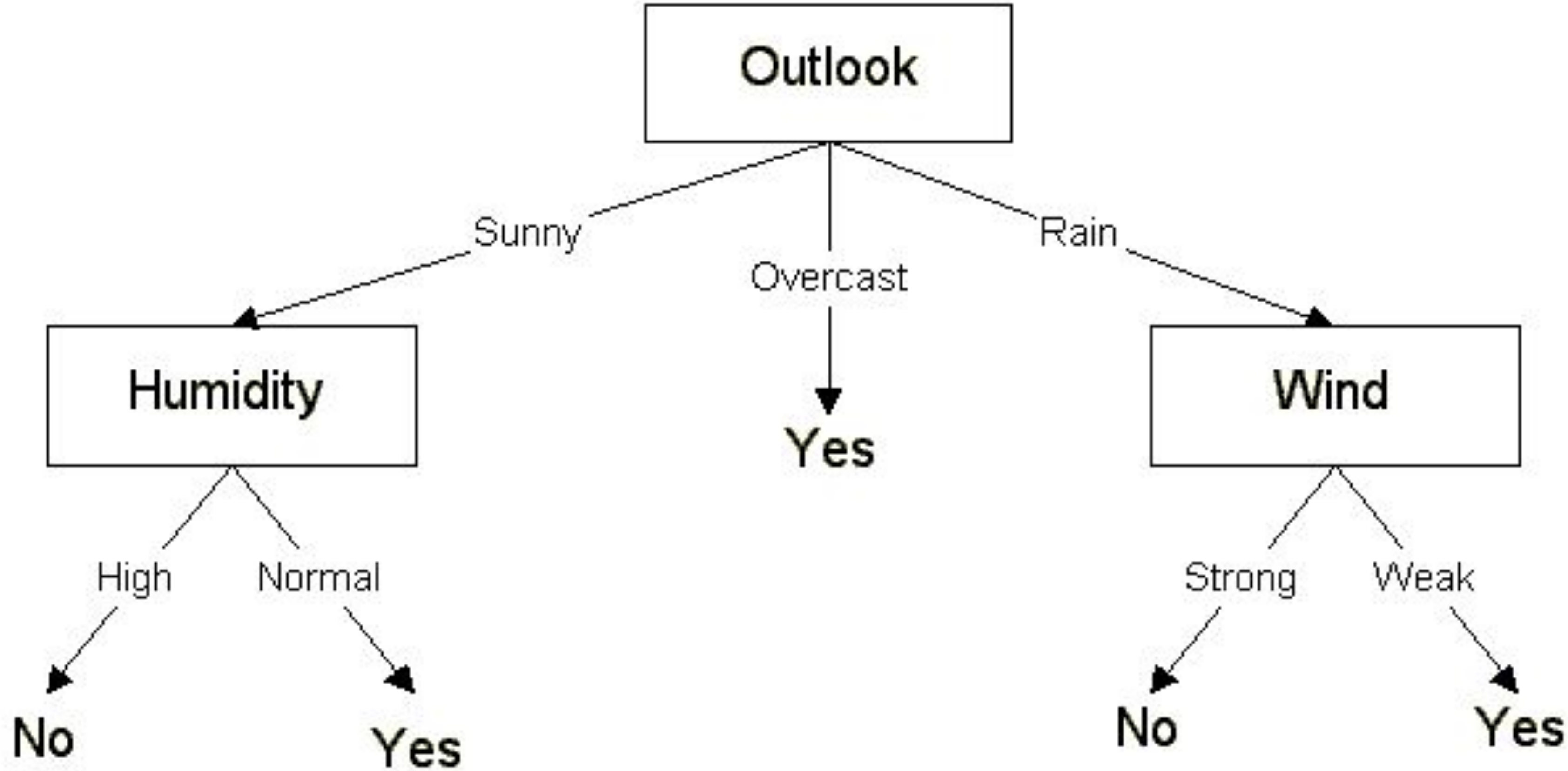
Arbori de decizie.
Algoritmul IDE3



$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

entropy(p_1, p_2, ... p_n) = -p_1 log p_1 - p_2 log p_2 ... - p_n log p_n

Ludwig Boltzmann
1844 - 1906 Viena
Austrian physicist and philosopher

The entropy law is sometimes refered to as the second law of thermodynamics

This second law states that for any irreversible process, entropy always increases.

Entropy is a measure of disorder.

!!! Since virtually all natural processes are irreversble, the entropy law implies that the universe is "running down"

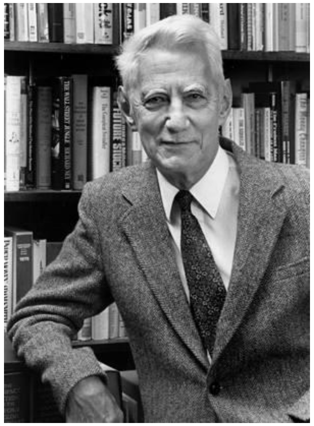

S = k. log W
LVDWIG
BOLTZMANN
1844 – 1906
HENRIETTE
BOLTZMANN
EB. EDLE VON AIGENTLER
1854 – 1938
DR PHIL. PAULA
BOLTZMANN
GEB. CHIARI
1891 – 1977
ARTHUR
BOLTZMANN
DIPL. ING. DR PHIL. HOFRAT
1881 – 1952
LVDWIG
BOLTZMANN
1923 – 1943
TZTER MÄNNLICHER NACHKOMME,
GEFALLEN BEI SMOLENSK

Entropy can be seen as a measure of the quality of energy:

Low entropy sources of energy are of high quality. Such energy sources have a high energy density.

High entropy sources of energy are closer to randomness and are therefore less available for use

Entropy in Information Theory



Claude Shannon
1916 –2001
American mathematician, electrical engineer, and cryptographer known as "the father of information theory".

- Claude Shannon transferred some of these ideas to the world of information processing. Information is associated by him with low entropy.
- Contrasted with information is "noise", randomness, high entropy.

- At the extreme of no information are random number.
- Of course, data may only look random.
  - But there may be hidden patterns, information in the data. The whole point of ML is to dig out the patterns. The descovered patterns are usually presented as rules or decision trees. Shannon's information theory can be used to construct decision trees.

A collection of random numbers has maximum entropy

Entropy in Information Theory

- Shannon defined the entropy *H* (Greek capital letter eta) of a discrete variable *X* with possible values {*x1, ..., xn*} and probability p as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

- whe

measure is the *bit.*

Example

•Presupunem evenimentul aruncării unui zar cu 6 fețe. Valorile variabilei *X* sunt {1,2,3,4,5,6} iar probabilitățile obținerii oricărei valori sunt egale.

-În acest caz entropia este:

$$-H(X) = -\sum_{i=1}^{6} (1/6)\log_2(1/6) = -6 * (1/6)\log_2(1/6) = -log_2(1/6) = 2.58$$

bits.

Which is the best attribute?

- Aim: to get the smallest tree
- Heuristic
  - *choose the attribute that produces the "purest" nodes*
  - *I.e. the greatest information gain*
- Information theory: measure information in bits

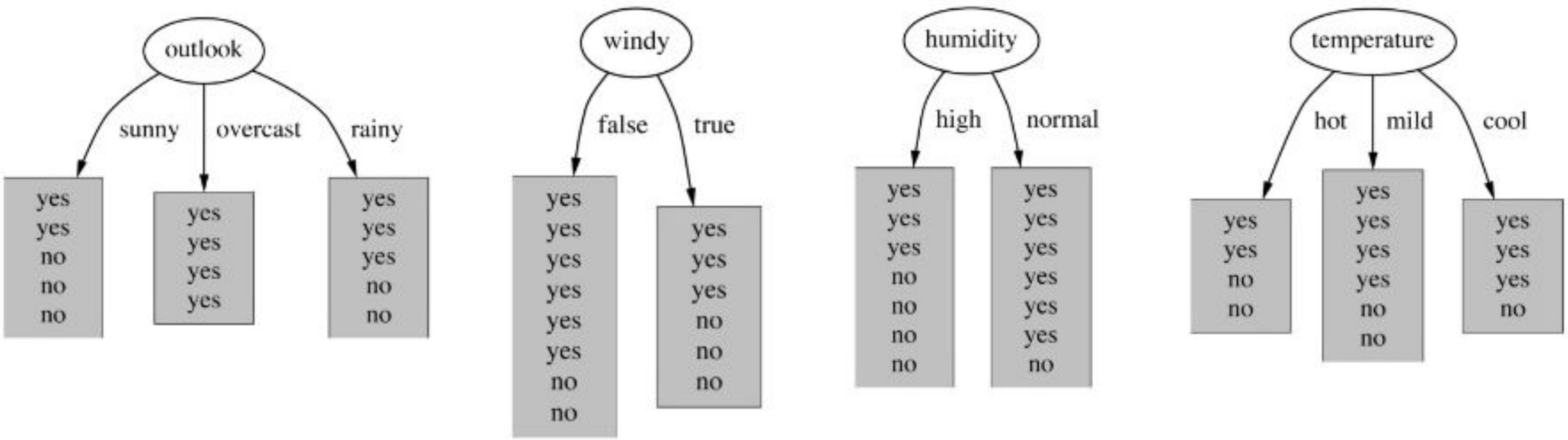
$$\text{entropy}(p_1, p_2, \ldots, p_n) = -p_1 \log p_1 - p_2 \log p_2 \ldots - p_n \log p_n$$

Information gain

- Amount of information gained by knowing the value of the attribute
- (Entropy of distribution before the split) – (entropy of distribution after it)

	outlook text	temperature text	humidity text	windy text	play text
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

Which attribute to select?



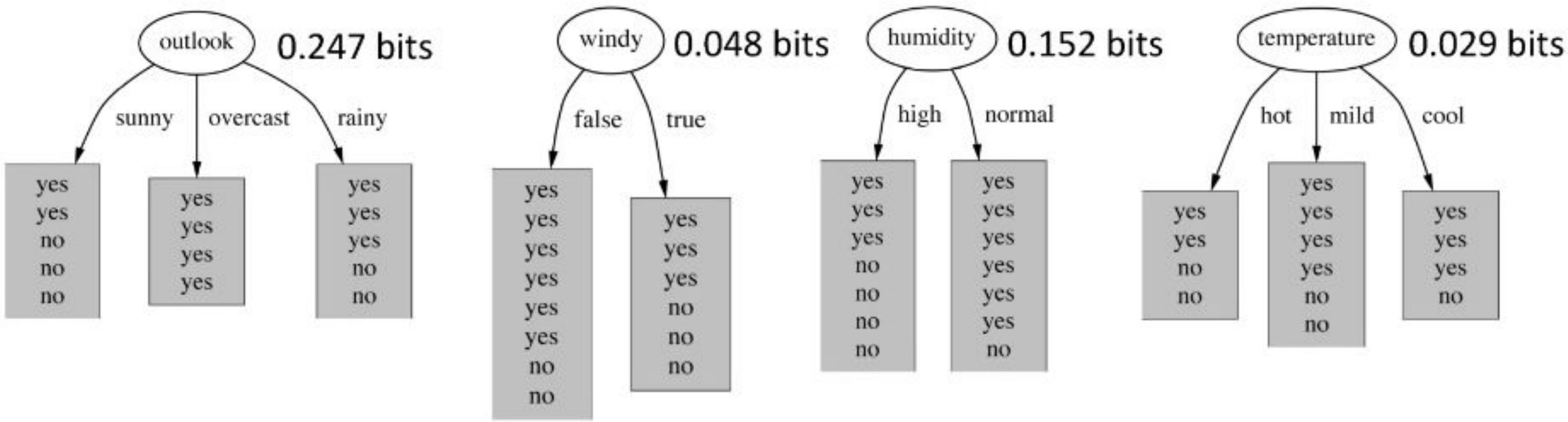
There are 14 instances: 9 yes, 5 no

TotalEntropy([9,5]) = 0.940 bits.

Entropy**outloo**k([2,3], [4,0],[3,2]) = (5/14) * 0.971 + (4/14) * 0.0 + (5/14) * 0.971 **= 0.693**bits

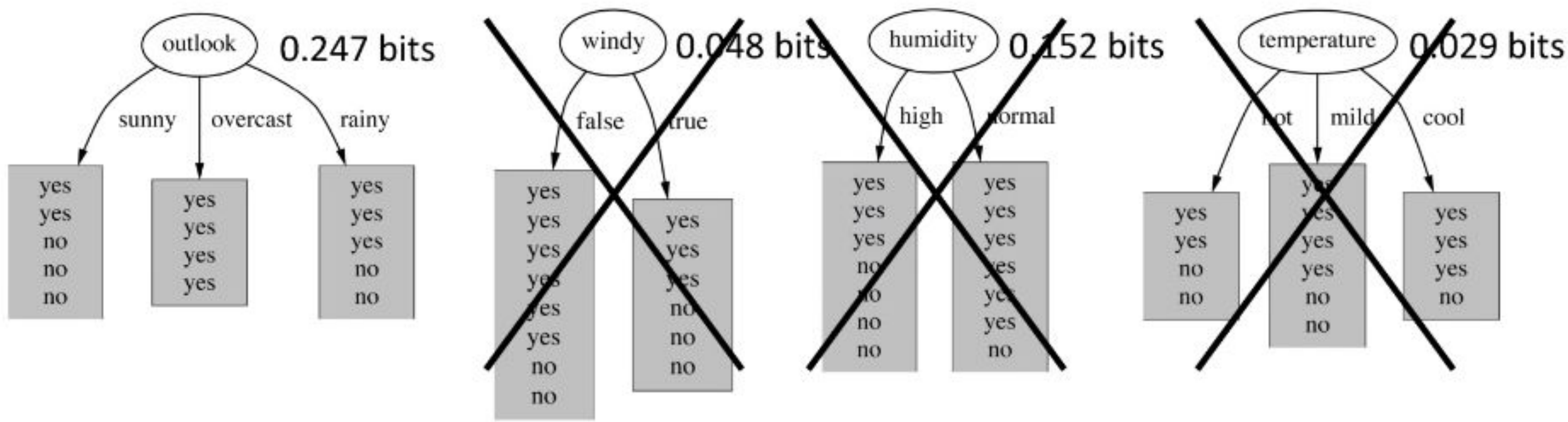
InfoGain = 0.940 – 0.693 = 0.247 bits of information

Which attribute to select?

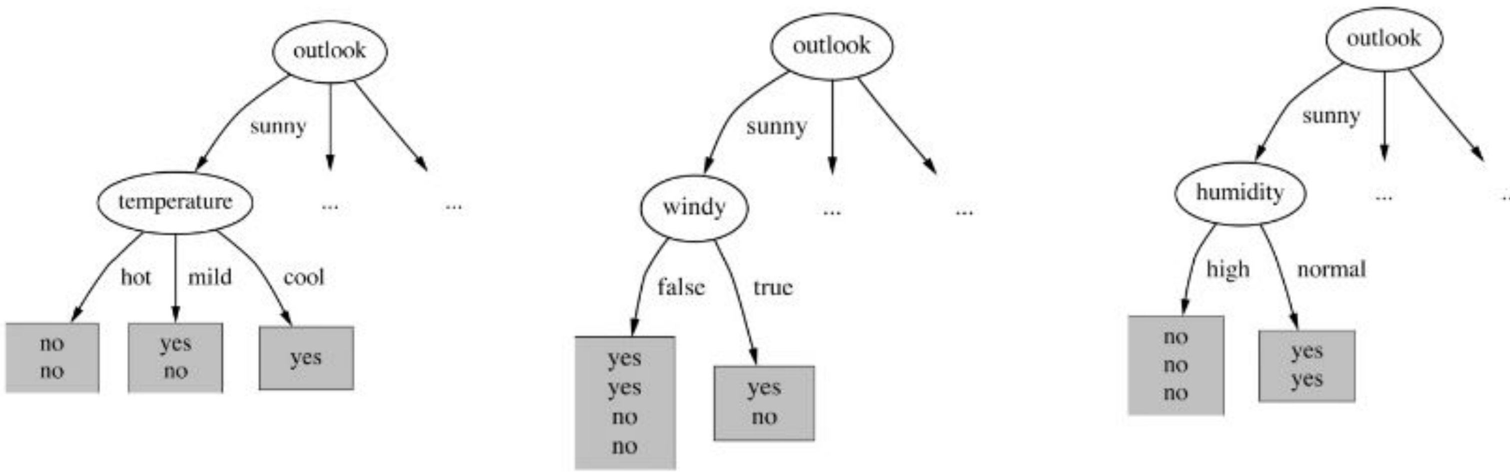


gain(outlook) = 0.940 - 0.693 = 0.247 bits of information.
gain(temperature) = 0.029 bits
gain(humidity) = 0.152 bits
gain(windy) = 0.048 bits

Which attribute to select?



Continue to split …

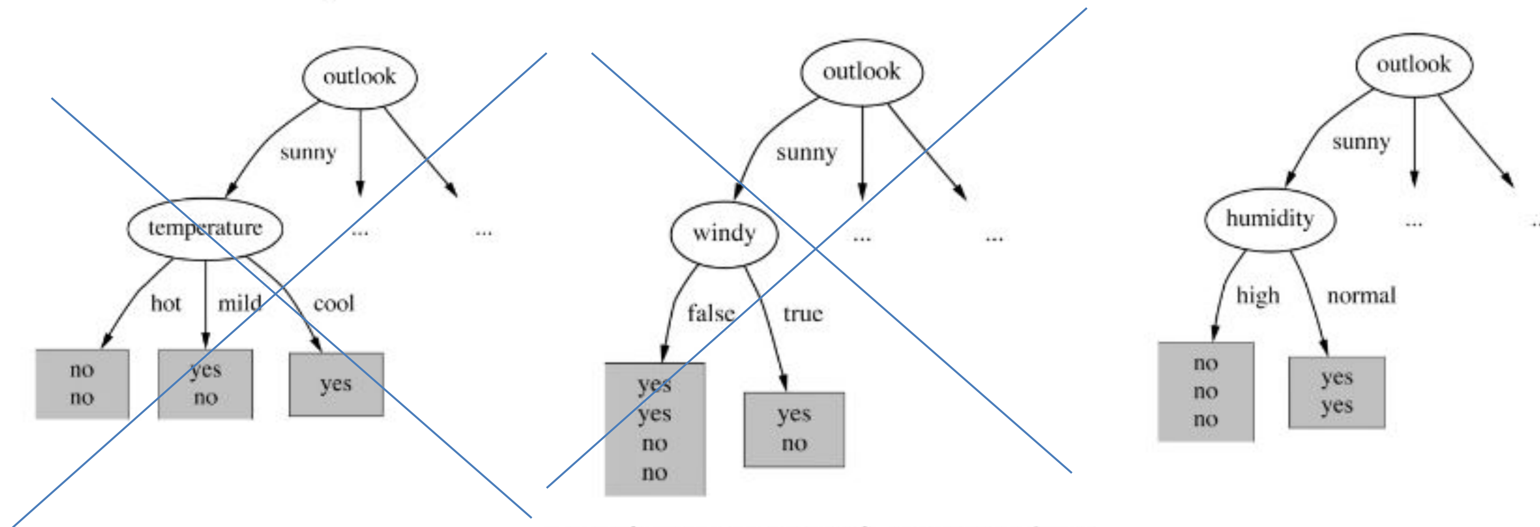


gain(*temperature*) = 0.571 bits

gain(*windy*) = 0.020 bits

gain(*humidity*) = 0.971 bits

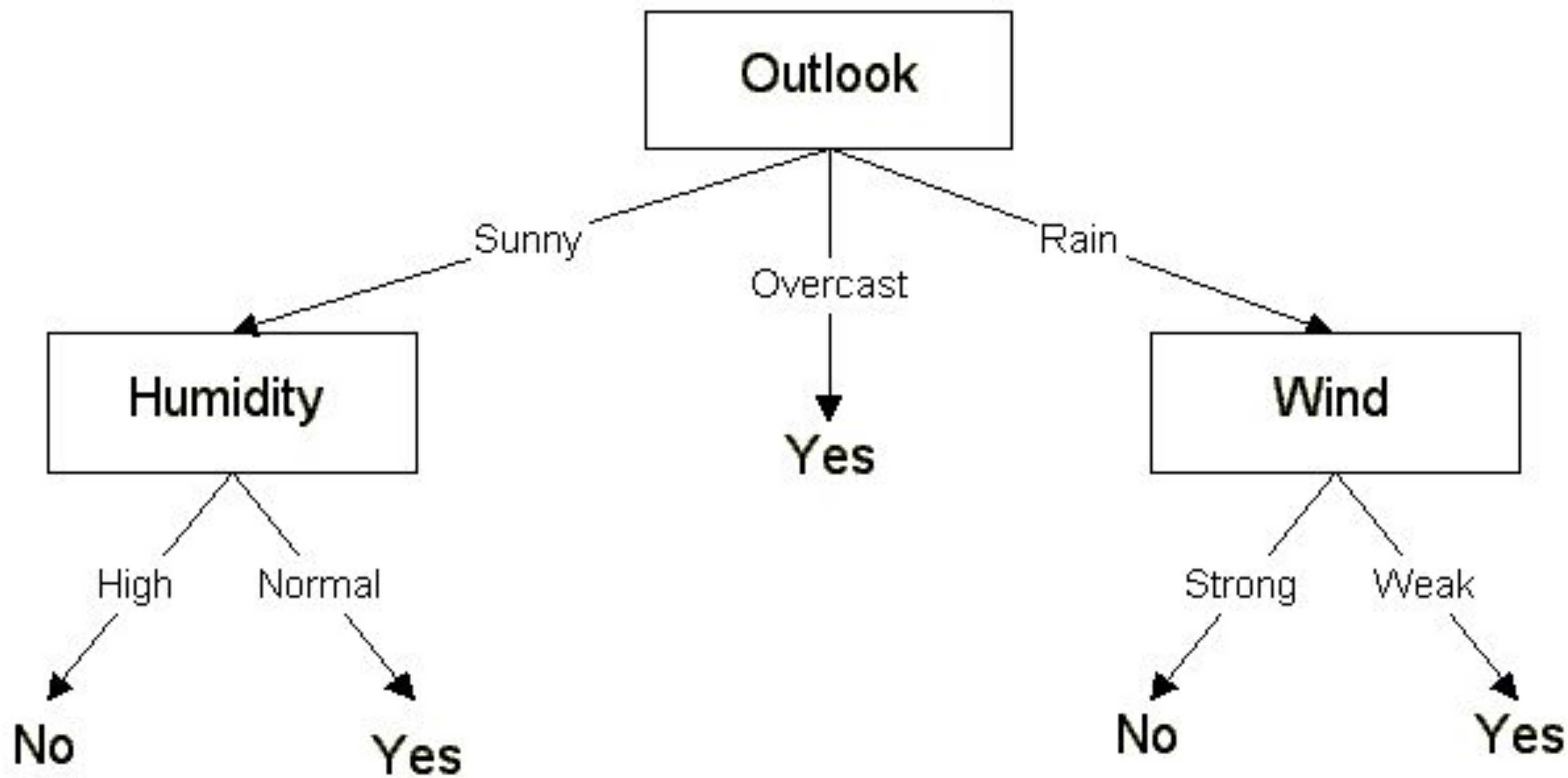
Continue to split …



gain(*temperature*) = 0.571 bits

gain(*windy*) = 0.020 bits

gain(*humidity*) = 0.971 bits



Outlook
Sunny
Overcast
Rain
Humidity
Yes
Wind
High
Normal
Strong
Weak
No
Yes
No
Yes

Thank you