Supervised Machine Learning

Decision Tree

Instructor, Nero Chan Zhen Yu

Forward School

How does human learn?

Observation

Past Experience

Forward School

How does machine learn?

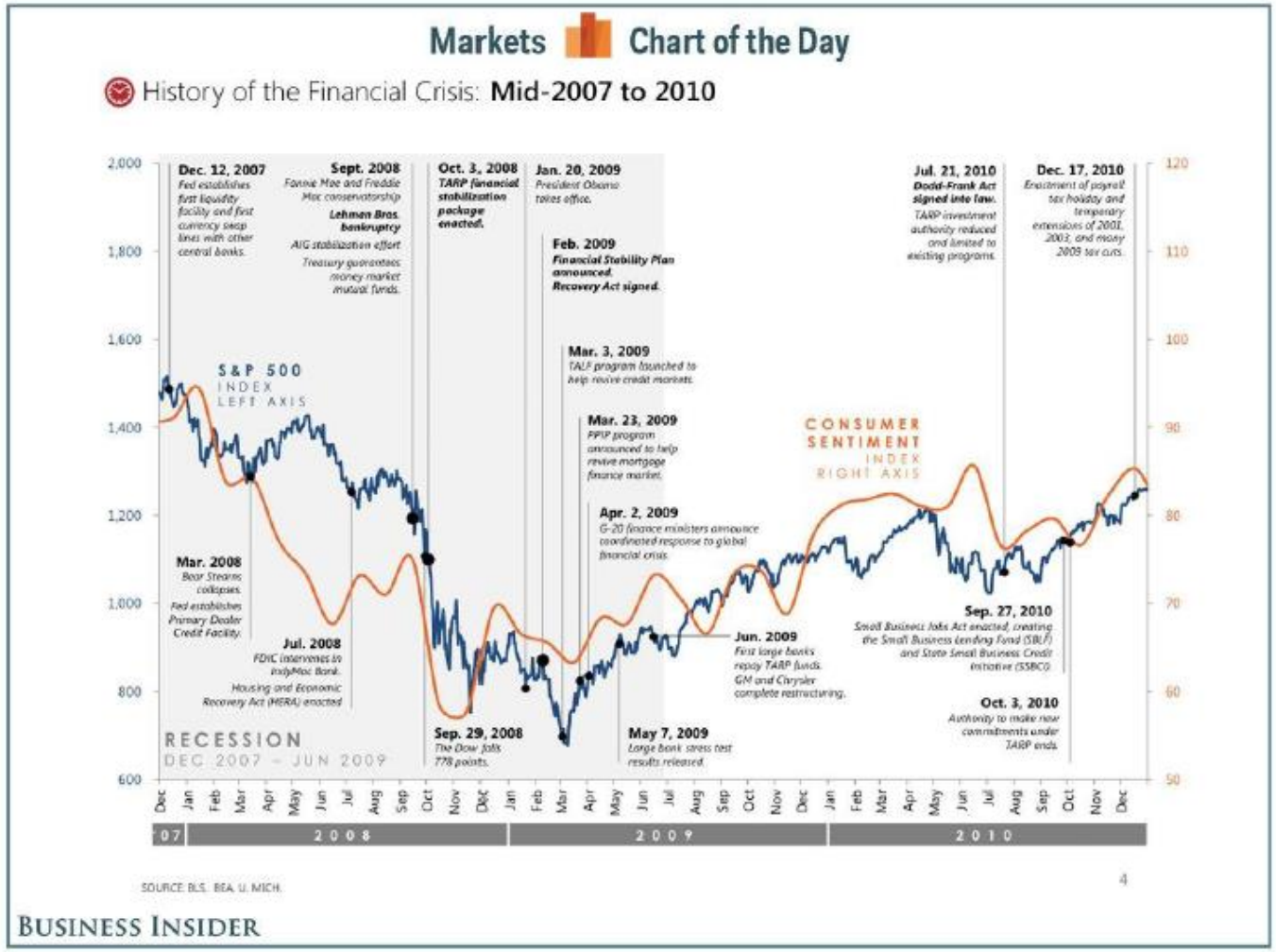
Forward School

Machine Learning: Definition

- Machine Learning is a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under certainty.

Forward School

Machine Learning (Example)



Forward School

Machine Learning (Example)

Google is Using Machine Learning to Predict the Likelihood of a Patient's Death – with 95% Accuracy!

PRANAV DAR, JUNE 19, 2018

Overview

- The AI research team at Google has developed a model that can predict the likelihood of a patient's death
- The AI is powered by neural networks and uses a ton of variables like the patient's old medical history, age and combines that with scribbled doctor's notes and PDFs
- Google tested the final model on 200,000+ patients and used over 46 billion data points
- The final model came up with an almost 95% accuracy when predicting patient outcomes

Sub

What are the possible features?

Forward School

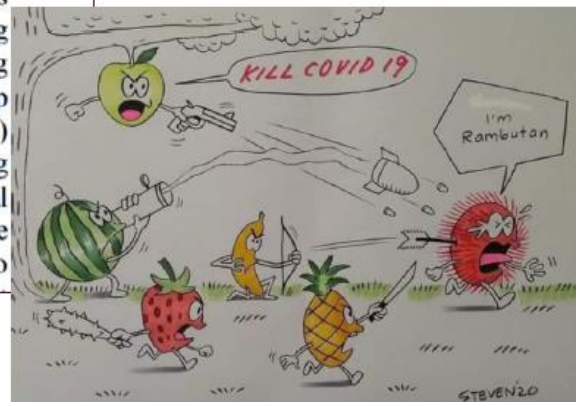
Machine Learning (Example)

Computers, Materials & Continua CMC, vol.63, no.1, pp.537-551, 2020

Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity

Xiangao Jiang[1], Megan Coffee[2, 3, *], Anasse Bari[4, *], Junzhang Wang[4], Xinyue Jiang[5], Jianping Huang[1], Jichan Shi[1], Jianyi Dai[1], Jing Cai[1], Tianxiao Zhang[6], Zhengxing Wu[1], Guiqing He[1] and Yitong Huang[7]

Abstract: The virus SARS-CoV2, which causes coronavirus disease (COVID-19) has become a pandemic and has spread to every inhabited continent. Given the increasing caseload, there is an urgent need to augment clinical skills in order to identify from among the many mild cases the few that will progress to critical illness. We present a first step towards building an artificial intelligence (AI) framework, with predictive analytics (PA) capabilities applied to real patient data, to provide rapid clinical decision-making support. COVID-19 has presented a pressing need as a) clinicians are still developing clinical acumen to this novel disease and b) resource limitations in a surging pandemic require difficult resource allocation decisions. The objectives of this research are: (1) to



Forward School

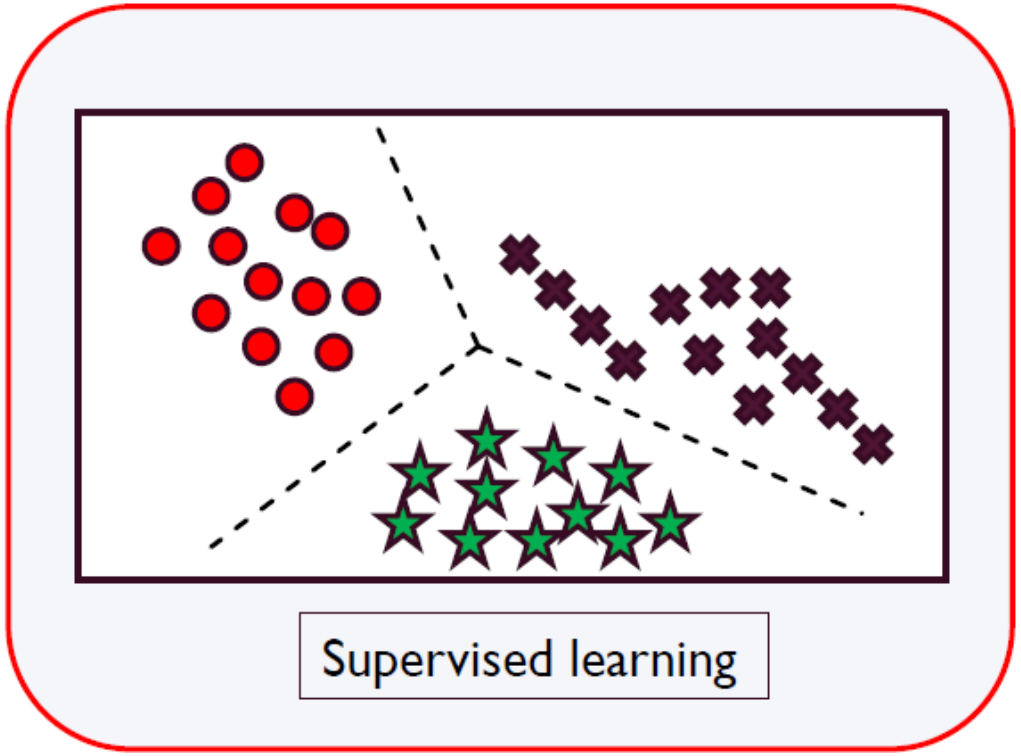
Types of Machine Learning

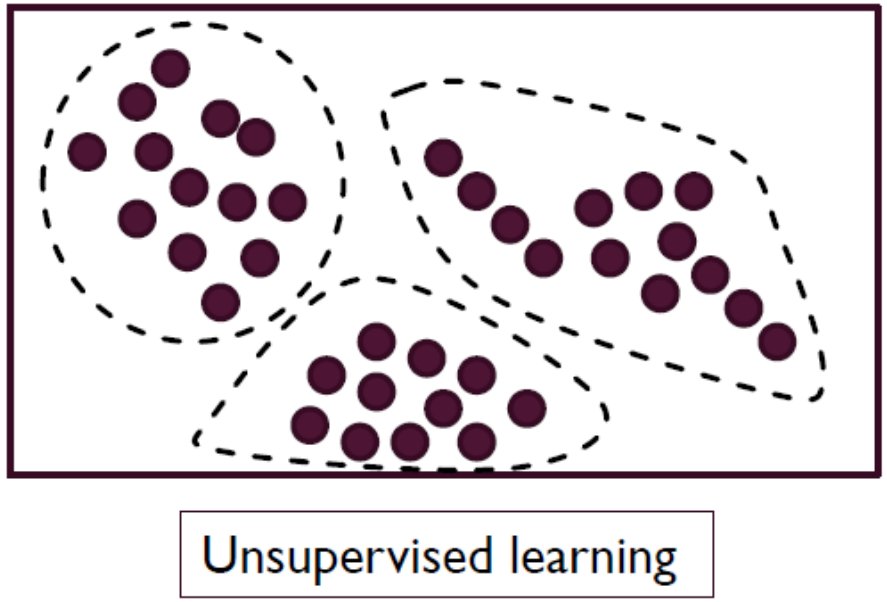
- Usually divided into two main types:
  - Supervised
  - Unsupervised

- Others:
  - Semi-supervised
  - Reinforcement learning

Forward School

Types of Machine Learning





Forward School

Machine Doesn’t Know What Fruits are these

X (Features)

y (Label)

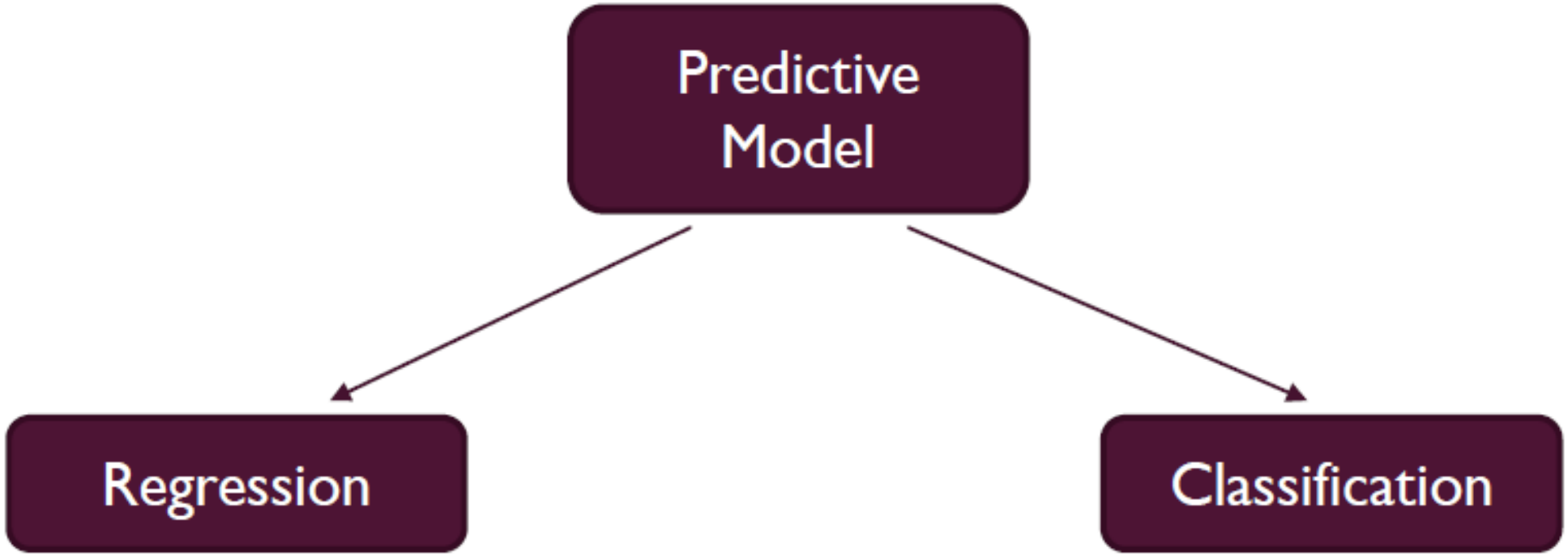
Features?

Label?

Color	Size	Fruit
Red	Big	Apple
Orange	Big	Orange
Red	Small	Grapes
Red	Big	Apple
Orange	Big	Orange

Forward School

Machine Learning



Forward School

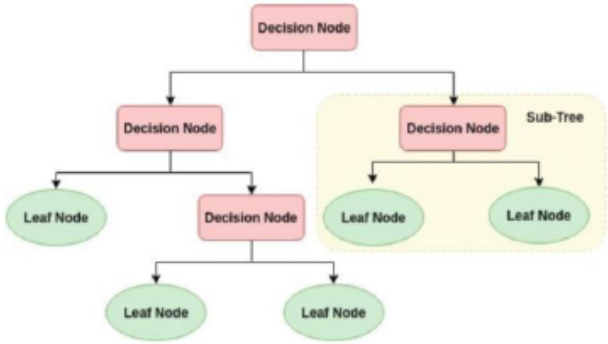
Assessing Classifier

- Contingency table or Confusion Matrix
- Accuracy, Precision and Recall
- ROC curves and Area Under the Curve
- Cross Validation

Forward School

Decision Tree

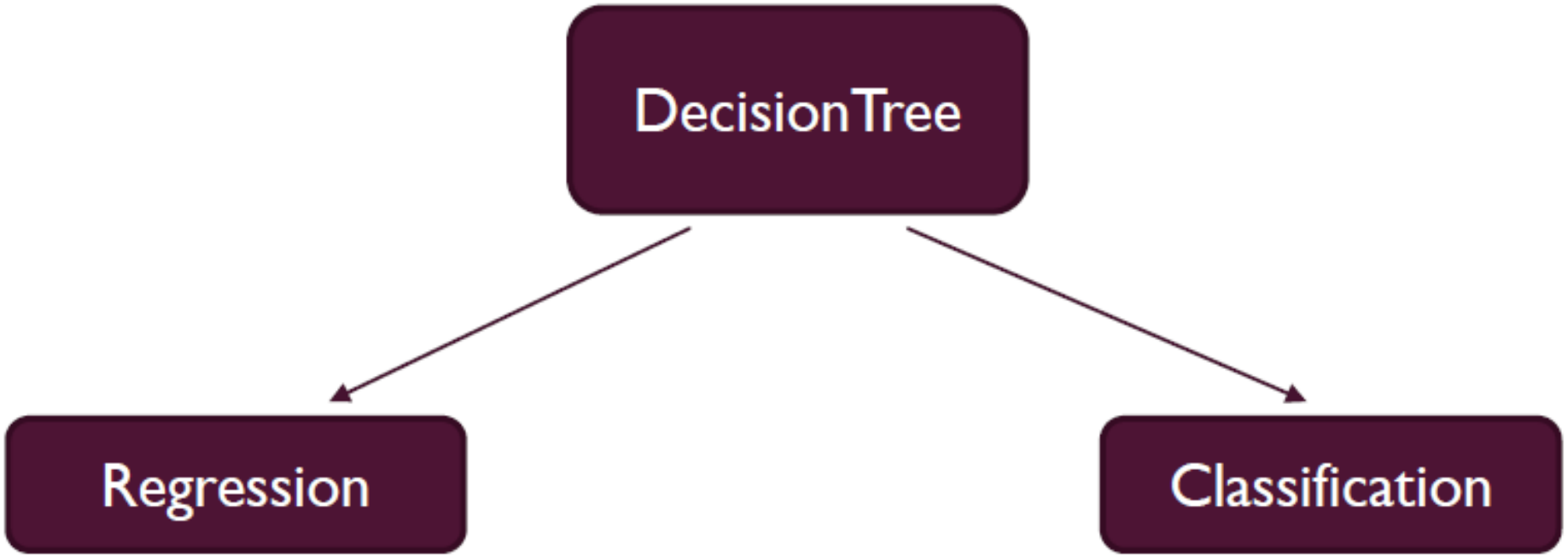
- A flow-chart-like tree structure
- Internal node denotes a test on an attribute
- Branch represents an outcome of the test
- Leaf nodes represent class labels or class distribution



- Decision tree generation consists of two phases
  - Tree construction
    - At start, all the training examples are at the root
    - Partition examples recursively based on selected attributes
  - Tree pruning
    - Identify and remove branches that reflect noise or outliers
- Use of decision tree: Classifying an unknown sample
  - Test the attribute values of the sample against the decision tree

Forward School

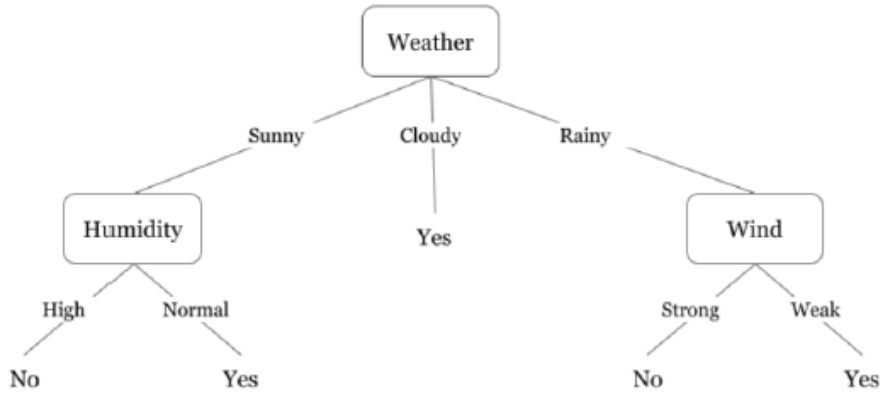
Decision Tree



Forward School

Decision Tree

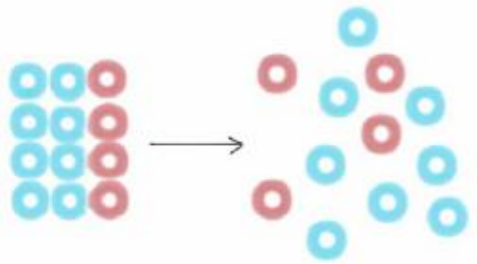
Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes

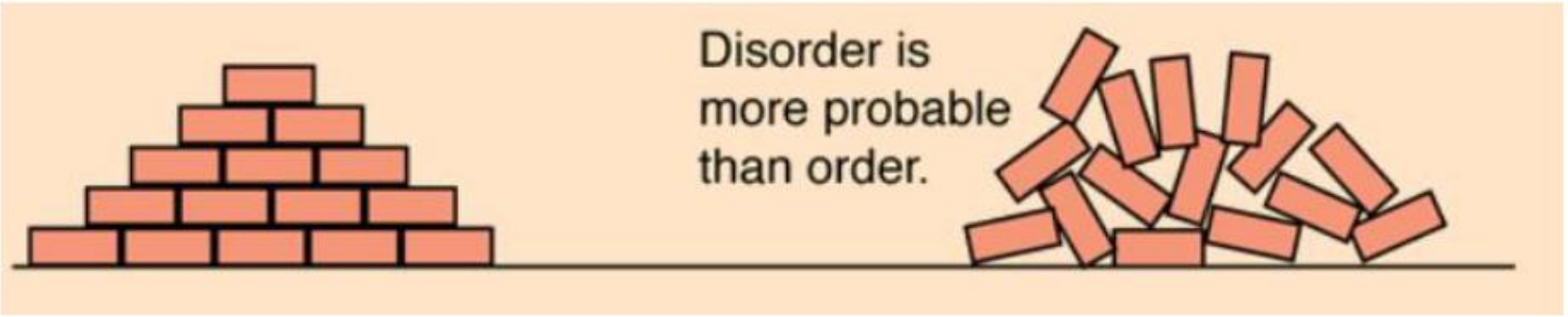


Forward School

Decision Tree (Splitting Criteria)

- Entropy $$H_i = -\sum_{\substack{k=1 \\ p_{i,k} \neq 0}}^{n} p_{i,k} \log\left(p_{i,k}\right)$$





- Information Gain
- Gini $$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^{2}$$

Forward School

Entropy???

Hard to predict
(High entropy)

Easier to predict
(Lower entropy)

Entropy is the measure of randomness in a dataset.

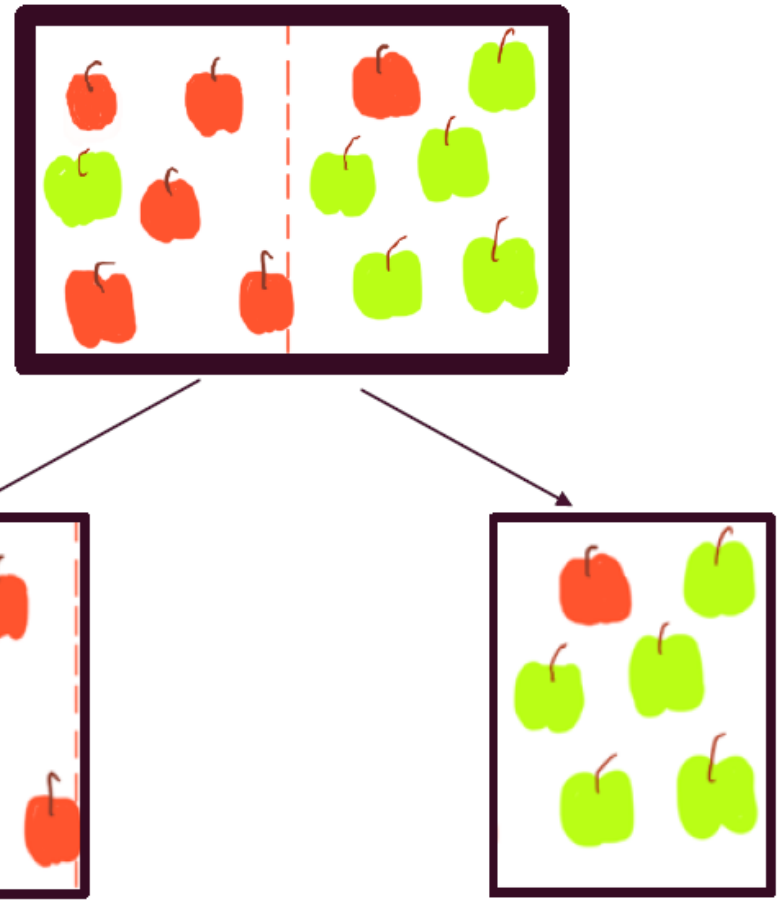
Aim of DT – split the data in a way that the entropy in the data decreases -> easier to make predictions

Forward School

Decision Tree - Entropy

Entropy

Measure of randomness and unpredictability



High Entropy

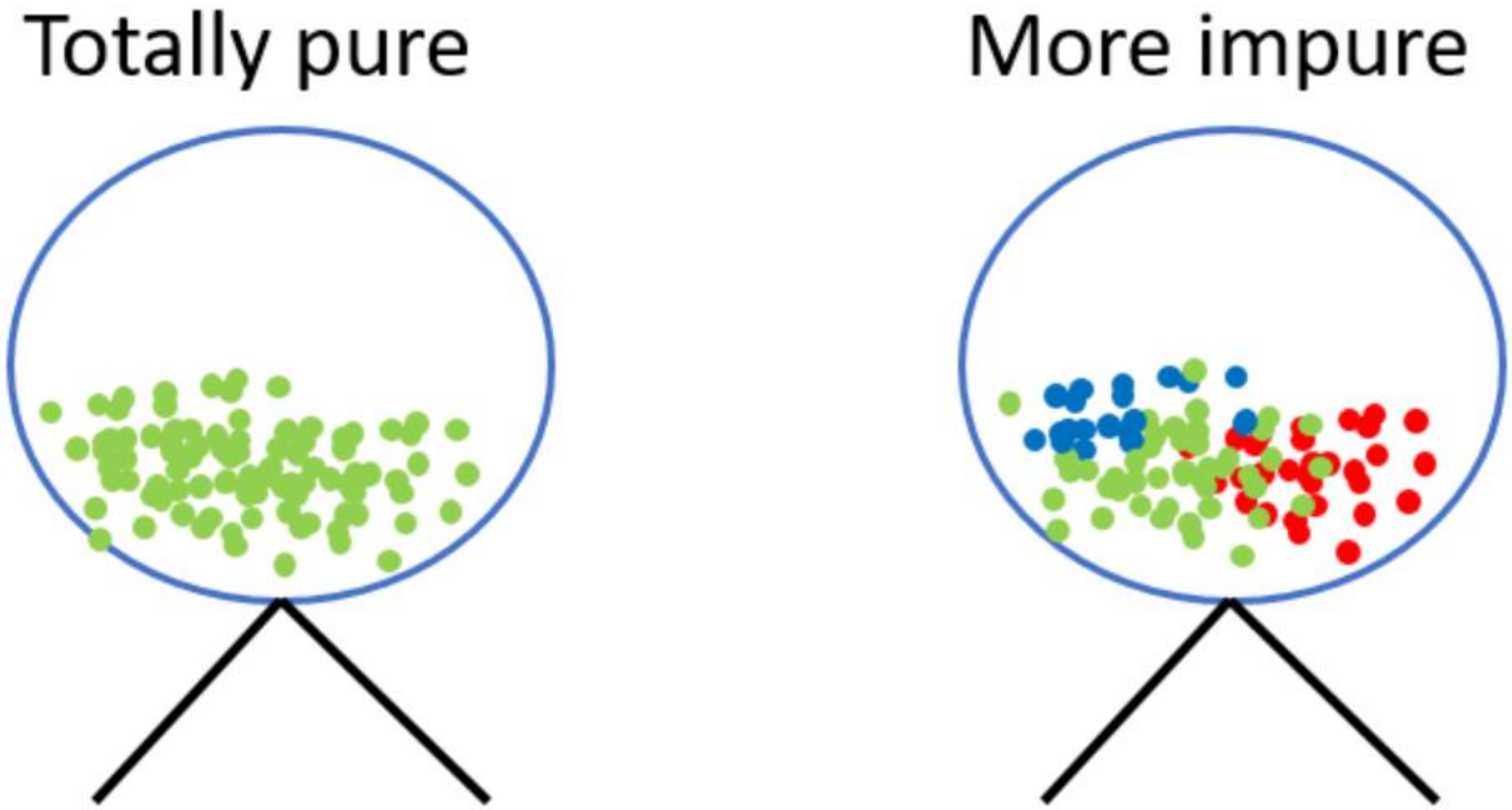
Low Entropy

Initial Dataset – all mixed up

Split – less random -> entropy decreases

Forward School

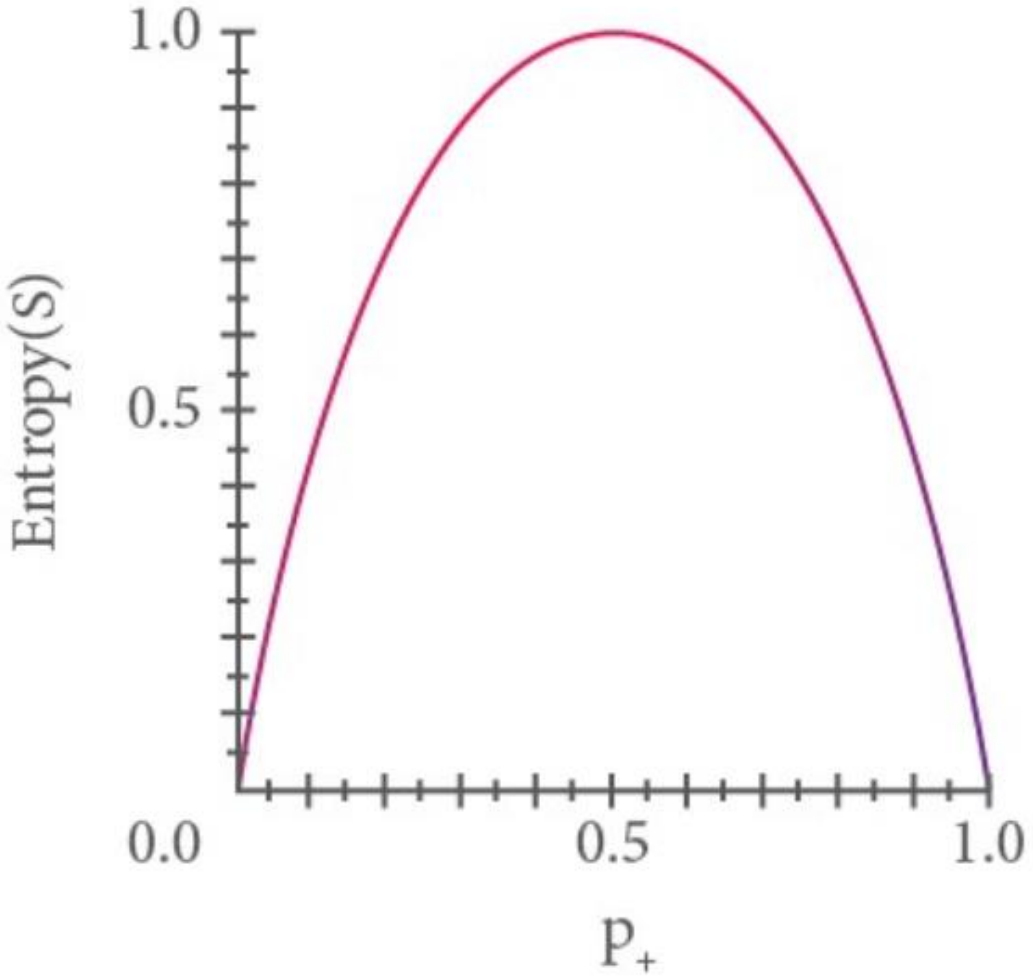
Entropy



Forward School

Entropy

$$\text{Entropy} = -p \log_2 p - q \log_2 q$$



Forward School

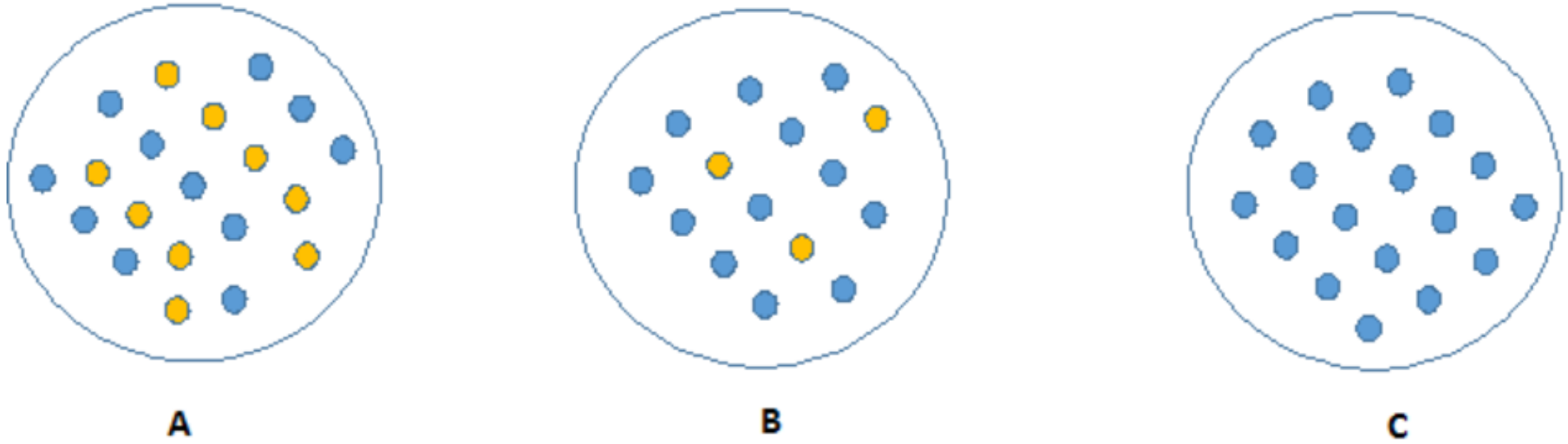
Information Gain

- Based on the decrease in entropy after a dataset is split on an attribute.
- Constructing a decision tree is all about finding attribute that returns the highest information to gain.
- Information gain computes the difference between entropy before split and average entropy after split of the dataset based on given attribute values

$$Information\ Gain = Entropy(parent\ node) — [Avg\ Entropy(children)]$$

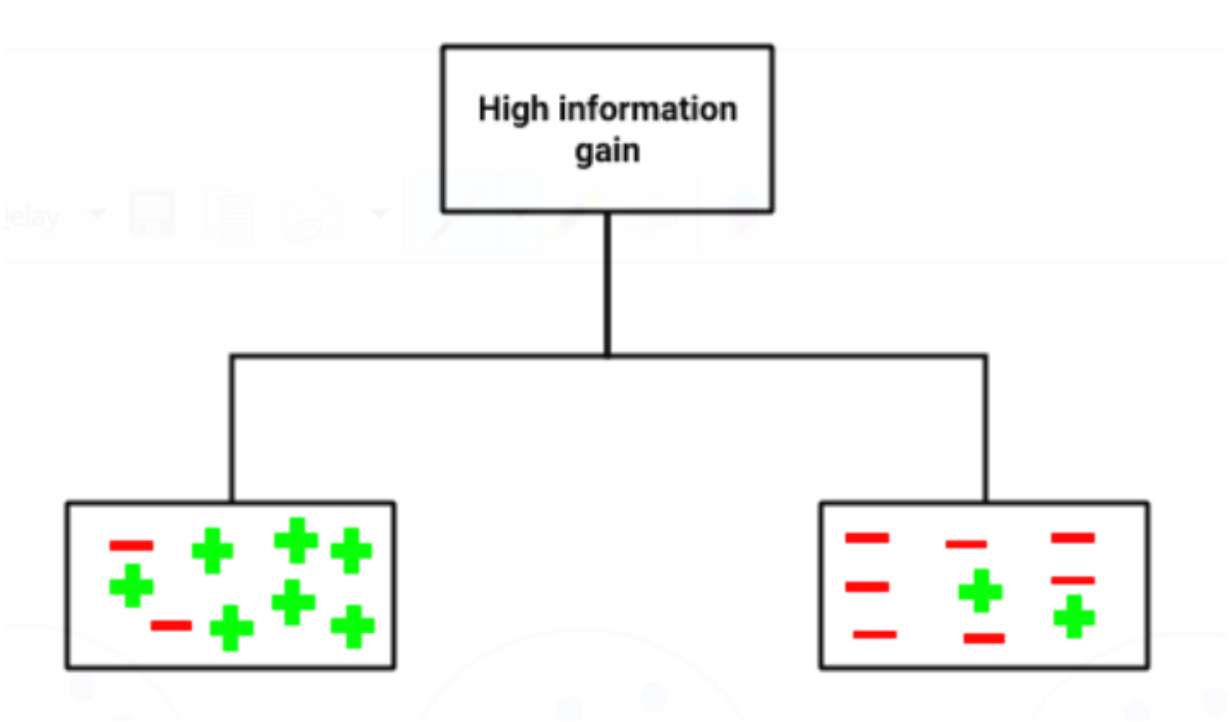
Forward School

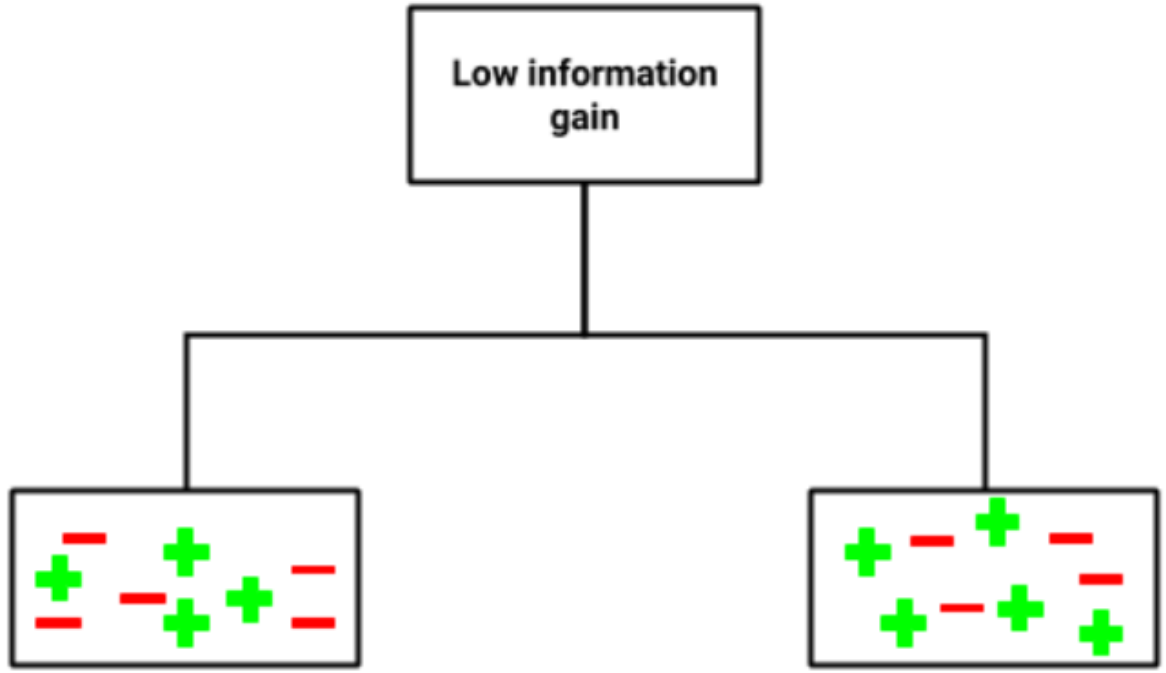
Information Gain



Forward School

High & Low Information Gain





Distribution of points in case of high and low information gain

Forward School

Steps to calculate entropy for a split:

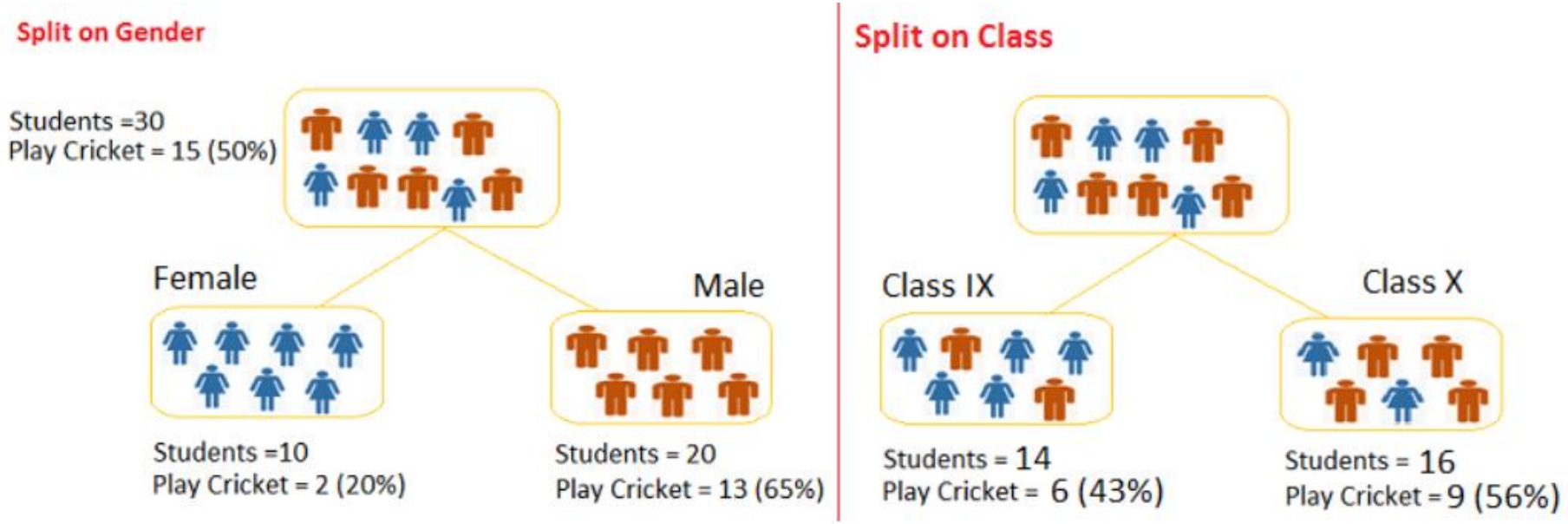
1. Calculate the entropy of the parent node

2. Calculate entropy of each individual node of split and calculate the weighted average of all sub-nodes available in a split.

Forward School

Example 1

Let's say we have a sample of 30 students with three variables Gender (Boy/ Girl), Class(IX/ X) and Height (5 to 6 ft). 15 out of these 30 play cricket in leisure time. Now, I want to create a model to predict who will play cricket during a leisure period? In this problem, we need to segregate students who play cricket in their leisure time based on highly significant input variable among all three.



Forward School

Example 1 Solution

Entropy for parent node = -(15/30) log2 (15/30) — (15/30) log2 (15/30) = **1**

For Split on gender:

Entropy for Female node = -(2/10) log2 (2/10) — (8/10) log2 (8/10) = 0.72 Entropy for Male node = -(13/20) log2 (13/20) — (7/20) log2 (7/20) = 0.93 Entropy for split Gender = (10/30)*0.72 + (20/30)*0.93 = **0.86**

Information Gain for split on gender = 1–0.86 = ***0.14***

For Split on Class:

Entropy for Class IX node = -(6/14) log2 (6/14) — (8/14) log2 (8/14) = 0.99 Entropy for Class X node = -(9/16) log2 (9/16) — (7/16) log2 (7/16) = 0.99 Entropy for split Class = (14/30)*0.99 + (16/30)*0.99 = **0.99**

Information Gain for split on Class = 1– 0.99 = ***0.01***

Forward School

GINI

Gini says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.

1. It works with categorical target variable “Success” or “Failure”.
2. It performs only Binary splits
3. Higher the value of Gini higher the homogeneity.
4. CART (Classification and Regression Tree) uses the Gini method to create binary splits.

Forward School

Steps to Calculate Gini for a split

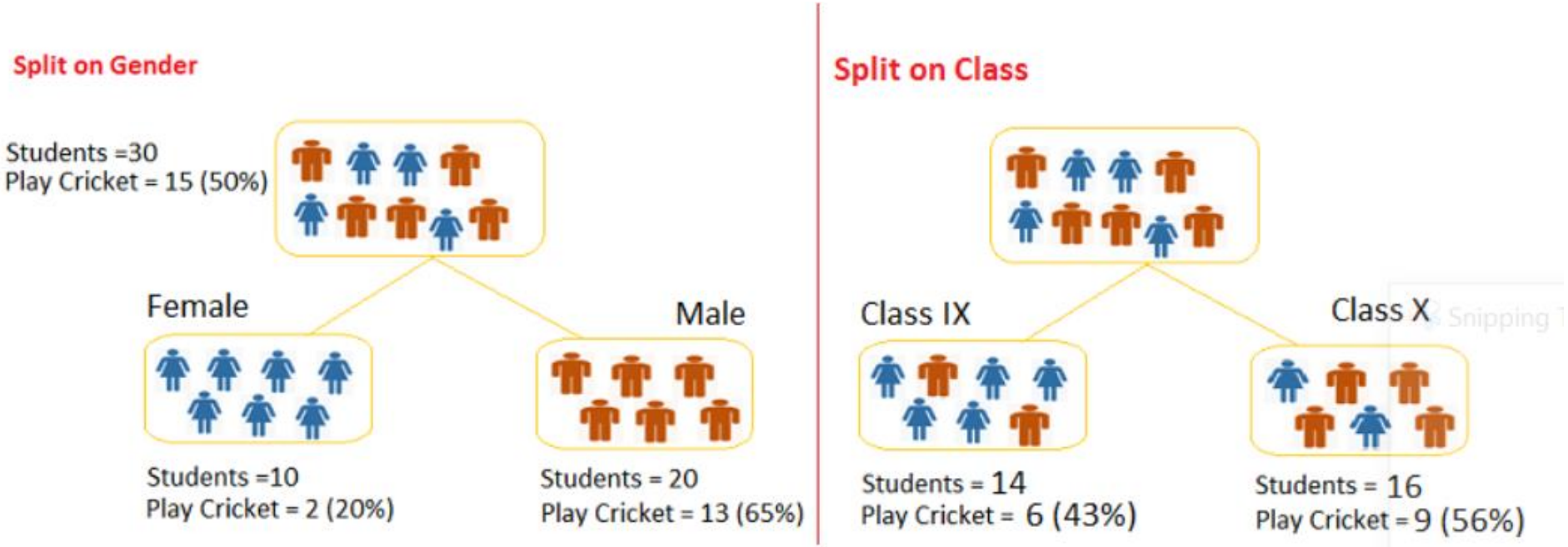
1. Calculate Gini for sub-nodes, using formula sum of the square of probability for success and failure (p^2+q^2).

2. Calculate Gini for split using weighted Gini score of each node of that split

Forward School

Example 2

Referring to the example used above, where we want to segregate the students based on the target variable (playing cricket or not). In the snapshot below, we split the population using two input variables Gender and Class. Now, I want to identify which split is producing more homogeneous sub-nodes using Gini.



Forward School

Example 2 Solution

Split on Gender:

1.Calculate, Gini for sub-node Female = (0.2)*(0.2)+(0.8)*(0.8)=0.68

2.Gini for sub-node Male = (0.65)*(0.65)+(0.35)*(0.35)=0.55

3.Calculate weighted Gini for Split Gender = (10/30)*0.68+(20/30)*0.55 = **0.59**

Similar for Split on Class:

1.Gini for sub-node Class IX = (0.43)*(0.43)+(0.57)*(0.57)=0.51

2.Gini for sub-node Class X = (0.56)*(0.56)+(0.44)*(0.44)=0.51

3.Calculate weighted Gini for Split Class = (14/30)*0.51+(16/30)*0.51 = **0.51**

Forward School

GINI

- Above, you can see that Gini score for *Split on Gender* is higher than *Split on Class,* hence, the node split will take place on Gender.

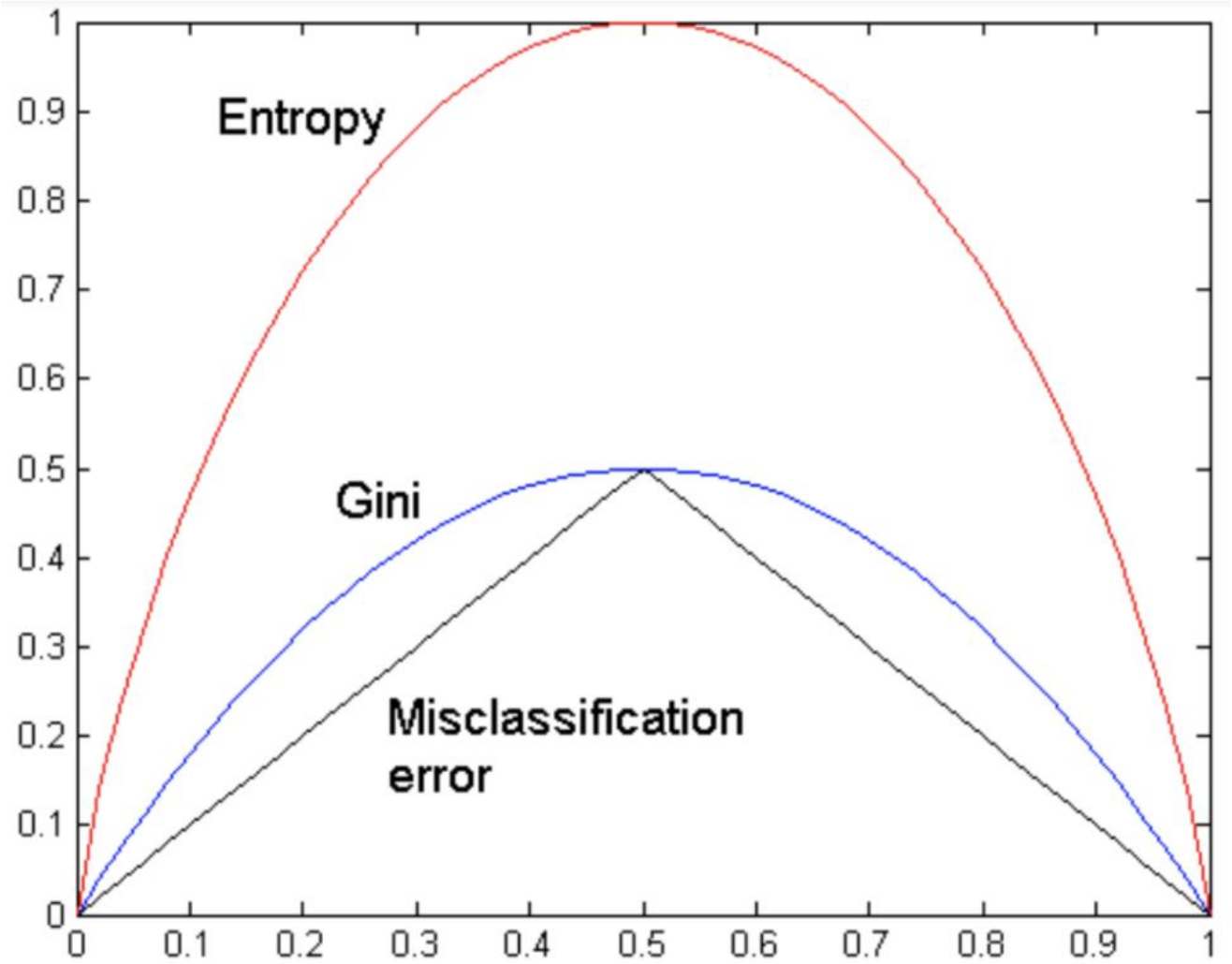
- You might often come across the term ‘Gini Impurity’ which is determined by subtracting the Gini value from 1. So mathematically we can say,

- *Gini Impurity = 1-Gini*

Forward School

Gini is default parameter in Decision Tree Algorithm.

Comparison between Entropy and GINI



Forward School

Decision Tree

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Forward School

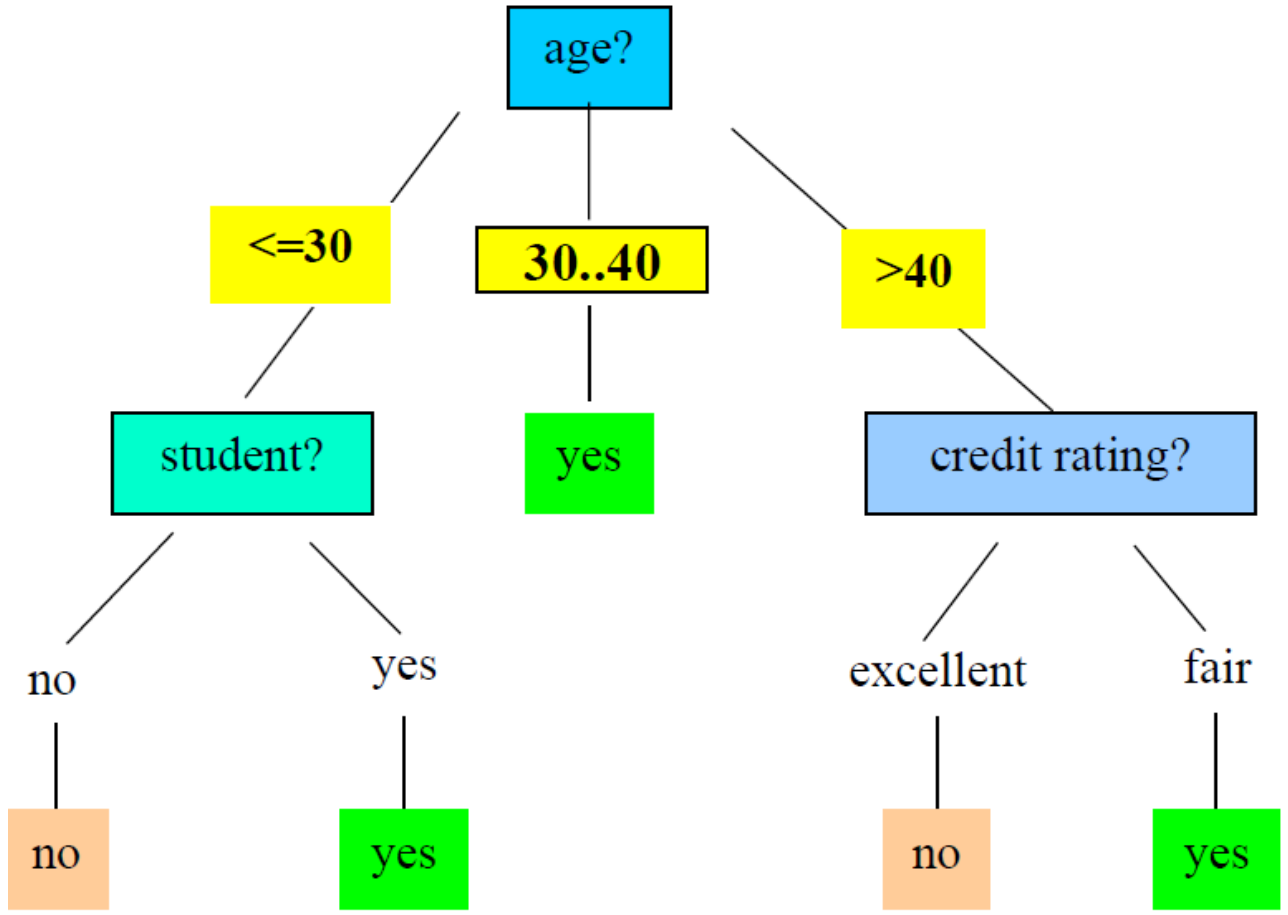
Decision Tree

- Three Data Sets formed after division at root node on the basis of “age” attribute.

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Forward School

Output Decision Tree for “buys_computer”



age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

On the basis of tree constructed in the manner described, classify a test sample
(age, student, creditrating, buys_computer)
(<=30, yes, excellent, ?)
-Will this student buy computer?

Forward School

Decision Tree

Step 1: Calculate entropy of target

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"
- $I(p,n) = I(9,5) = 0.940$

$$I(p,n) = -\left(\frac{p}{p+n}\log_2\frac{p}{p+n}\right) - \left(\frac{n}{p+n}\log_2\frac{n}{p+n}\right)$$

Step 2: Calculate entropy of attributes

- Compute the entropy for age

$$E(A) = \sum_{i=1}^{v}\frac{p_i+n_i}{p+n}I(p_i,n_i)$$

$$E(age) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.69$$

Step 3: Calculate information gain

Hence,

$$Gain(A) = I(p,n) - E(A)$$

$$Gain(age) = I(p,n) - E(age)$$
$$= 0.940 - 0.69 = 0.25$$

Similarly,

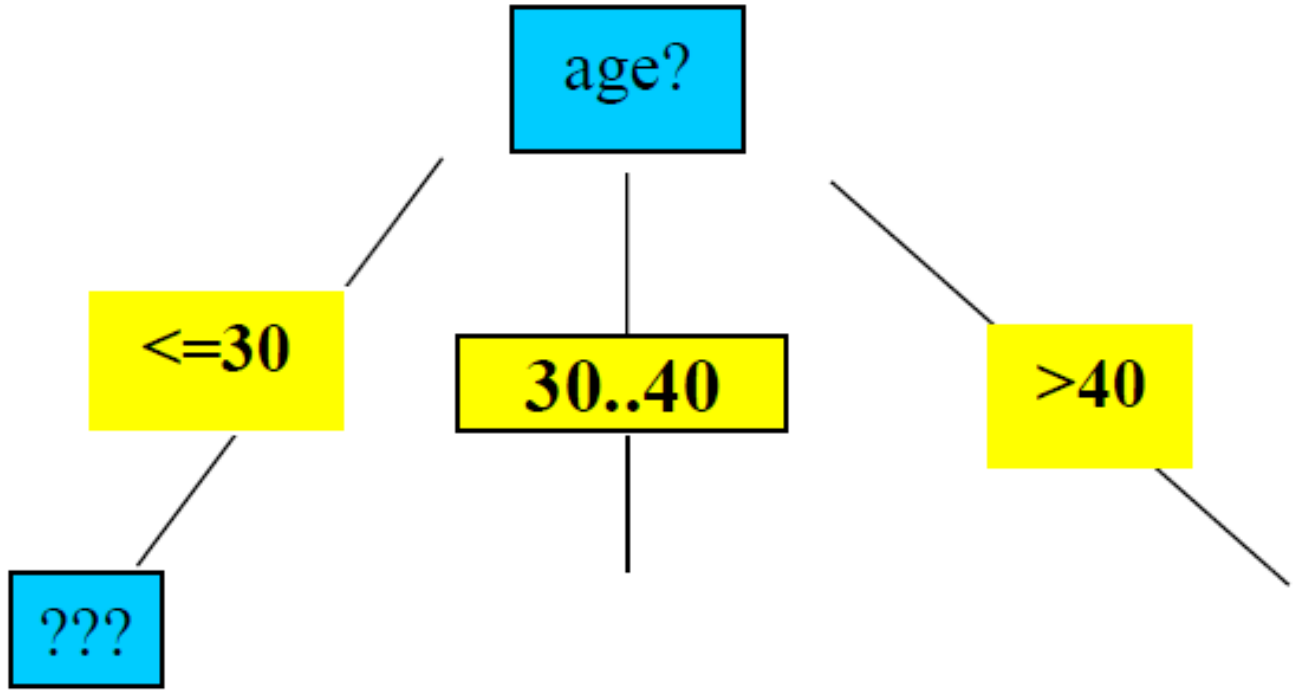
$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

Forward School

Decision Tree

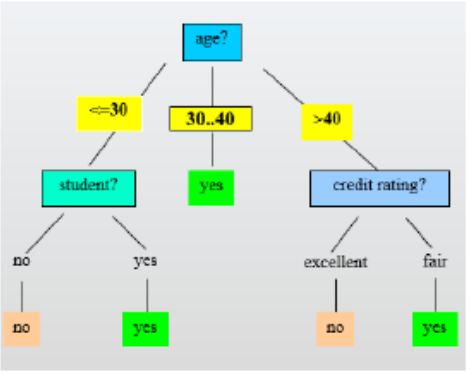


Repeat the processes of finding the attribute with the highest information gain.

Forward School

Extracting Classification Rules from Trees

- Represent the knowledge in the form of **IF-THEN** rules
- One rule is created for each path from the root to a leaf
- Each attribute-value pair along a path forms a conjunction
- The leaf node holds the class prediction
- Rules are easier for humans to understand
- Example



IF *age* = "<=30" AND *student* = "*no*" THEN *buys_computer* = "*no*"

IF *age* = "<=30" AND *student* = "*yes*" THEN *buys_computer* = "*yes*"

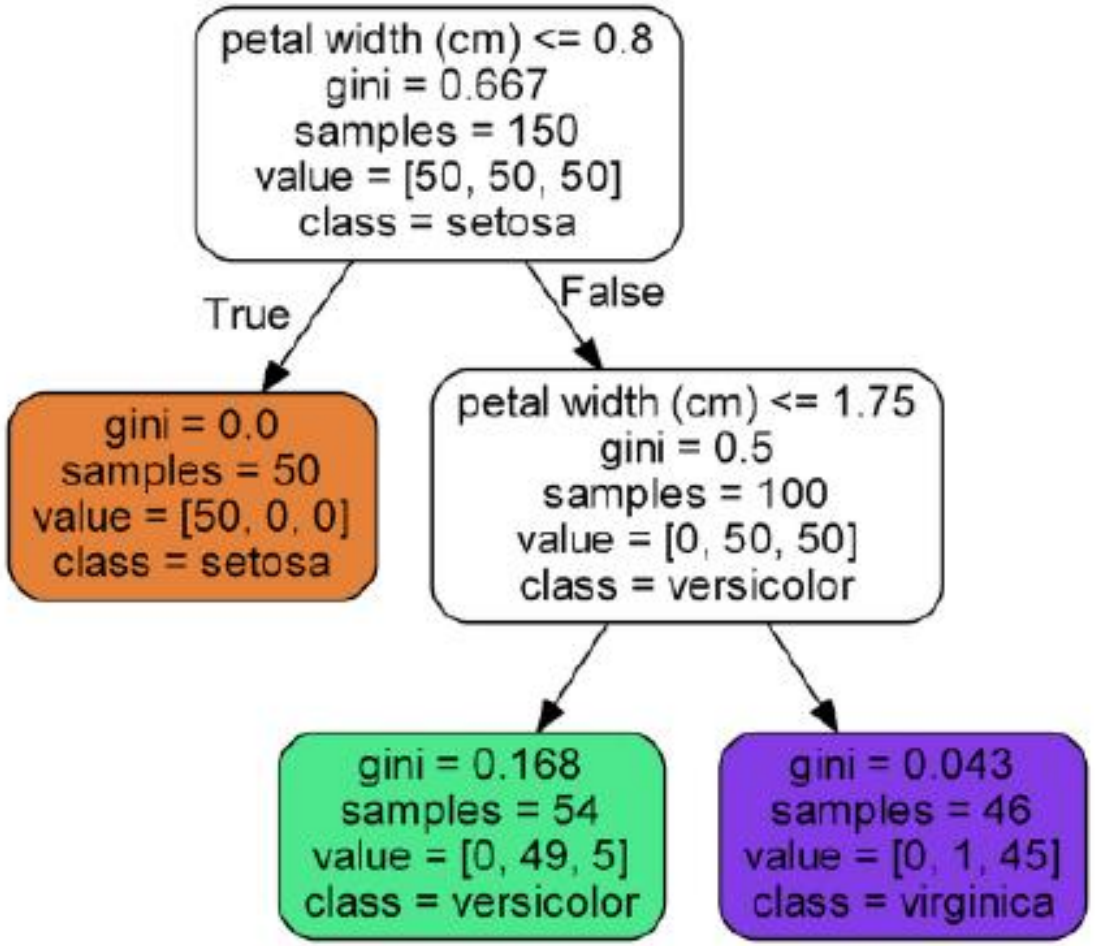
IF *age* = "31…40" THEN *buys_computer* = "*yes*"

IF *age* = ">40" AND *credit_rating* = "*excellent*" THEN *buys_computer* = "*yes*"

IF *age* = ">40" AND *credit_rating* = "*fair*" THEN *buys_computer* = "*no*"

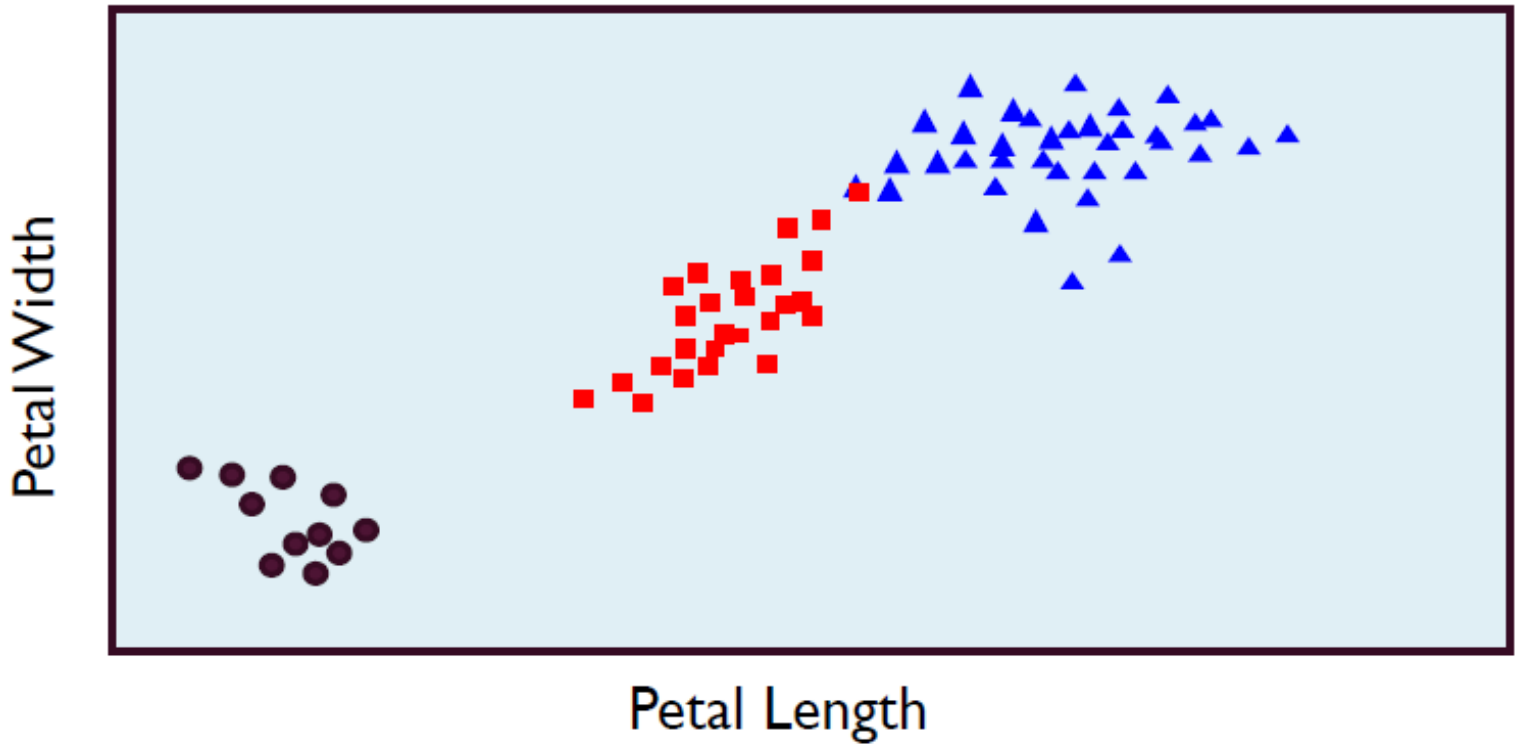
Forward School

Decision Tree Sample



Forward School

Decision Tree Boundaries



Forward School

Avoid Overfitting in Classification

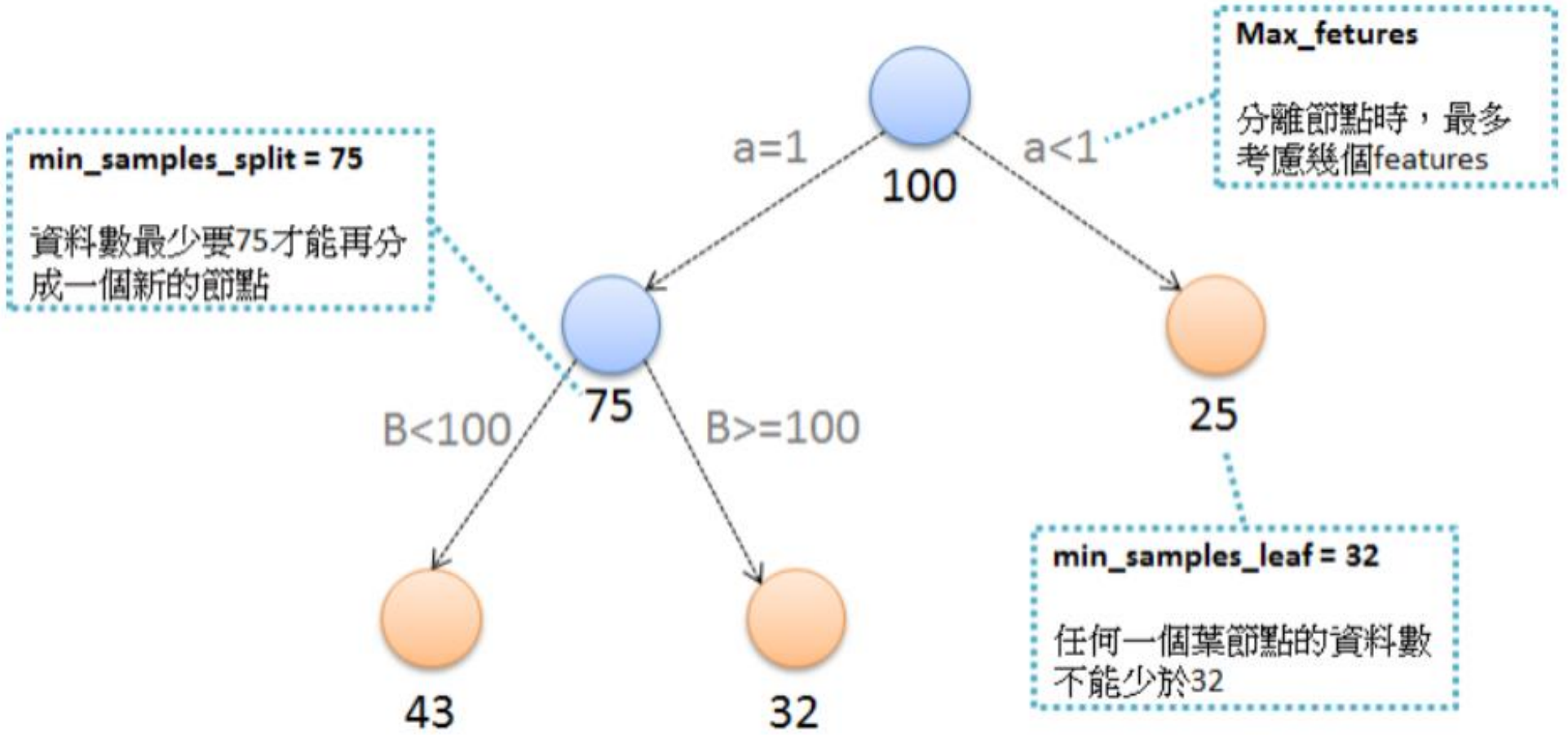
- The generated tree may overfit the training data
  - Too many branches, some may reflect anomalies due to noise or outliers
  - Result is in poor accuracy for unseen samples
- Two approaches to avoid overfitting
  - Prepruning: Halt tree construction early—do not split a node if this would result in the goodness measure falling below a threshold
    - Difficult to choose an appropriate threshold
  - Postpruning: Remove branches from a "fully grown" tree—get a sequence of progressively pruned trees
    - Use a set of data different from the training data to decide which is the "best pruned tree"

Regularization Hyperparameters

- Scikit-Learn
  - mx_depth　　#限制樹的高度最多幾層
  - min_samples_split　　#資料數目不得小於多少才能再產生新節點
  - min_samples_leaf　　#要成為葉節點，最少需要多少資料
  - min_weight_fraction_leaf　　#要成為葉節點，最少需要多少比重
  - max_leaf_nodes　　#限制最終葉節點的數目
  - max_features　#在分離節點時，最多考慮幾種特徵值

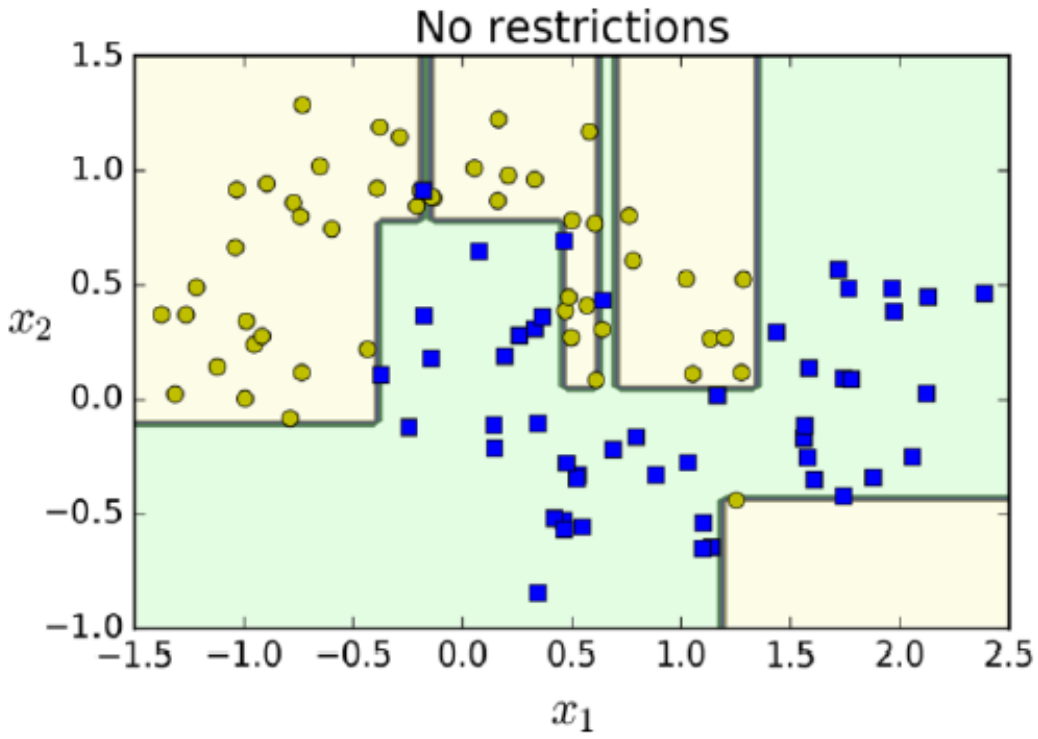
Forward School

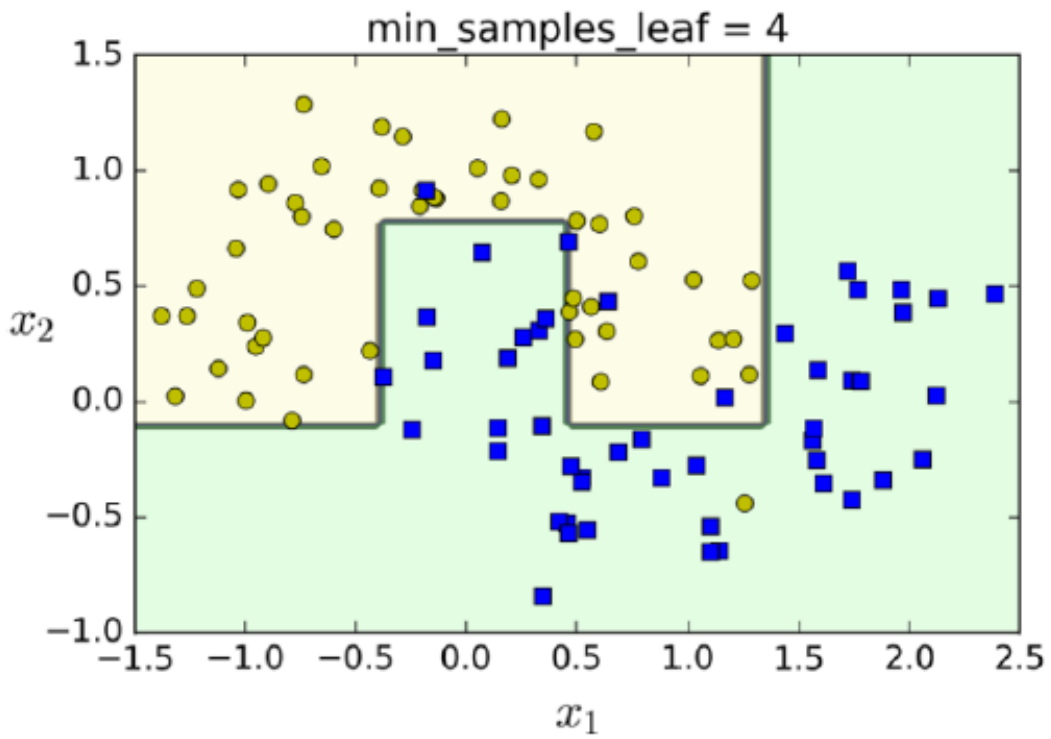
Regularization Hyperparameters



Forward School

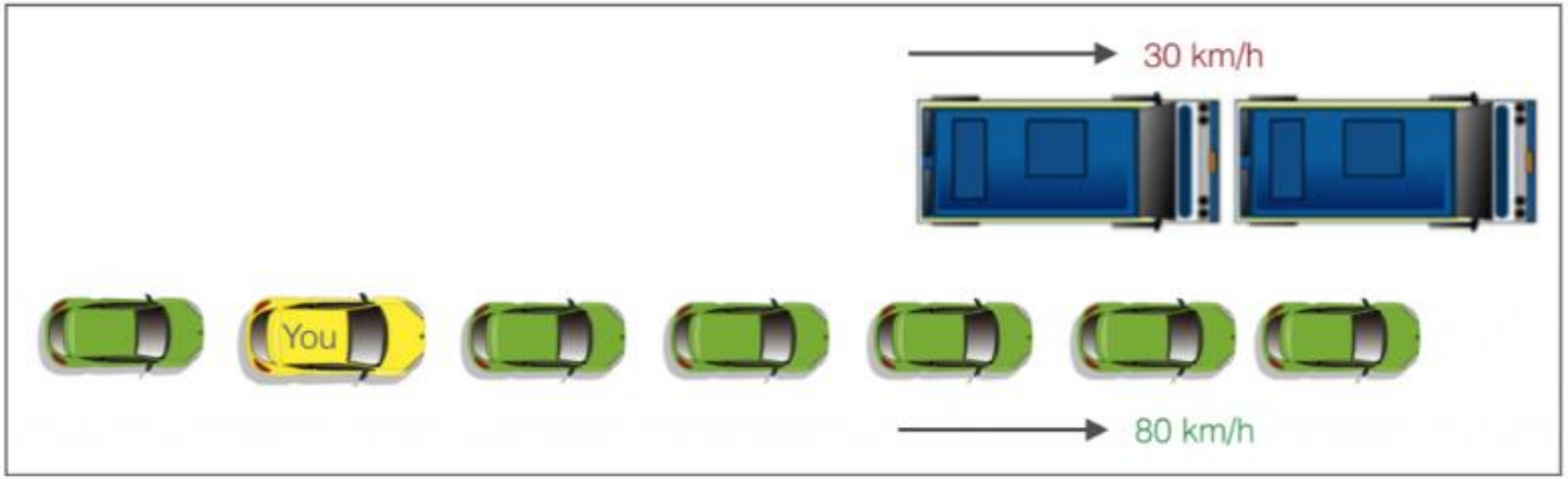
Regularization using min_samples_leaf





Forward School

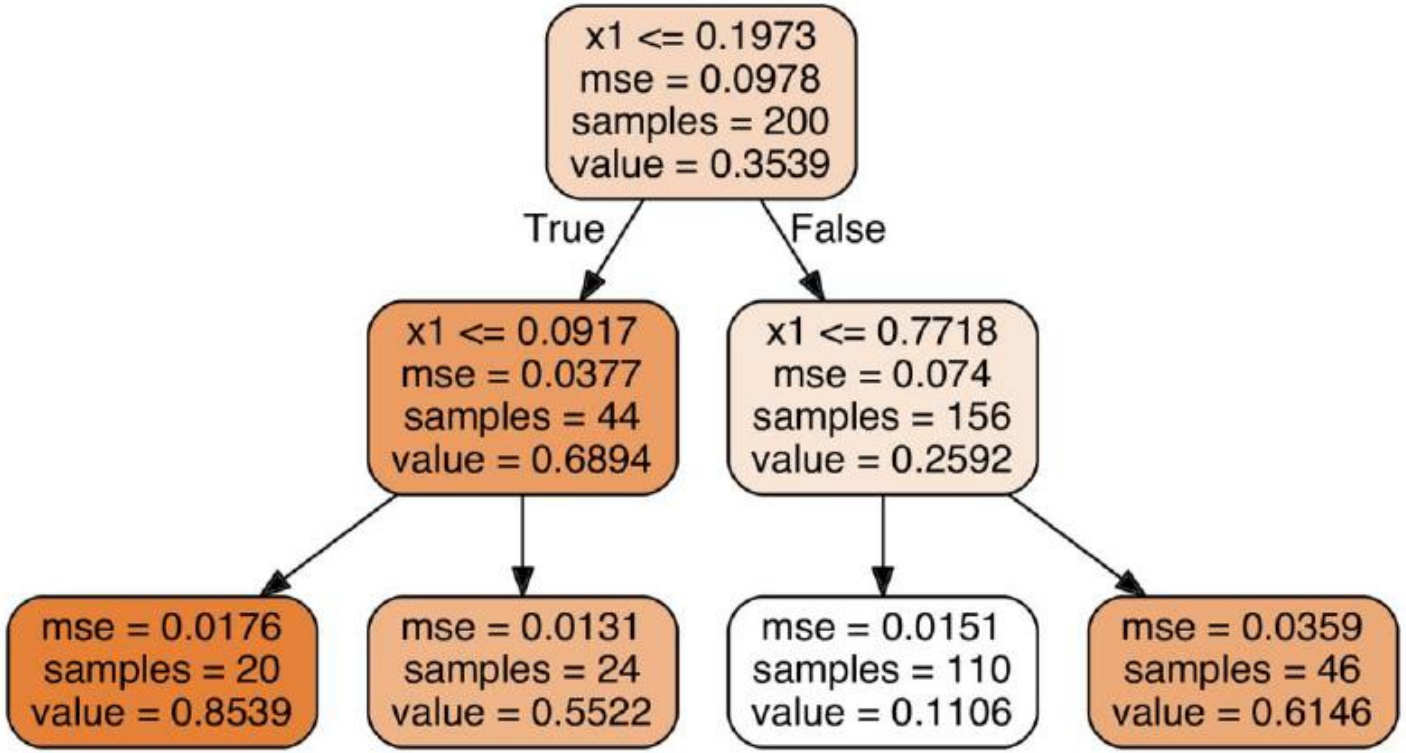
Tree Pruning



Forward School

DT for Regression

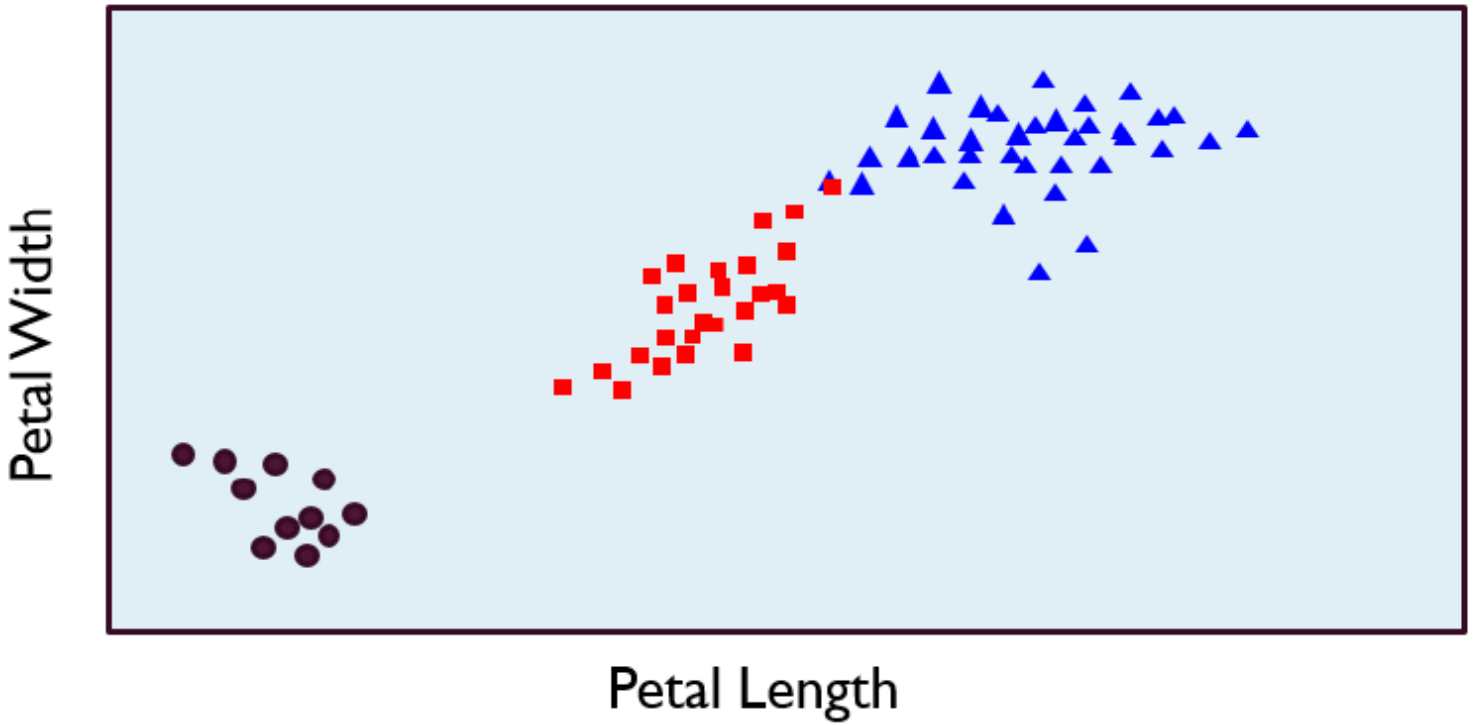
```
from sklearn.tree import DecisionTreeRegressor

tree_reg = DecisionTreeRegressor(max_depth=2)
tree_reg.fit(X, y)
```



Forward School

Limitation

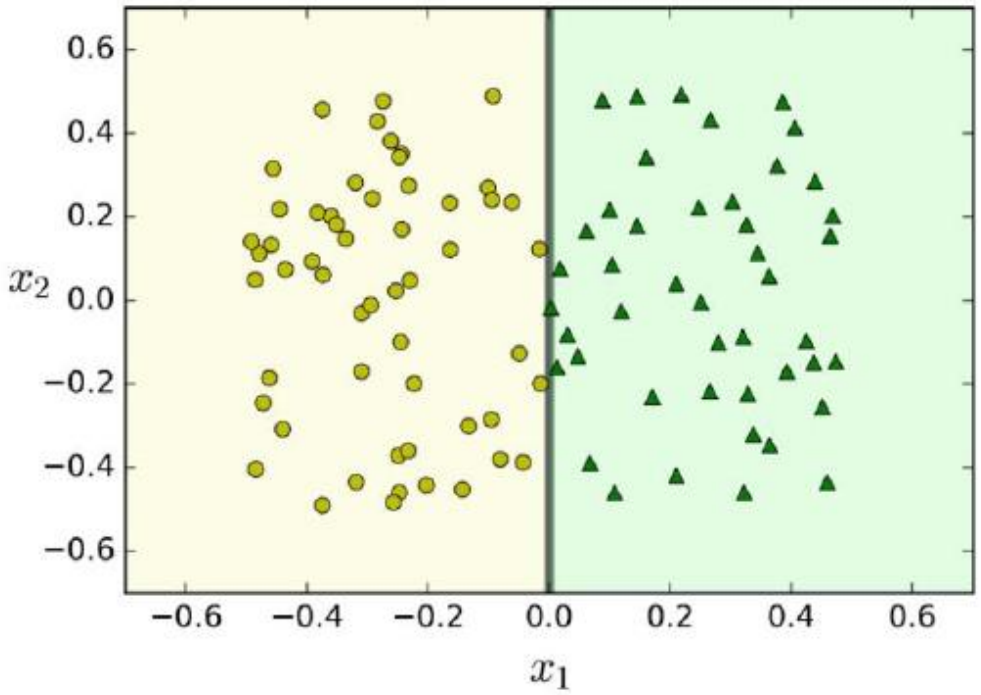
- Sensitive to small variations in training data.

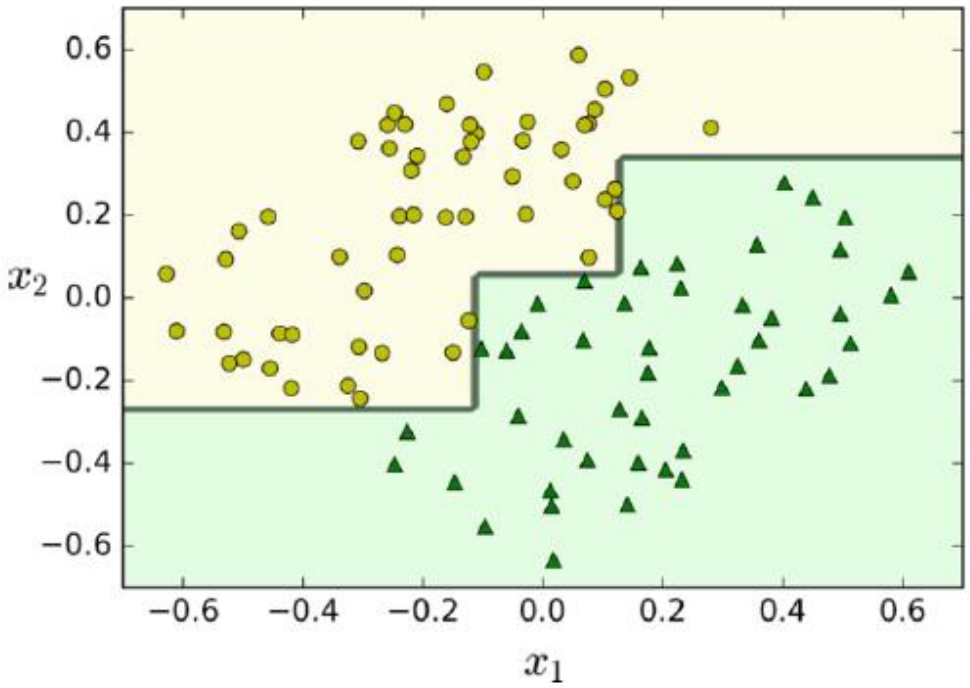


Forward School

Limitation

- Sensitive to rotation





Forward School

Example 3

Use Decision Tree to classify the content in different photos. (Hint: There are 7 contents: Dessert, GroupPhotos, Barbecue, Badminton,Baseball, Basketball, PingPong)

Desserts點心 (99張)

GroupPhoto團體照(271張)

Forward School

Example 3 - Solution

Feature Extraction

1. ColorStats

```
def describe(image):

    (means, stds) = cv2.meanStdDev(cv2.cvtColor(image, cv2.COLOR_BGR2HSV))
colorStats = np.concatenate([means, stds]).flatten()
```

2. Texture

```
haralick = mahotas.features.haralick(gray).mean(axis=0)
return np.hstack([colorStats, haralick])
```

Forward School

Example 3 - Solution

INFORMATION GAIN (entropy) :

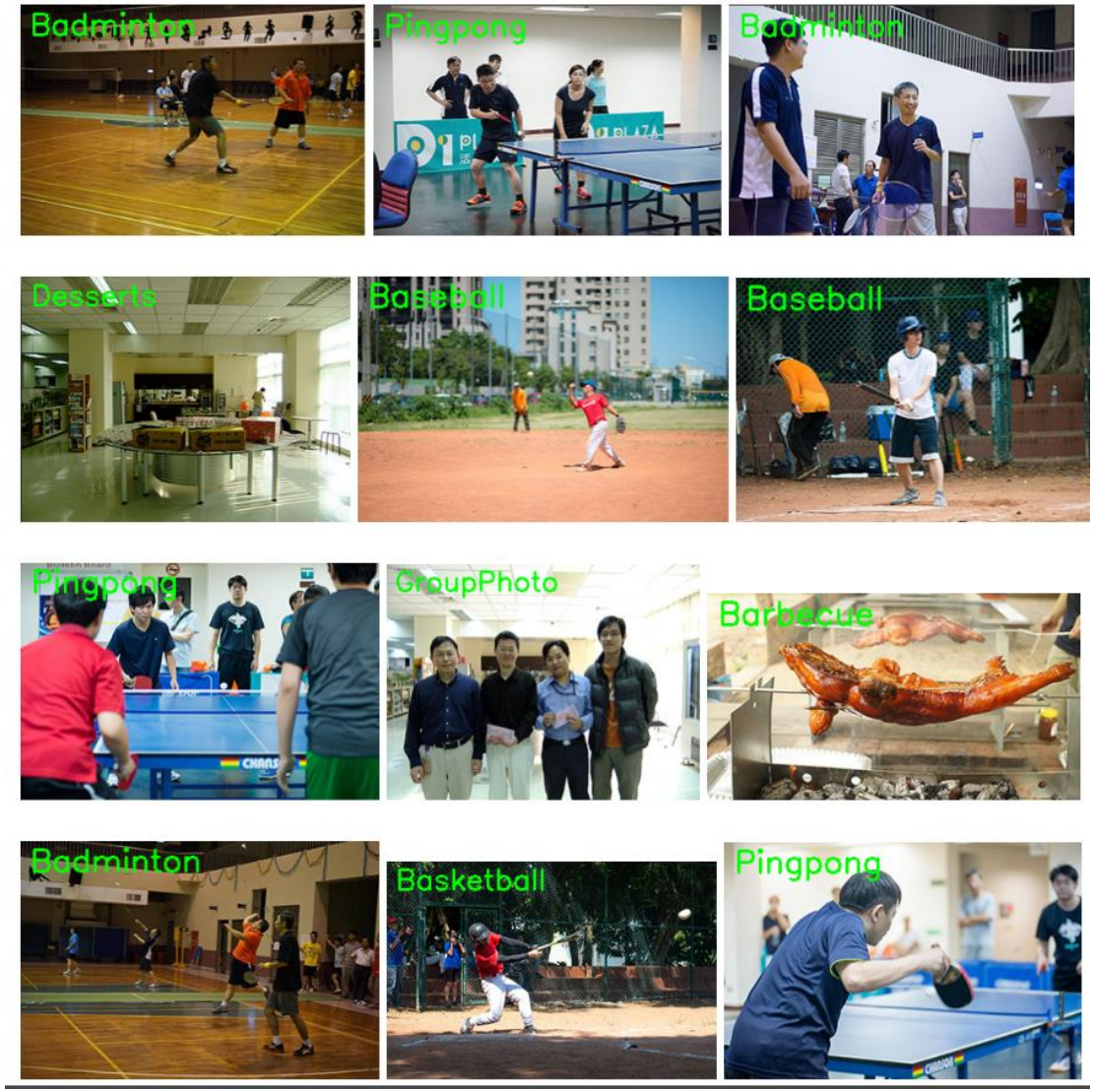
	precision	recall	f1-score	support
Badminton	0.75	0.82	0.78	62
Barbecue	0.65	0.62	0.63	21
Baseball	0.89	0.73	0.80	64
Basketball	0.62	0.53	0.57	15
Desserts	0.73	0.35	0.47	23
GroupPhoto	0.66	0.85	0.75	48
Pingpong	0.73	0.85	0.79	39
avg / total	0.75	0.74	0.73	272

Gini Index (gini) :

	precision	recall	f1-score	support
Badminton	0.86	0.81	0.83	62
Barbecue	0.65	0.71	0.68	21
Baseball	0.84	0.83	0.83	64
Basketball	0.33	0.27	0.30	15
Desserts	0.33	0.17	0.23	23
GroupPhoto	0.61	0.75	0.67	48
Pingpong	0.69	0.79	0.74	39
avg / total	0.70	0.71	0.70	272

Forward School

Example 3 - Solution



Forward School

Thank You

Forward School