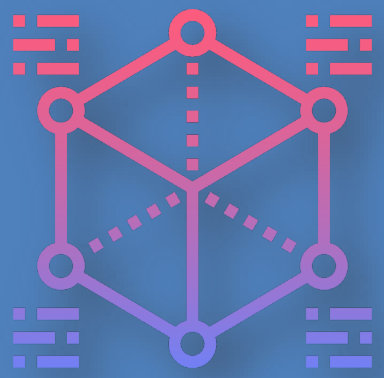
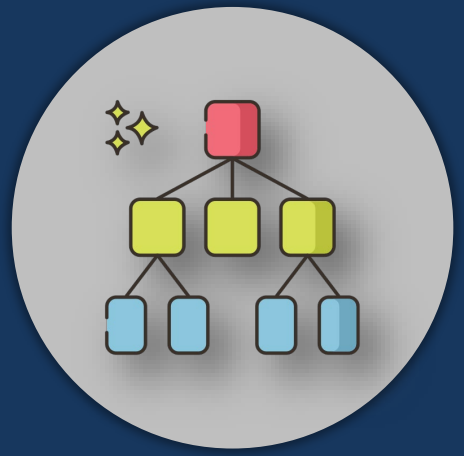
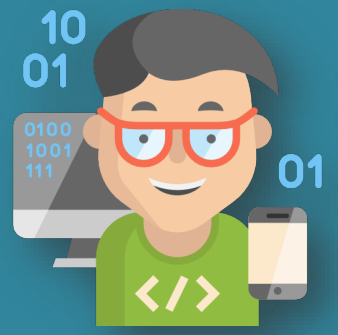
greatlearning
Power Ahead





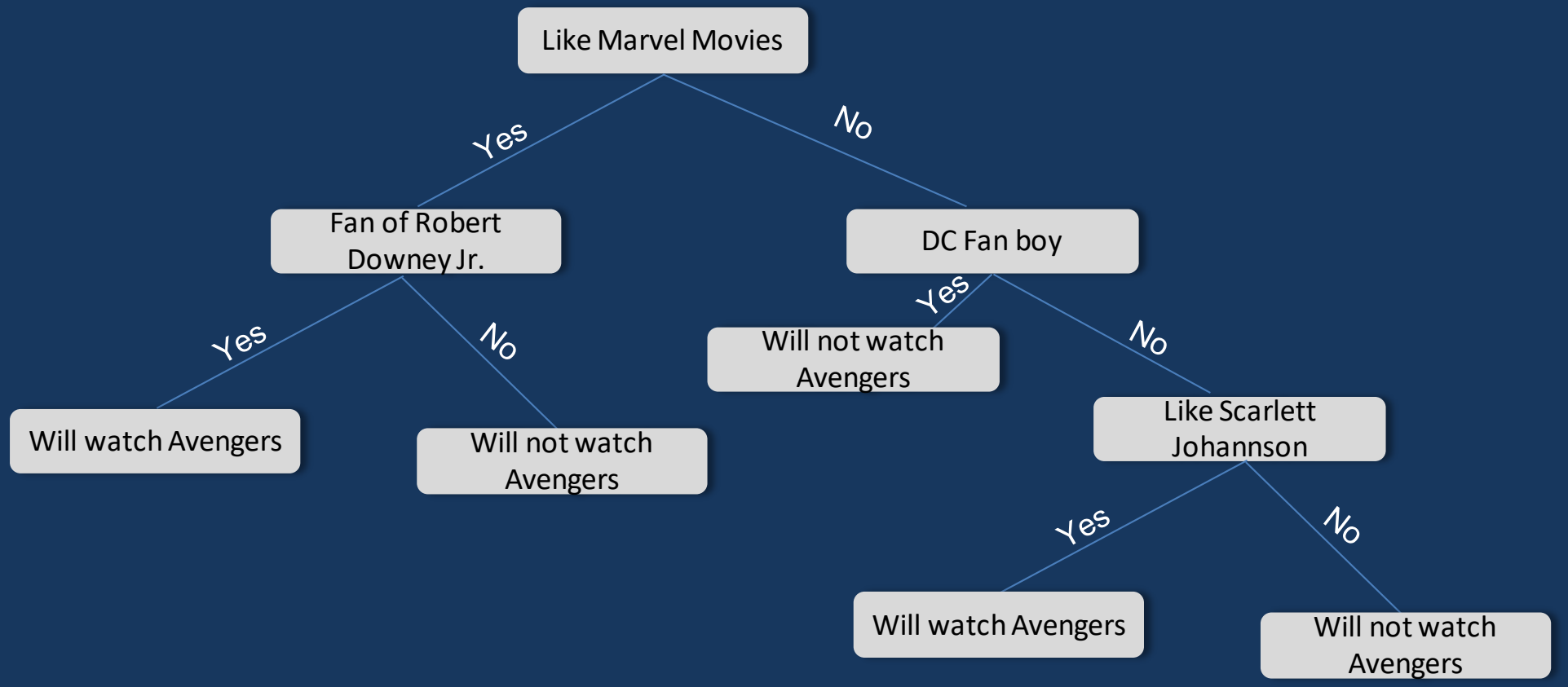
# Classification with Tree Based Models



# Decision Tree Algorithm

greatlearning
Power Ahead

Decision Tree Algorithm is a supervised learning method used for both classification and regression



# Decision Tree Algorithm-CART

greatlearning
Power Ahead

| X1 | X2 | Y |
|---|---|---|
| 0.2 | 0.3 | Good |
| 0.4 | 0.3 | Bad |
| 0.2 | 0.1 | Good |
| 0.6 | 0.5 | Bad |
| 0.5 | 0.5 | Good |

Categorical in Nature

- Spam/Not Spam
- Tumor/No Tumor
- Lend Money/Deny

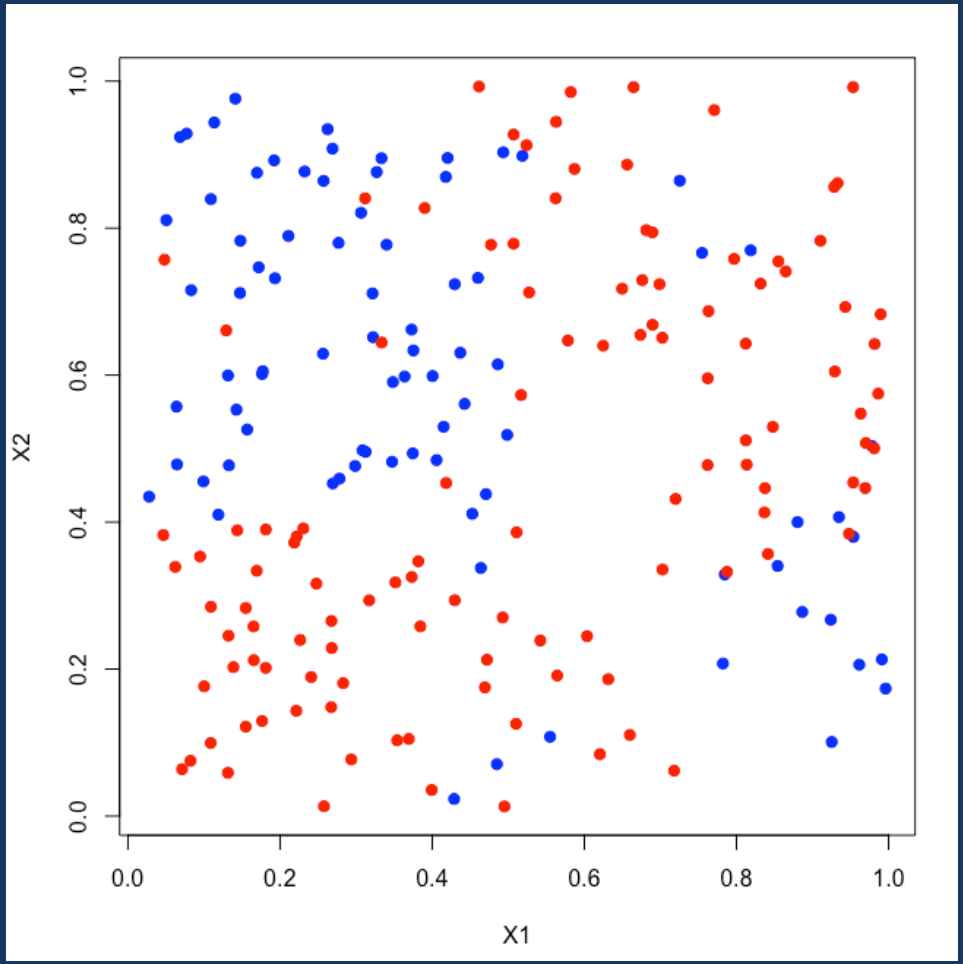
| X1 | X2 | Y |
|---|---|---|
| 0.2 | 0.3 | 56 |
| 0.4 | 0.3 | 34 |
| 0.2 | 0.1 | 76 |
| 0.6 | 0.5 | 12 |
| 0.5 | 0.5 | 45 |

Numerical in Nature

- Predict Stock Returns
- Predicting Sports Scores
- Pricing a house

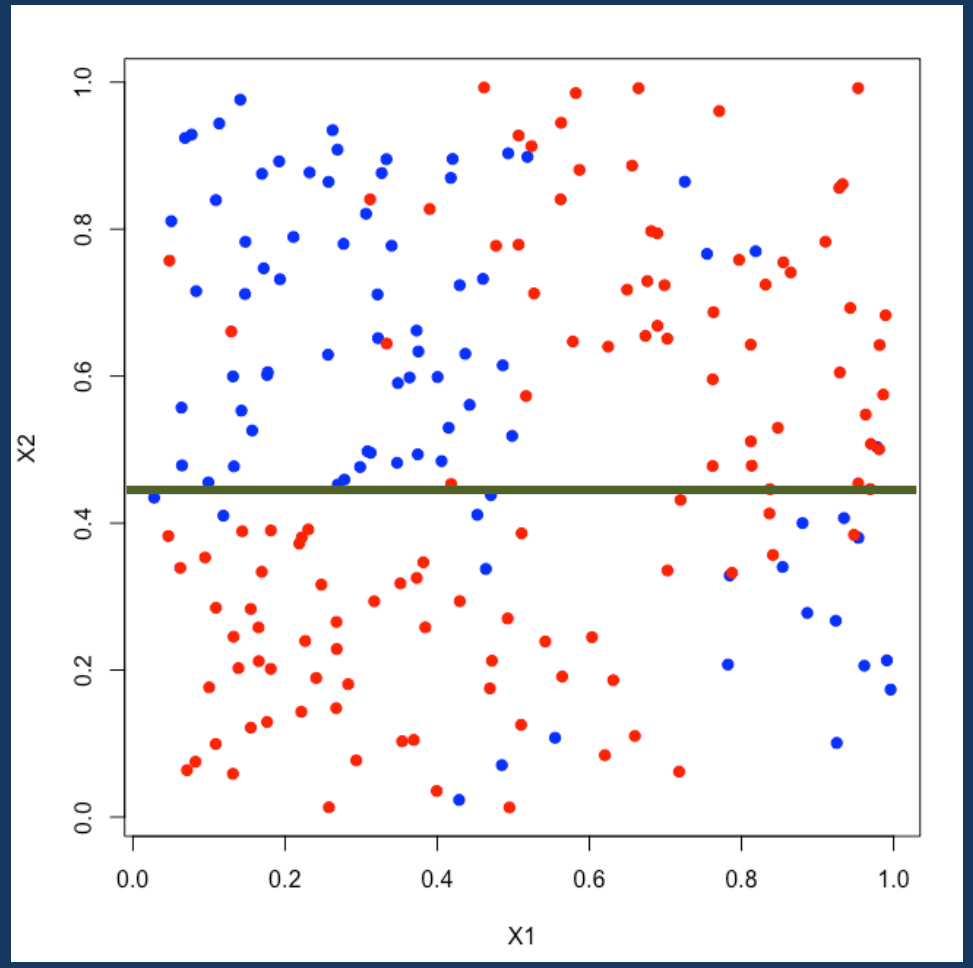
# How are decision trees built?

The general idea is that we will segment the space into a number of simple regions



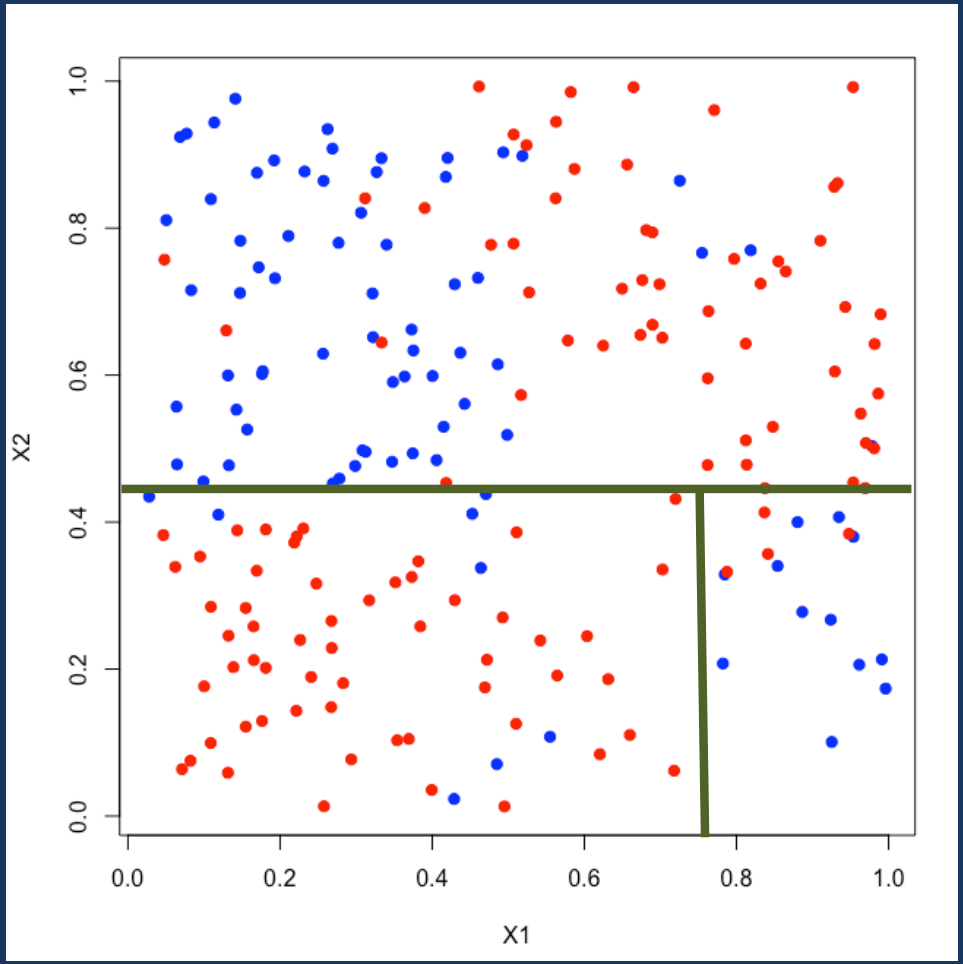
# How are decision trees built?

The general idea is that we will segment the space into a number of simple regions



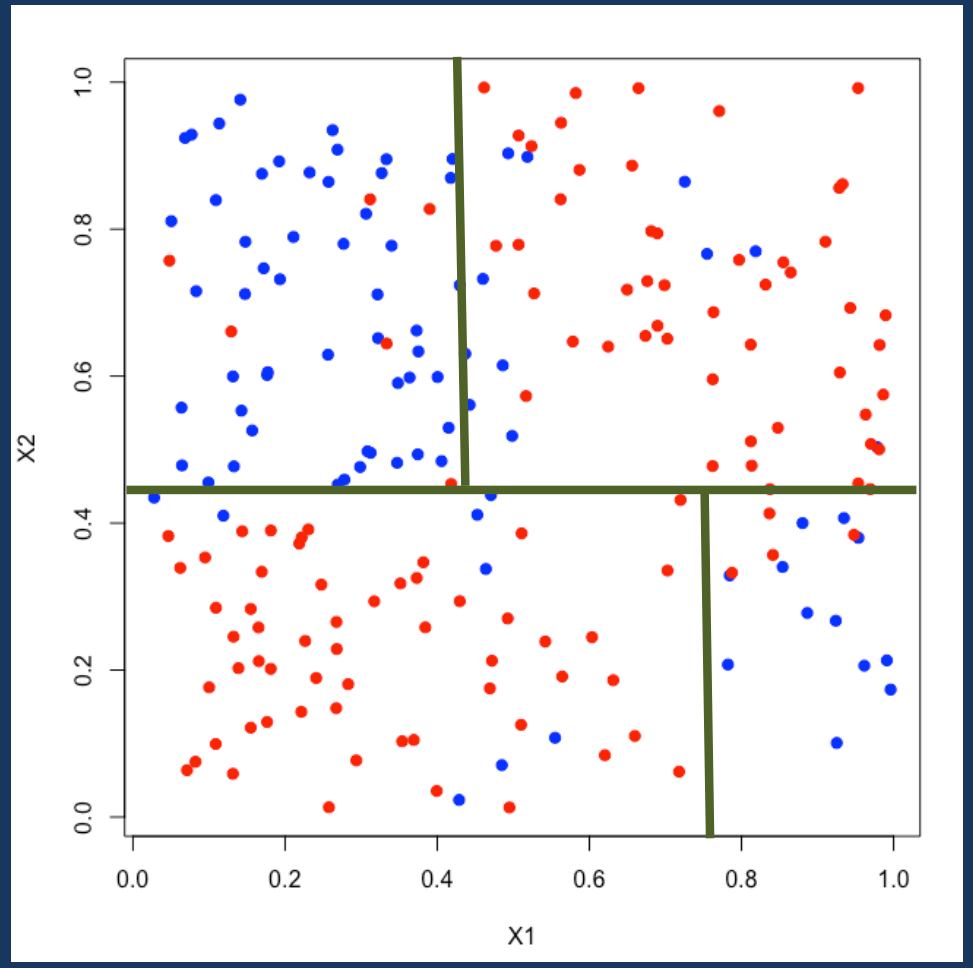
# How are decision trees built?

The general idea is that we will segment the space into a number of simple regions

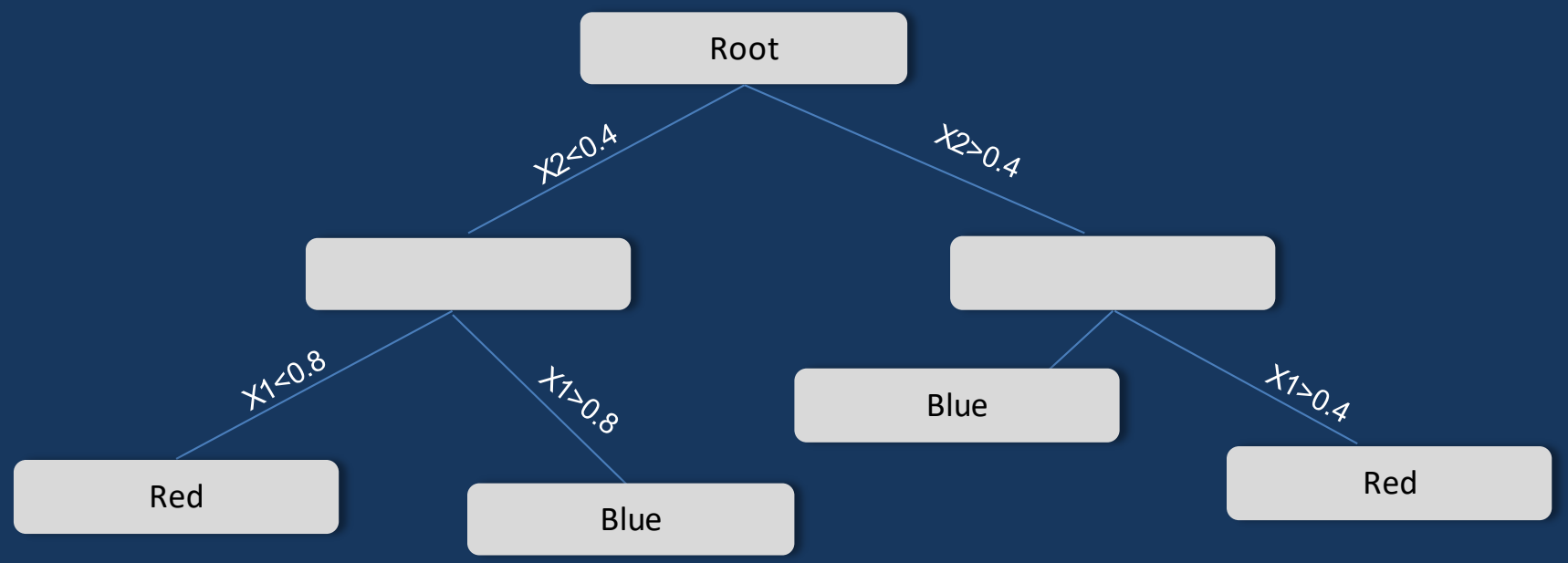


# How are decision trees built?

The general idea is that we will segment the space into a number of simple regions



# How are decision trees built?



# Measures of Impurity

These metrics measure how similar a region or a node is. They are said to measure the impurity of a region

Larger these impurity metrics the larger the “dissimilarity” of a nodes/regions

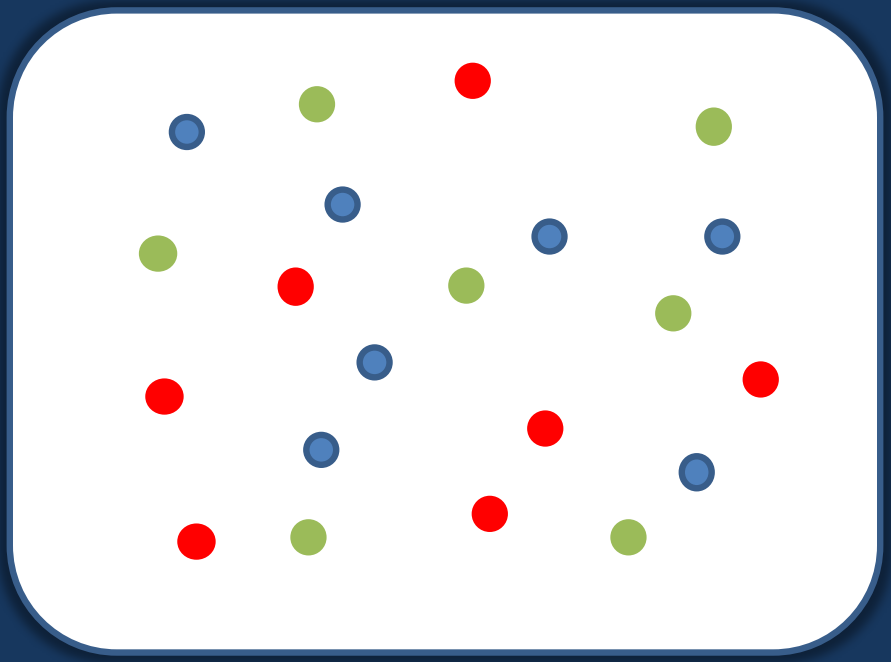
Gini Impurity

Entropy

Variance

# Gini Impurity

Gini Impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset



| P1 | P2 | P3 |
|---|---|---|

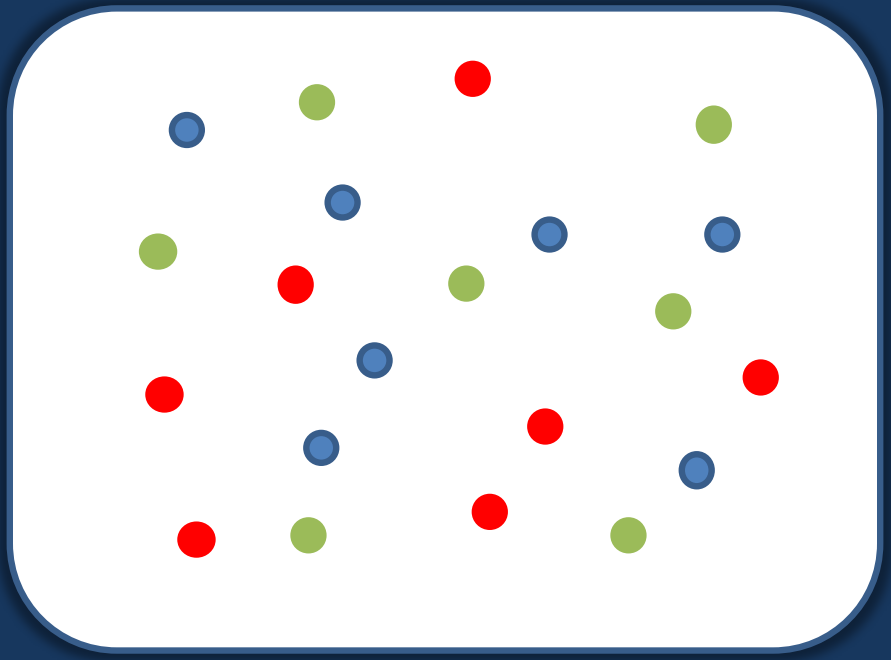
p1(1-p1) → P1p2 + p1p3

p2(1-p2) → P2p1 + p2p3

p3(1-p3) → P3p1 + p3p2

# Gini Impurity

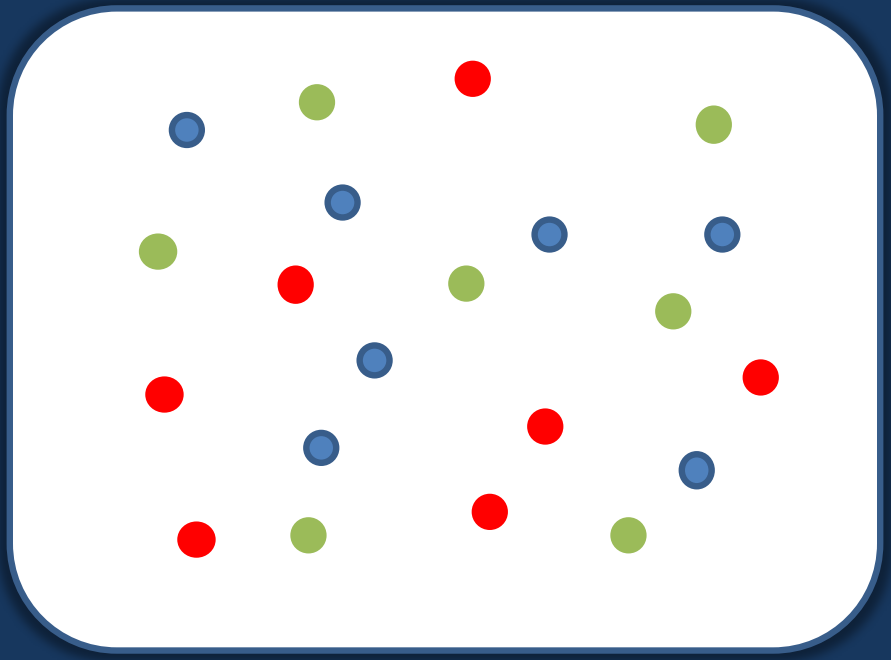
Gini Impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset

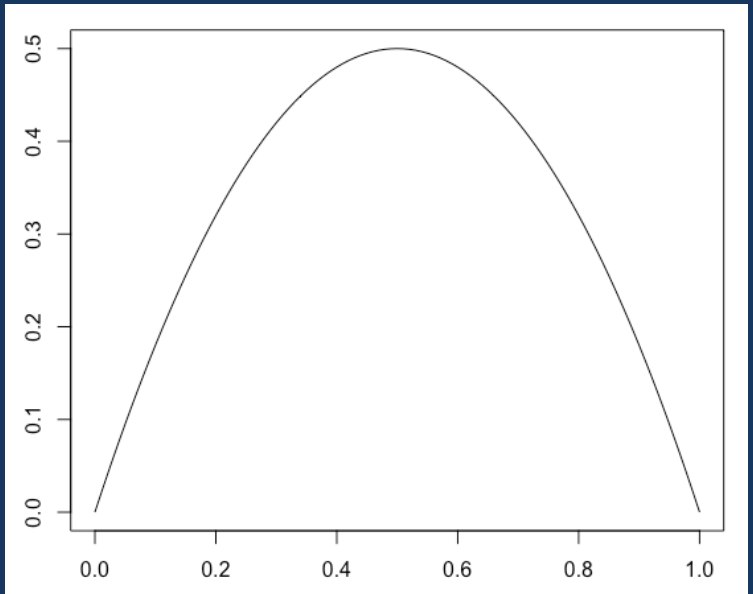


$$1 - \sum_{i=1}^{J} p_i^2$$

# Gini Impurity

Gini Impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset





# Tree to a Forest

Decision trees are very sensitive to even small changes in the data - usually called unstable

Can we get a whole bunch of decision trees to work together to yield a better and more robust prediction?

Then for prediction we could use the mean for regression trees and mode for classification trees

# Bagging and Random Forest