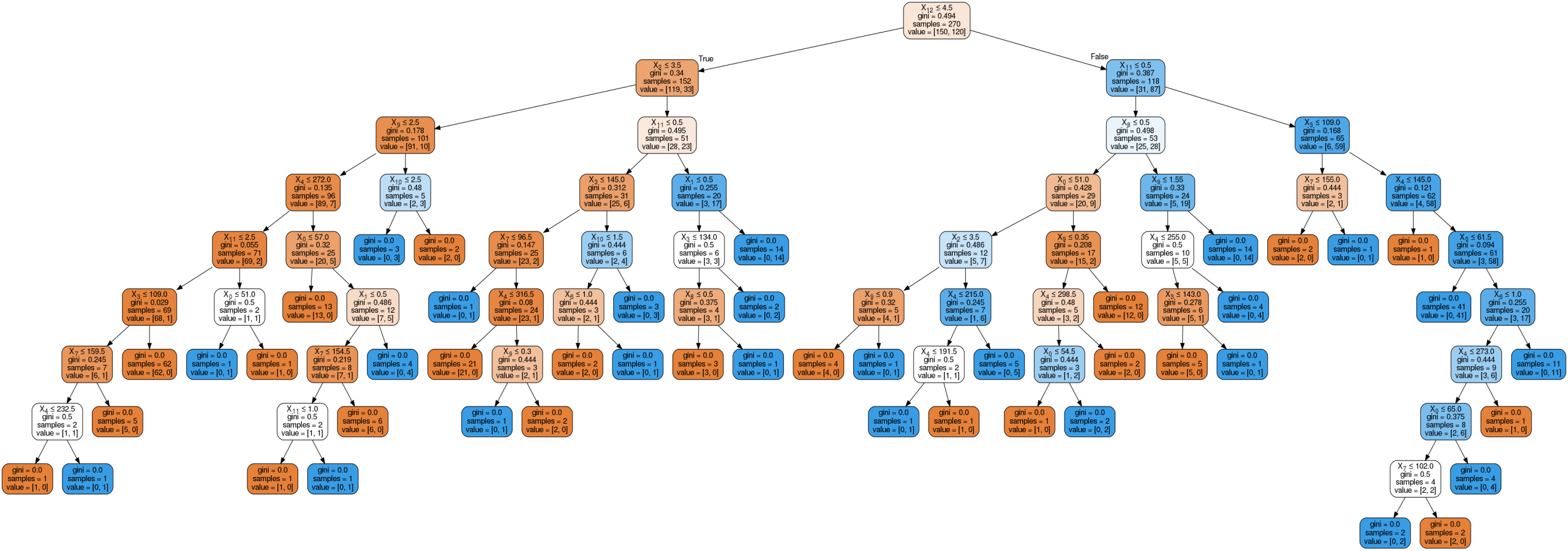
Decision Trees Classifier

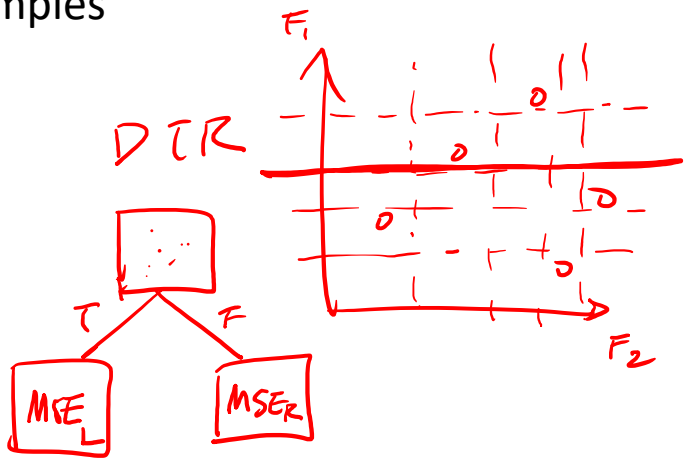
Decision Tree Classifier

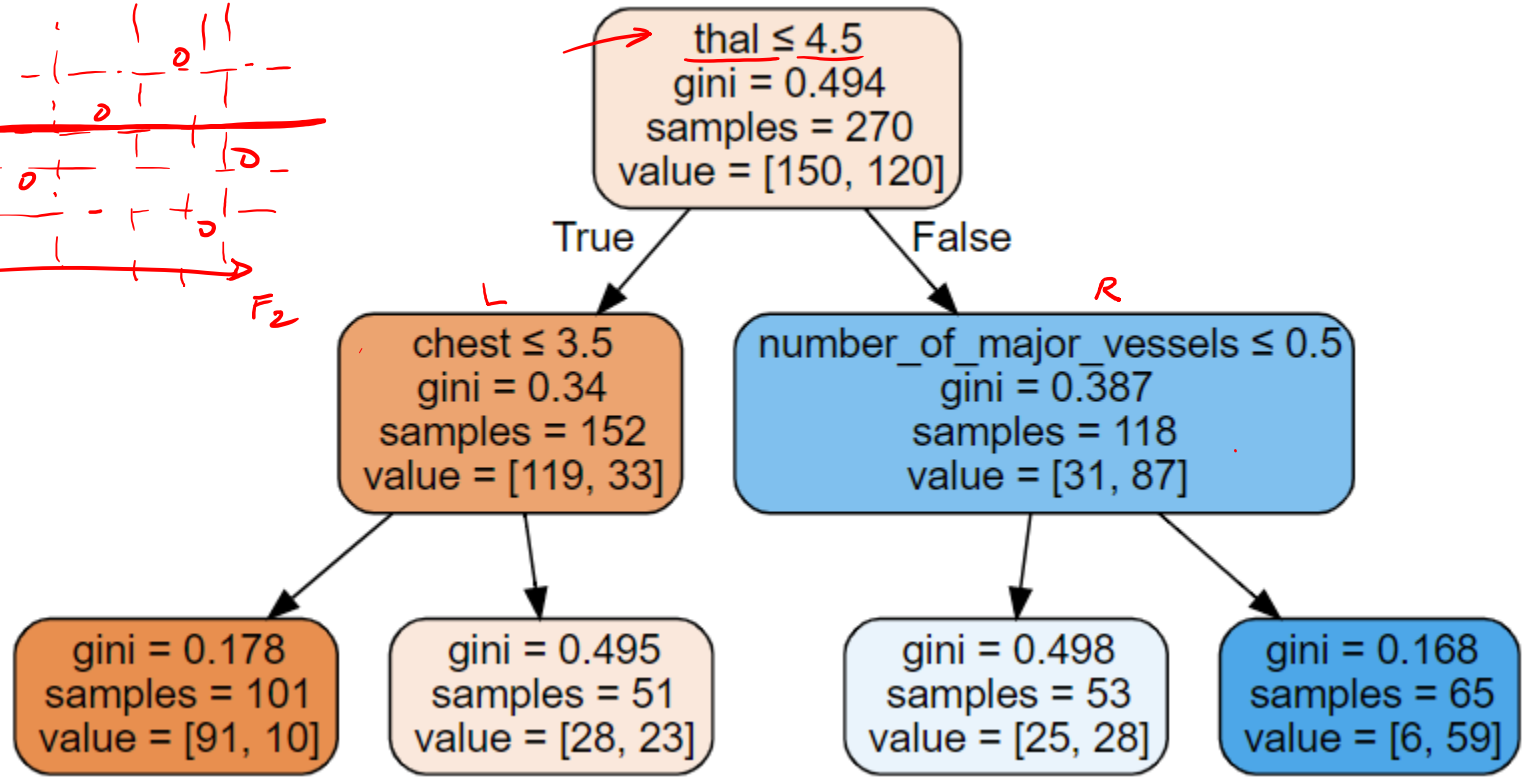
Heart Dataset (statlog version)
13 features
270 samples



Decision Tree Classifier

Heart Dataset (statlog version)
13 features
270 samples





Decision Tree Split Criteria

Regression Tree

MSE (RSS)

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} (y_i - \bar{y}_m)^2$$

MAE

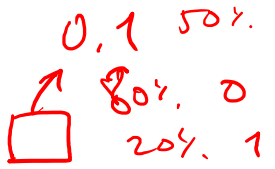
$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} |y_i - \bar{y}_m|$$

Classification Tree

Gini

impurity

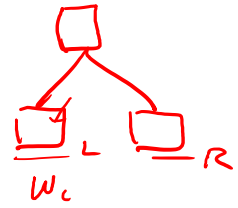
$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$



Entropy

uncertainty

$$H(X_m) = -\sum_k p_{mk} \log(p_{mk})$$

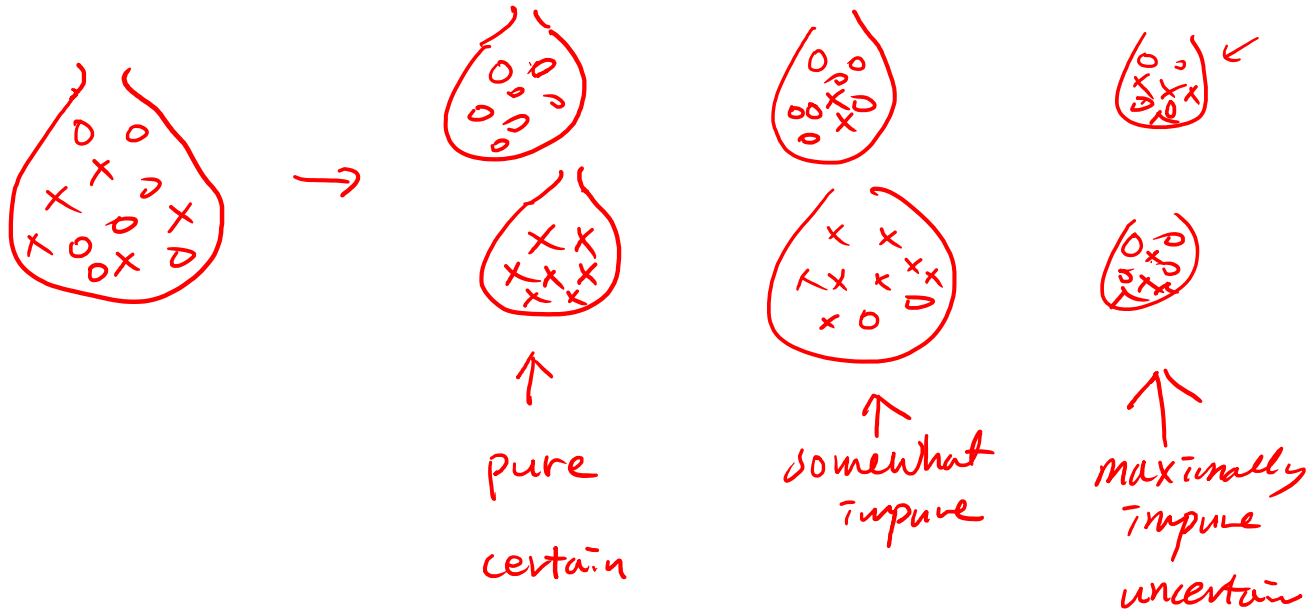


Information Gain = E(parent)-E(children)

Decision Tree Split Criteria

Gini

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

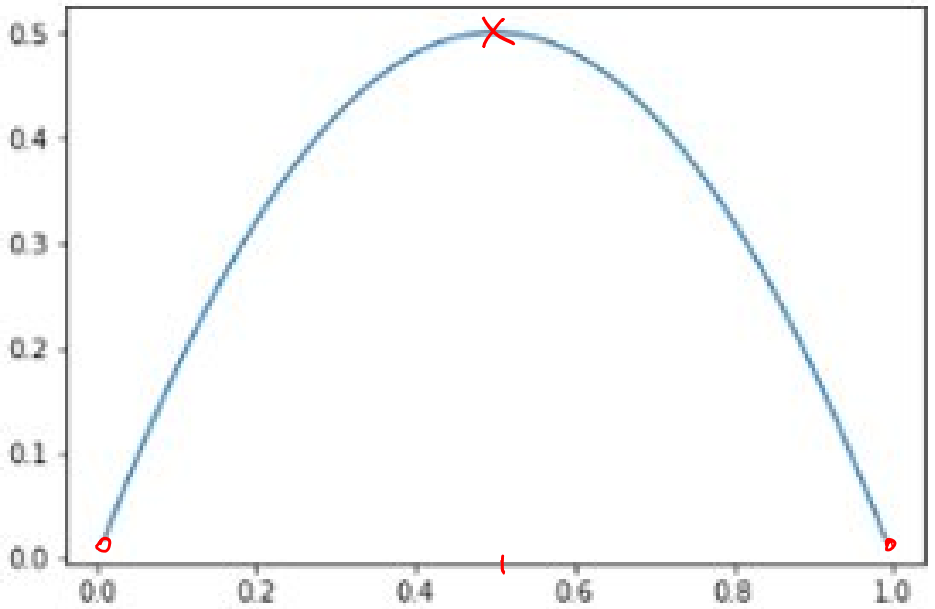


Entropy

$$H(X_m) = -\sum_k p_{mk} \log(p_{mk})$$

Split criterion- Gini index

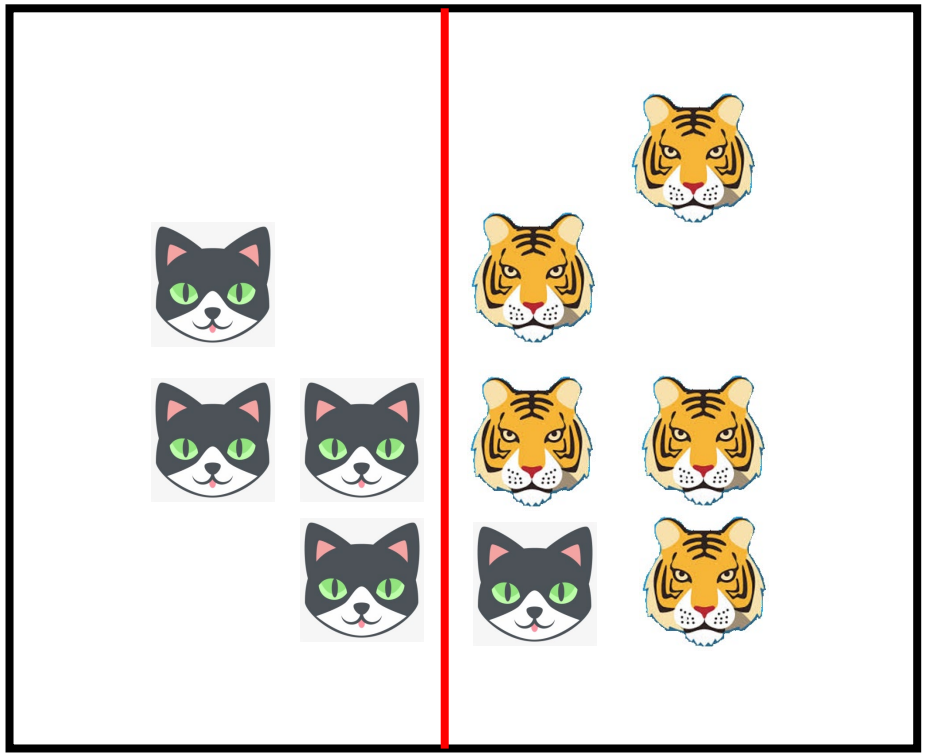
```
a = np.arange(0.01,1,0.01)
plt.plot(a,2*a*(1-a));
```



$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

What is the Gini of this box?

Gini: $H(X_m) = \sum_k p_{mk}(1 - p_{mk})$



D

0.5

$\frac{1}{2} \cdot (\frac{1}{2}) + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$

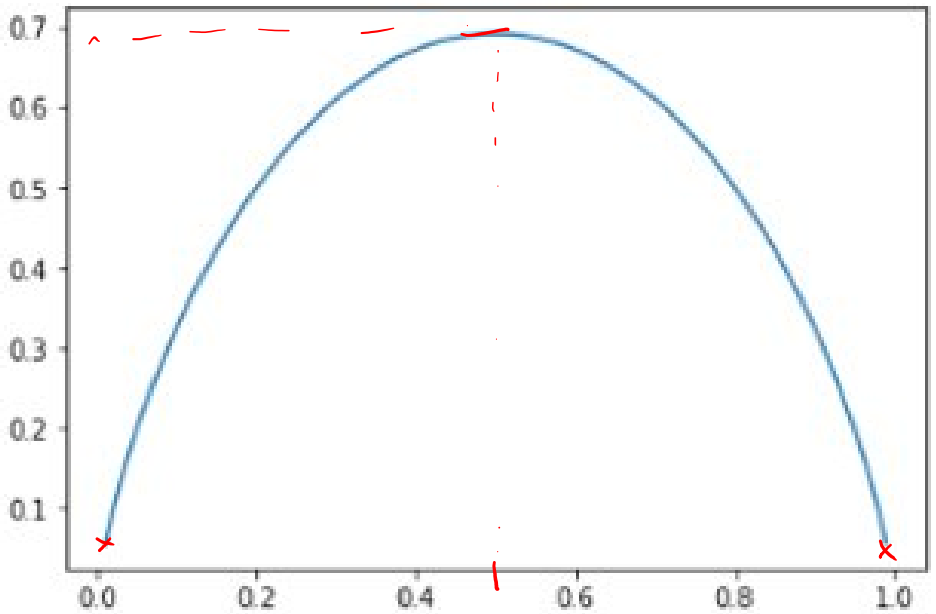
$P_c = \frac{1}{6}$

$P_t = 5/6$

$\frac{1}{6} \cdot \frac{5}{6} + \frac{5}{6} \frac{1}{6} = \frac{5}{18}$

Split criterion- Entropy

```
a = np.arange(0.01,1,0.01)
plt.plot(a,-a*np.log(a)-(1-a)*np.log(1-a));
```



$\log_2$ $\log_{10}$ $\log$

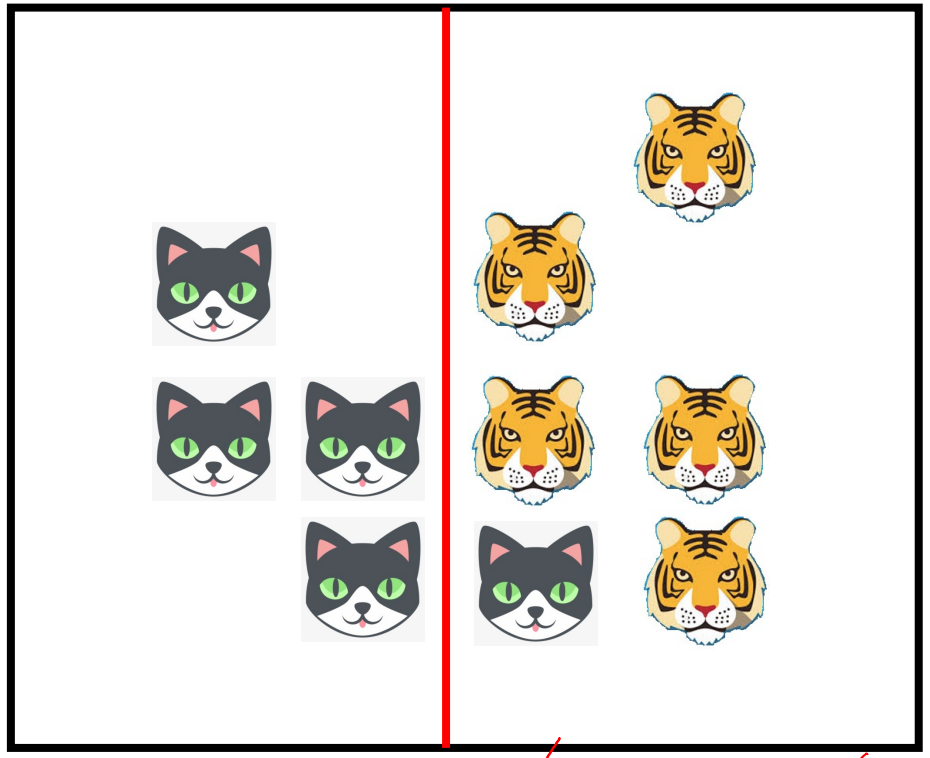
$$H(X_m) = -\sum_k p_{mk} \log(p_{mk})$$

Split criterion- Information gain

Information Gain = Reduction in Entropy

$$-\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right)=1$$

$$-\sum_i P_i \log P_i$$

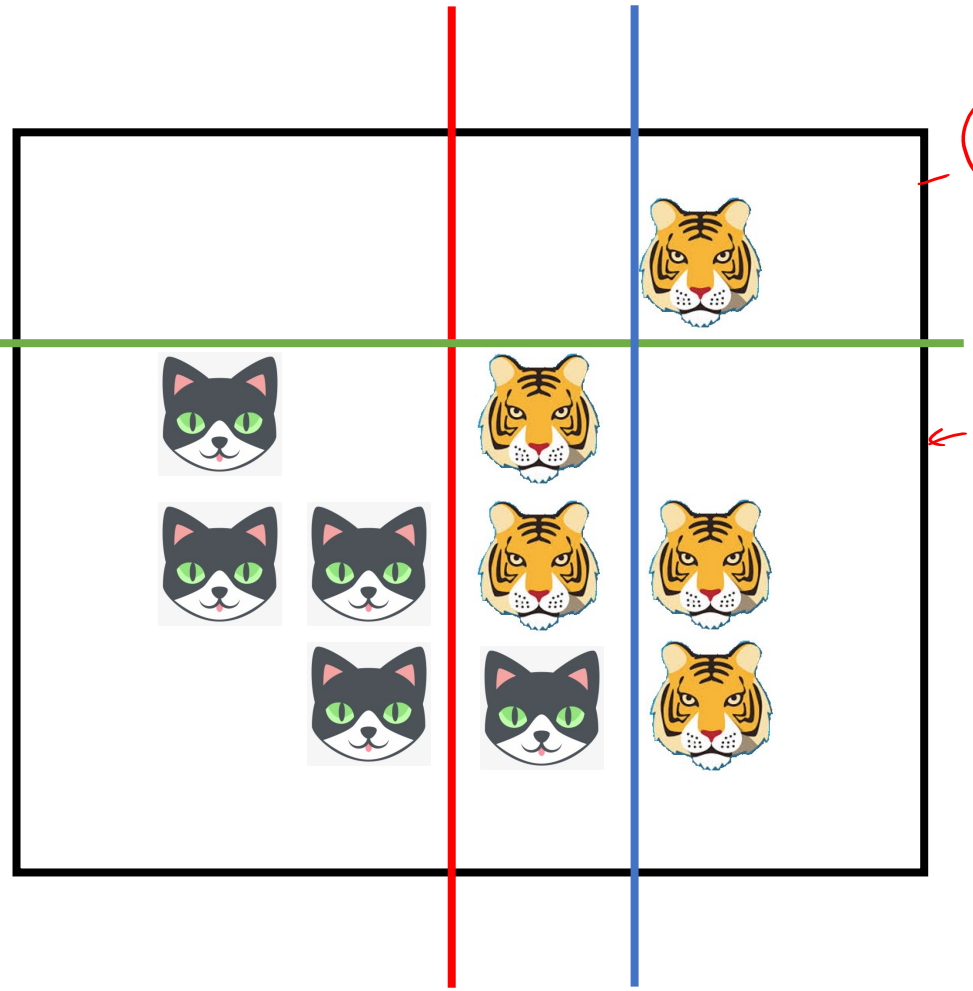


0

$$-\left(\frac{1}{6}\log_2\frac{1}{6}+\frac{5}{6}\log_2\frac{5}{6}\right)=0.65$$

Information Gain = 1 - 0.4*0 - 0.6*0.65 = 0.61

Which split gives the maximum information gain?



Decision Tree Split Criteria

Regression Tree

MSE (RSS)

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} (y_i - \bar{y}_m)^2$$

MAE

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} |y_i - \bar{y}_m|$$

Classification Tree

Gini

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

Entropy

$$H(X_m) = -\sum_k p_{mk} \log(p_{mk})$$

Information Gain = E(parent)-E(children)