Example: Play Golf dataset - available on Kaggle

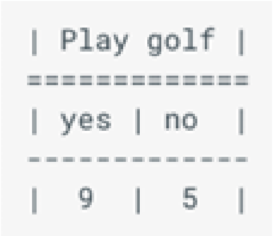
Attributes

class

	Outlook	Temperature	Humidity	Windy	Play Golf
1	Rainy	Hot	High	FALSE	No
2	Rainy	Hot	High	TRUE	No
3	Overcast	Hot	High	FALSE	Yes
4	Sunny	Mild	High	FALSE	Yes
5	Sunny	Cool	Normal	FALSE	Yes
6	Sunny	Cool	Normal	TRUE	No
7	Overcast	Cool	Normal	TRUE	Yes
8	Rainy	Mild	High	FALSE	No
9	Rainy	Cool	Normal	FALSE	Yes
10	Sunny	Mild	Normal	FALSE	Yes
11	Rainy	Mild	Normal	TRUE	Yes
12	Overcast	Mild	High	TRUE	Yes
13	Overcast	Hot	Normal	FALSE	Yes
14	Sunny	Mild	High	TRUE	No

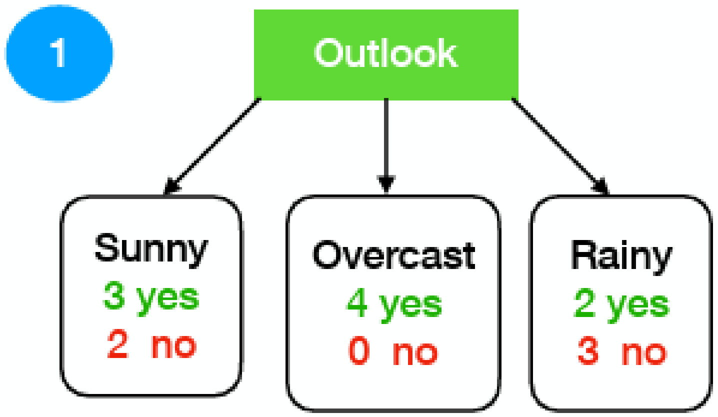
- 4 features:
  - **outlook**: *rainy, overcast, sunny*
  - **temperature**: *cool, mild, hot*
  - **humidity**: *normal, high*
  - **windy**: *false, true*
- Possible outcomes (play golf?):
  - **false**
  - **true**

Frequency Table



```
| Play golf |
=============
| yes | no  |
-------------
|  9  |  5  |
```

Potential Splits on X (attributes)



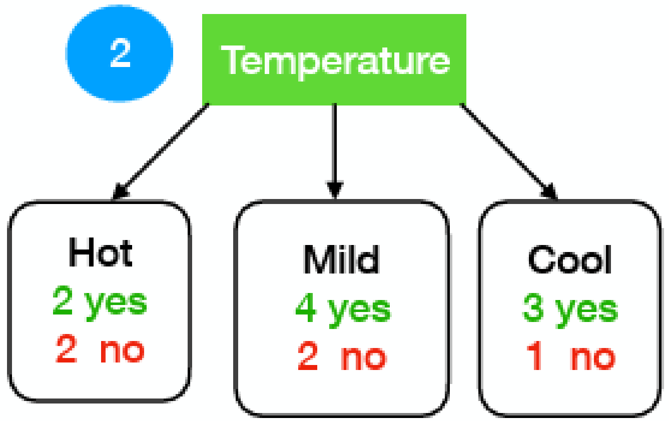
1

Outlook

Sunny
3 yes
2 no

Overcast
4 yes
0 no

Rainy
2 yes
3 no



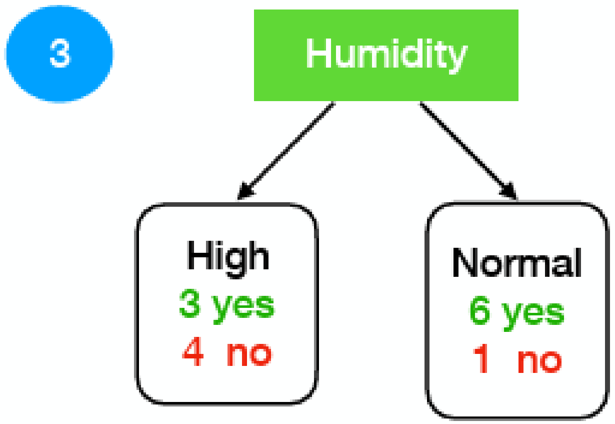
2

Temperature

Hot
2 yes
2 no

Mild
4 yes
2 no

Cool
3 yes
1 no

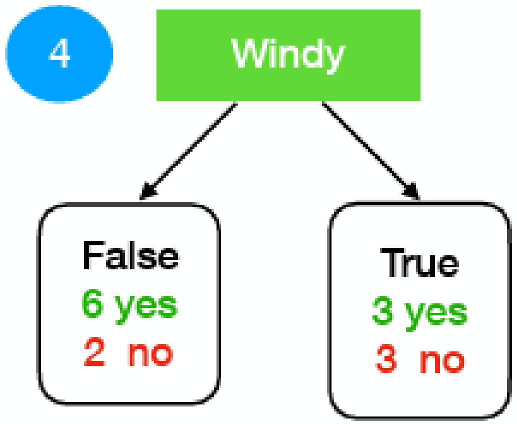


3

Humidity

High
3 yes
4 no

Normal
6 yes
1 no



4

Windy

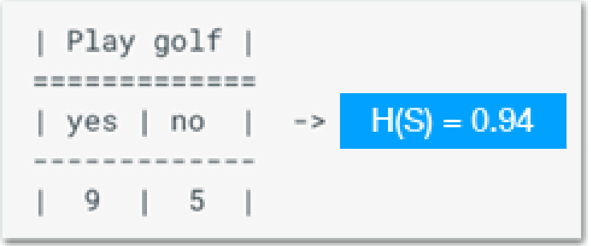
False
6 yes
2 no

True
3 yes
3 no

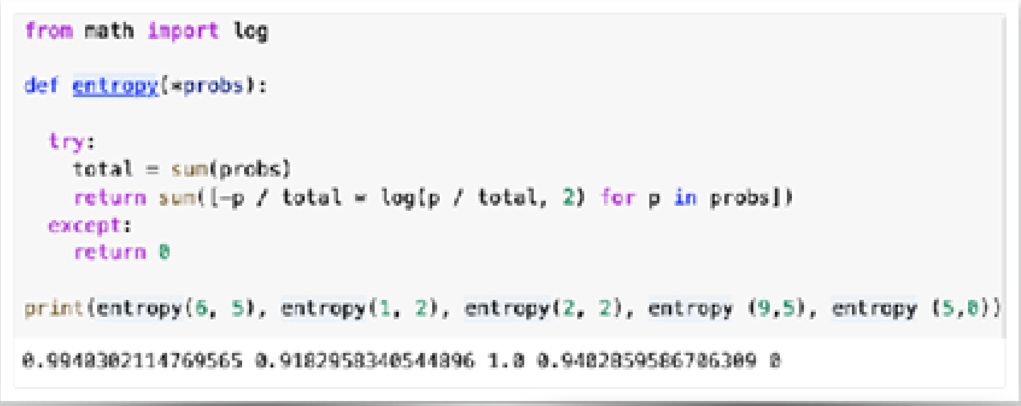
Play not play Tree!

Let S be the set of training samples with c possible classes, thus $S = \{S_1, S_2, ..., S_n\}$

$$\text{Entropy: } H(S) = -\sum_{i=1}^{C} p_i \cdot \log(p_i) = -\sum_{i=1}^{C} \frac{|S_i|}{|S|} \cdot \log(\frac{|S_i|}{|S|})$$



```
| Play golf |
=============
| yes | no  |  ->  H(S) = 0.94
-------------
|  9  |  5  |
```

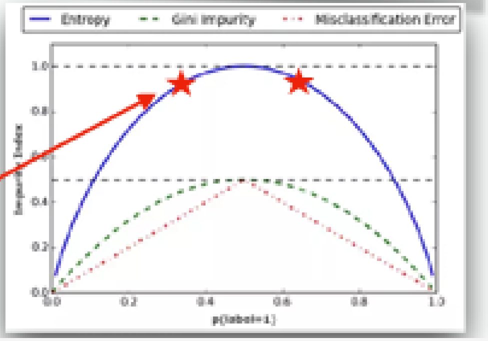


```
from math import log

def entropy(*probs):

  try:
    total = sum(probs)
    return sum([-p / total * log(p / total, 2) for p in probs])
  except:
    return 0

print(entropy(6, 5), entropy(1, 2), entropy(2, 2), entropy (9,5), entropy (5,0))
```

```
0.9940302114769565 0.9182958340544896 1.0 0.9402859586706309 0
```

Entropy [my data = entropy (9,5)]:

$$H(S) = -\frac{9}{14}\log(\frac{9}{14}) - \frac{5}{14}\log(\frac{5}{14}) = \boxed{0.94}$$

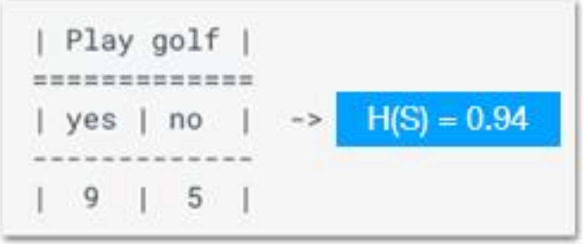


Play not play Tree!

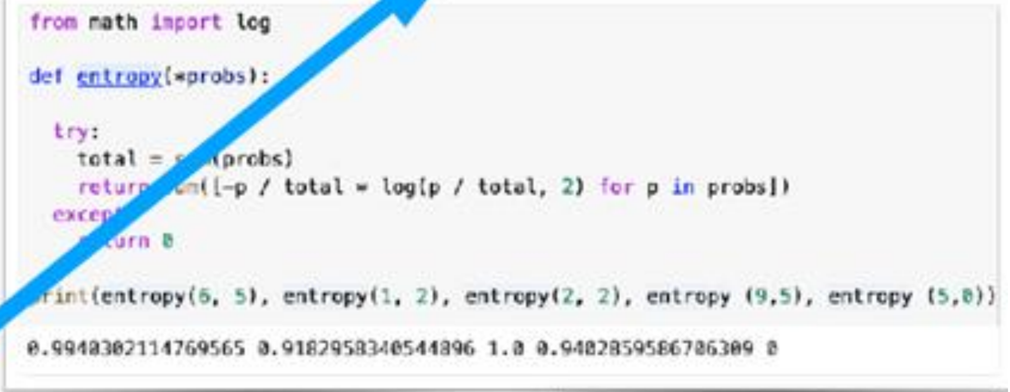
Let S be the set of training samples with c possible classes, thus $S = \{S_1, S_2, ..., S_n\}$

{number of observations of class 1 (i) over the total number of observations}

$$\text{Entropy: } H(S) = -\sum_{i=1}^{C} p_i \cdot \log(p_i) = -\sum_{i=1}^{C} \frac{|S_i|}{|S|} \cdot \log(\frac{|S_i|}{|S|})$$

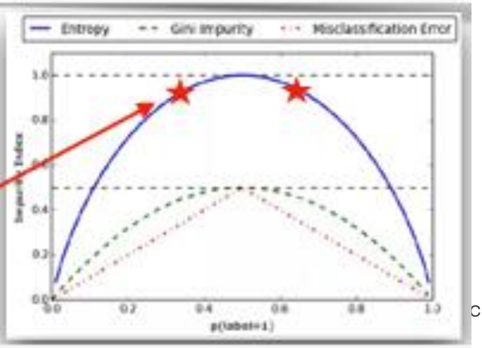


```
| Play golf |
=============
| yes | no  |  ->  H(S) = 0.94
-------------
|  9  |  5  |
```



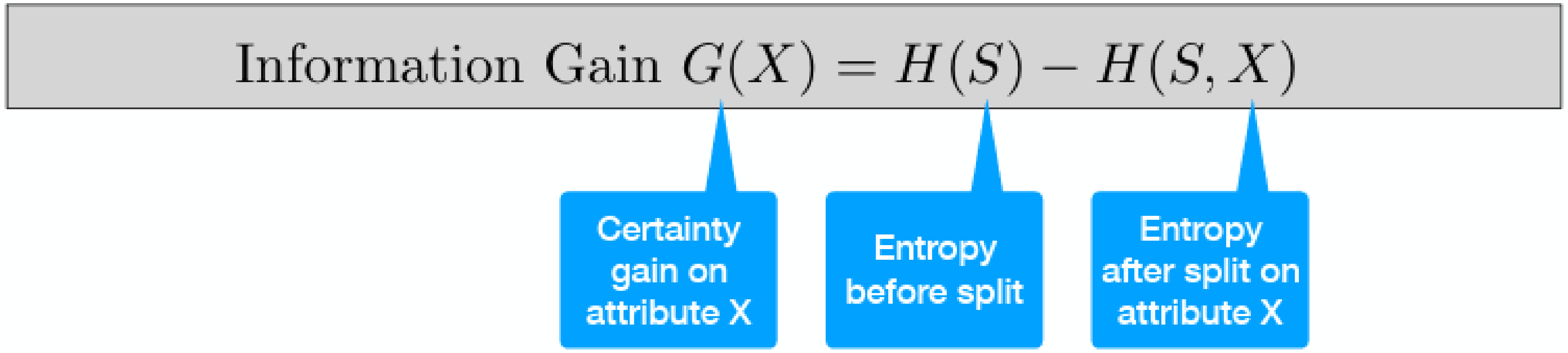
```
from math import log

def entropy(*probs):

  try:
    total = sum(probs)
    return sum([-p / total * log(p / total, 2) for p in probs])
  except:
    return 0

print(entropy(6, 5), entropy(1, 2), entropy(2, 2), entropy (9,5), entropy (5,0))
```

```
0.9940302114769565 0.9182958340544896 1.0 0.9402859586706309 0
```



Entropy [my data = entropy (9,5)]:

$$H(S) = -\frac{9}{14}\log(\frac{9}{14}) - \frac{5}{14}\log(\frac{5}{14}) = 0.94$$

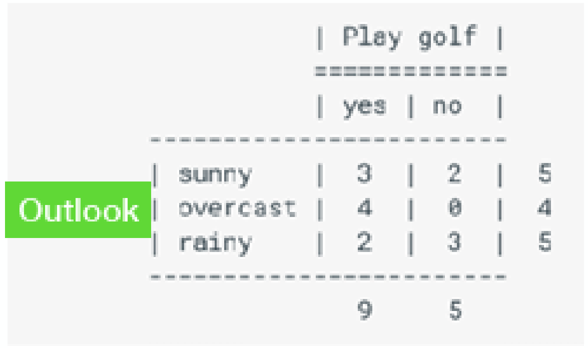


Information Gain $G(X) = H(S) - H(S,X)$

Certainty gain on attribute X

Entropy before split

Entropy after split on attribute X



```
                    | Play golf |
                    =============
                    | yes | no  |
         --------------------------
         | sunny    |  3  |  2  |  5
 Outlook | overcast |  4  |  0  |  4
         | rainy    |  2  |  3  |  5
         --------------------------
                       9     5
```

entropy(3, 2), 0, entropy(2, 3)

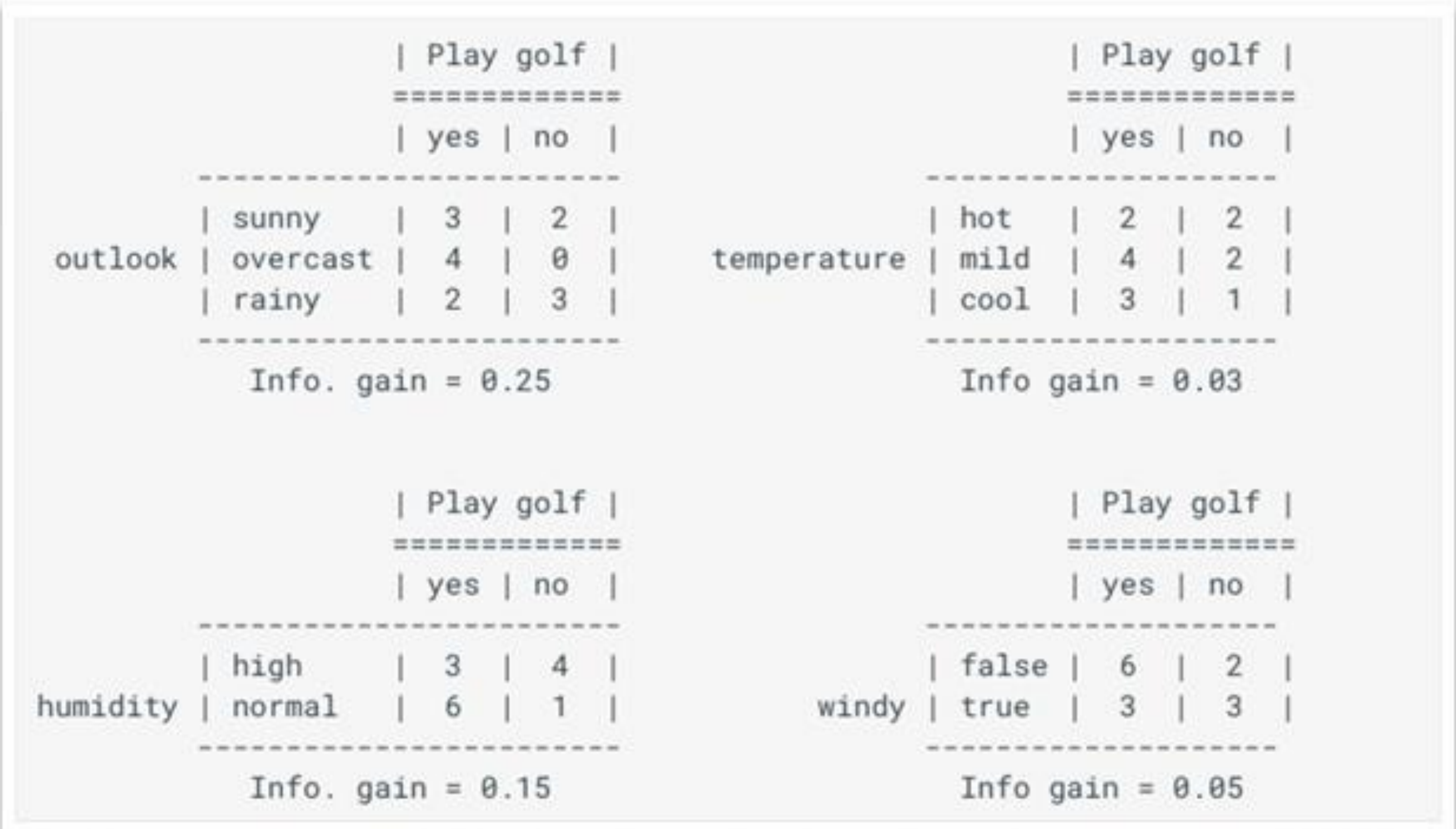
$H(\text{sunny}) = 0.97$
$H(\text{overcast}) = 0$
$H(\text{rainy}) = 0.97$

$$H(S,\text{outlook}) = P(\text{sunny}) \cdot H(\text{sunny}) + P(\text{overcast}) \cdot H(\text{overcast}) + P(\text{rainy}) \cdot H(\text{rainy})$$

$$= \frac{5}{14} \cdot 0.97 + \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot 0.97 = 0.69$$

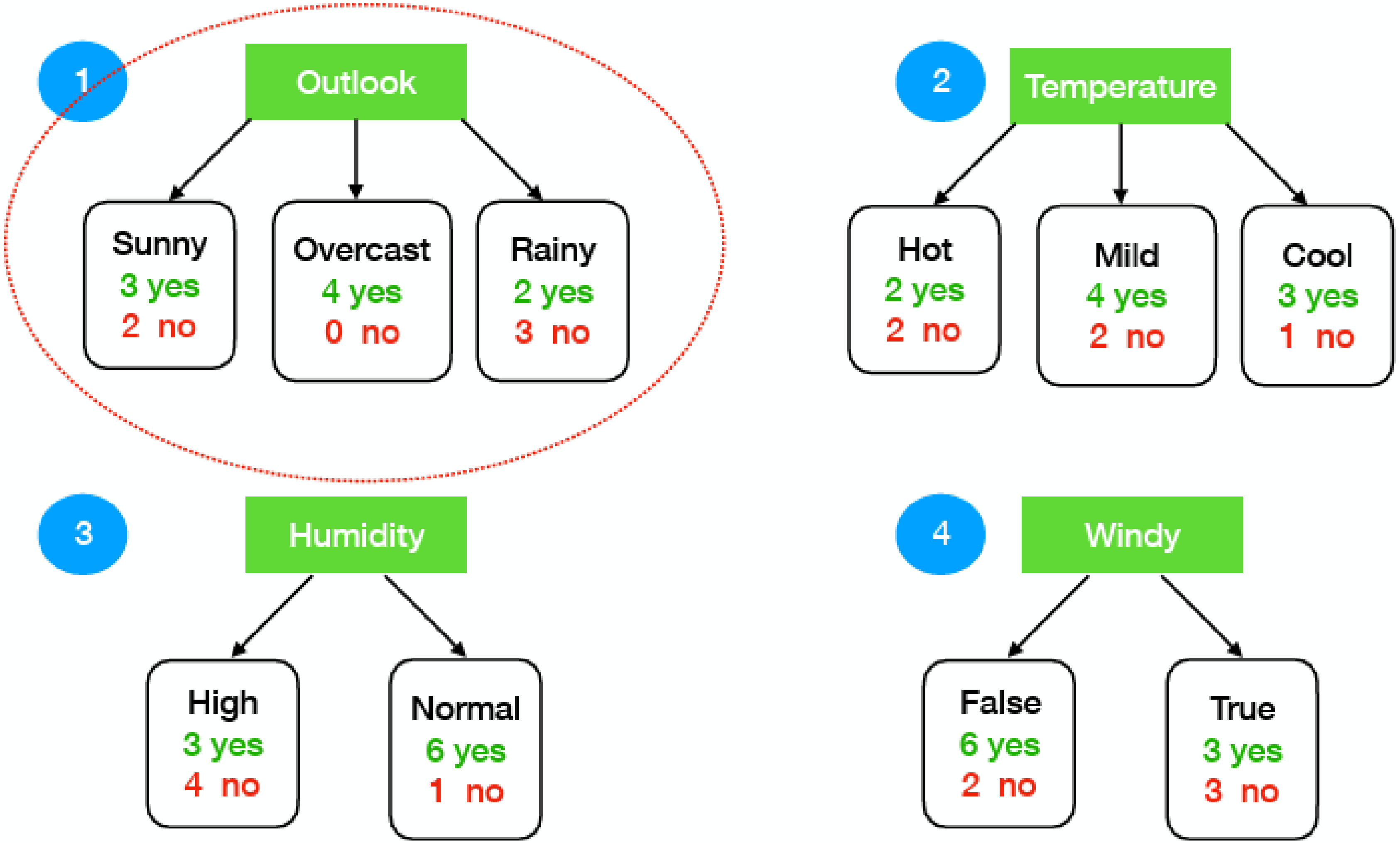
Information Gain $G(\text{outlook}) = H(S) - H(S,\text{outlook}) = 0.94 - 0.69 = 0.25$

Potential Splits on X (attributes)



```
                    | Play golf |                           | Play golf |
                    =============                           =============
                    | yes | no  |                           | yes | no  |
         ------------------------                   ---------------------
         | sunny    |  3  |  2  |                   | hot   |  2  |  2  |
 outlook | overcast |  4  |  0  |       temperature | mild  |  4  |  2  |
         | rainy    |  2  |  3  |                   | cool  |  3  |  1  |
         ------------------------                   ---------------------
            Info. gain = 0.25                          Info gain = 0.03


                    | Play golf |                           | Play golf |
                    =============                           =============
                    | yes | no  |                           | yes | no  |
         ------------------------                   ---------------------
         | high     |  3  |  4  |                   | false |  6  |  2  |
humidity | normal   |  6  |  1  |             windy | true  |  3  |  3  |
         ------------------------                   ---------------------
            Info. gain = 0.15                          Info gain = 0.05
```

Potential Splits on X (attributes)



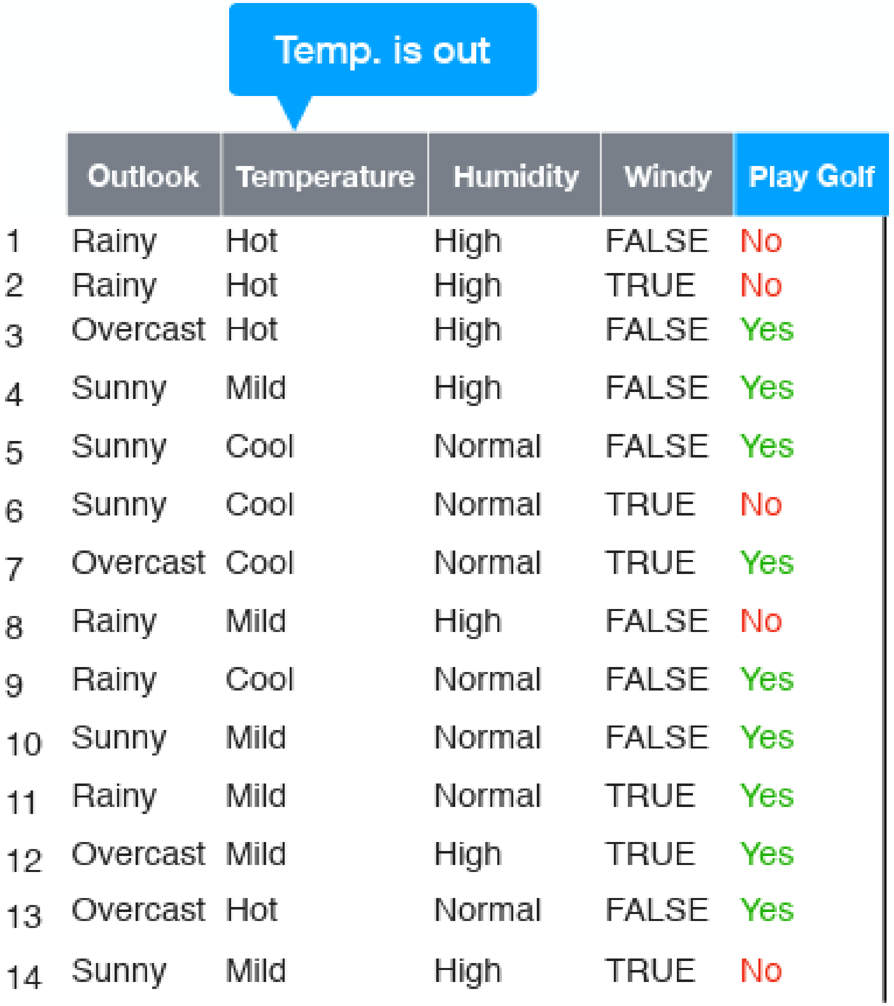
1 Outlook
- Sunny: 3 yes, 2 no
- Overcast: 4 yes, 0 no
- Rainy: 2 yes, 3 no

2 Temperature
- Hot: 2 yes, 2 no
- Mild: 4 yes, 2 no
- Cool: 3 yes, 1 no

3 Humidity
- High: 3 yes, 4 no
- Normal: 6 yes, 1 no

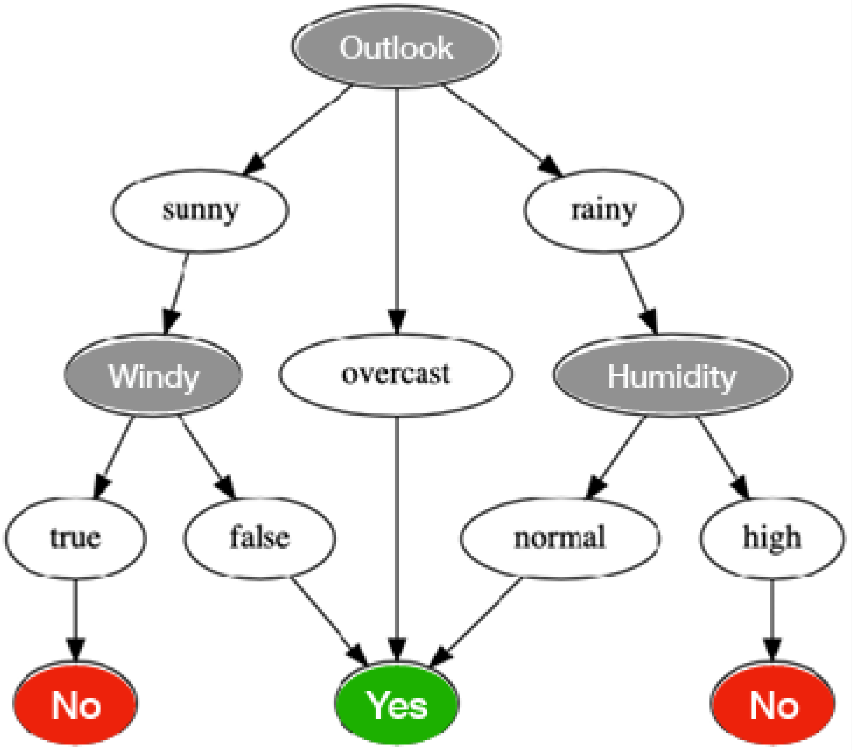
4 Windy
- False: 6 yes, 2 no
- True: 3 yes, 3 no

Our tree then is:



Temp. is out

	Outlook	Temperature	Humidity	Windy	Play Golf
1	Rainy	Hot	High	FALSE	No
2	Rainy	Hot	High	TRUE	No
3	Overcast	Hot	High	FALSE	Yes
4	Sunny	Mild	High	FALSE	Yes
5	Sunny	Cool	Normal	FALSE	Yes
6	Sunny	Cool	Normal	TRUE	No
7	Overcast	Cool	Normal	TRUE	Yes
8	Rainy	Mild	High	FALSE	No
9	Rainy	Cool	Normal	FALSE	Yes
10	Sunny	Mild	Normal	FALSE	Yes
11	Rainy	Mild	Normal	TRUE	Yes
12	Overcast	Mild	High	TRUE	Yes
13	Overcast	Hot	Normal	FALSE	Yes
14	Sunny	Mild	High	TRUE	No


Outlook
sunny
rainy
Windy
overcast
Humidity
true
false
normal
high
No
Yes
No