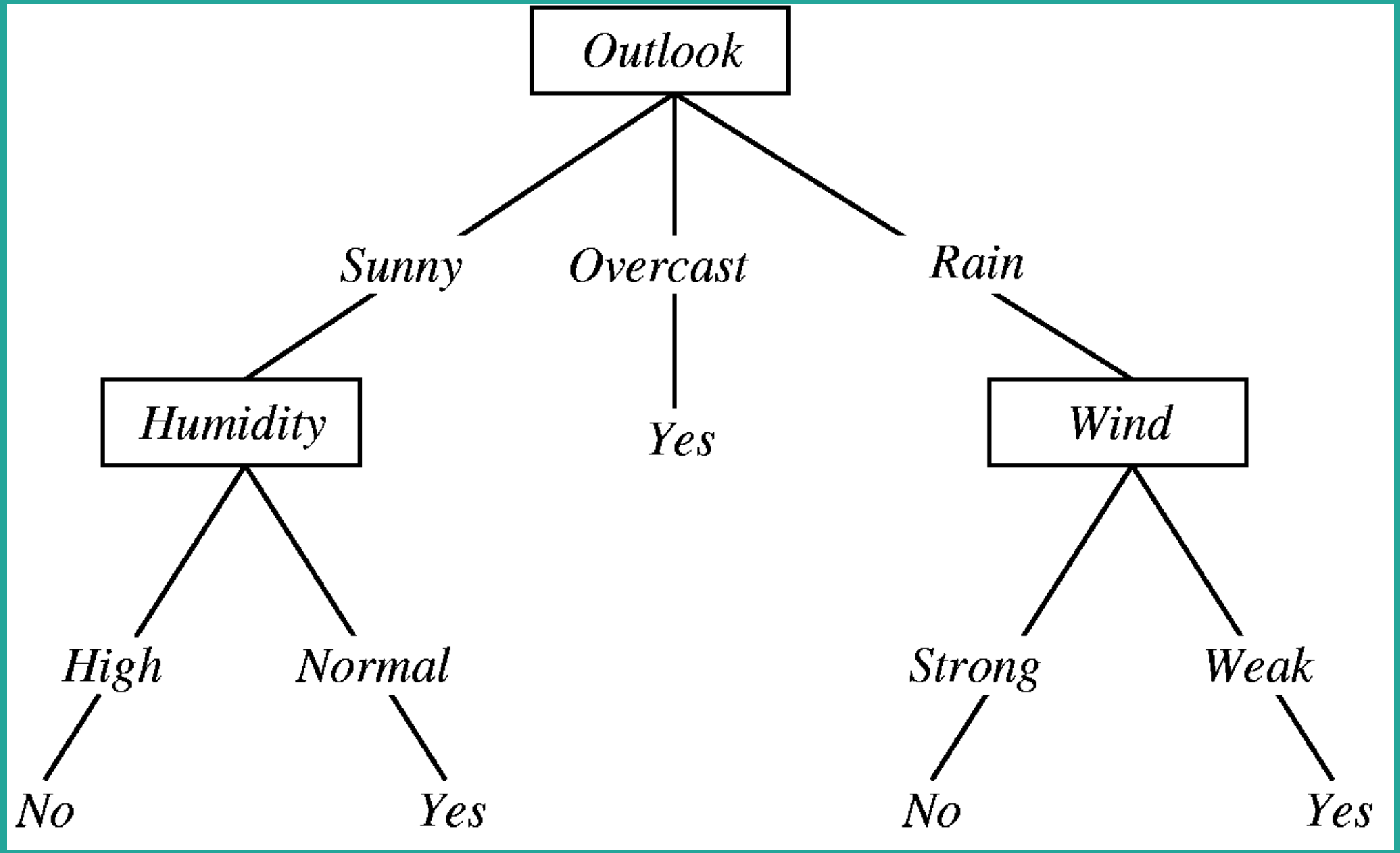
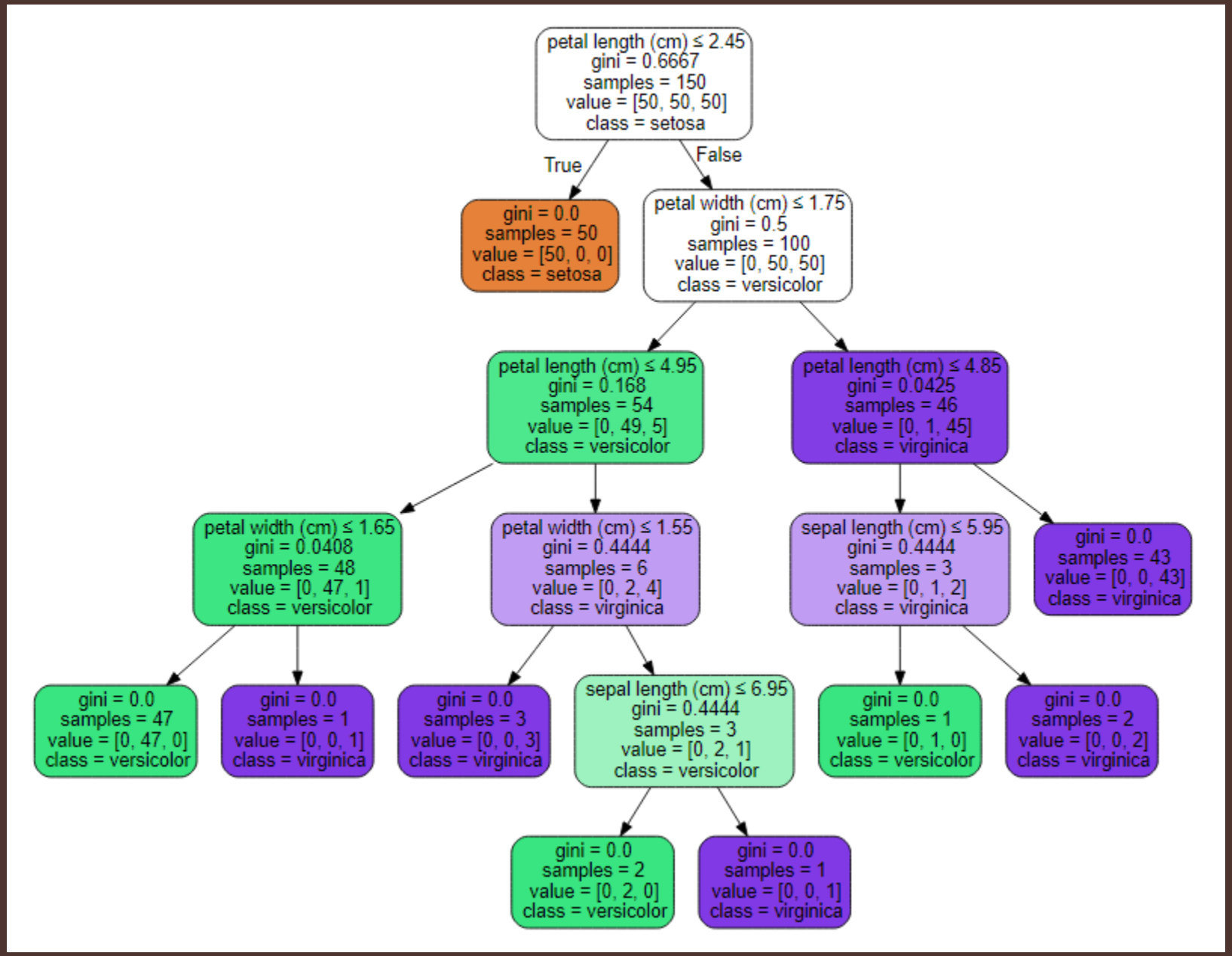
Decision Trees


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


petal length (cm) ≤ 2.45
gini = 0.6667
samples = 150
value = [50, 50, 50]
class = setosa
True
False
gini = 0.0
samples = 50
value = [50, 0, 0]
class = setosa
petal width (cm) ≤ 1.75
gini = 0.5
samples = 100
value = [0, 50, 50]
class = versicolor
petal length (cm) ≤ 4.95
gini = 0.168
samples = 54
value = [0, 49, 5]
class = versicolor
petal length (cm) ≤ 4.85
gini = 0.0425
samples = 46
value = [0, 1, 45]
class = virginica
petal width (cm) ≤ 1.65
gini = 0.0408
samples = 48
value = [0, 47, 1]
class = versicolor
petal width (cm) ≤ 1.55
gini = 0.4444
samples = 6
value = [0, 2, 4]
class = virginica
sepal length (cm) ≤ 5.95
gini = 0.4444
samples = 3
value = [0, 1, 2]
class = virginica
gini = 0.0
samples = 43
value = [0, 0, 43]
class = virginica
gini = 0.0
samples = 47
value = [0, 47, 0]
class = versicolor
gini = 0.0
samples = 1
value = [0, 0, 1]
class = virginica
gini = 0.0
samples = 3
value = [0, 0, 3]
class = virginica
sepal length (cm) ≤ 6.95
gini = 0.4444
samples = 3
value = [0, 2, 1]
class = versicolor
gini = 0.0
samples = 1
value = [0, 1, 0]
class = versicolor
gini = 0.0
samples = 2
value = [0, 0, 2]
class = virginica
gini = 0.0
samples = 2
value = [0, 2, 0]
class = versicolor
gini = 0.0
samples = 1
value = [0, 0, 1]
class = virginica

ID3 (X, y, f, F)

Create a root node R for the tree

If all $y^{(i)} = c \; in \; X$ return node with label c

If $F = \emptyset$, then Return node with label = most common value of $y^{(i)}$

Else

$a \leftarrow$ best attribute classifier

Decision Tree attribute for $R = f$.

For each possible value, v_k, of f,

Create a new branch b_k below R, corresponding to $X_f = v_k$.

Let $E_k \subset X$ where E_k is the set of all examples where $X_f = v_k$

If $E_k = \emptyset$

Then add a node with label = most common value of $y^{(i)}$

Else add the subtree ID3 $(E_k, y, a, F - \{f\})$ below

Return R

How do you find the best feature classifier for a set of examples?

Shape	Color	Size	Fruit
Round	Red	Big	APPLE
Round	Orange	Big	ORANGE
Round	Red	Big	APPLE
Round	Orange	Medium	ORANGE

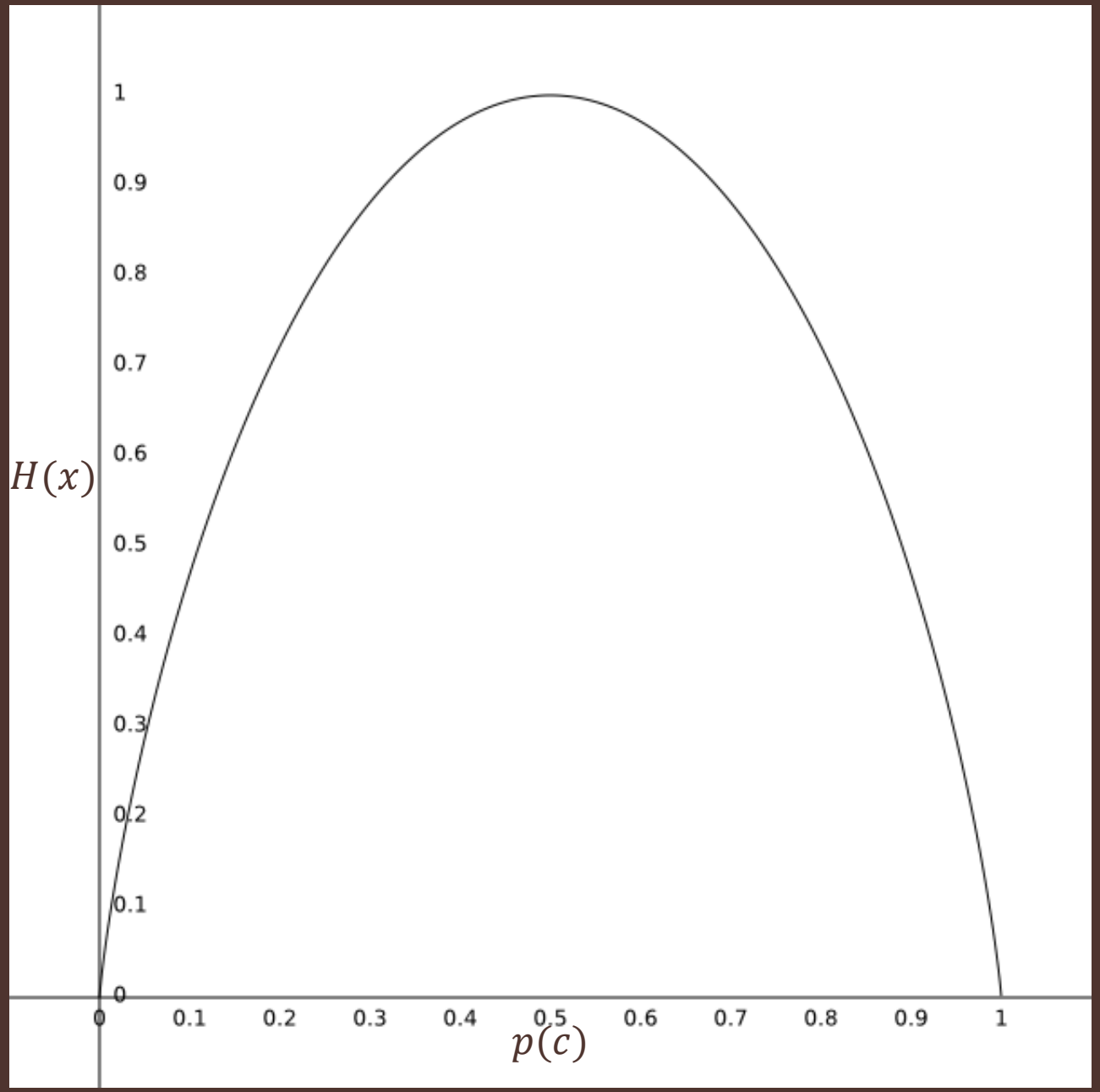
Color is the best classifier since different values for color split the examples into uniform subsets of the examples

The splitting ability of a feature can be measured by the uniformity of the subsets

The uniformity of a subset can be measured using the measure of set uncertainty: **ENTROPY** ($H(S)$)

$$H(X) = \sum_{c \in C} -p(c) \log_2 p(c)$$

where:
$H(X)$ is the information gain of set X
C is set of possible values for y
$p(c) = \frac{N_c}{|S|}$ where N_c is the number of elements in S where $y = c$


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
$H(x)$
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
$p(c)$

The more uniform the subset is, the lower the entropy

Using the measure of entropy, we can choose the best feature using the **INFORMATION GAIN** function ($IG(f, X)$)

$$IG(f,X) = H(X) - \sum_{t \in T} p(t)H(t)$$

where:

$IG(f,X)$ is the information gain of splitting X using feature f

T is set of subsets of X split using f

$p(t) = \frac{|t|}{|s|}$

The higher the information gain the better the feature is at splitting X

outlook	temp	humidity	windy	play
sunny	hot	high	false	NO
sunny	hot	high	true	NO
overcast	hot	high	false	YES
rainy	mild	high	false	YES
rainy	cool	normal	false	YES
rainy	cool	normal	true	NO
overcast	cool	normal	true	YES
sunny	mild	high	false	NO
sunny	cool	normal	false	YES
rainy	mild	normal	false	YES
sunny	mild	normal	true	YES
overcast	mild	high	true	YES
overcast	hot	normal	false	YES
rainy	mild	high	true	NO

$$
\begin{aligned}
&IG(\text{outlook}, X) \\
&= H(X) - (p(\text{sunny})H(\text{sunny}) \\
&+ p(\text{overcast})H(\text{overcast}) \\
&+ p(\text{rainy})H(\text{rainy}))
\end{aligned}
$$

$$H(X) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14}$$
$$= -(0.36)(-1.49) - (0.64)(-0.64)$$
$$= 0.946$$

$$H(\text{sunny}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$
$$= -(0.6)(-0.74) - (0.4)(-1.32)$$
$$= 0.972$$

$$H(\text{sunny}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$
$$= -(0.6)(-0.74) - (0.4)(-1.32)$$
$$= 0.972$$

$$H(\text{overcast}t) = -\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4}$$
$$= -(0) - (1)(0)$$
$$= 0$$

$$H(\text{rainy}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$
$$= -(0.75)(-0.42) - (0.25)(-2)$$
$$= 0.815$$

$$IG(\text{outlook}, X)$$
$$= 0.946$$
$$- \left(\frac{5}{14}(0.972) + \frac{4}{14}(0) + \frac{5}{14}(0.815) \right)$$
$$= 0.307$$

Overfit Decision Trees

Decision Trees can be prone to overfitting. There are many ways to regularize overfit decision trees:

- Pruning to reduce the size of trees
- Increasing entropy threshold for building leafs
- Creating forest of decision trees instead of one tree to classify

Regression Decision Trees

To convert decision trees for regression, instead of calculating the entropy, heterogeneity is used. Heterogeneity can be calculated using standard deviation. If the standard deviation is small the heterogeneity is low.