Implementation

We select the best attribute at each node by computing its information gain, which is the decrease in entropy of the dataset after it has been split on that attribute.

$$\text{Gain(Attribute)} = I(p,n) - \left[\frac{p_0+n_0}{p+n}\, I(p_0,n_0) + \frac{p_1+n_1}{p+n}\, I(p_1,n_1)\right]$$

$$\begin{aligned}
p &= \text{Number of positive examples before split} \\
n &= \text{Number of negative examples before split} \\
p_k &= \text{Number of positive examples with attribute} = k \\
n_k &= \text{Number of negative examples with attribute} = k \\
I(p,n) &= -\frac{p}{p+n}\log\frac{p}{p+n} - \frac{n}{p+n}\log\frac{n}{p+n}
\end{aligned}$$

We could have used other metrics such as the Gini impurity:

$$I(p,n) = \frac{p}{p+n}\left(1-\frac{p}{p+n}\right) + \frac{n}{p+n}\left(1-\frac{n}{p+n}\right)$$

which should give similar results to entropy since $-x\log x \approx x(1-x)$ when $x \approx 0$.

Cross validation:
1. Shuffle the dataset and split it into $K = 10$ parts
2. For each $k \in \{1,\ldots,K\}$ we train the decision tree on the dataset excluding part k and then test the tree on part k.

Evaluation

Confusion matrix:

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger						
Disgust						
Fear						
Happiness						
Sadness						
Surprise						

Precision
Recall
F_1 score

Miscellaneous

Noisy-Clean Datasets Question

The noisy dataset has lower performance.

Ambiguity Question

Pruning Question