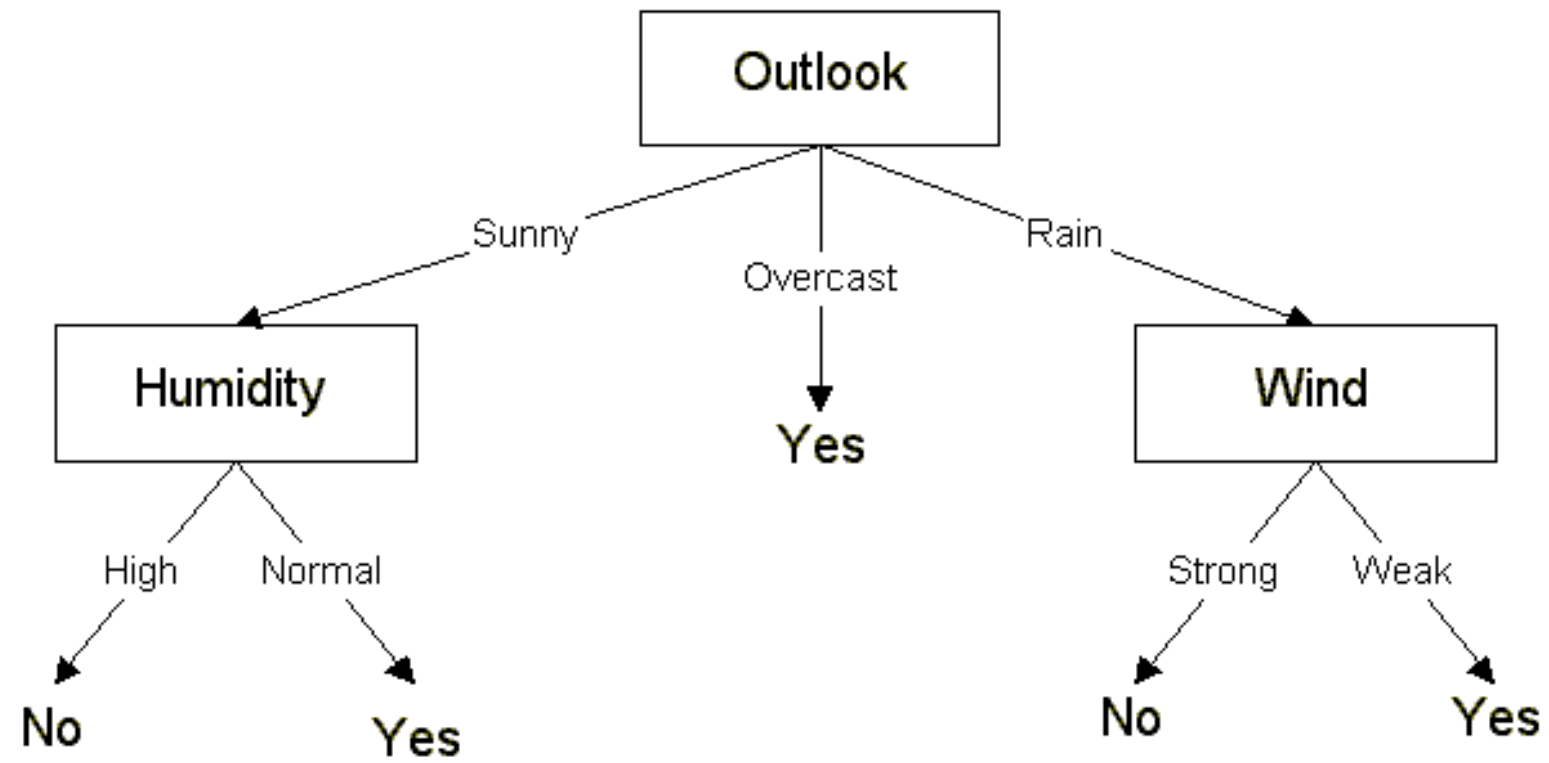


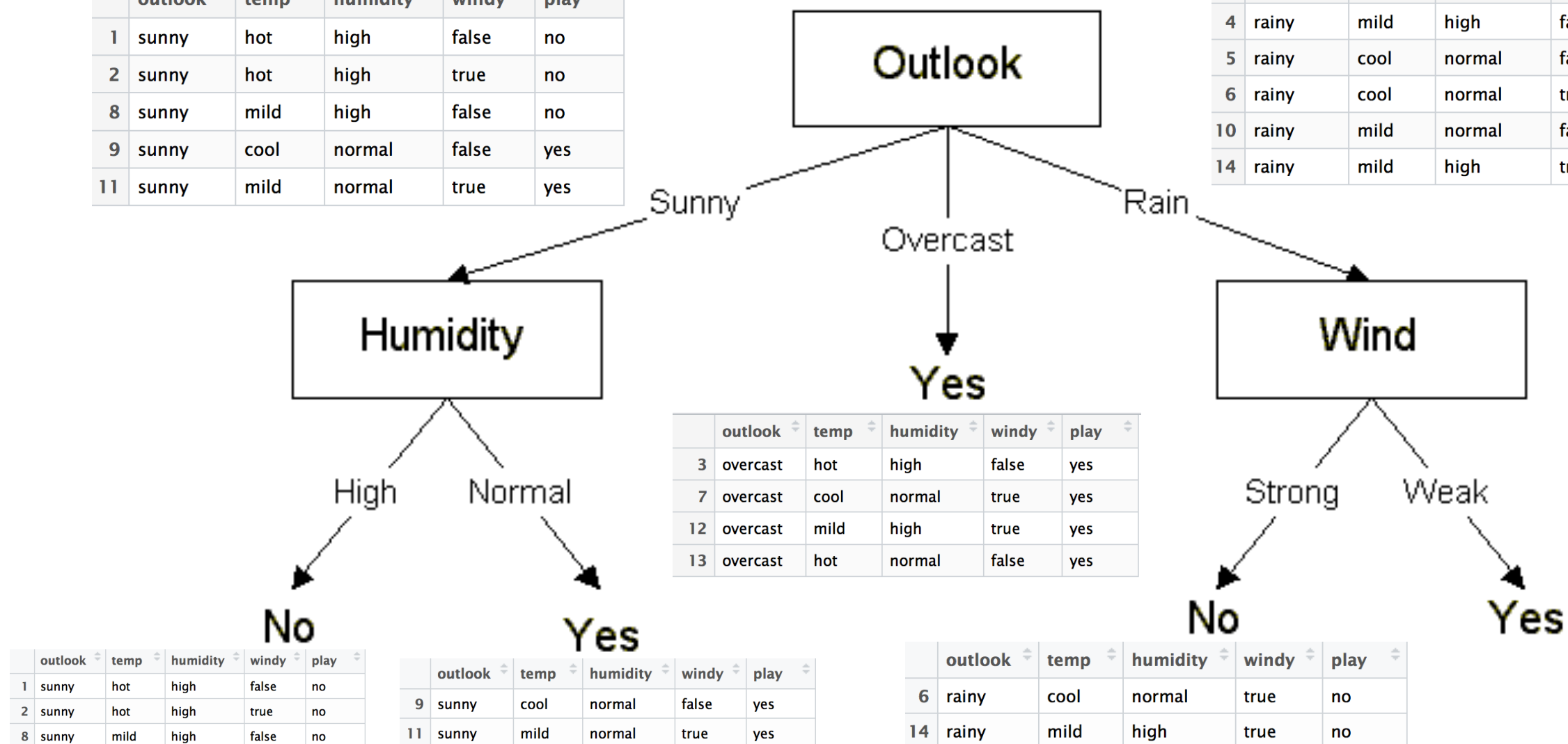
# Decision Trees

	outlook ↕	temp ↕	humidity ↕	windy ↕	play ↕
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no



	outlook ↕	temp ↕	humidity ↕	windy ↕	play ↕
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
11	sunny	mild	normal	true	yes

	outlook ↕	temp ↕	humidity ↕	windy ↕	play ↕
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
10	rainy	mild	normal	false	yes
14	rainy	mild	high	true	no



# Split Column Selection

$$H(S) = - p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$$

# Split Column Selection

$$H(S) = - p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$$

	outlook ↕	temp ↕	humidity ↕	windy ↕	play ↕
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
11	sunny	mild	normal	true	yes

# Split Column Selection

$$H(S) = - p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$$

	outlook ↕	temp ↕	humidity ↕	windy ↕	play ↕
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
11	sunny	mild	normal	true	yes

0.9709506

$$-(2/5) \log_2 (2/5) - (3/5) \log_2 (3/5)$$

# Split Column Selection

$$H(S) = - p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$$

$$-(5/5) * \log_2(5/5) - (0/5) * \log_2(0/5) = 0$$

$$-(3/6) * \log_2(3/6) - (3/6) * \log_2(3/6) = 1$$



- + good for huge datasets
- + ignores redundant features
- + small trees easy to interpret

- - large trees hard to interpret
- - poor prediction performance

# Summary