Random Forests

Jean-Claude Walser

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

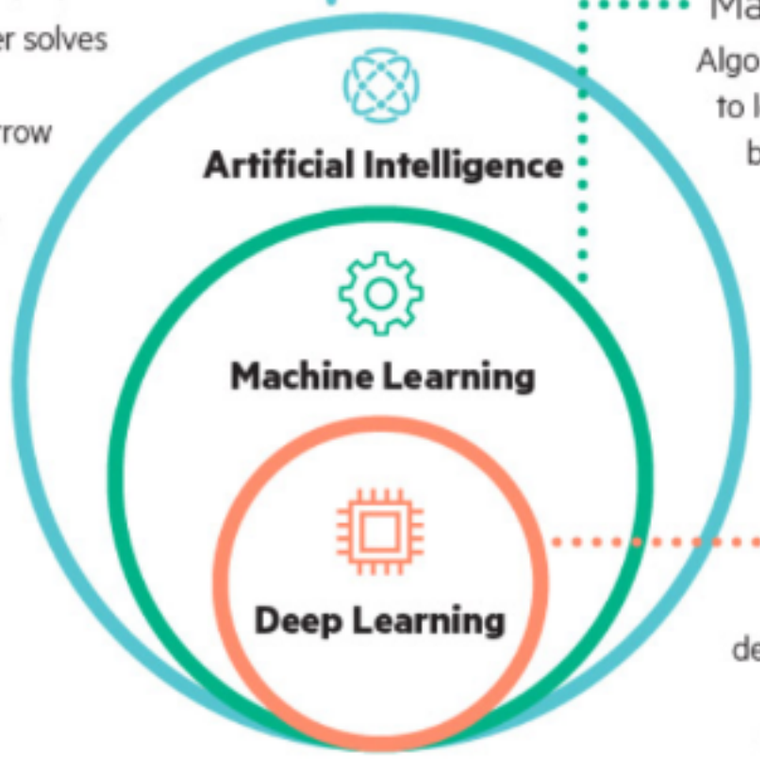
GDC Genetic Diversity Centre Zurich

What Makes a Machine Intelligent?

While AI is the headliner, there are actually subsets of the technology which can be applied to solving human problems in different ways.

Artificial Intelligence (AI)

A process where a computer solves a task in a way that mimics human behavior. Today, narrow AI—when a machine is trained to do one particular task—is becoming more widely used, from virtual assistants to self-driving cars to automatic tagging your friends in your photos on Facebook.

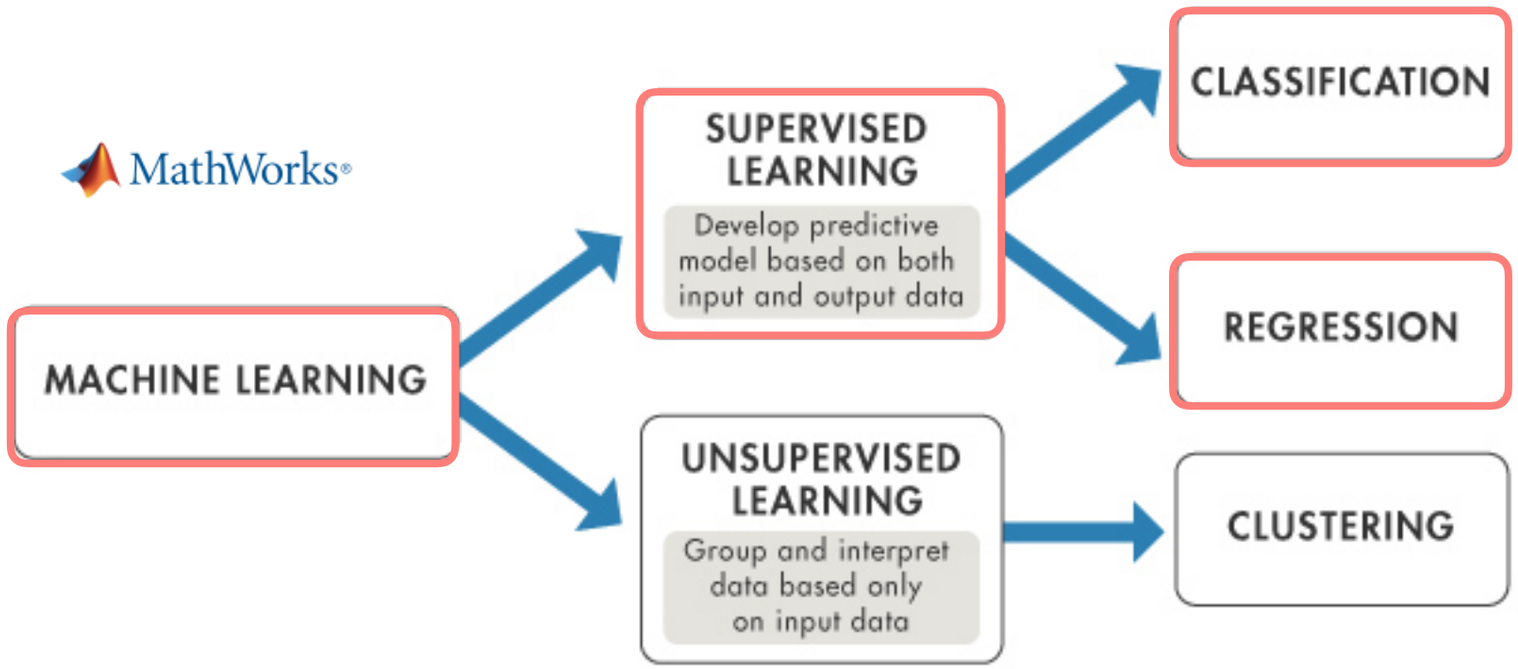


Machine Learning (ML)

Algorithms that allow computers to learn from examples without being explicitly programmed.

Deep Learning (DL)

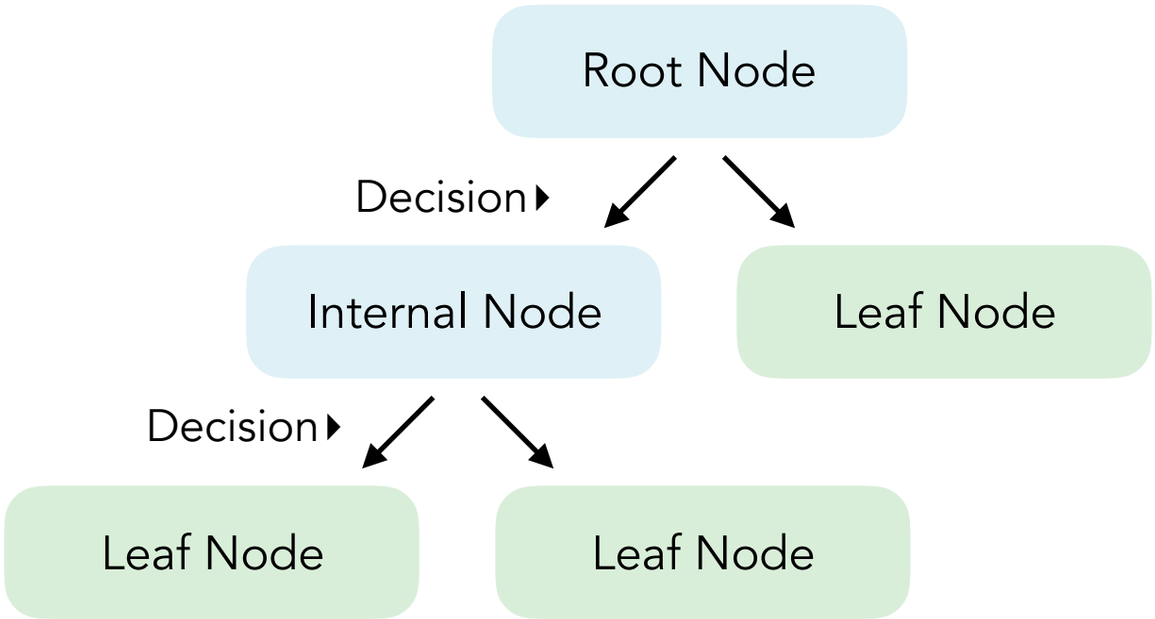
A subset of ML which uses deep artificial neural networks as models and does not require feature engineering.



- Decision tree
- Ensembles: Bagging / Boosting / **Random Forest**

- **Supervised Learning (input-output pairs, direct feedback, predict outcome)**
- Unsupervised Learning (no feedback, find hidden structure)
- Reinforcment Learning (reward system, decision process)

Decision Trees



Desicion Types

Binary: TRUE / FALSE

Threshold: e.g. bigger x

Factor: blue, yellow, green

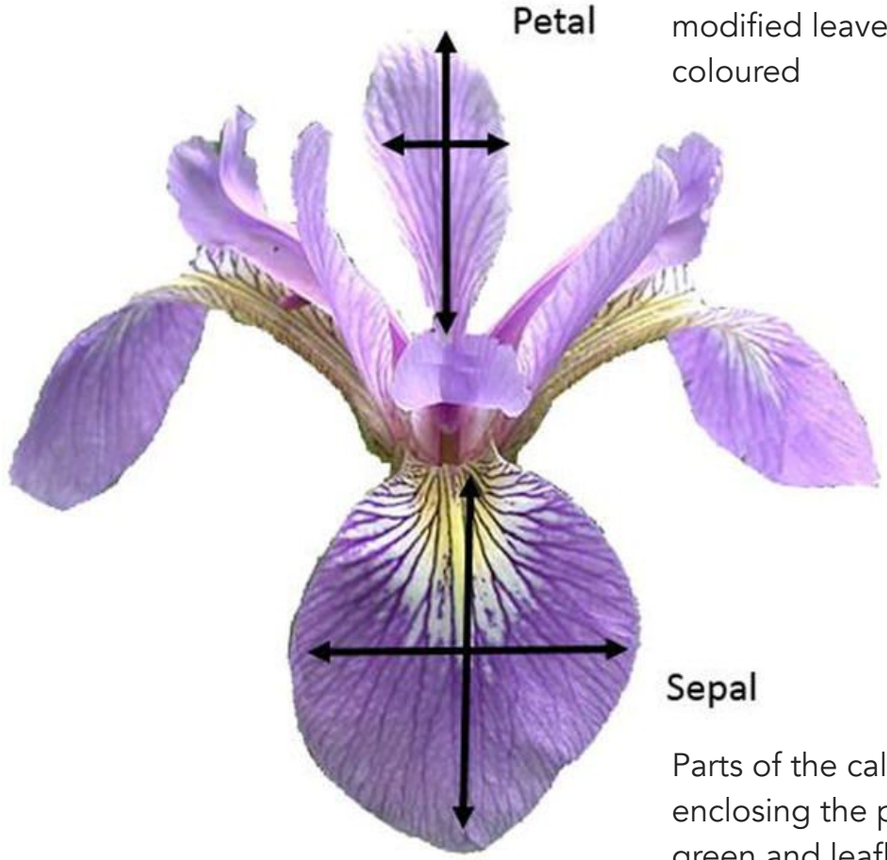
1. Find root node by comparing Gini impurity scores.

2. Calcualte Gini impurity scores.

3. If the node itself has the lowest score it becomes a leaf node.

4. If the node has a higher score it becomes a internal node and the separtions with the lowest score will be the next node.


```
library(datasets)
data(iris)
summary(iris)
```

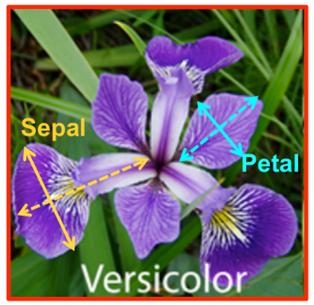


Corolla of a flower, which are modified leaves and are typically coloured

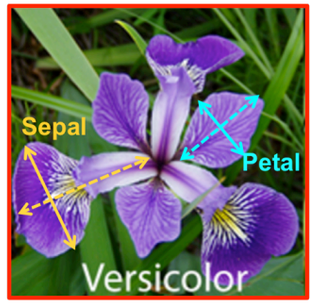
Parts of the calyx of a flower, enclosing the petals and typically green and leaflike.

Predictors **Response**

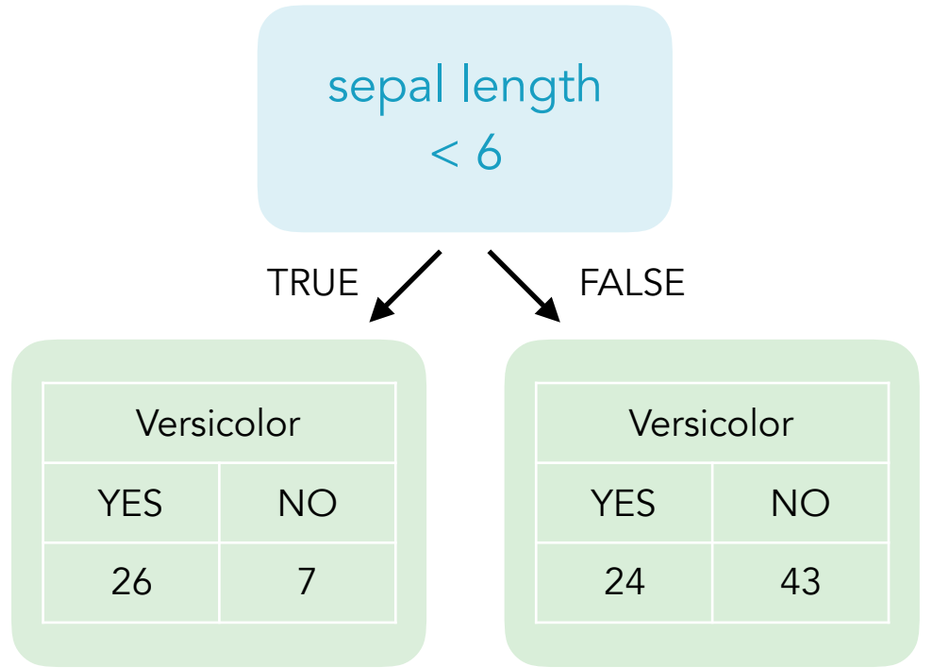
Sepal Length	Sepal Width	Petal Length	Petal Width	Species
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
…	…	…	…	…
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
6.5	3.0	5.8	2.2	virginica
…	…	…	…	…

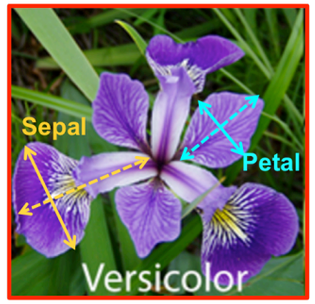


We use a slightly modified data set based on the famous (Fisher's or Anderson's) **iris data** set. It contains measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

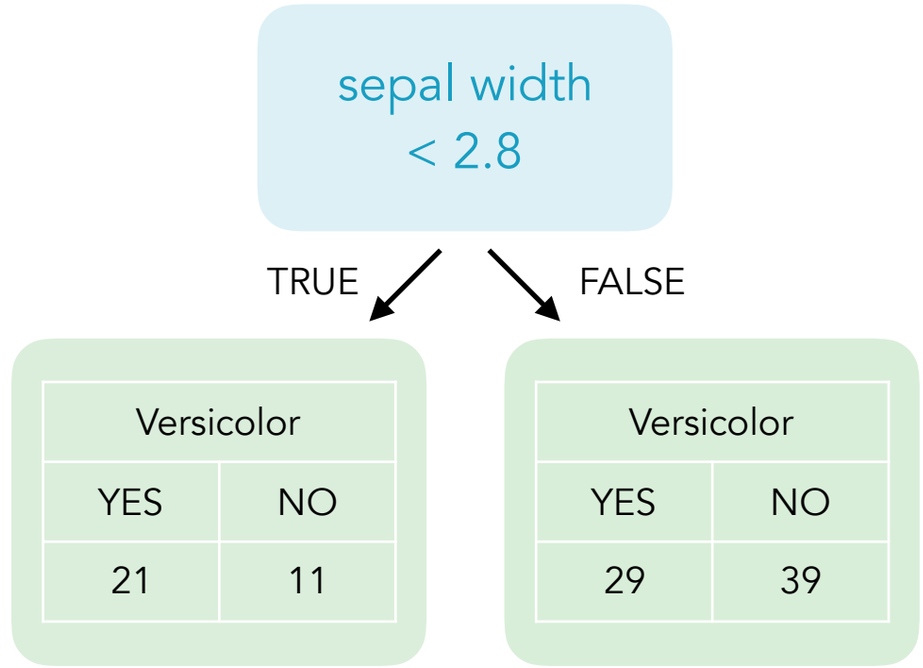


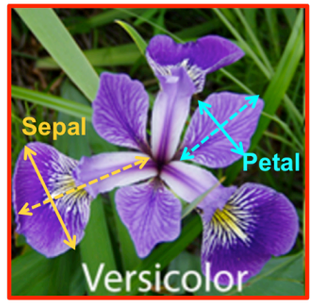
Sepal Length	Sepal Width	Petal Length	Petal Width	Species
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
…	…	…	…	…
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
6.5	3.0	5.8	2.2	virginica
…	…	…	…	…



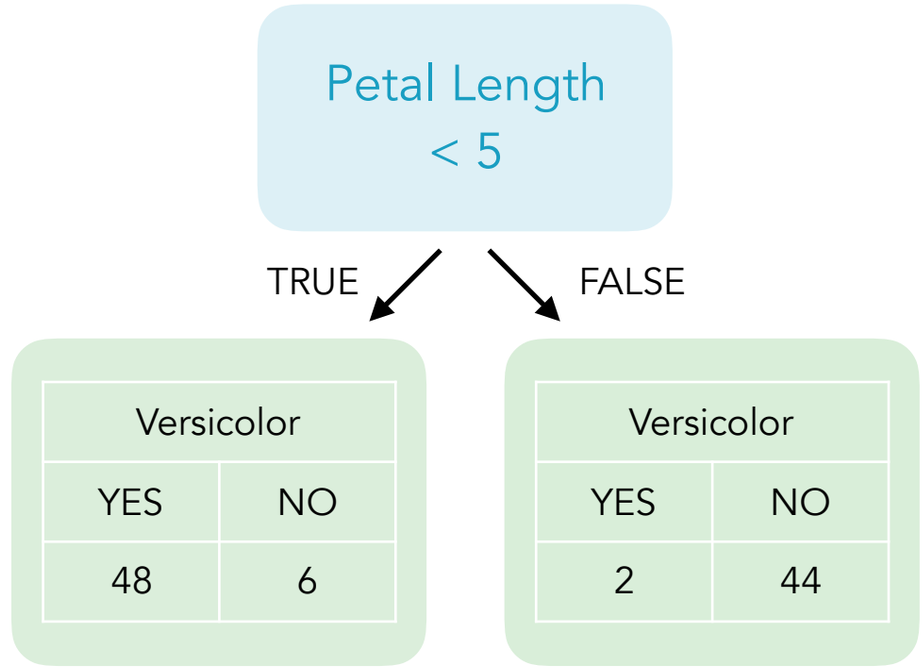


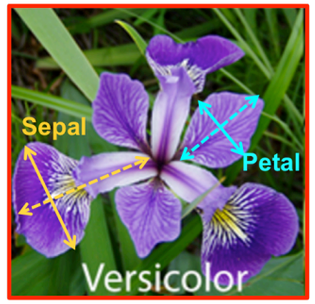
Sepal Length	Sepal Width	Petal Length	Petal Width	Species
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
…	…	…	…	…
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
6.5	3.0	5.8	2.2	virginica
…	…	…	…	…



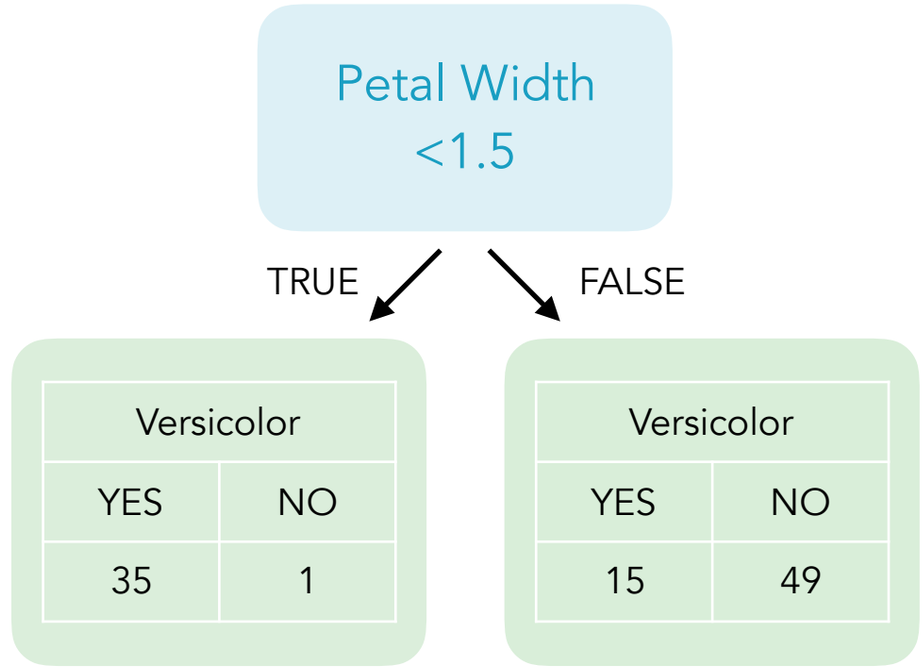


Sepal Length	Sepal Width	Petal Length	Petal Width	Species
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
…	…	…	…	…
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
6.5	3.0	5.8	2.2	virginica
…	…	…	…	…





Sepal Length	Sepal Width	Petal Length	Petal Width	Species
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
…	…	…	…	…
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
6.5	3.0	5.8	2.2	virginica
…	…	…	…	…



GDC Genetic Diversity Centre Zurich

sepal length <6

TRUE

Versicolor	
YES	NO
26	7

FALSE

Versicolor	
YES	NO
24	43

petal length <5

TRUE

Versicolor	
YES	NO
48	6

FALSE

Versicolor	
YES	NO
2	44

sepal width <2.8

TRUE

Versicolor	
YES	NO
21	11

FALSE

Versicolor	
YES	NO
29	39

petal width <1.5

TRUE

Versicolor	
YES	NO
35	1

FALSE

Versicolor	
YES	NO
15	49

Gini Impurity Measure

$$gini = 1 - \left(\frac{\text{propability of TRUE}}{\text{number of cases}}\right)^2 - \left(\frac{\text{propability of FALSE}}{\text{number of cases}}\right)^2$$

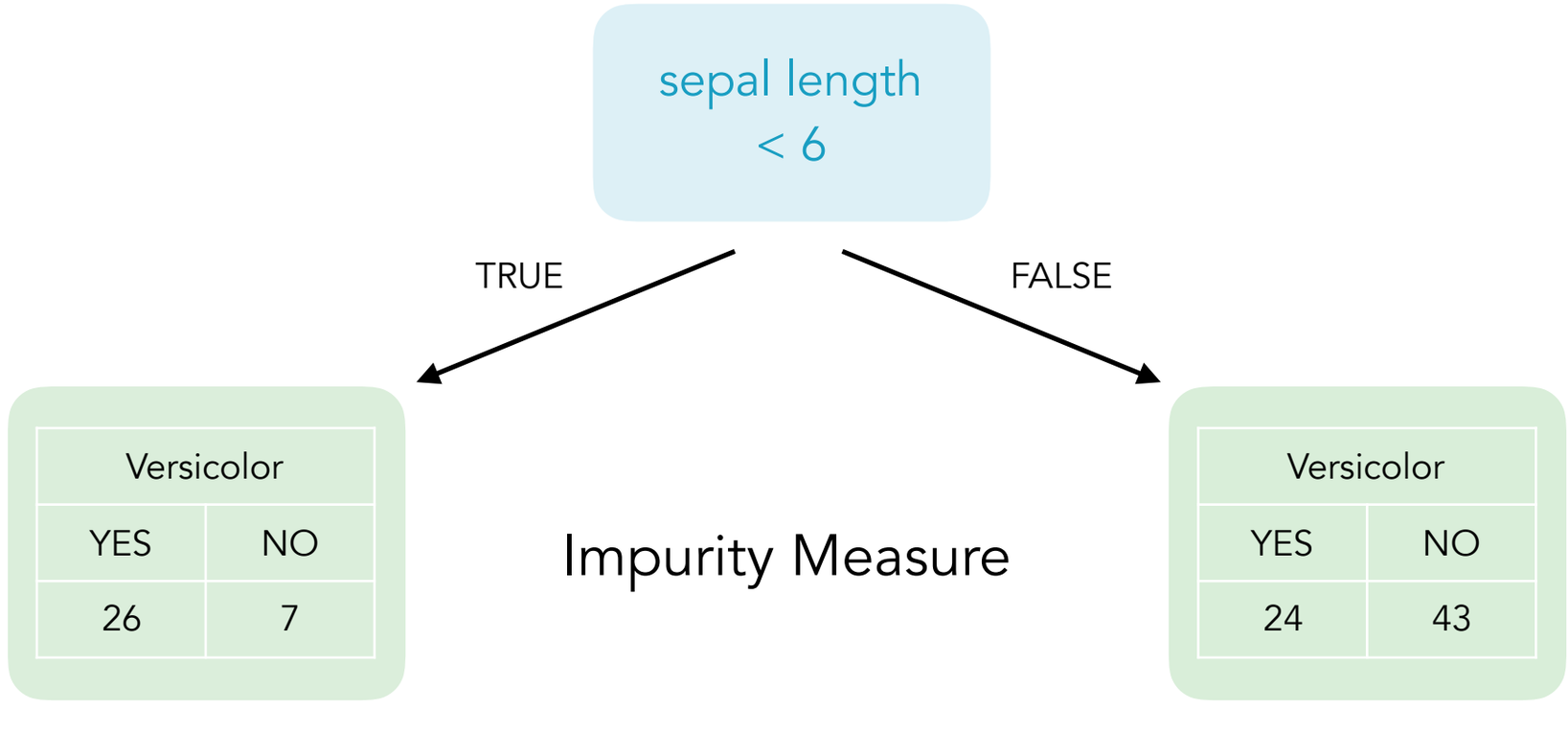
YES	NO
100	0

$$gini = 1 - \left(\frac{100}{100}\right)^2 - \left(\frac{0}{100}\right)^2 = 0$$

YES	NO
50	50

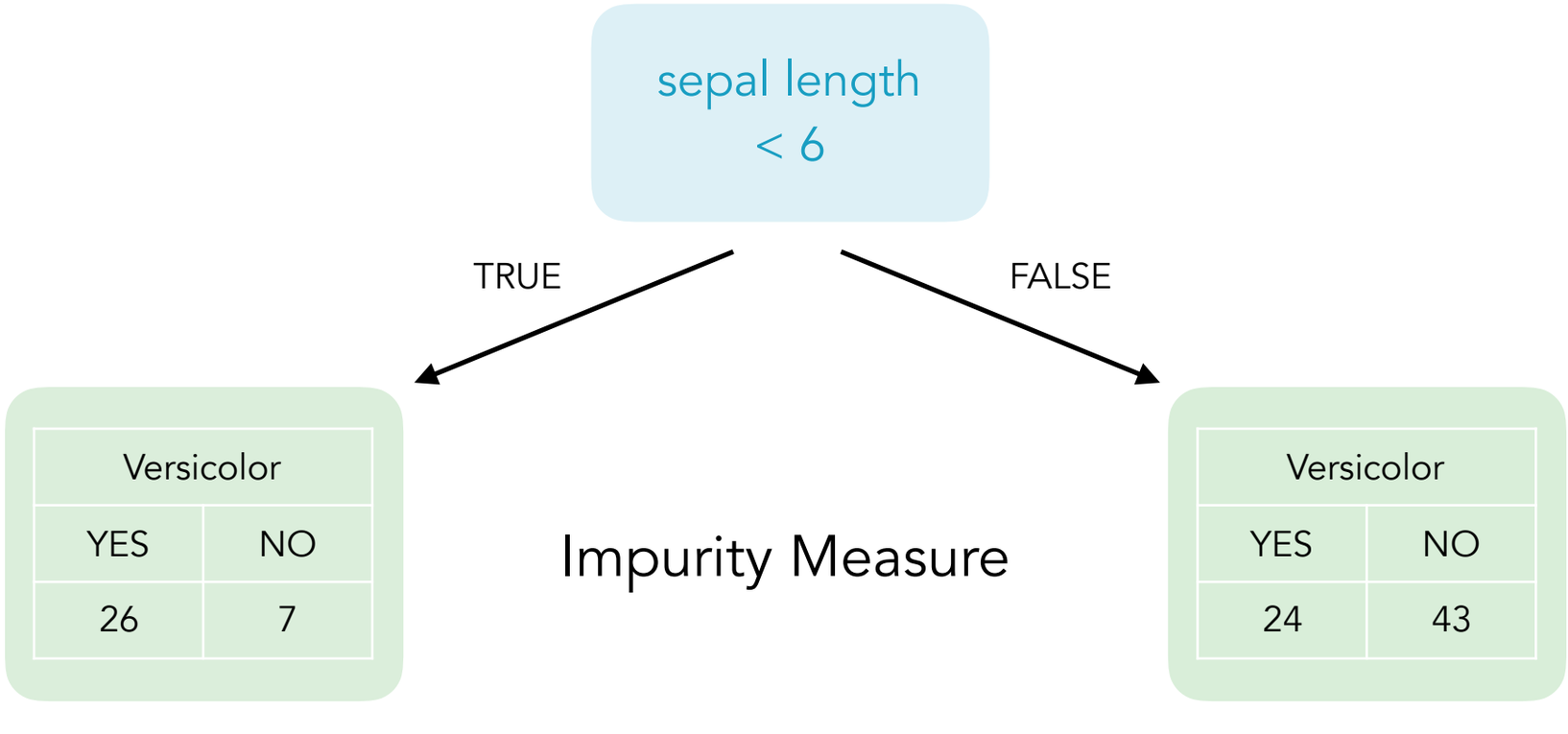
$$gini = 1 - \left(\frac{50}{100}\right)^2 - \left(\frac{50}{100}\right)^2 = 0.5$$

*The **node impurity** is a measure of the **homogeneity** of the labels at the node. The current implementation provides two impurity measures for classification (Gini impurity and entropy) and one impurity measure for regression (variance).



$$gini = 1 - \left(\frac{26}{26+7}\right)^2 - \left(\frac{7}{26+7}\right)^2 = 0.334$$

$$gini = 1 - \left(\frac{24}{24+43}\right)^2 - \left(\frac{43}{24+43}\right)^2 = 0.460$$



$$gini = 1 - \left(\frac{26}{26+7}\right)^2 - \left(\frac{7}{26+7}\right)^2 = 0.334$$

$$gini = 1 - \left(\frac{24}{24+43}\right)^2 - \left(\frac{43}{24+43}\right)^2 = 0.460$$

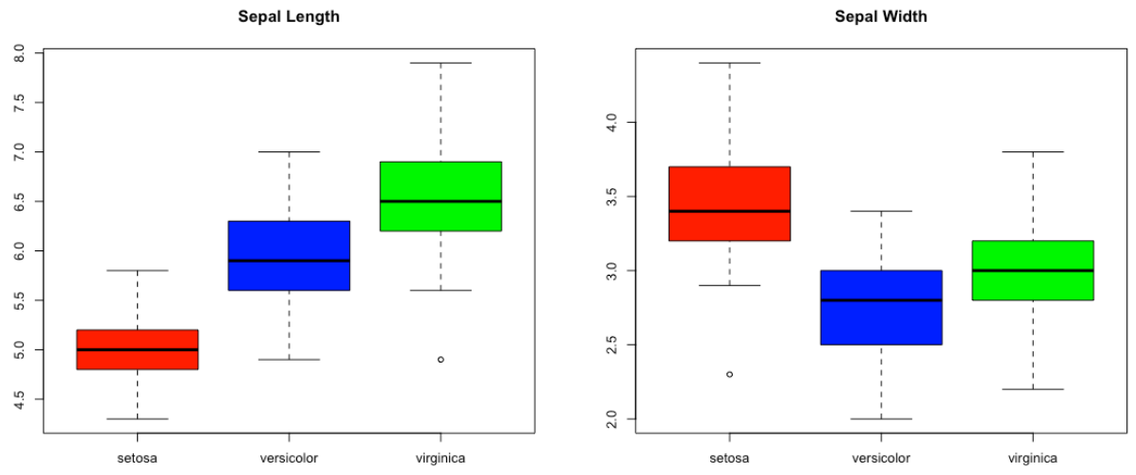
$$\text{weighted mean} = \frac{33}{33+67} * 0.334 + \frac{67}{33+67} * 0.460 = \underline{\underline{0.418}}$$

sepal length
< 6

gini impurity = 0.418

sepal width
< 2.8

gini impurity = 0.477

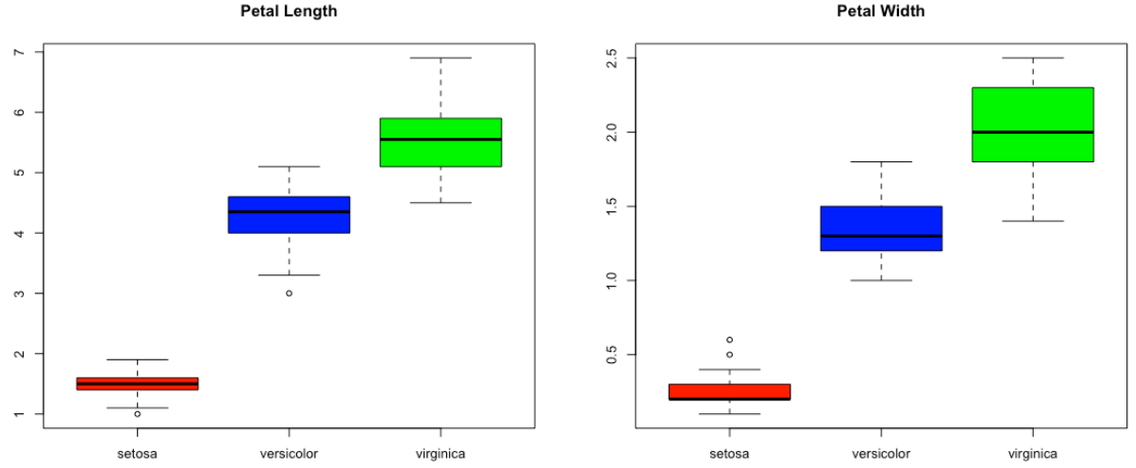


petal length
< 5

gini impurity = 0.145

petal width
< 1.5

gini impurity = 0.249



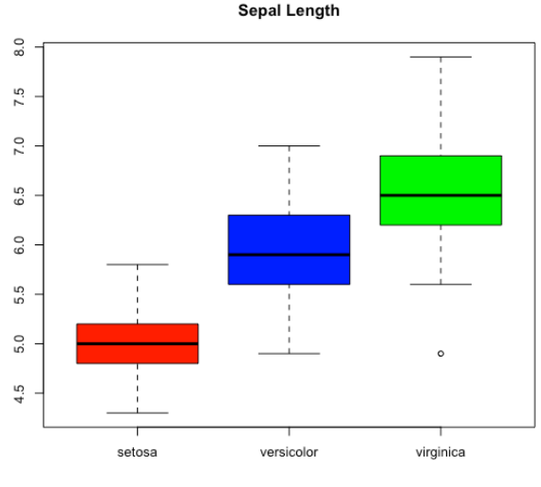
GDC Genetic Diversity Centre Zurich

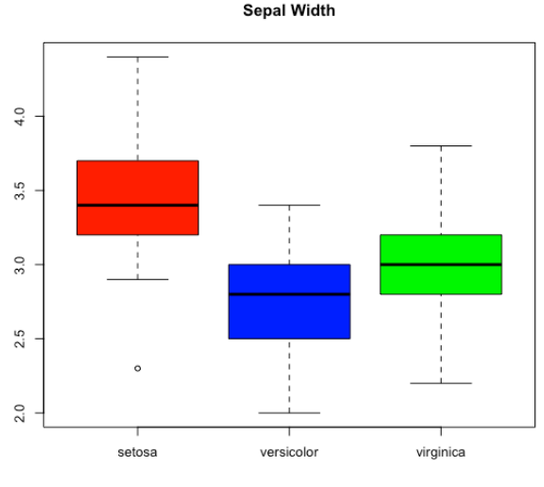
sepal length
< 6

gini impurity = 0.418

sepal width
< 2.8

gini impurity = 0.477



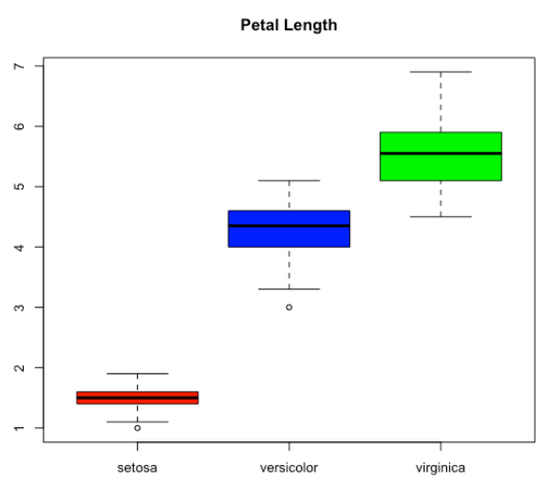


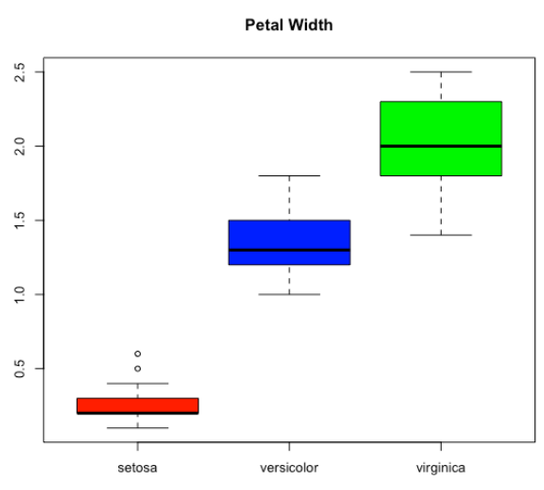
petal length
< 5

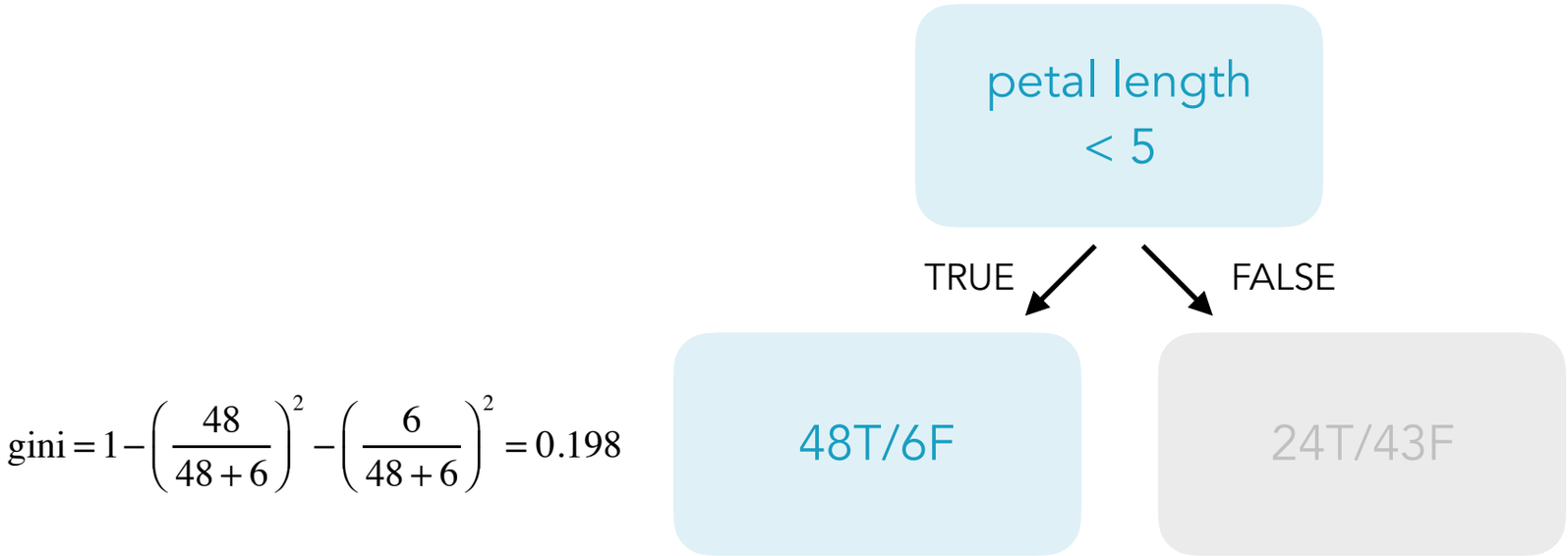
gini impurity = 0.145

petal width
< 1.5

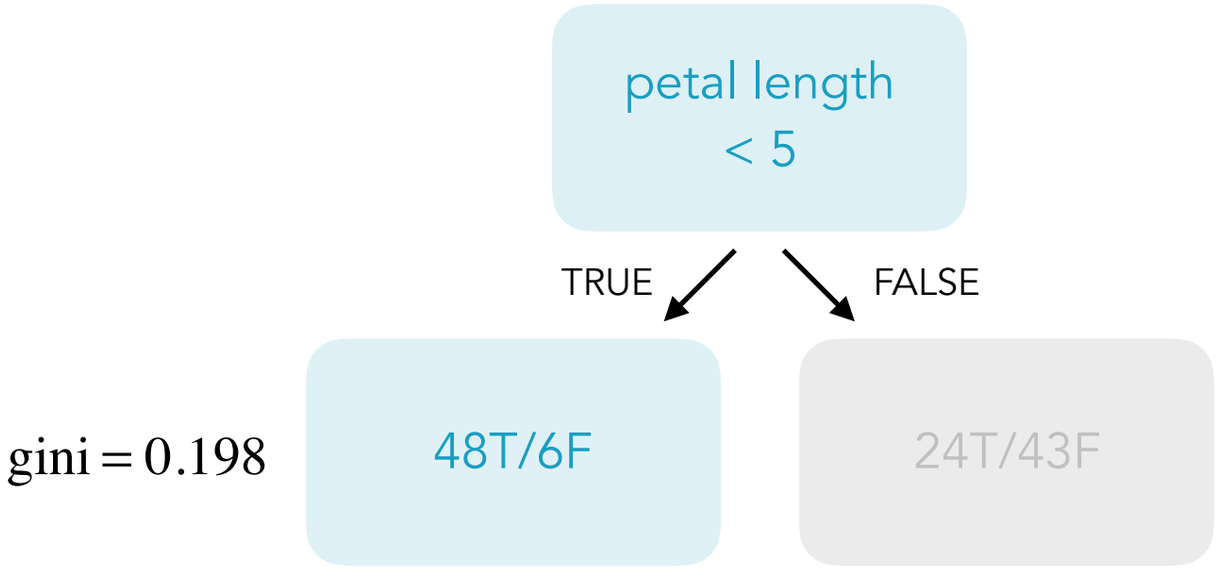
gini impurity = 0.249



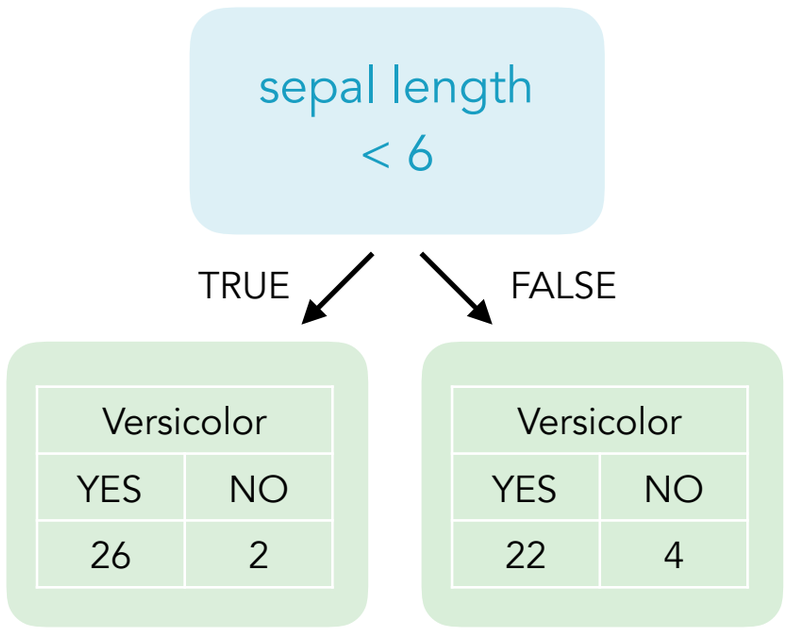




$$\text{gini} = 1 - \left(\frac{48}{48+6}\right)^2 - \left(\frac{6}{48+6}\right)^2 = 0.198$$



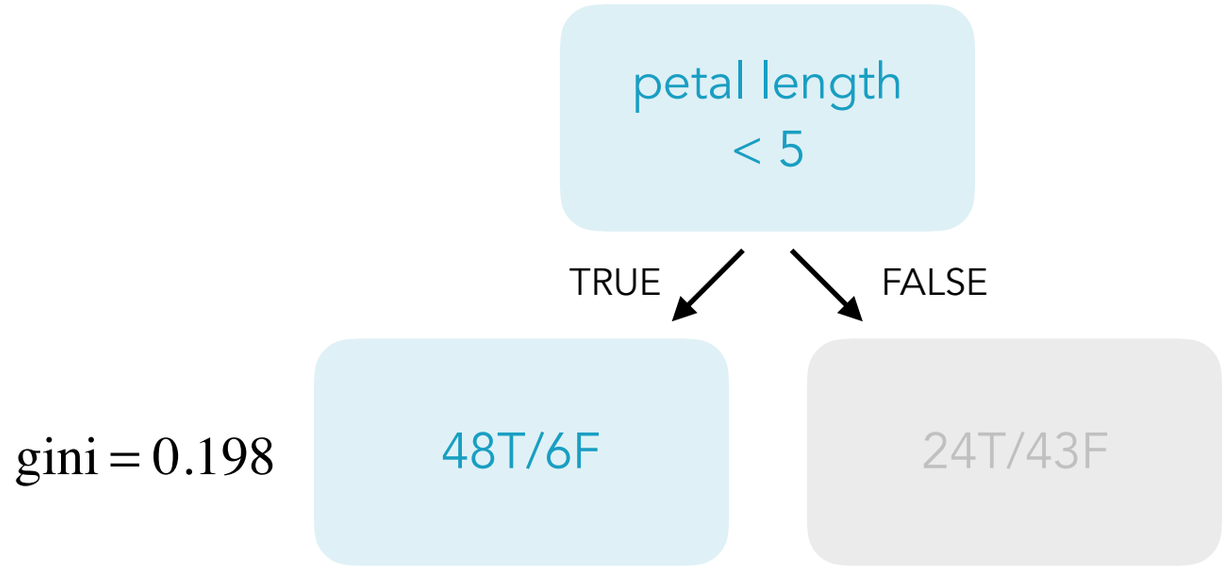
$\text{gini} = 0.198$



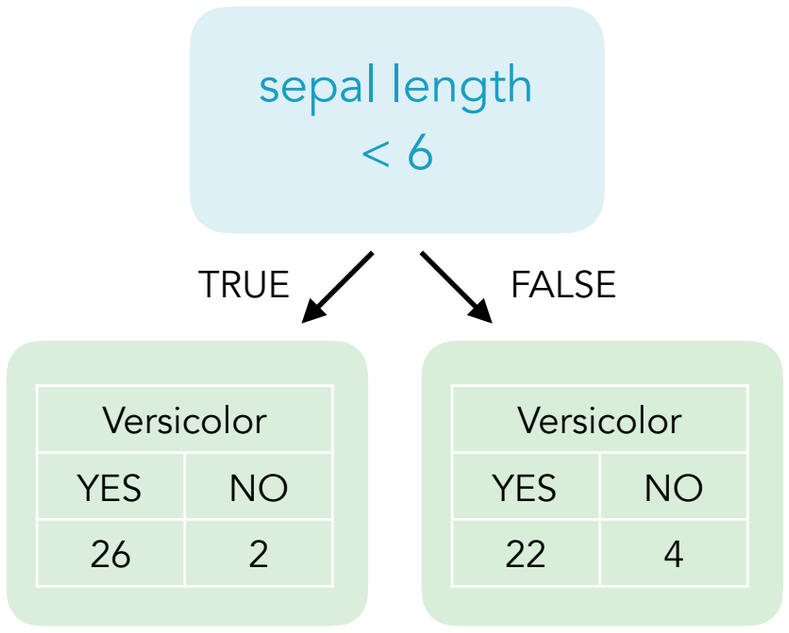
Versicolor	
YES	NO
26	2

Versicolor	
YES	NO
22	4

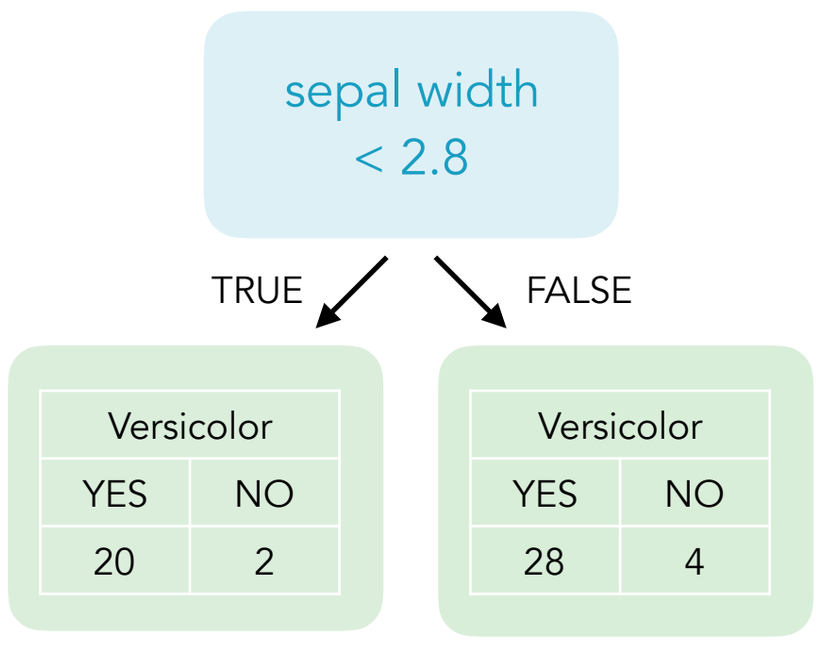
$\text{gini} = 0.194$



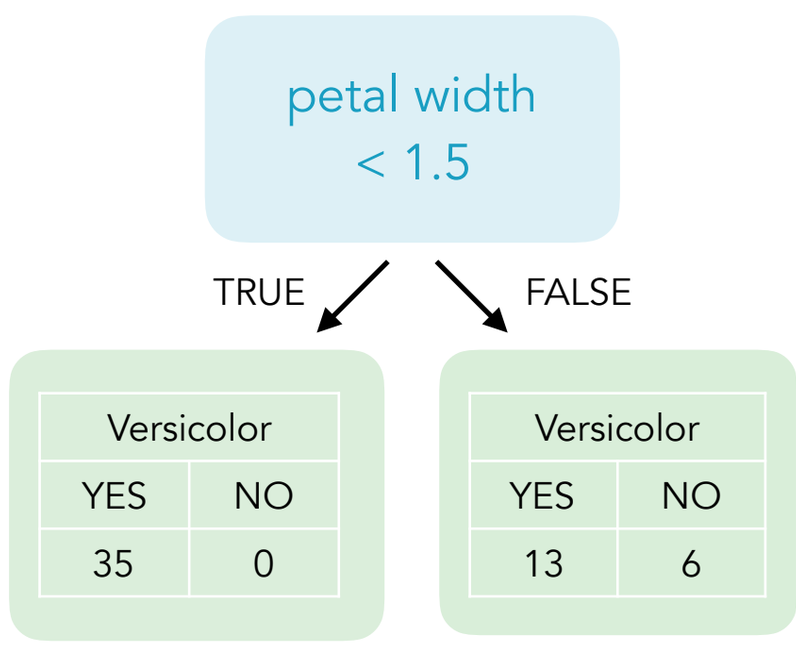
$\text{gini} = 0.198$



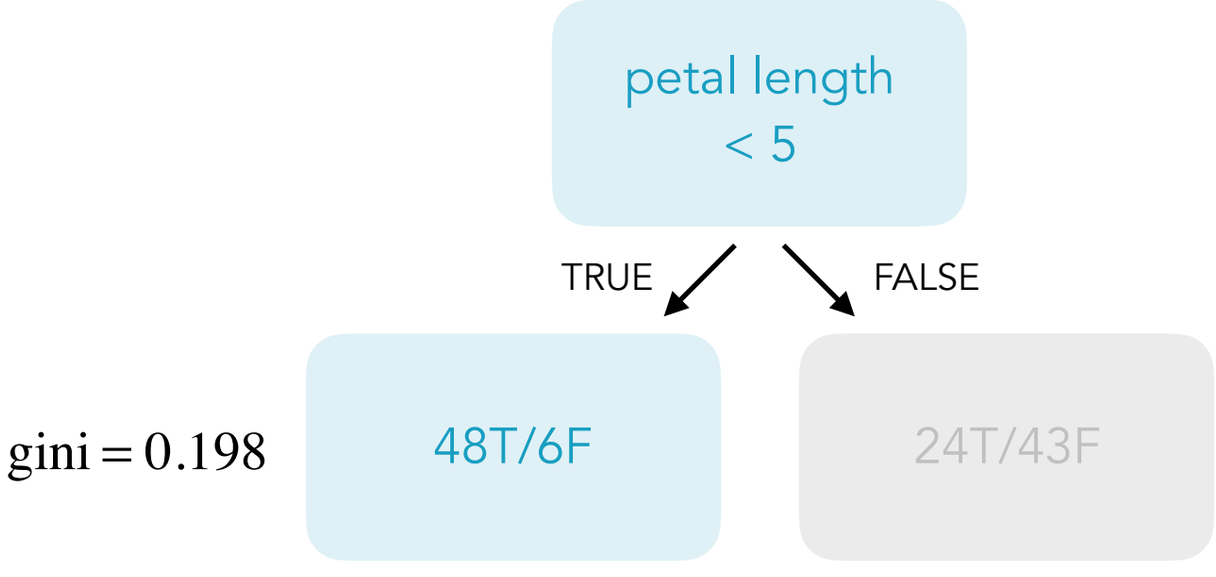
$\text{gini} = 0.194$



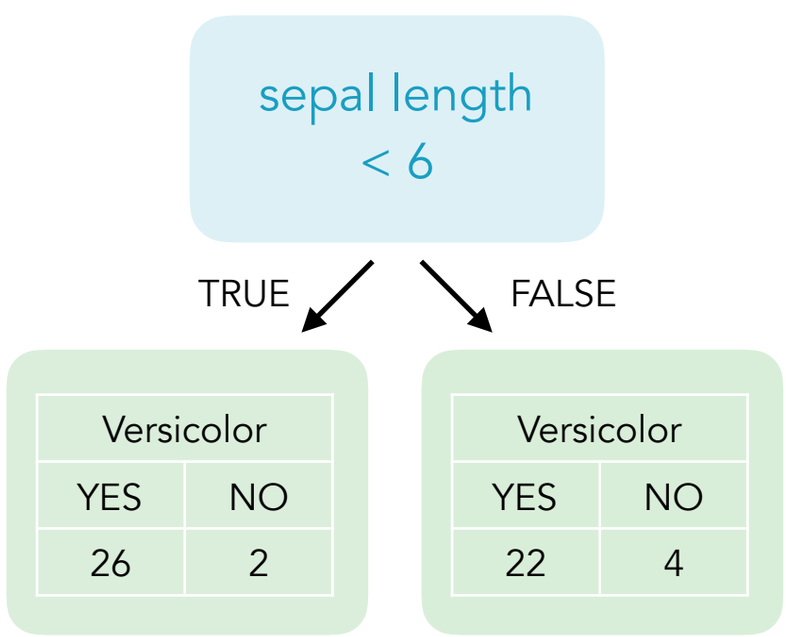
$\text{gini} = 0.197$



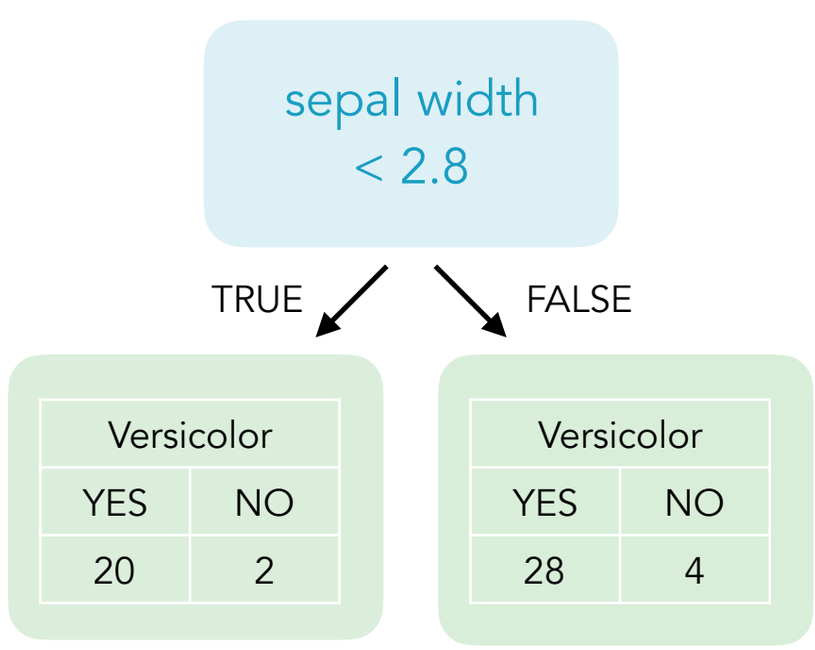
$\text{gini} = 0.104$



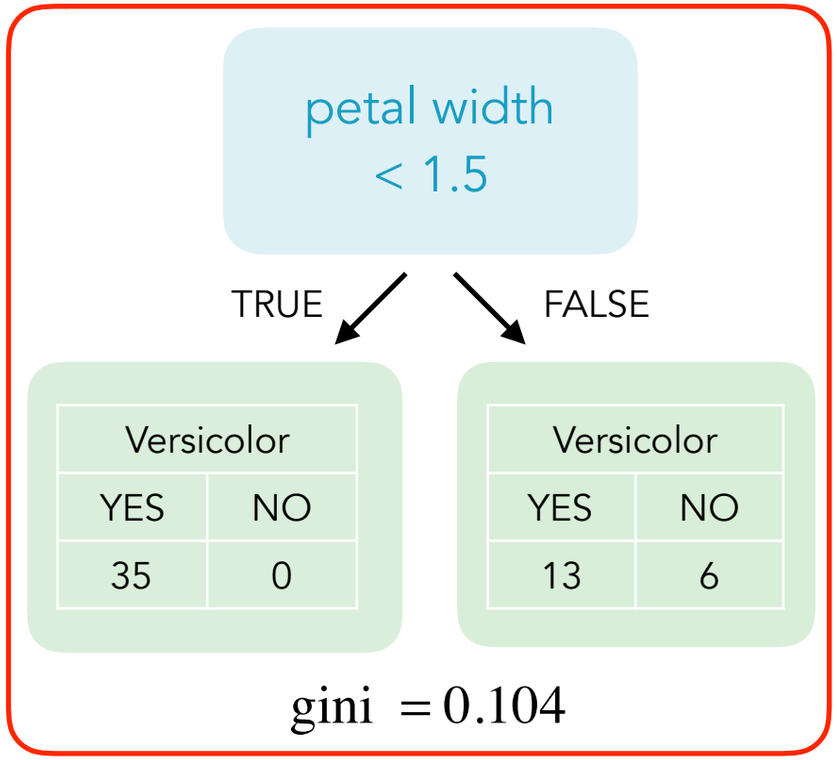
$\text{gini} = 0.198$



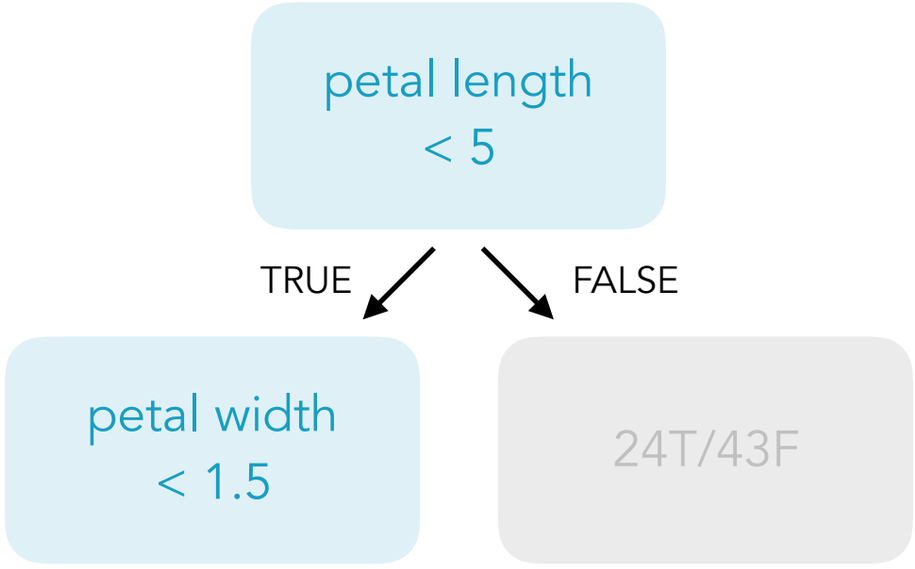
$\text{gini} = 0.194$

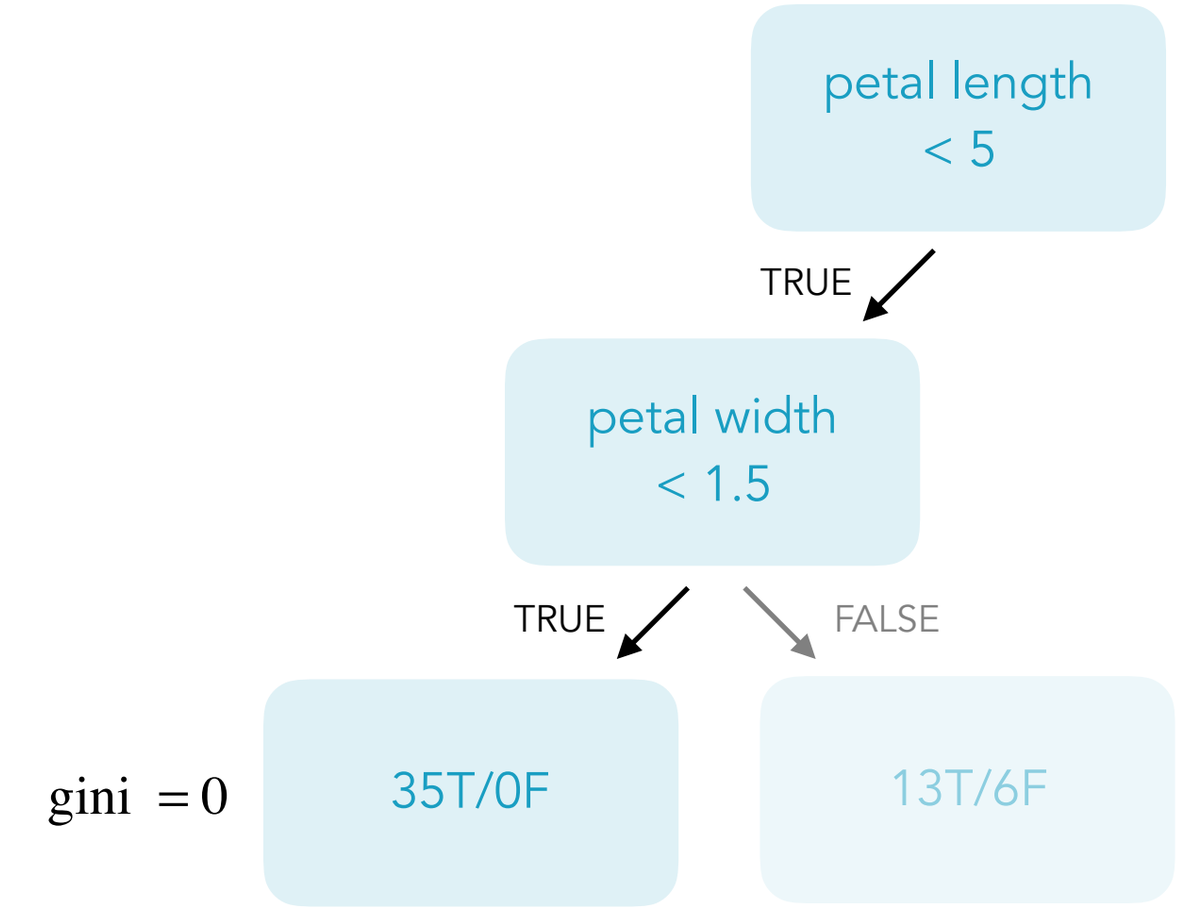


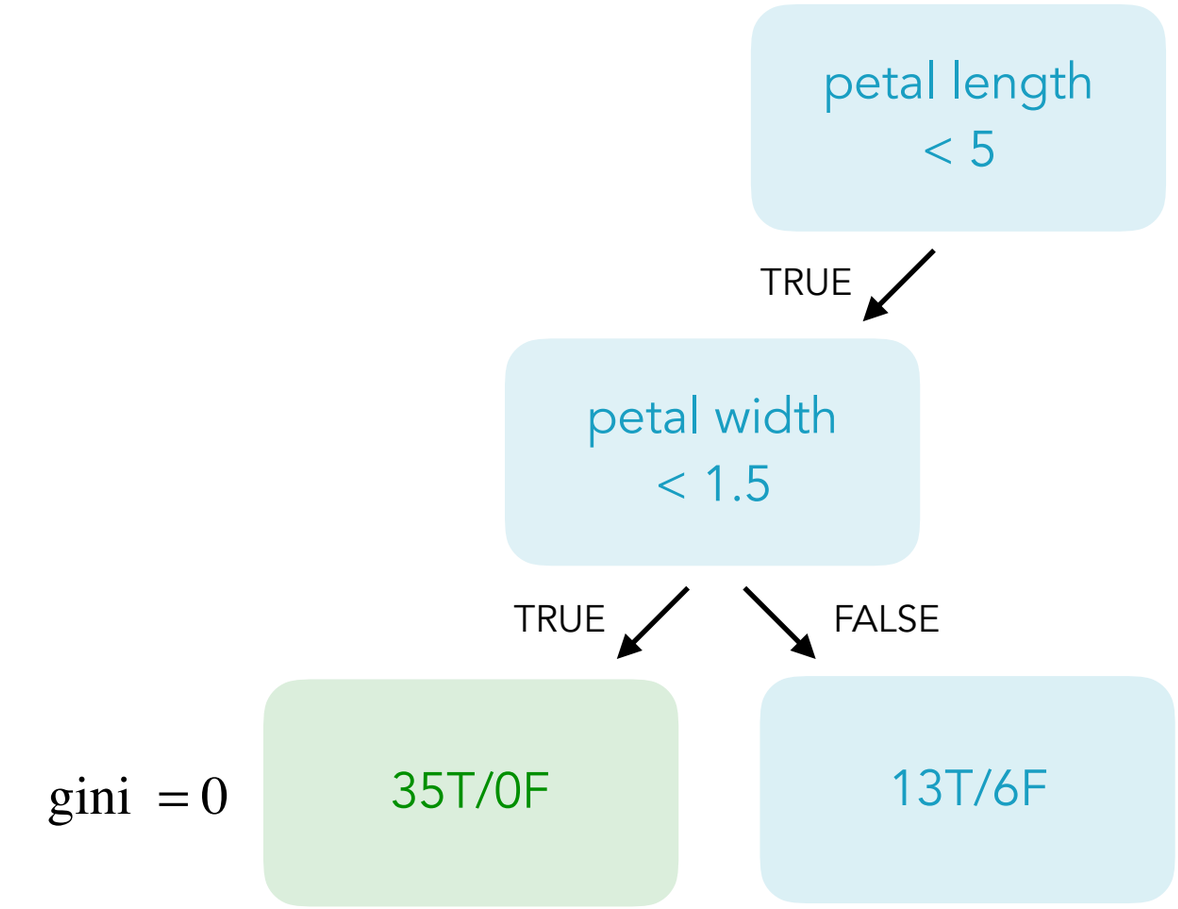
$\text{gini} = 0.197$

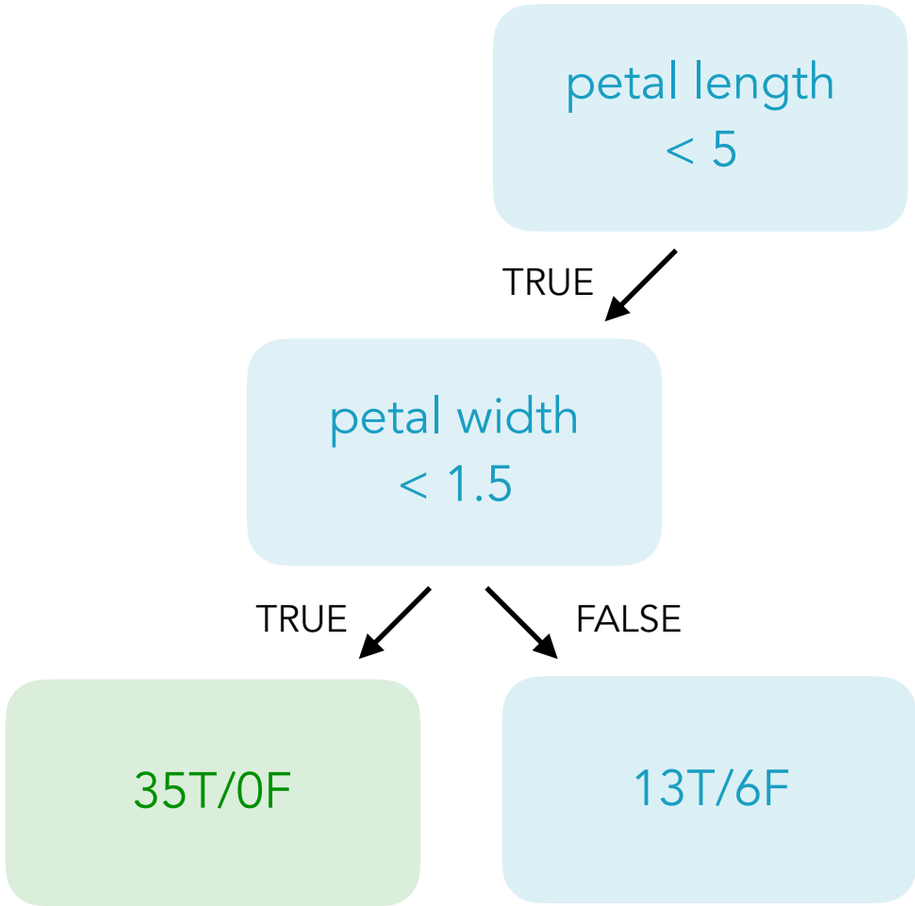


$\text{gini} = 0.104$


petal length
< 5
TRUE
FALSE
petal width
< 1.5
24T/43F

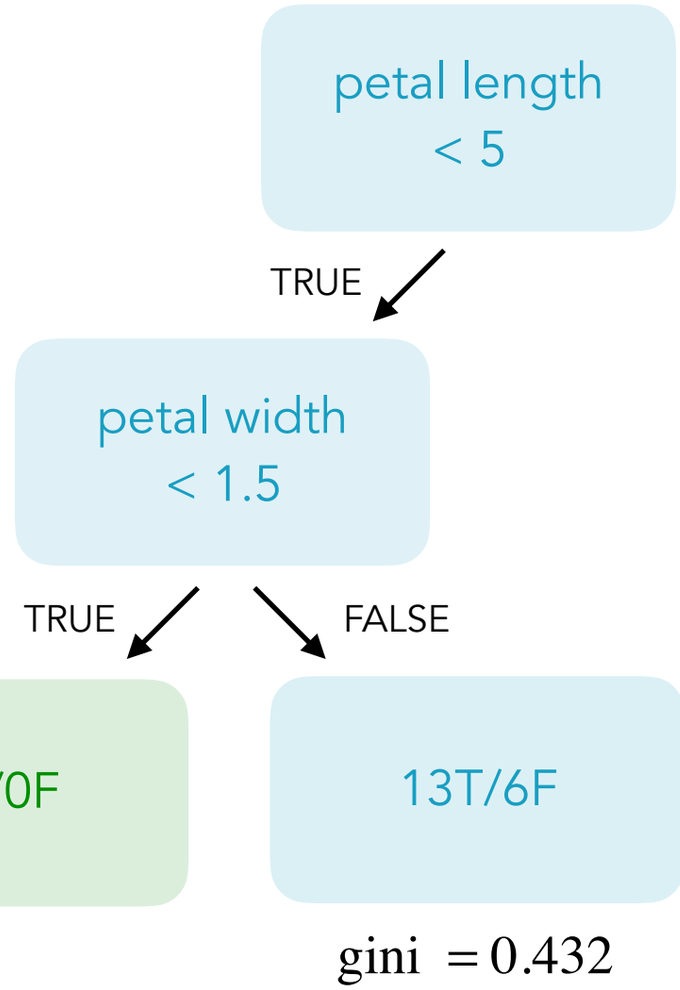

petal length
< 5
TRUE
petal width
< 1.5
TRUE
FALSE
gini = 0
35T/0F
13T/6F

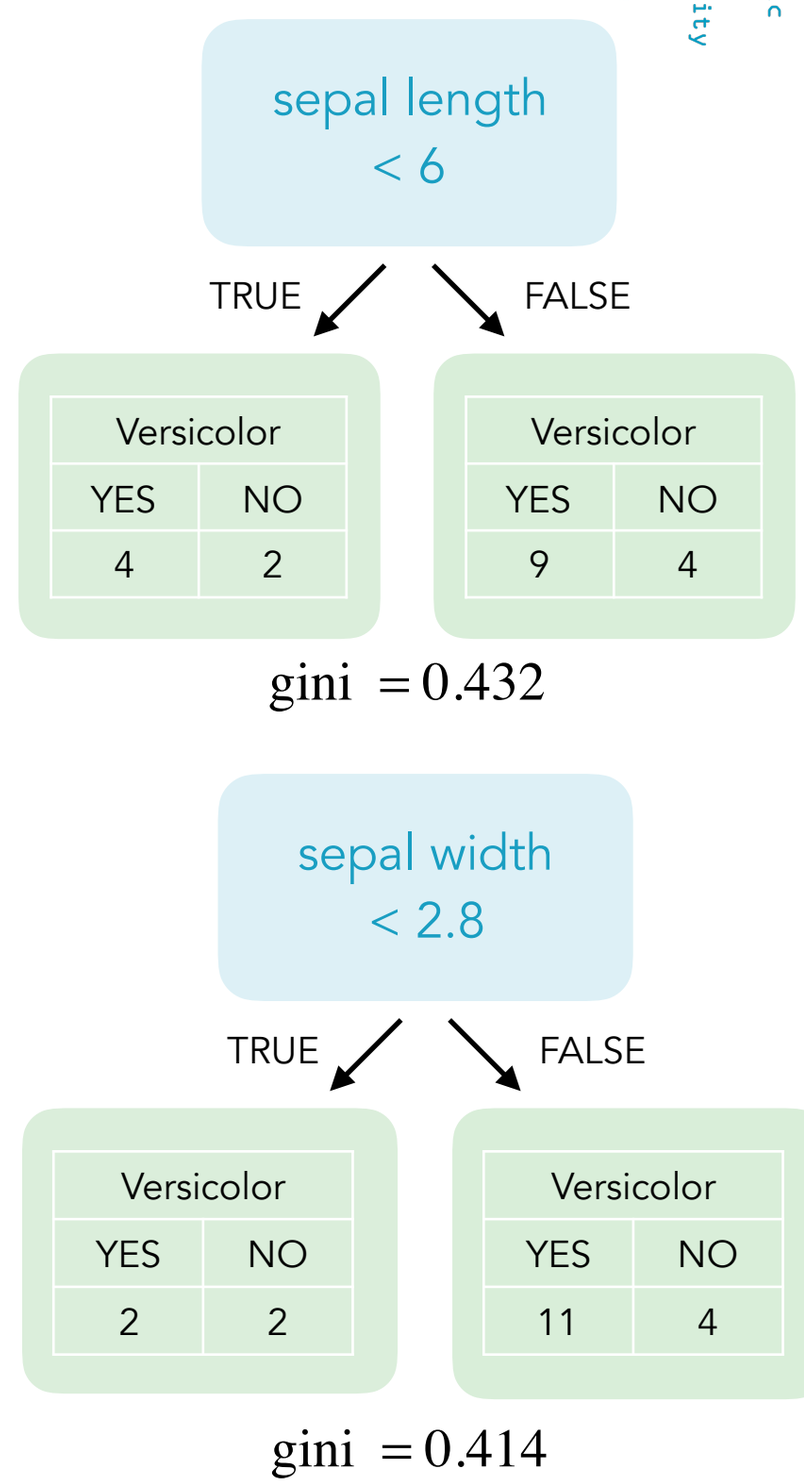

petal length
< 5
TRUE
petal width
< 1.5
TRUE
FALSE
gini = 0
35T/0F
13T/6F

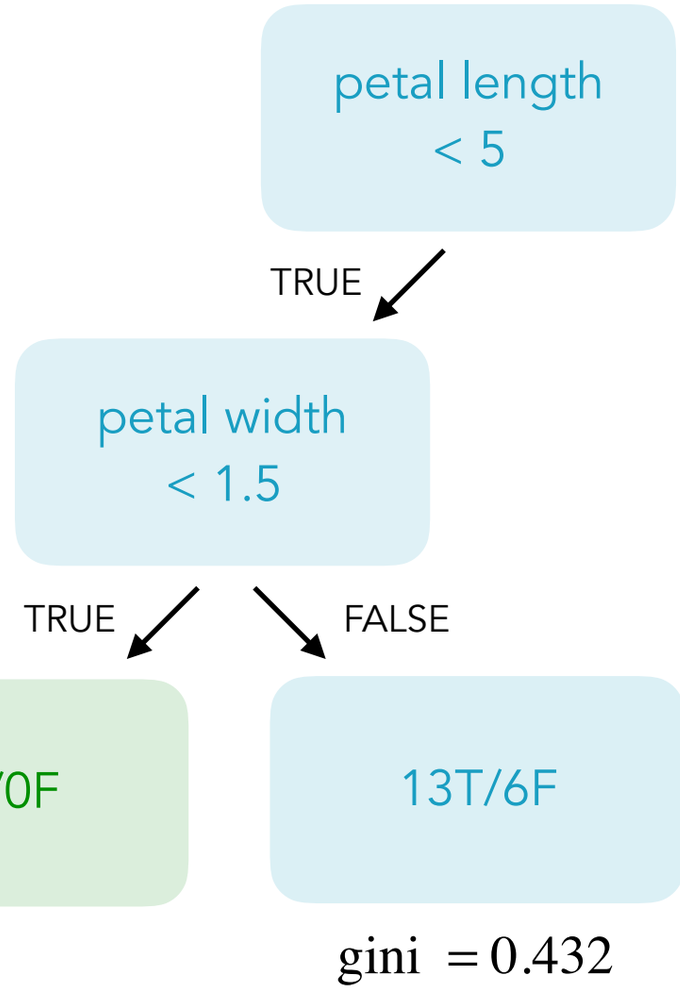


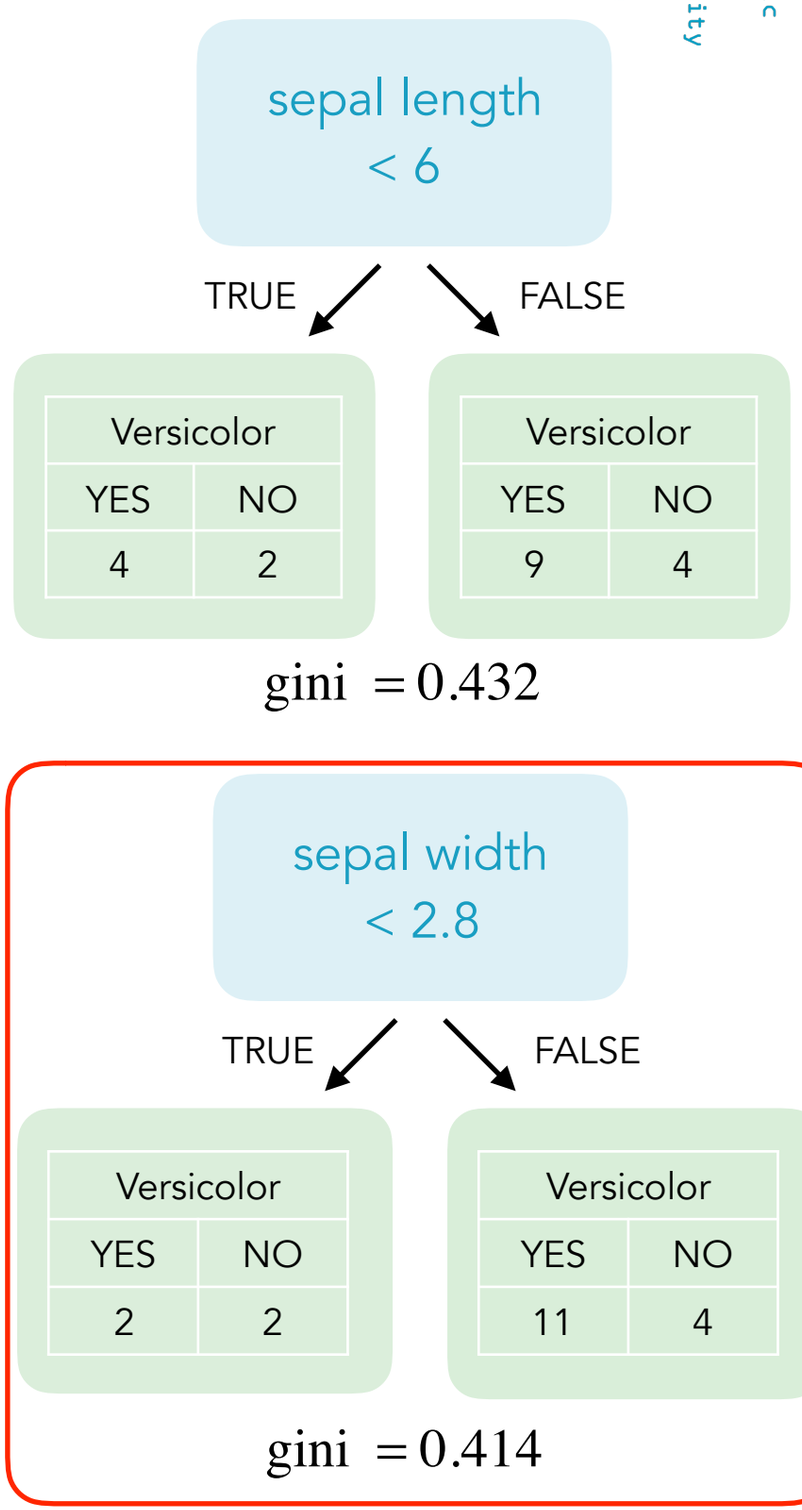
$$\text{gini} = 1 - \left(\frac{13}{13+6}\right)^2 - \left(\frac{6}{13+6}\right)^2 = 0.432$$

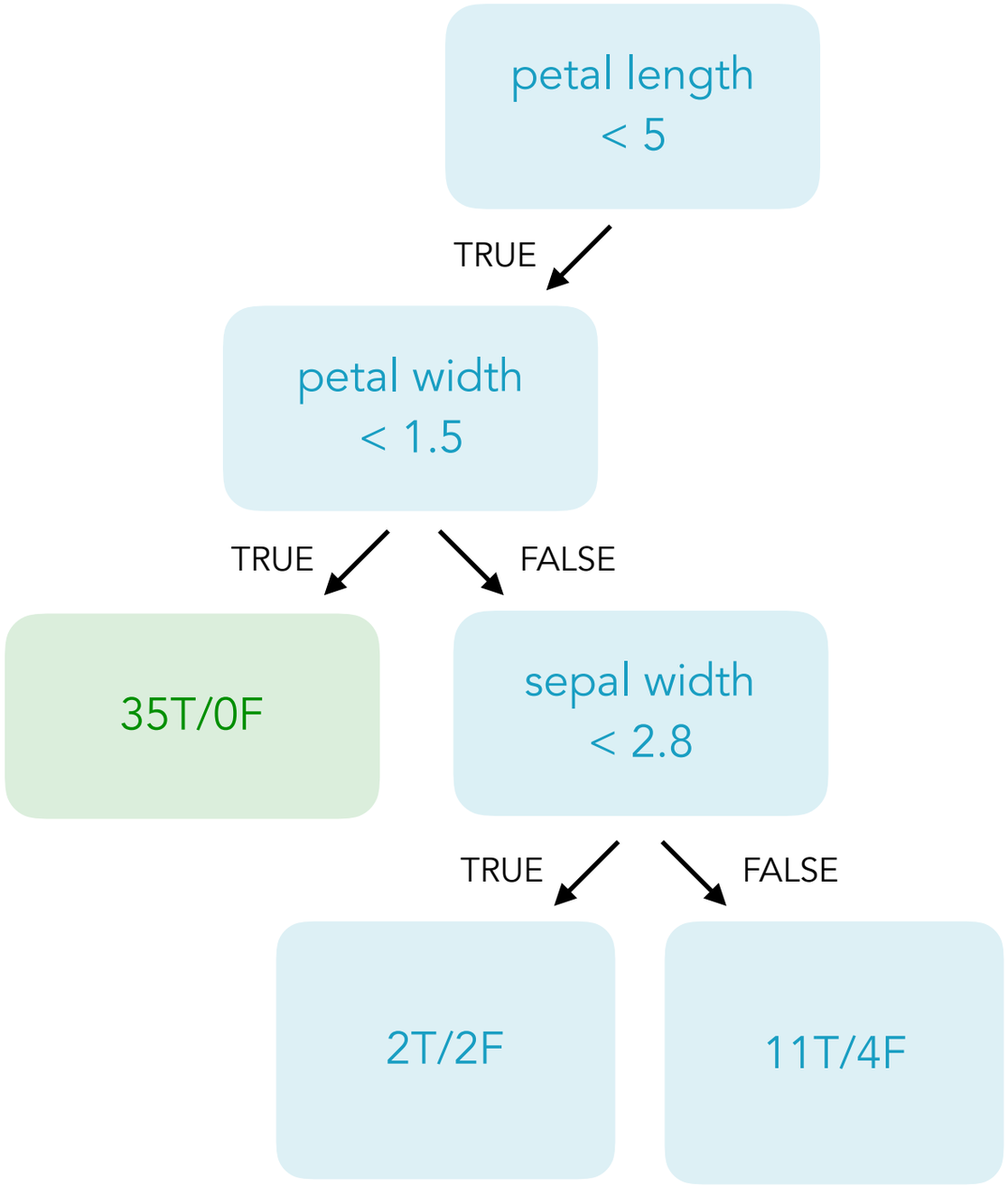
GDC Genetic Diversity Centre Zurich

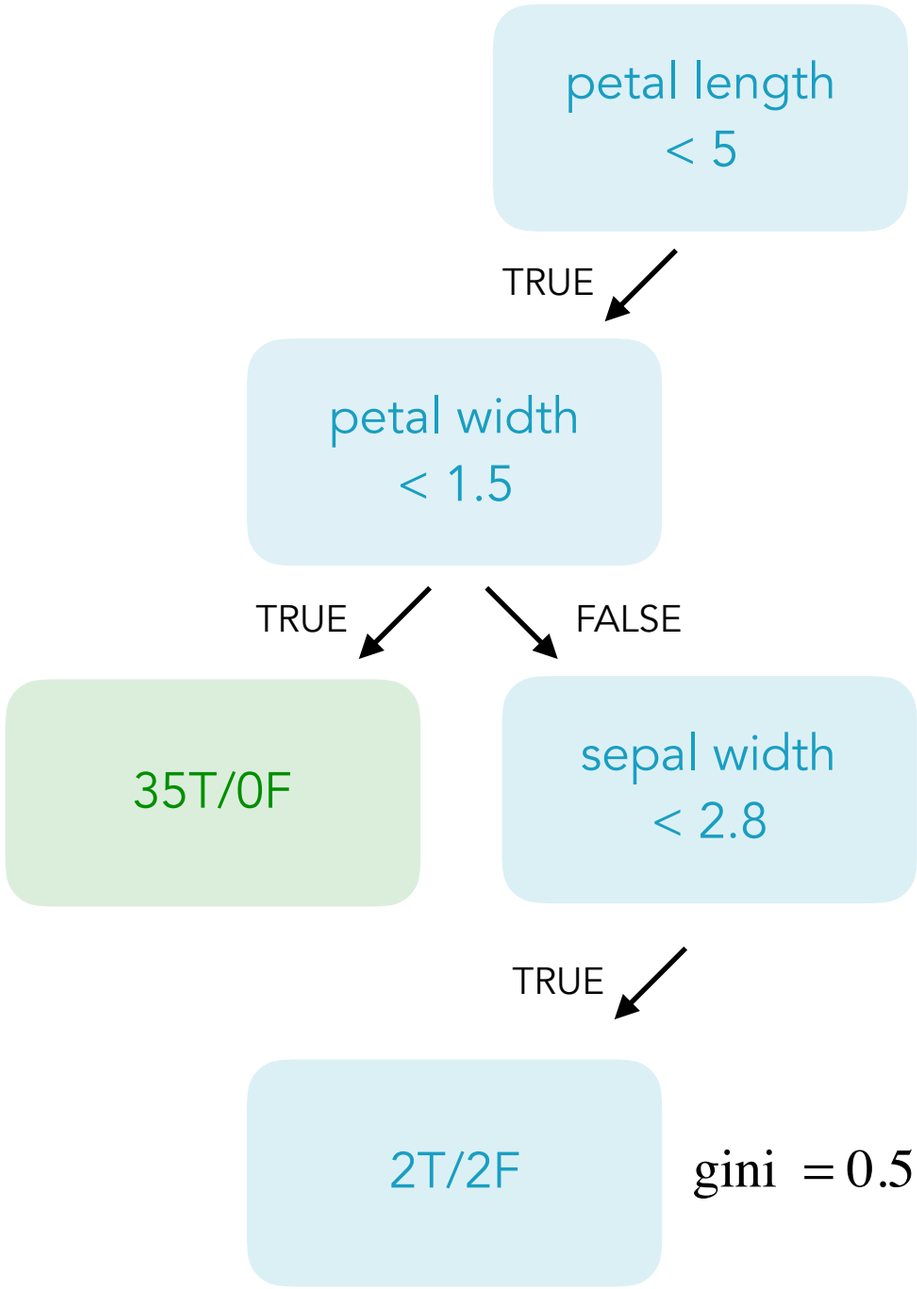


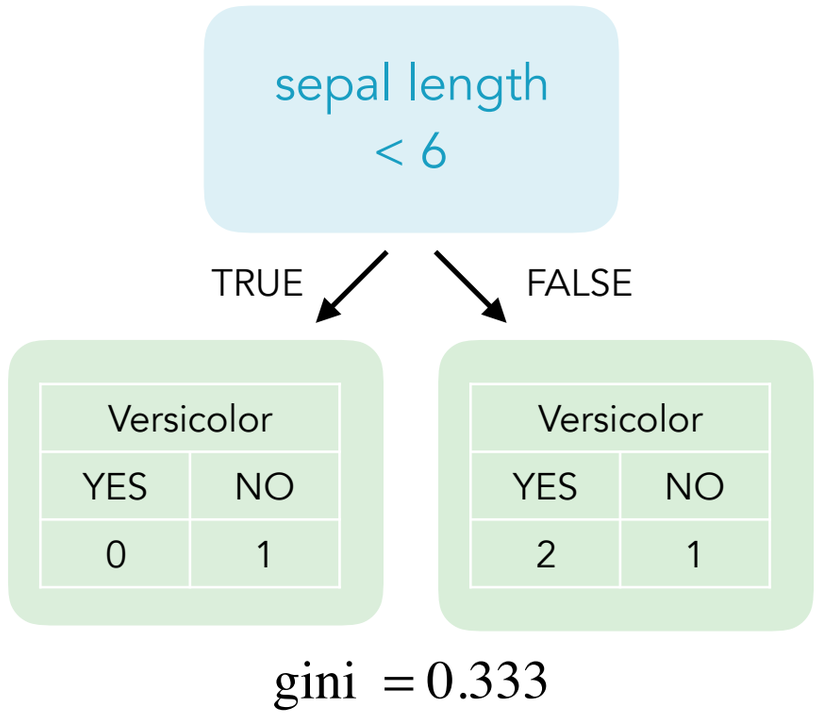


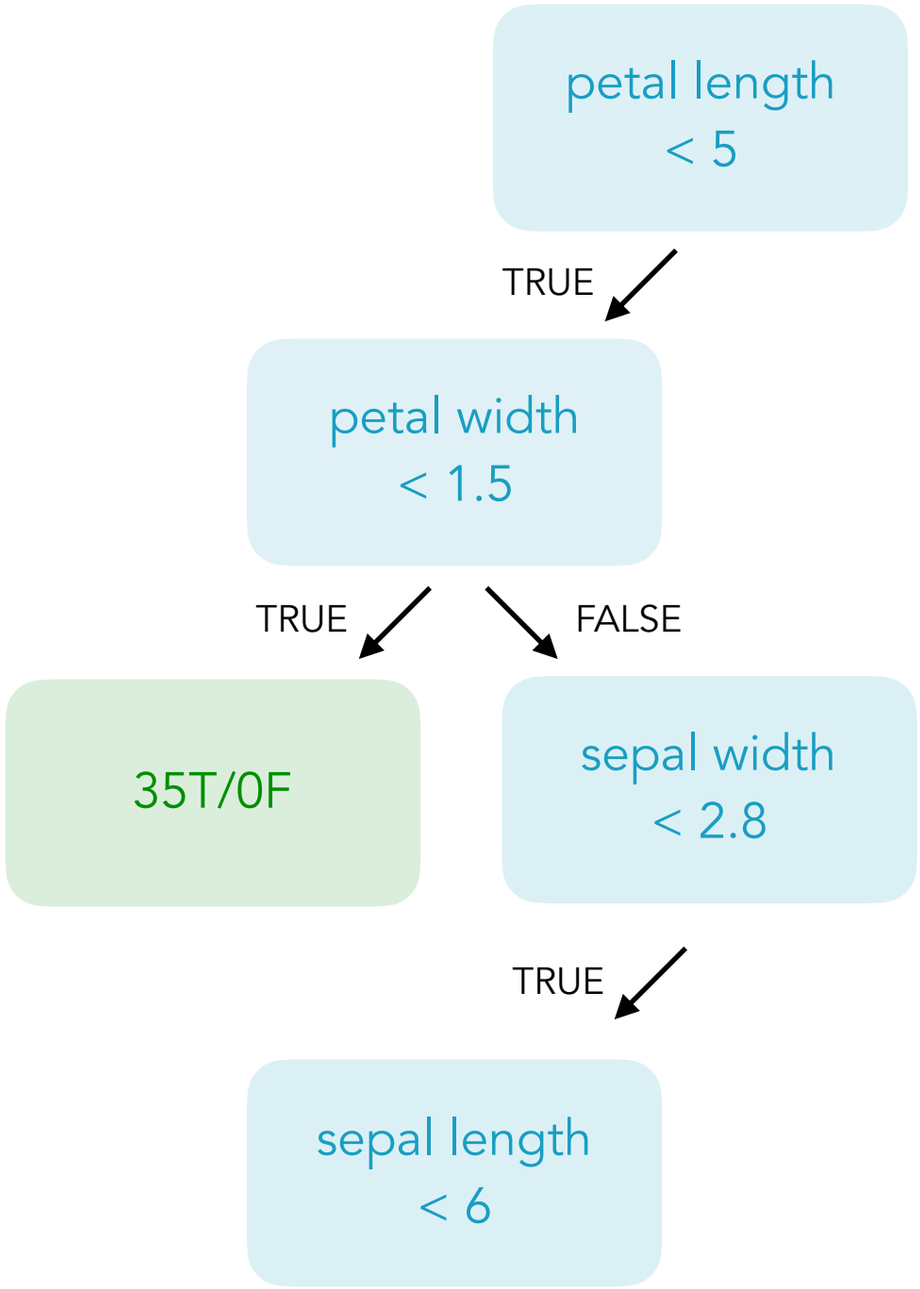

petal length
< 5
TRUE
petal width
< 1.5
TRUE
FALSE
35T/0F
13T/6F
gini = 0.432

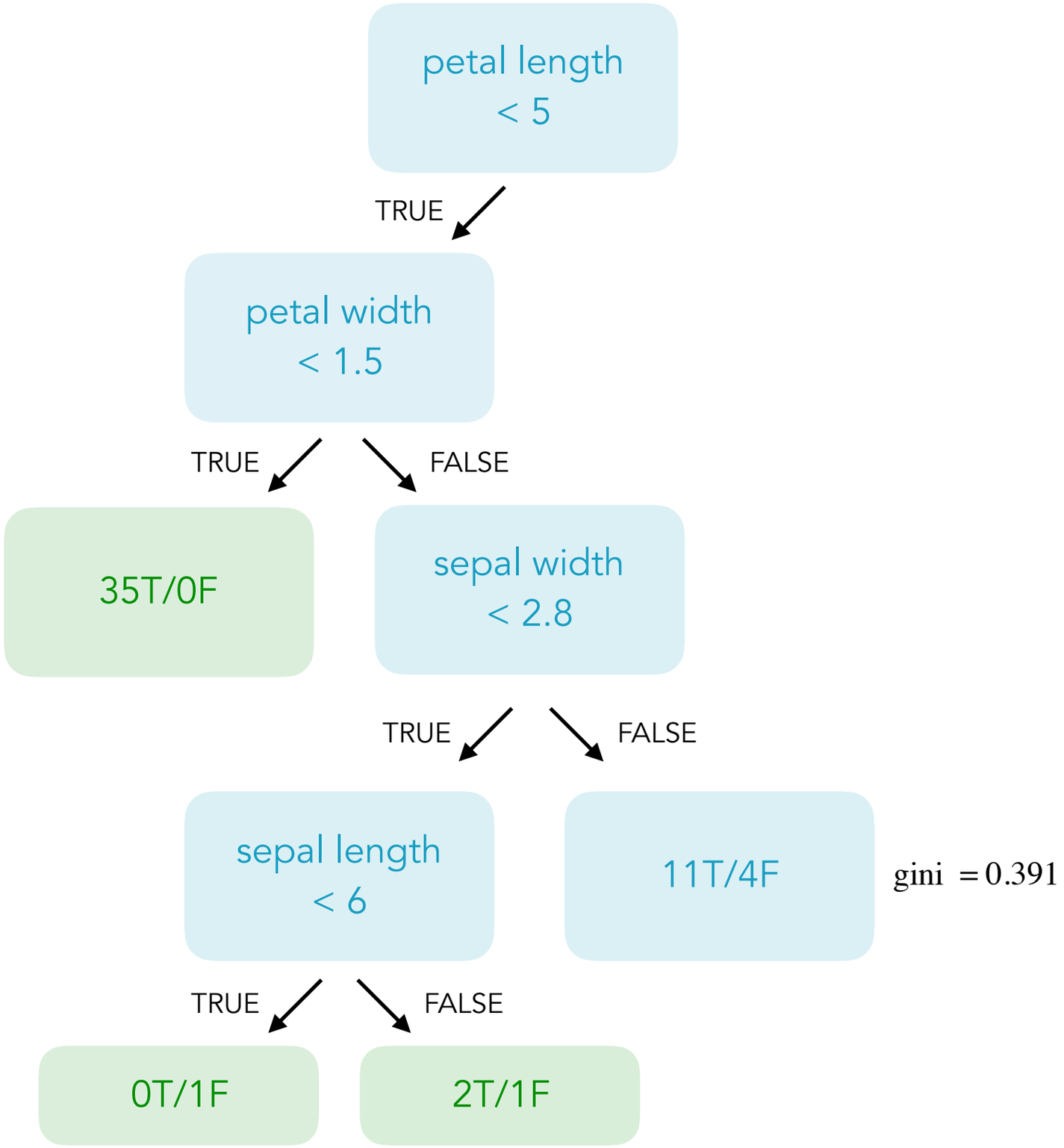

sepal length
< 6
TRUE
FALSE
Versicolor
YES
NO
4
2
Versicolor
YES
NO
9
4
gini = 0.432
sepal width
< 2.8
TRUE
FALSE
Versicolor
YES
NO
2
2
Versicolor
YES
NO
11
4
gini = 0.414


petal length
< 5
TRUE
petal width
< 1.5
TRUE
FALSE
35T/0F
sepal width
< 2.8
TRUE
FALSE
2T/2F
11T/4F


petal length
< 5
TRUE
petal width
< 1.5
TRUE
FALSE
35T/0F
sepal width
< 2.8
TRUE
2T/2F
gini = 0.5


sepal length
< 6
TRUE
FALSE
Versicolor
YES
NO
0
1
Versicolor
YES
NO
2
1
gini = 0.333


petal length
< 5
TRUE
petal width
< 1.5
TRUE
FALSE
35T/0F
sepal width
< 2.8
TRUE
sepal length
< 6



$$\text{gini} = 0.391$$

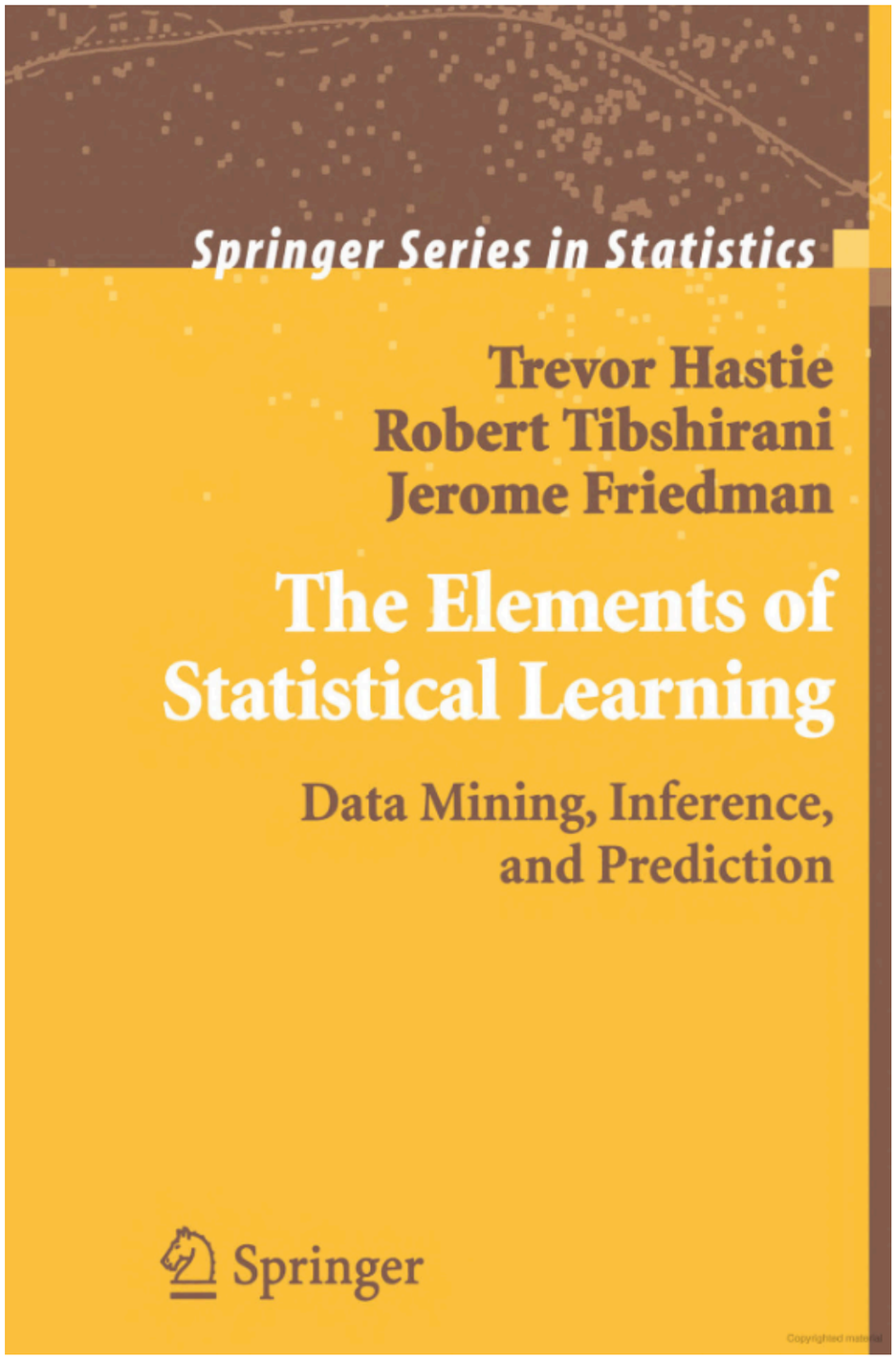
sort ↓

	Sepal.Length	Sepal.Width	Species
1	4.3	3.0	setosa
2	4.4	2.9	setosa
3	4.4	3.0	setosa
4	4.4	3.2	setosa
5	4.5	2.3	setosa
6	4.6	3.1	setosa
7	4.6	3.4	setosa
8	4.6	3.6	setosa
9	4.6	3.2	setosa
10	4.7	3.2	setosa
11	4.7	3.2	setosa
12	4.8	3.4	setosa
13	4.8	3.0	setosa
14	4.8	3.4	setosa
15	4.8	3.1	setosa
16	4.8	3.0	setosa
17	4.9	3.0	setosa
18	4.9	3.1	setosa
…	…	…	…

gini impurity for $4.35 = ?$

gini impurity for $4.4 = ?$

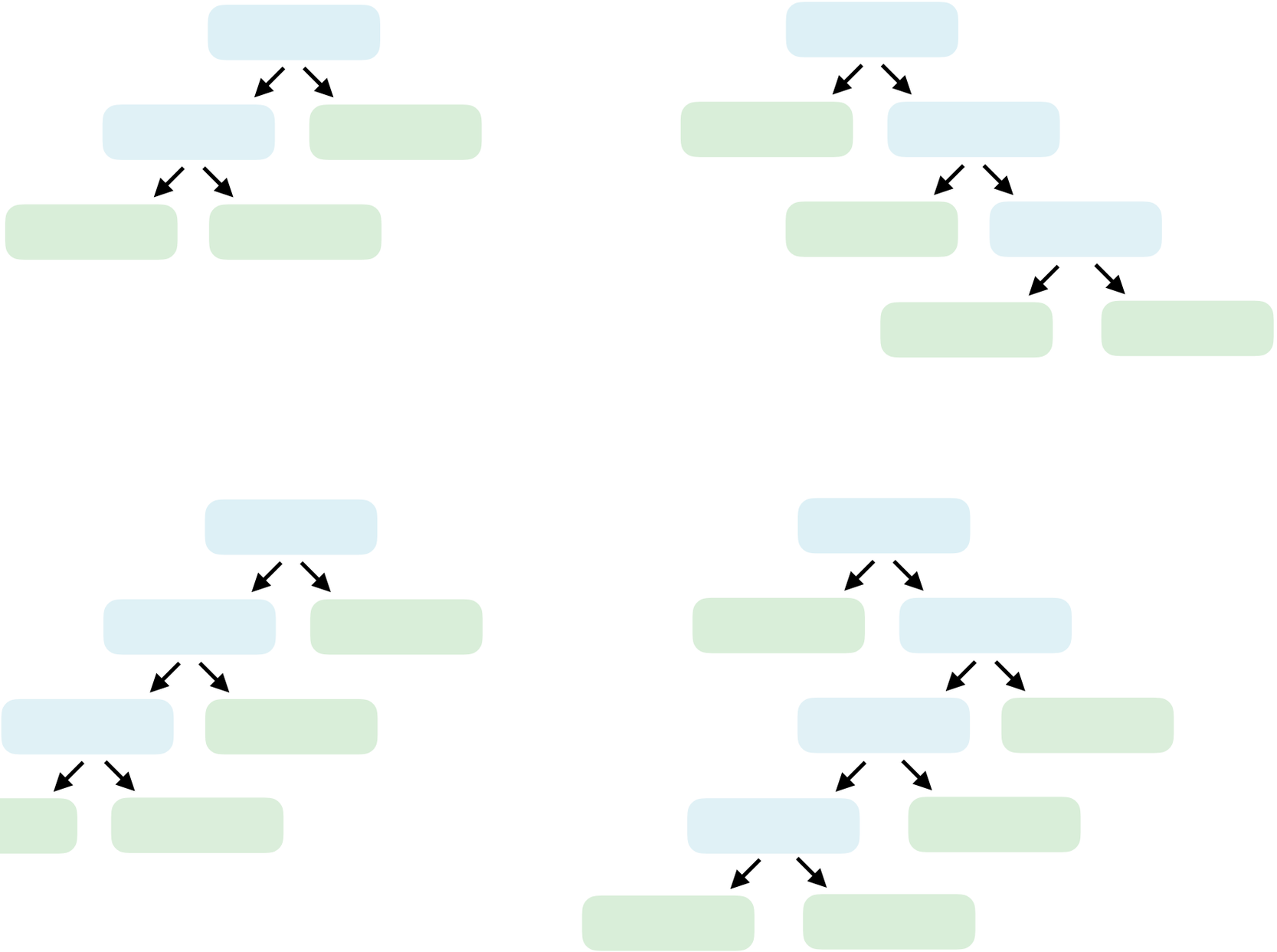
Calculate impurity values for each mean sepal length and chose the mean with the lowest Gini value as threshold.



Trees have one aspect that prevents them from being the ideal tool for predictive lerning, namely **inaccuracy**. They seldom provide predictive accuracy comparable to the best that can be achieved with the data a hand.

➜ Boosting or bagging decision trees can improves accuracy.

Random Forests



Original Dataset (N=10)

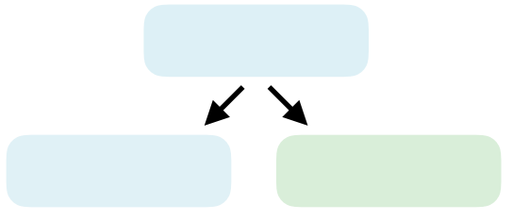
Sepal Length	Sepal Width	Petal Length	Petal Width	Species
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
6.5	3.0	5.8	2.2	virginica

1. Create a bootstrap dataset

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
6.9	3.1	4.9	1.5	versicolor
6.5	2.8	4.6	1.5	versicolor
7.1	3.0	5.9	2.1	virginica
7.1	3.0	5.9	2.1	virginica
6.9	3.1	4.9	1.5	versicolor
6.5	2.8	4.6	1.5	versicolor
5.8	2.7	5.1	1.9	virginica
6.4	3.2	4.5	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.3	2.9	5.6	1.8	virginica

GDC Genetic Diversity Centre Zurich

2. Create a decision tree but only use a random subset of variables (mtry) at each node.



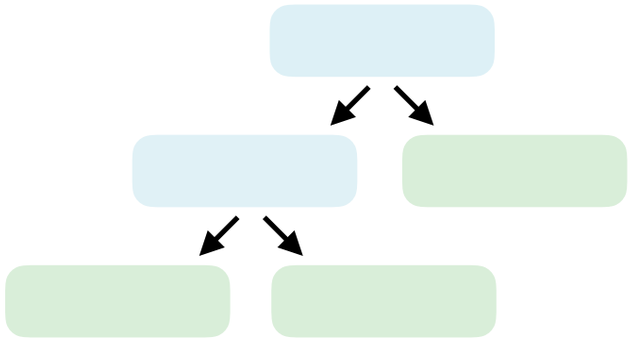
1. Create a bootstrap dataset

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
6.9	3.1	4.9	1.5	versicolor
6.5	2.8	4.6	1.5	versicolor
7.1	3.0	5.9	2.1	virginica
7.1	3.0	5.9	2.1	virginica
6.9	3.1	4.9	1.5	versicolor
6.5	2.8	4.6	1.5	versicolor
5.8	2.7	5.1	1.9	virginica
6.4	3.2	4.5	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.3	2.9	5.6	1.8	virginica

mtry=2

starting point: mtry = $\sqrt{N_{variables}}$

2. Create a decision tree but only use a random subset of variables (mtry) at each node.



1. Create a bootstrap dataset

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
6.9	3.1	4.9	1.5	versicolor
6.5	2.8	4.6	1.5	versicolor
7.1	3.0	5.9	2.1	virginica
7.1	3.0	5.9	2.1	virginica
6.9	3.1	4.9	1.5	versicolor
6.5	2.8	4.6	1.5	versicolor
5.8	2.7	5.1	1.9	virginica
6.4	3.2	4.5	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.3	2.9	5.6	1.8	virginica

1. Create a bootstrap dataset

2. Create a decision tree but only use a random subset of variables (mtry) at each node.

3. Repeat (ntree)

New Sample

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.70	2.15	3.98	1.31	???

Versicolor	
YES	NO
987	13

Bagging

Bootstrapping means taking a sample of a population by drawing with replacement. It is one of the main ideas behind Bagging (which stands for Bootstrap AGGregatING).

Estimate accuracy of the random forest

Original Dataset (N=10)

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
6.5	3.0	5.8	2.2	virginica

1. Create a bootstrap dataset

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
6.9	3.1	4.9	1.5	versicolor
6.5	2.8	4.6	1.5	versicolor
7.1	3.0	5.9	2.1	virginica
7.1	3.0	5.9	2.1	virginica
6.9	3.1	4.9	1.5	versicolor
6.5	2.8	4.6	1.5	versicolor
5.8	2.7	5.1	1.9	virginica
6.4	3.2	4.5	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.3	2.9	5.6	1.8	virginica

About 1/3 of the original data does is not being used in the bootstrap dataset. **> Out-of-Bag Dataset**

Out-of-Bag (OOB) Dataset

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
7.0	3.2	4.7	1.4	versicolo
6.3	3.3	6.0	2.5	virginica
6.5	3.0	5.8	2.2	virginica

Estimate accuracy of the random forest

Versicolor	
YES	NO
601	399

Versicolor	
YES	NO
644	356

Versicolor	
YES	NO
468	532

OOB Error Rate ▸ mtry

Random Forests with R

```
library(randomForest); packageVersion("randomForest")
# 4.6.14
set.seed(190717)
new.iris.rf <- randomForest(Species ~ ., data = new.iris,
                            mtry = 2,
                            ntree = 1000)
print(new.iris.rf)
```

```
Call:
 randomForest(formula = Species ~ ., data = new.iris, mtry = 2, ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 7%
Confusion matrix:
           versicolor virginica class.error
versicolor         47         3        0.06
virginica           4        46        0.08
```

Predict Outcome

```
new.iris[c(1,100),]
#     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
# 1            7.0         3.2          4.7         1.4 versicolor
# 100          5.9         3.0          5.1         1.8  virginica

new.data      <- data.frame(new.iris[c(1,100),])
new.data.pred <- predict(new.iris.rf, new.data)
table(observed = new.data$Species, predicted = new.data.pred)
#            predicted
# observed   versicolor virginica
# versicolor          1         0
# virginica           0         1
```

Random Forests with Phyloseq Objects

```
predictors  <- t(otu_table(phyloseq.data))
response    <- as.factor(sample_data(phyloseq.data)$?)
phyloseq.df <- data.frame(response, predictors)
phyloseq.rf <- randomForest(response~., data = phyloseq.df, ntree = 100)
print(phyloseq.rf)
```