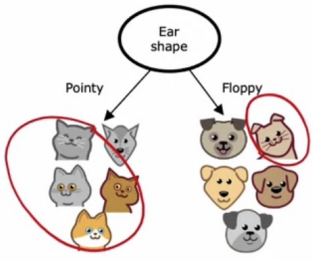
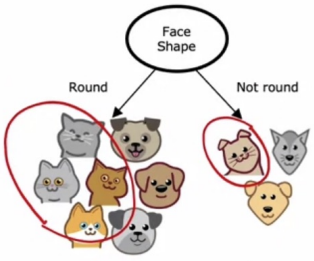
Suppose, we have to choose between these 3 features and check which one gives us the lowest entropy.



$p_1 = 4/5 = 0.8$ $p_1 = 1/5 = 0.2$

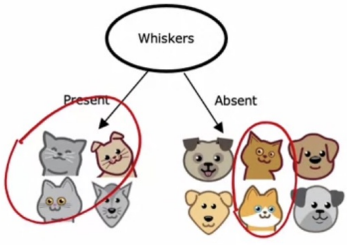
$H(0.8) = 0.72$ $H(0.2) = 0.72$

entropy of left side — entropy of right side



$p_1 = 4/7 = 0.57$ $p_1 = 1/3 = 0.33$

$H(0.57) = 0.99$ $H(0.33) = 0.92$



$p_1 = 3/4 = 0.75$ $p_1 = 2/6 = 0.33$

$H(0.75) = 0.81$ $H(0.33) = 0.92$

To choose the best feature, we calculate the "information gain".

Information gain:

$$= H(p_1^{root}) - \left(w^{left} H(p_1^{left}) + w^{right} H(p_1^{right}) \right)$$

p_1^{root} – fraction of cats (pure class) in the root node.

w^{left} – fraction of examples that went to the left side and w^{right} – fraction of examples that went to the right side.

p_1^{left} – fraction of cats on the left side and p_1^{right} fraction of cats on the right side.

Q. How did we come up with that?

A. It wouldn't be smart to just add the entropy on the left and entropy on the right and judge which feature gives least entropy (impurity) because cases where entropy is high and no. of examples is also high is worse than a set with less examples and high entropy.

Therefore, we choose a weighted average:-

$$w^{left} H(p_1^{left}) + w^{right} H(p_1^{right})$$

But we don't just want the weighted average, we want reduction in entropy.

↓

subtract weighted average from entropy of p_1^{root}

Q. Why do we want reduction in entropy?

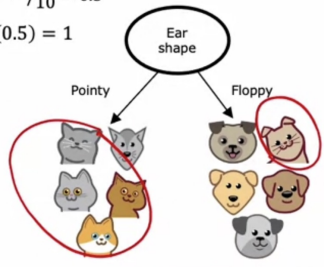
A. One of major criteria to decide when to stop splitting further is checking reduction in entropy. If reduction in entropy is too small, we stop splitting.

Information gain helps us to find the feature which ensures maximum purity on subset of both sides.

cat example :-

$p_1 = {}^{5}/_{10} = 0.5$

$H(0.5) = 1$



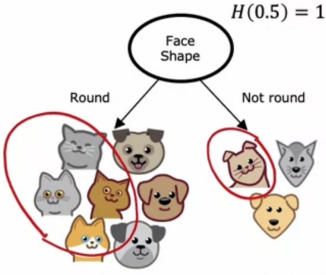
$p_1 = {}^{4}/_{5} = 0.8$ $p_1 = {}^{1}/_{5} = 0.2$

$H(0.8) = 0.72$ $H(0.2) = 0.72$

$$H(0.5) - \left(\frac{5}{10}H(0.8) + \frac{5}{10}H(0.2)\right)$$

$= 0.28$

$H(0.5) = 1$



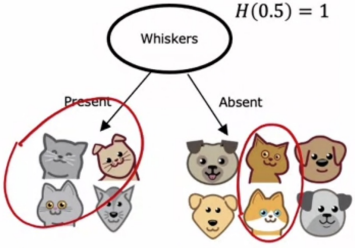
$p_1 = {}^{4}/_{7} = 0.57$ $p_1 = {}^{1}/_{3} = 0.33$

$H(0.57) = 0.99$ $H(0.33) = 0.92$

$$H(0.5) - \left(\frac{7}{10}H(0.57) + \frac{3}{10}H(0.33)\right)$$

$= 0.03$

$H(0.5) = 1$



$p_1 = {}^{3}/_{4} = 0.75$ $p_1 = {}^{2}/_{6} = 0.33$

$H(0.75) = 0.81$ $H(0.33) = 0.92$

$$H(0.5) - \left(\frac{4}{10}H(0.75) + \frac{6}{10}H(0.33)\right)$$

$= 0.12$

Information gain