Mohamed Noordeen Alaudeen

Random Forest

Ensemble

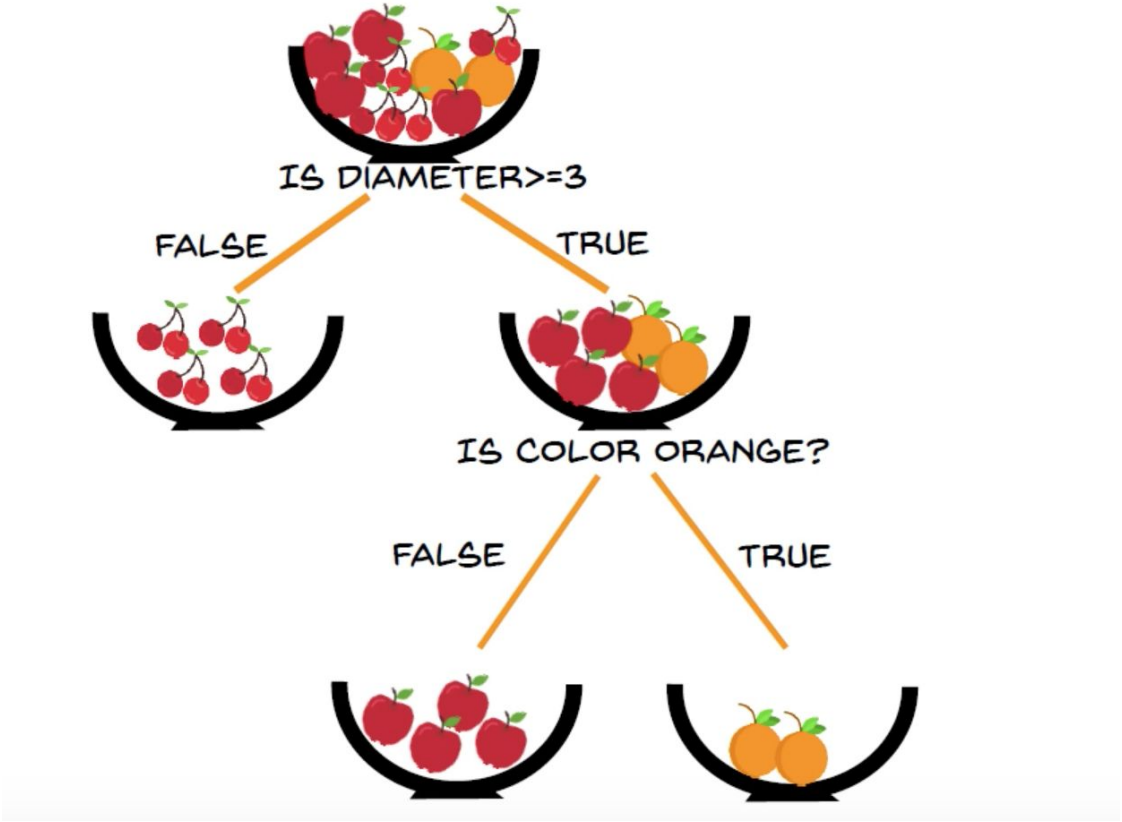
Ensemble

Machine learning paradigm which combine weak learners to become a strong learner

Decision Tree Summary

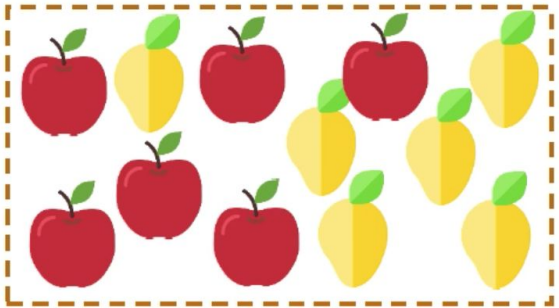
- Rule based system
- Set of rules
- Calculating the nodes
- Information gain and Gini index

Decision Tree is a tree shaped diagram used to determine a course of action. Each branch of the tree represents a possible decision, occurrence or reaction



Entropy

Entropy is the measure of randomness or unpredictability in the dataset

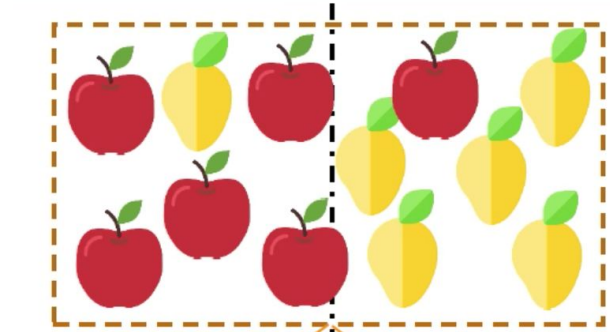


High entropy

E1

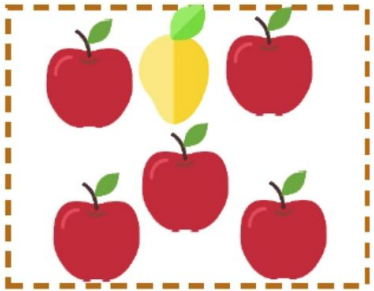
Entropy

Entropy is the measure of randomness or unpredictability in the dataset

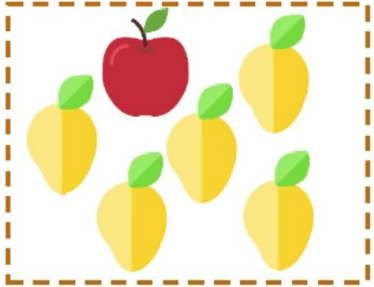


Initial Dataset

Decision Split



Set 1



Set 2

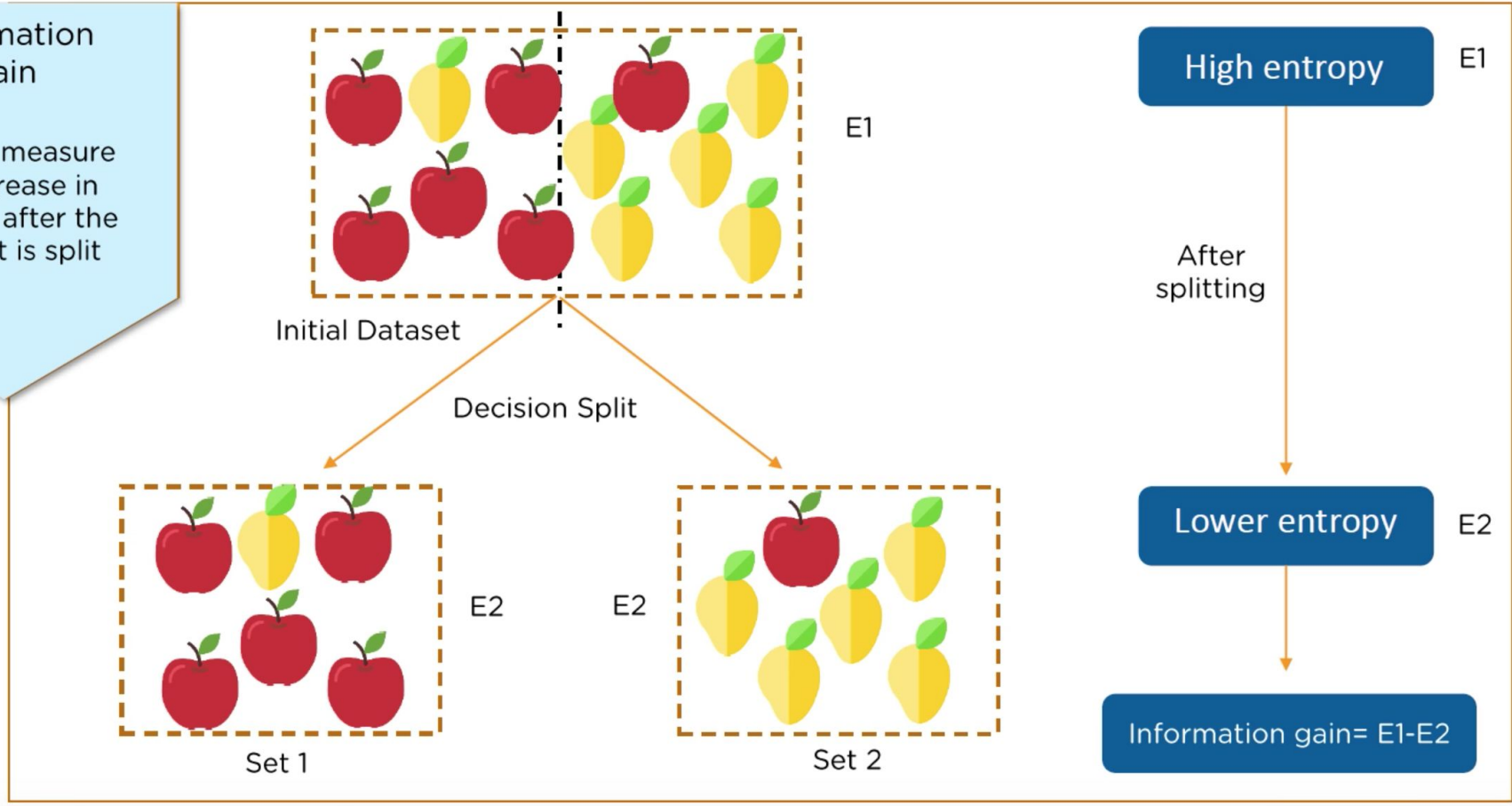
High entropy E1

After Splitting

Lower entropy E2

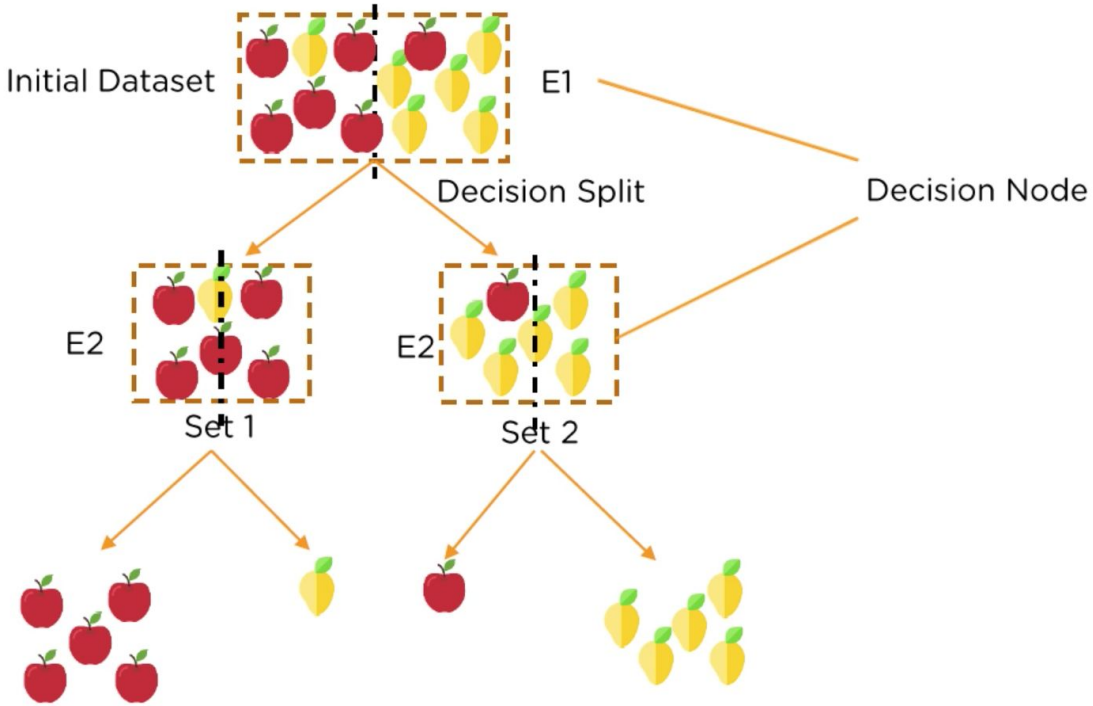
Information gain

It is the measure of decrease in entropy after the dataset is split



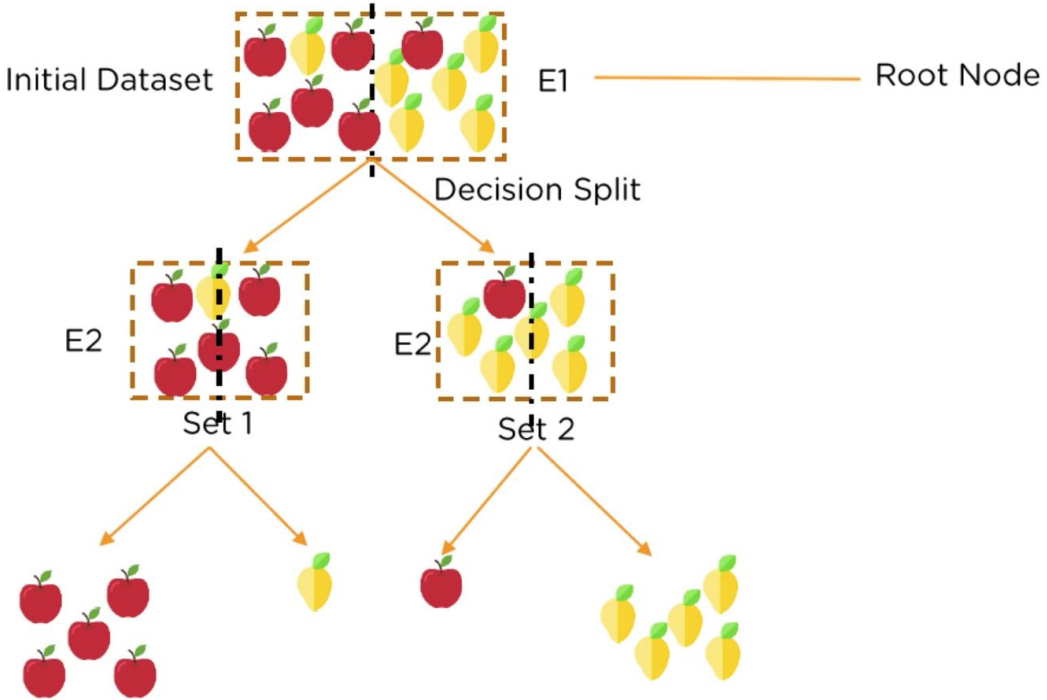
Decision Node

Decision node has two or more branches



Root Node

The top most Decision node is known as the Root node



PROBLEM STATEMENT

TO CLASSIFY THE DIFFERENT TYPES OF FRUITS IN THE BOWL BASED ON DIFFERENT FEATURES

THE DATASET(BOWL) IS LOOKING QUITE MESSY AND THE ENTROPY IS HIGH IN THIS CASE

TRAINING DATASET

COLOR	DIAMETER	LABEL
RED	3	APPLE
YELLOW	3	LEMON
PURPLE	1	GRAPES
RED	3	APPLE
YELLOW	3	LEMON
PURPLE	1	GRAPES

HOW TO SPLIT THE DATA

WE HAVE TO FRAME THE CONDITIONS THAT SPLIT THE DATA IN SUCH A WAY THAT THE INFORMATION GAIN IS THE HIGHEST

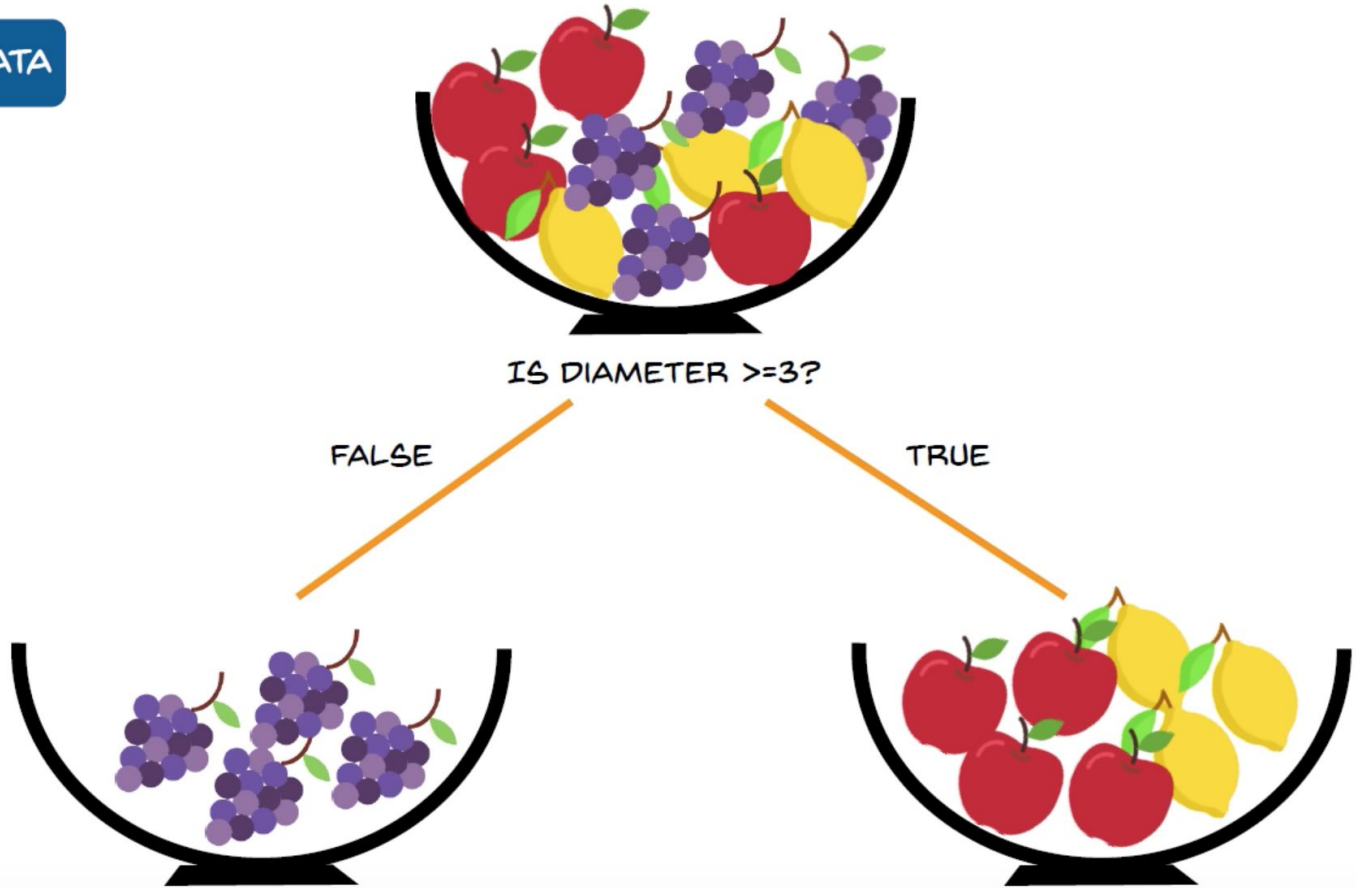
HOW TO SPLIT THE DATA

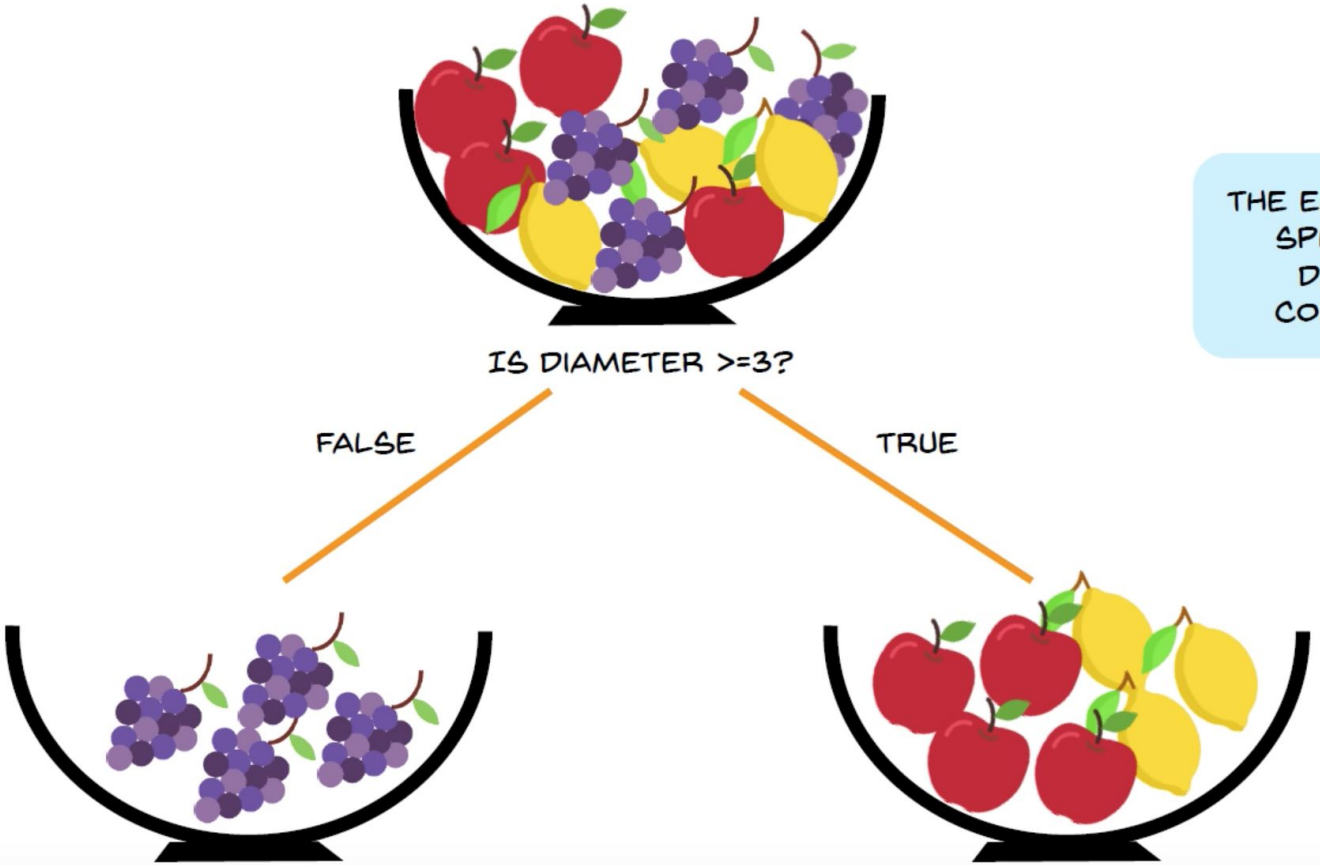
WE HAVE TO FRAME THE CONDITIONS THAT SPLIT THE DATA IN SUCH A WAY THAT THE INFORMATION GAIN IS THE HIGHEST

NOTE

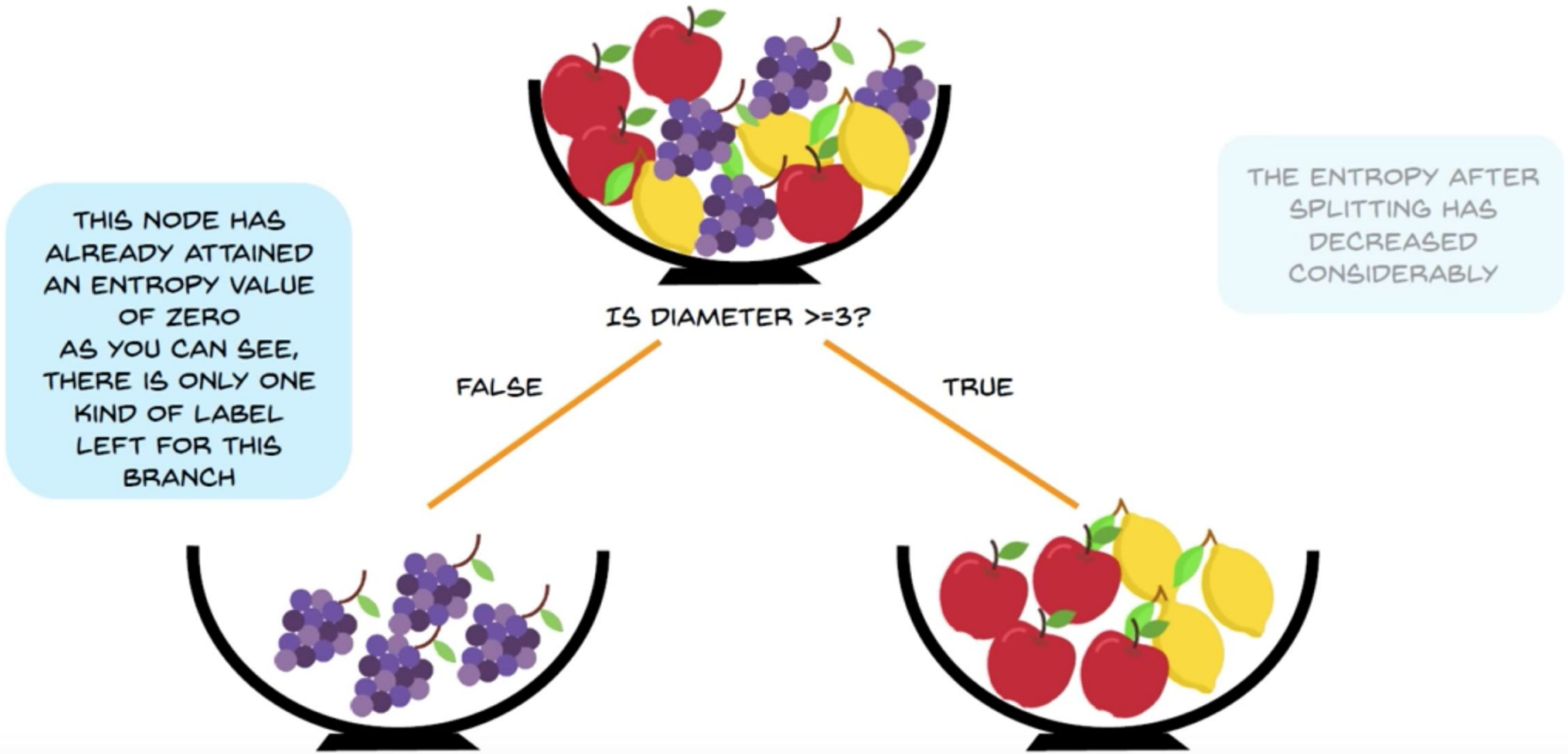
GAIN IS THE MEASURE OF DECREASE IN ENTROPY AFTER SPLITTING

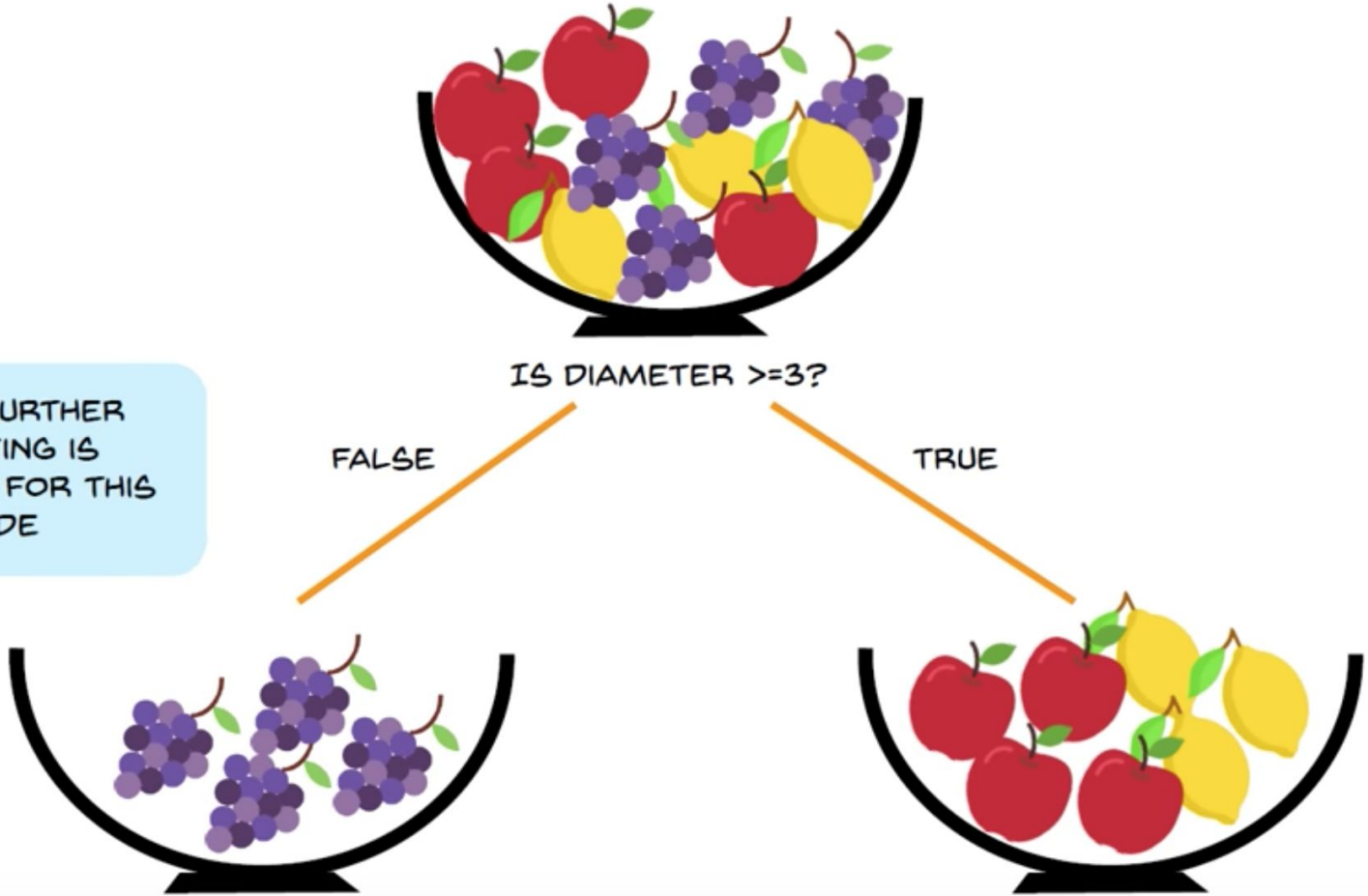
WE SPLIT THE DATA

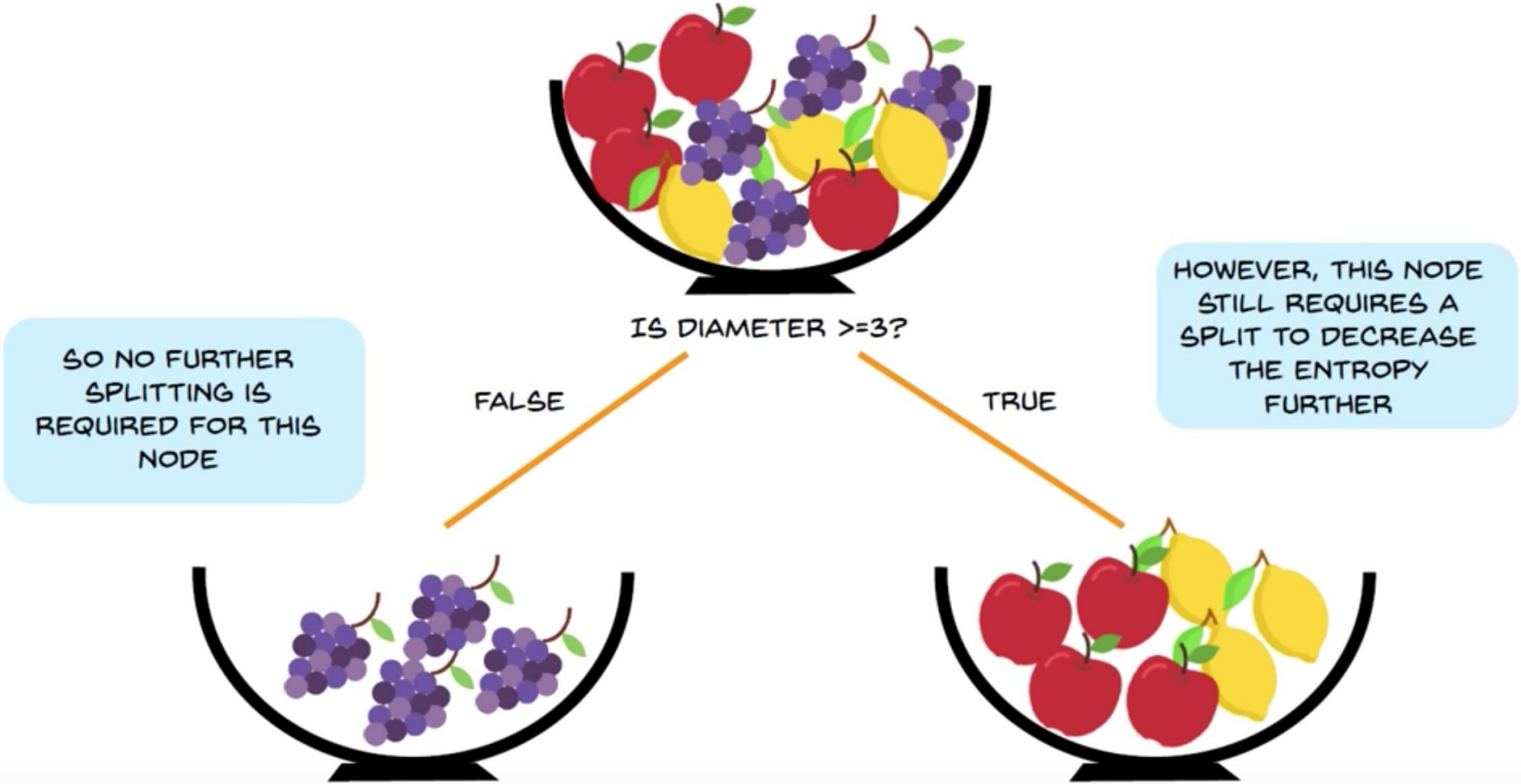


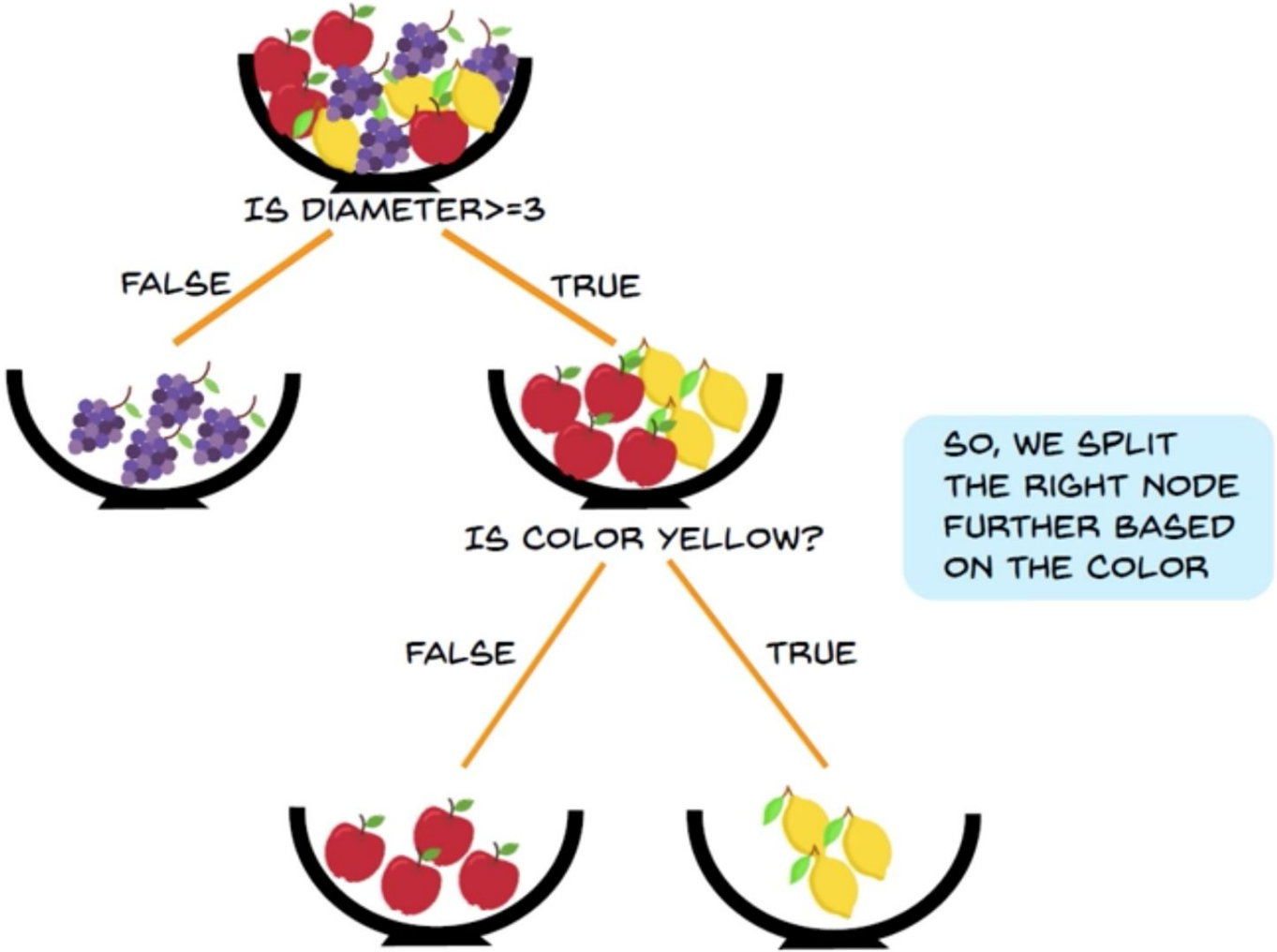


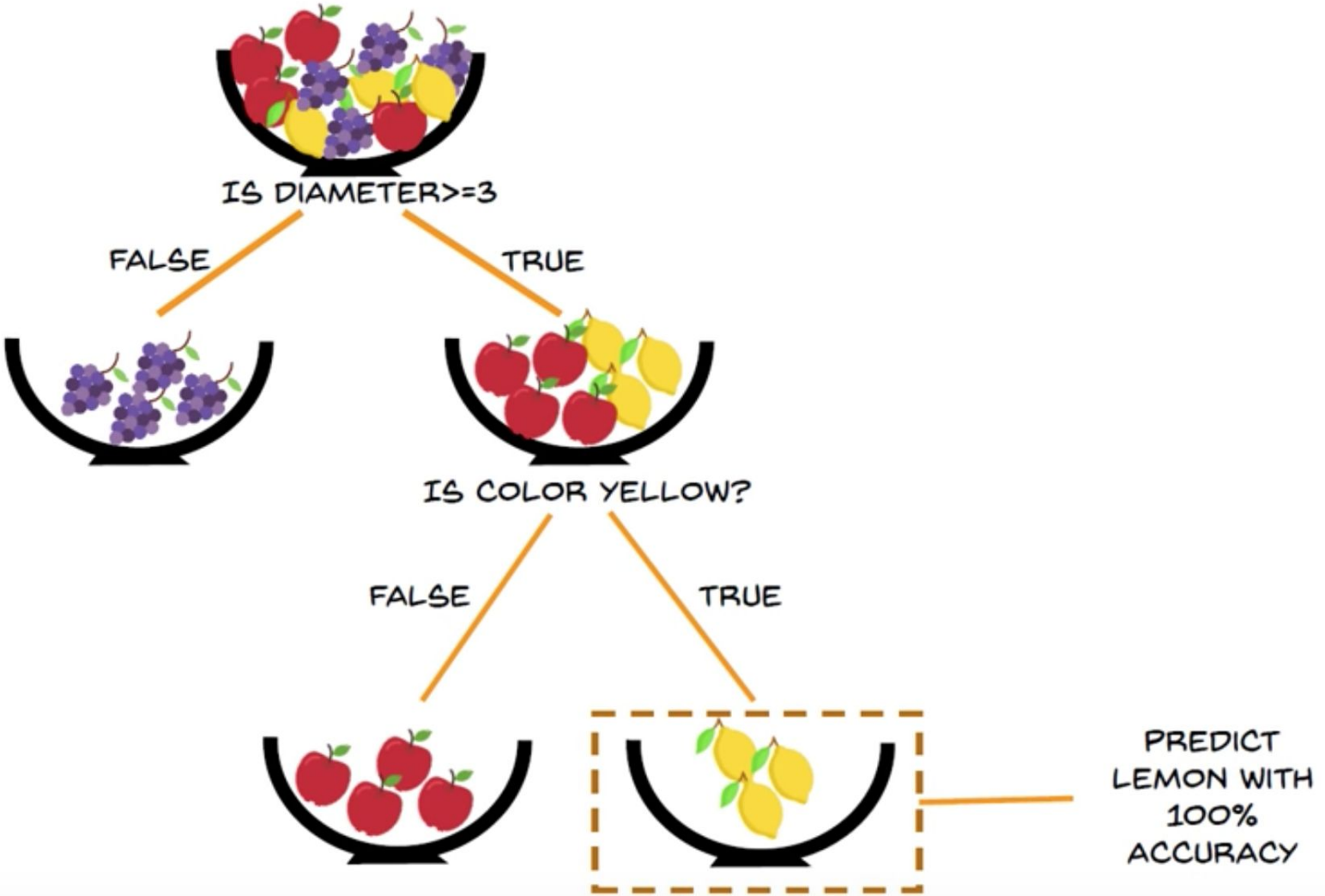
THE ENTROPY AFTER SPLITTING HAS DECREASED CONSIDERABLY

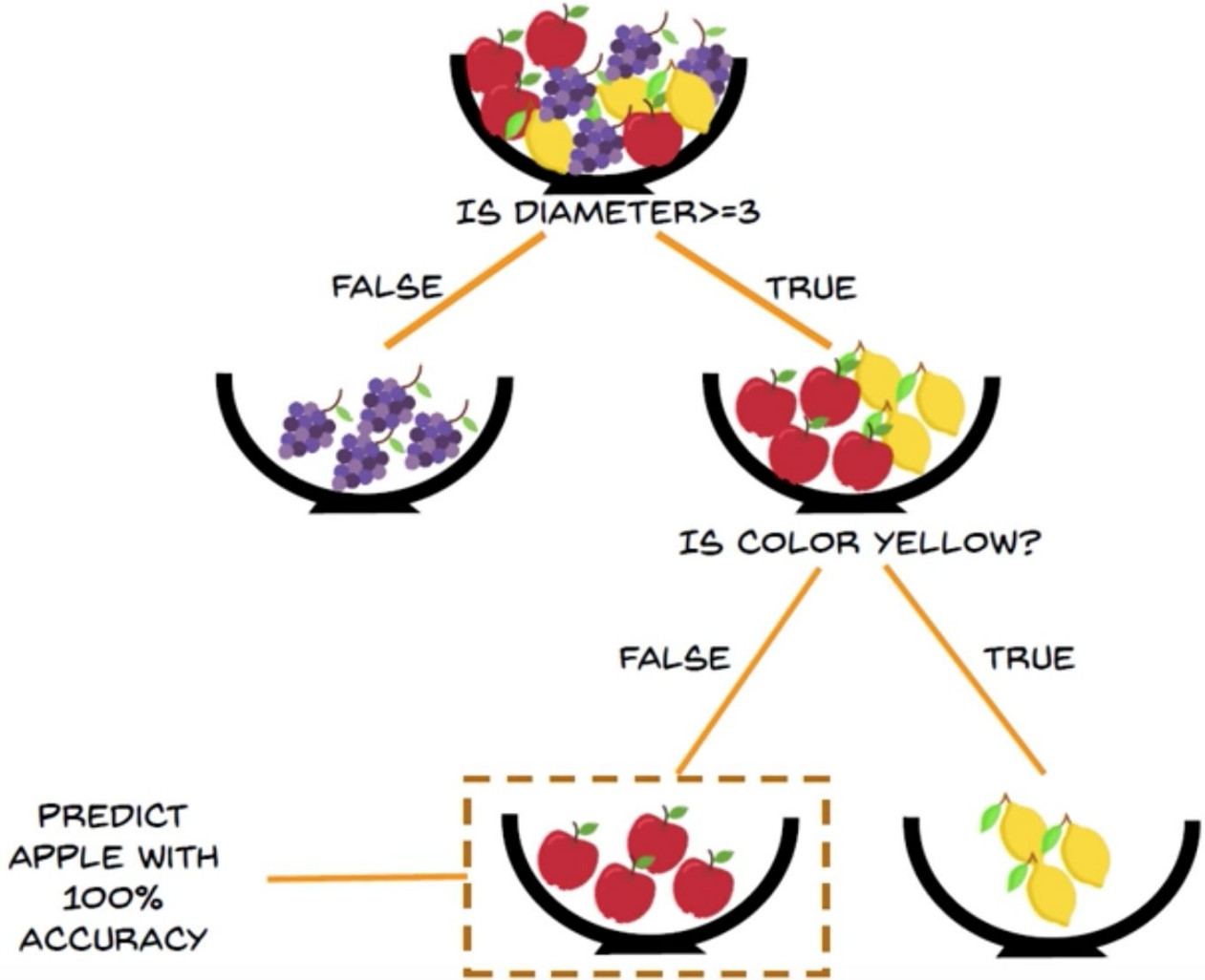

THIS NODE HAS ALREADY ATTAINED AN ENTROPY VALUE OF ZERO
AS YOU CAN SEE, THERE IS ONLY ONE KIND OF LABEL LEFT FOR THIS BRANCH
IS DIAMETER >=3?
FALSE
TRUE
THE ENTROPY AFTER SPLITTING HAS DECREASED CONSIDERABLY

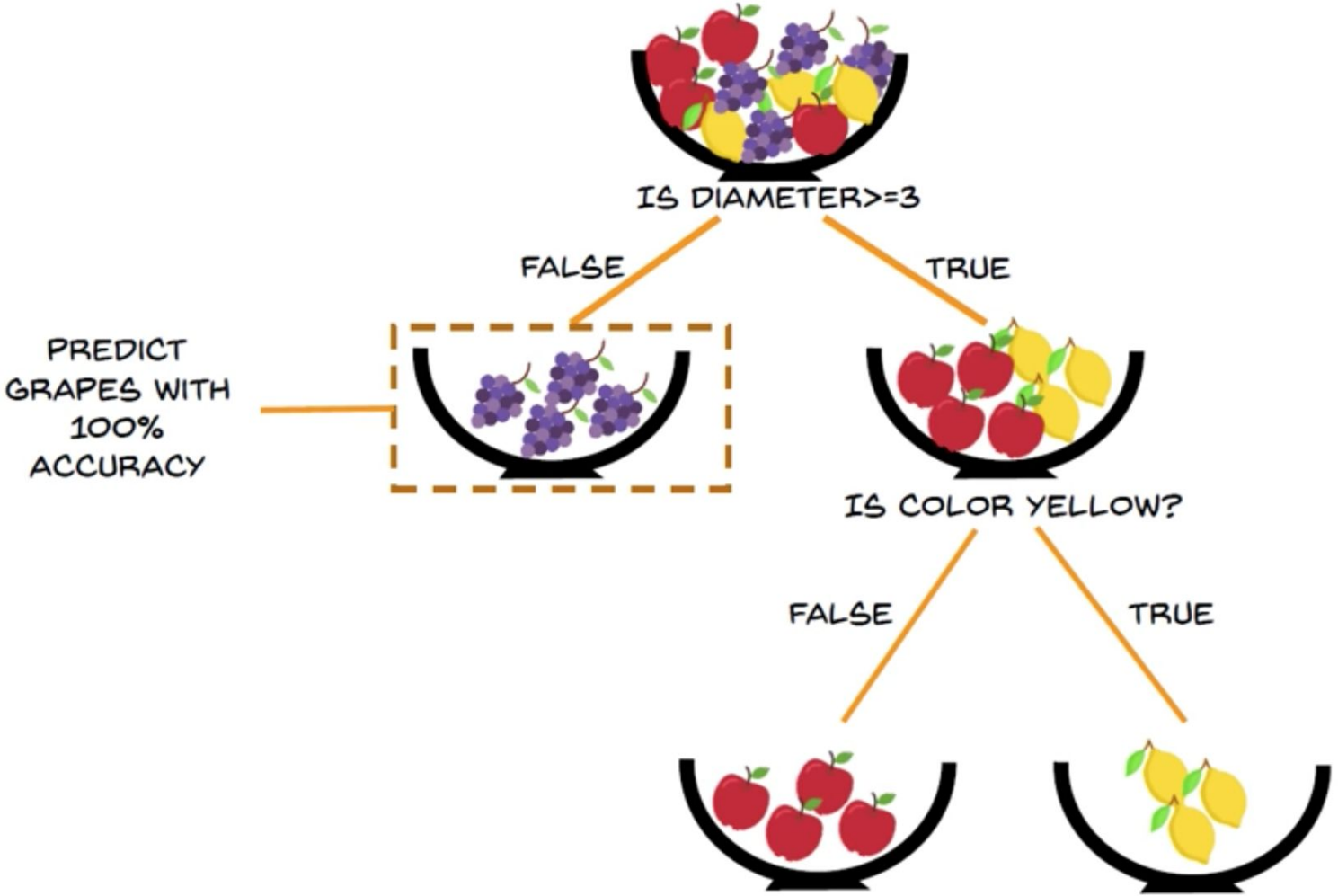

IS DIAMETER >=3?
FALSE
TRUE
SO NO FURTHER SPLITTING IS REQUIRED FOR THIS NODE


IS DIAMETER >=3?
FALSE
TRUE
SO NO FURTHER SPLITTING IS REQUIRED FOR THIS NODE
HOWEVER, THIS NODE STILL REQUIRES A SPLIT TO DECREASE THE ENTROPY FURTHER


IS DIAMETER>=3
FALSE
TRUE
IS COLOR YELLOW?
FALSE
TRUE
SO, WE SPLIT THE RIGHT NODE FURTHER BASED ON THE COLOR


IS DIAMETER>=3
FALSE
TRUE
IS COLOR YELLOW?
FALSE
TRUE
PREDICT LEMON WITH 100% ACCURACY


IS DIAMETER>=3
FALSE
TRUE
IS COLOR YELLOW?
FALSE
TRUE
PREDICT APPLE WITH 100% ACCURACY


IS DIAMETER>=3
FALSE
TRUE
PREDICT GRAPES WITH 100% ACCURACY
IS COLOR YELLOW?
FALSE
TRUE

Problem with Decision Trees?

- Greedy Algorithm
- Overfitting
- Low prediction accuracy
- Calculations can become complex when there are many class labels.

Bagging

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the **stability** and **accuracy** of machine learning algorithms used in statistical classification and regression.

It also reduces variance and helps to avoid overfitting.

Why Random Forest?

No overfitting

Use of multiple trees reduce the risk of overfitting

Training time is less

High accuracy

Runs efficiently on large database

For large data, it produces highly accurate predictions

Estimates missing data

Random Forest can maintain accuracy when a large proportion of data is missing

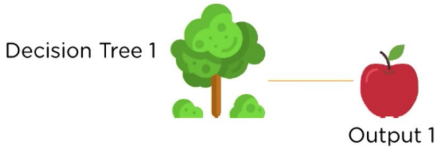
Real Life Example

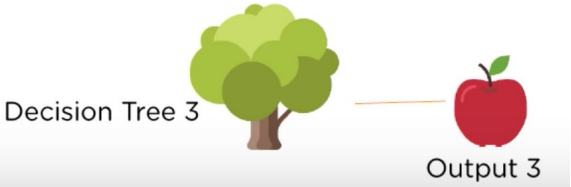
- You have decided to go for a tour
- But you need to choose which place to go

- You have set of attributes like
- Budget
- Type of Place (City, Village, Hill station)
- Domestic or International

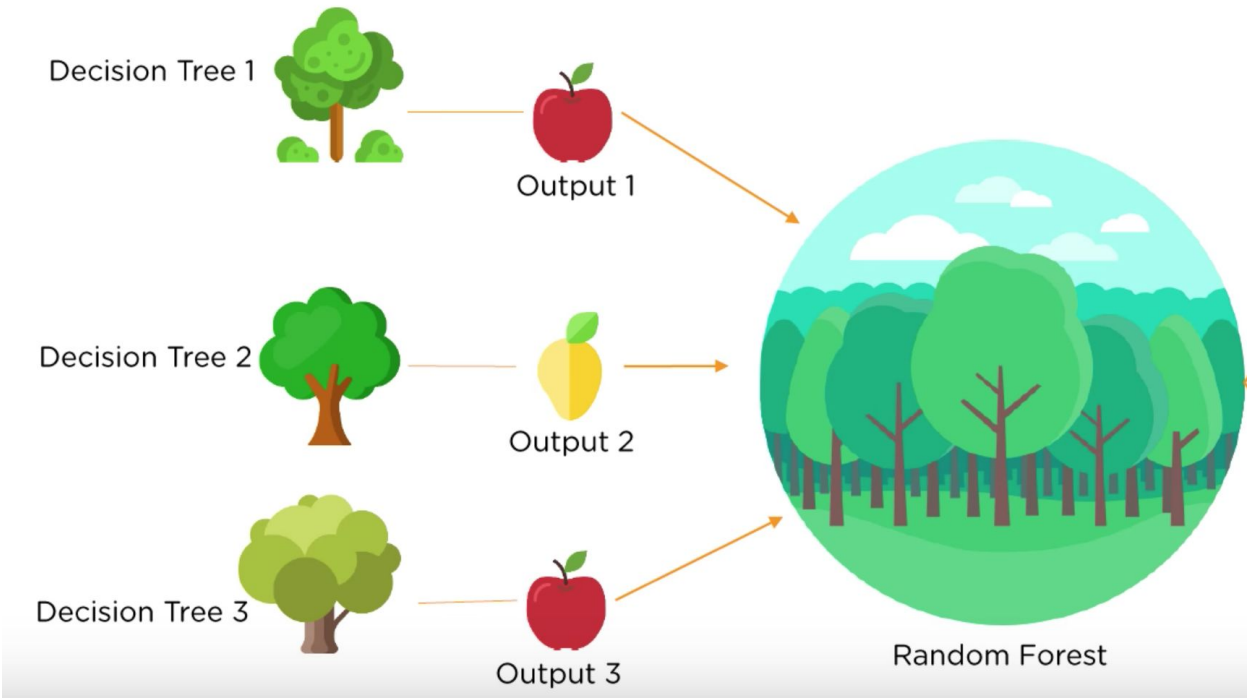
- Your option are (Ooty, Shimla, Switzerland, coorg, Newyork,etc)
- You go and ask many people who have been to the place with any two conditions
- Finally after getting the result you decide based on maximum votes

Random forest or Random Decision Forest is a method that operates by constructing multiple Decision Trees during training phase.
The Decision of the majority of the trees is chosen by the random forest as the final decision

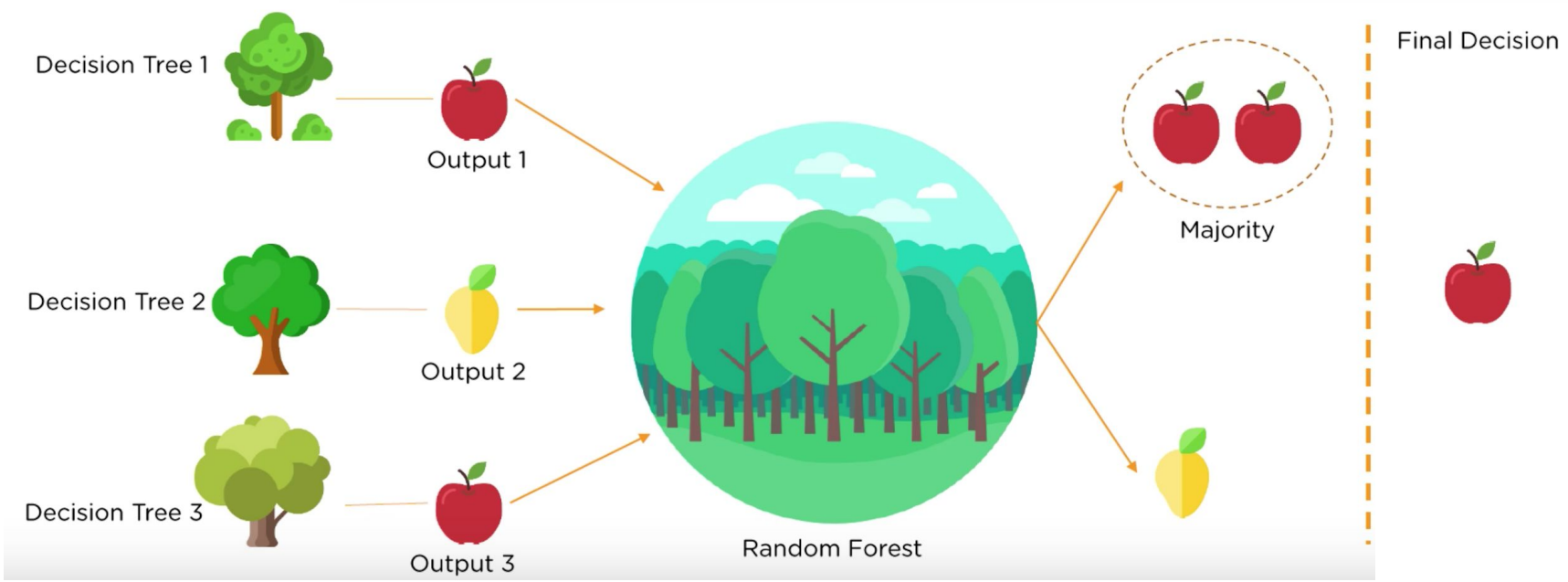




Random forest or Random Decision Forest is a method that operates by constructing multiple Decision Trees during training phase.
The Decision of the majority of the trees is chosen by the random forest as the final decision



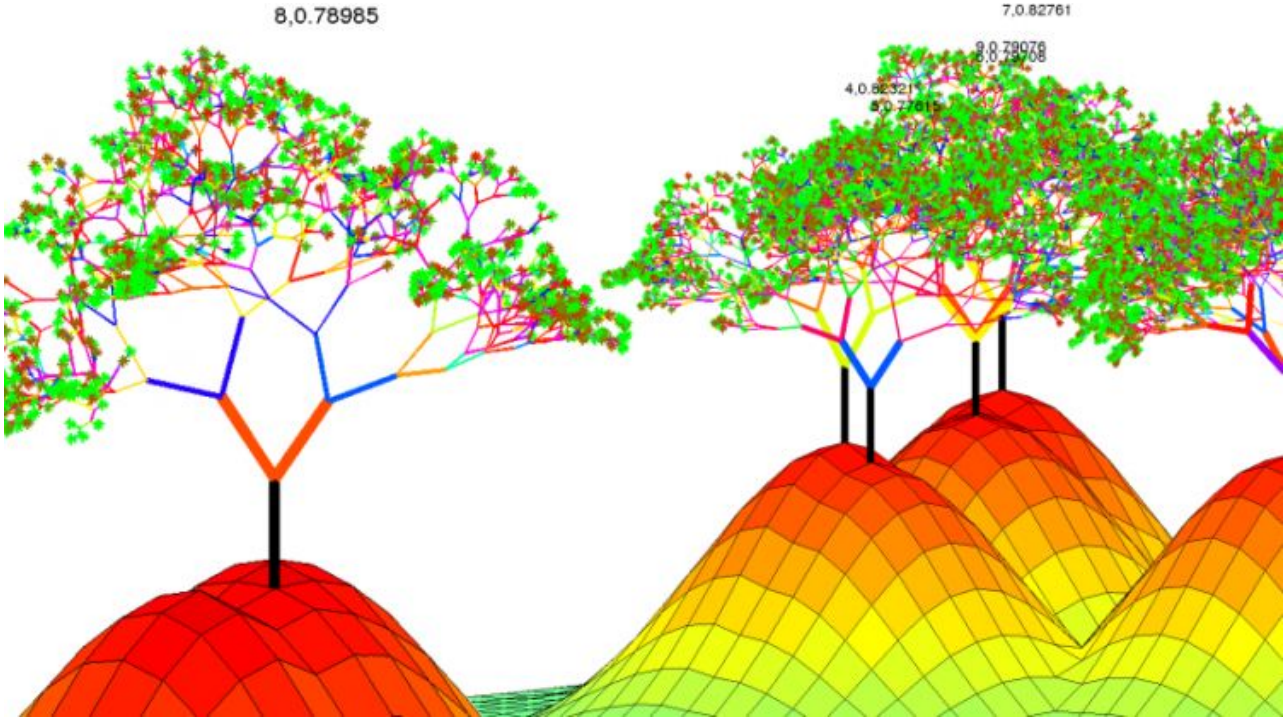
Random forest or Random Decision Forest is a method that operates by constructing multiple Decision Trees during training phase.
The Decision of the majority of the trees is chosen by the random forest as the final decision

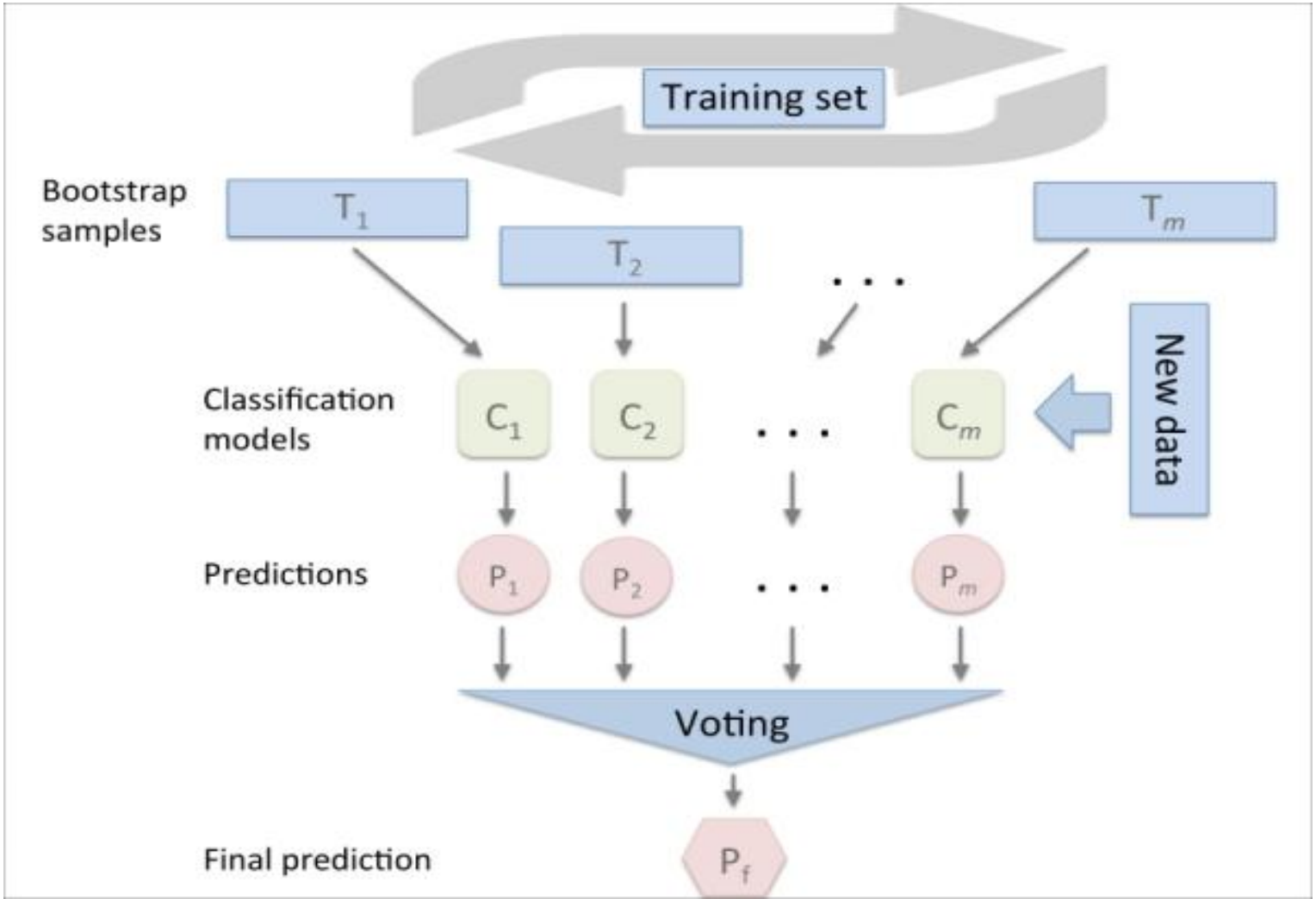


How Random Forest Work?

Most used algorithm- Bagging Technique

What insight can u get?




Training set
Bootstrap samples
T_1
T_2
T_m
Classification models
C_1
C_2
C_m
New data
Predictions
P_1
P_2
P_m
Voting
Final prediction
P_f

Random Forest Pseudocode

(a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.

(b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

i. Select m variables at random from the p variables.

ii. Pick the best variable/split-point among the m.

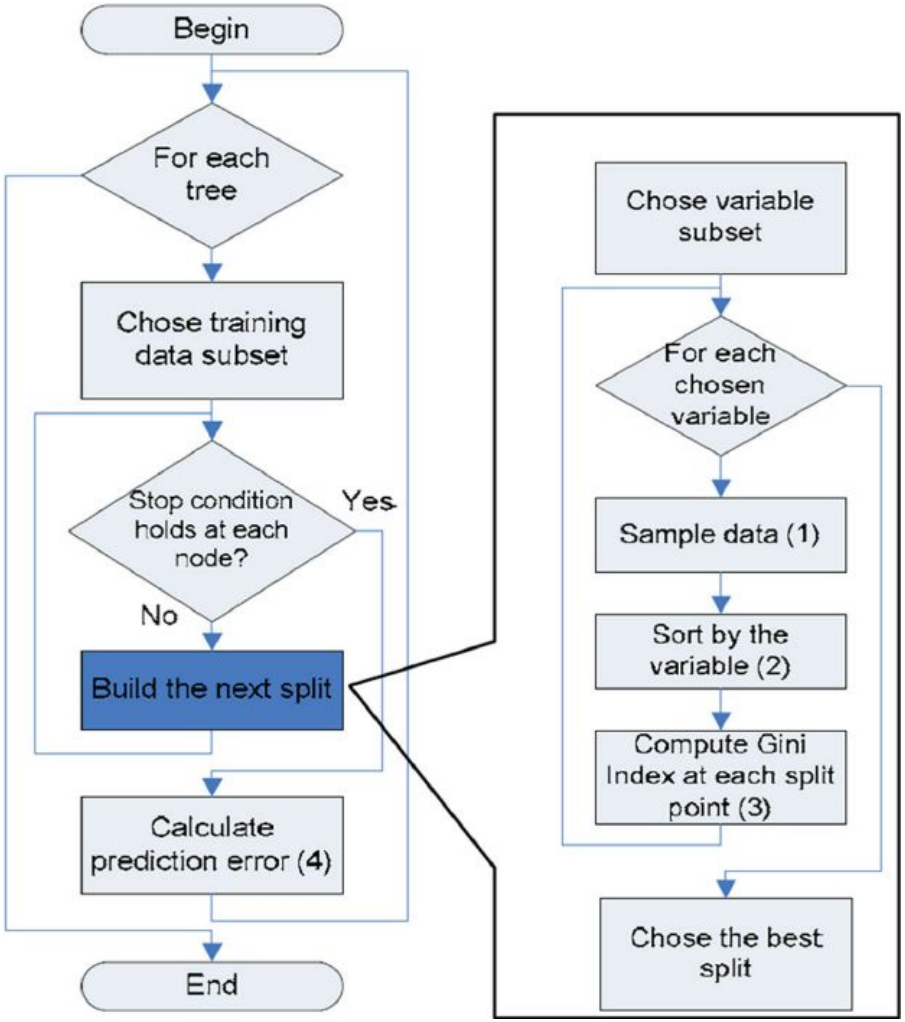
iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest


Begin
For each tree
Chose training data subset
Stop condition holds at each node?
Yes
No
Build the next split
Calculate prediction error (4)
End
Chose variable subset
For each chosen variable
Sample data (1)
Sort by the variable (2)
Compute Gini Index at each split point (3)
Chose the best split

CONDITIONS

COLOR== PURPLE?

DIAMETER=3

COLOR== YELLOW?

COLOR== RED?

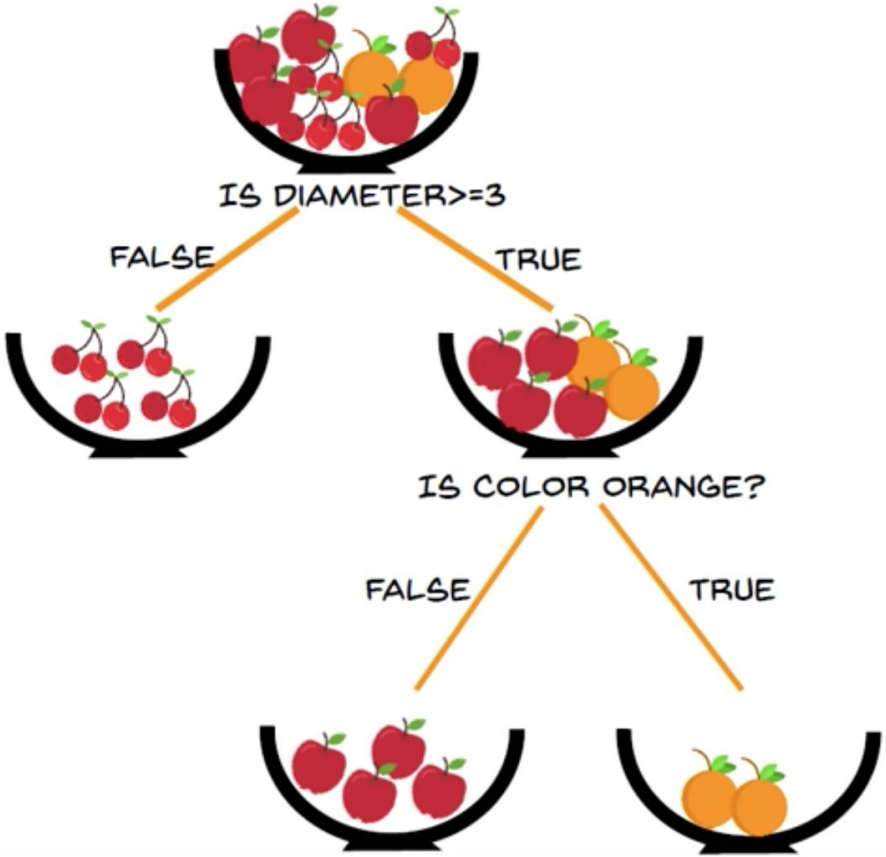
DIAMETER=1

LET'S SAY THIS CONDITION GIVES US THE MAXIMUM GAIN

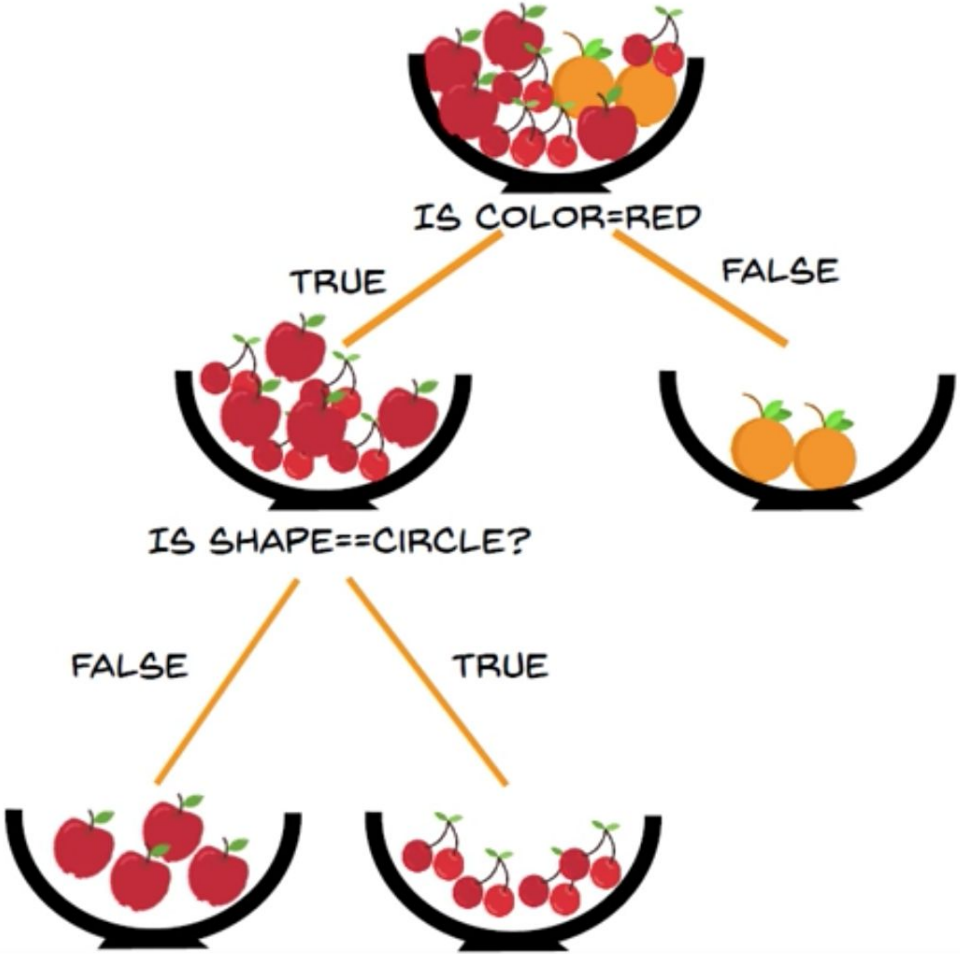
TRAINING DATASET

COLOR	DIAMETER	LABEL
RED	3	APPLE
YELLOW	3	LEMON
PURPLE	1	GRAPES
RED	3	APPLE
YELLOW	3	LEMON
PURPLE	1	GRAPES

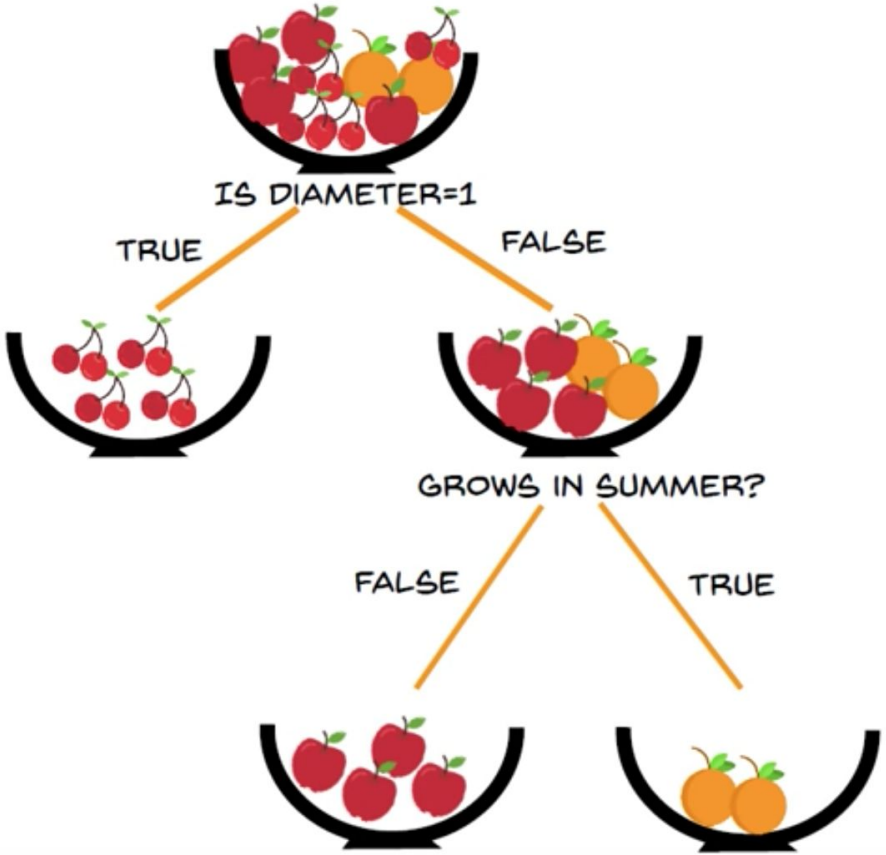
LET THIS BE TREE 1



LET THIS BE TREE 2



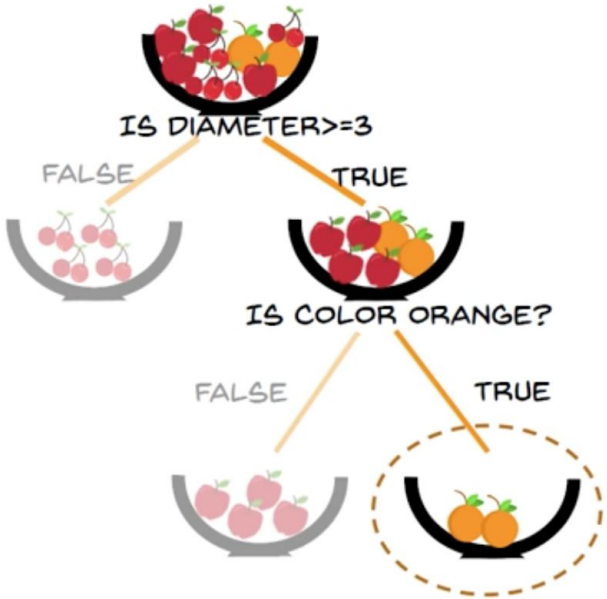
LET THIS BE TREE 3



NOW LETS TRY TO CLASSIFY THIS FRUIT

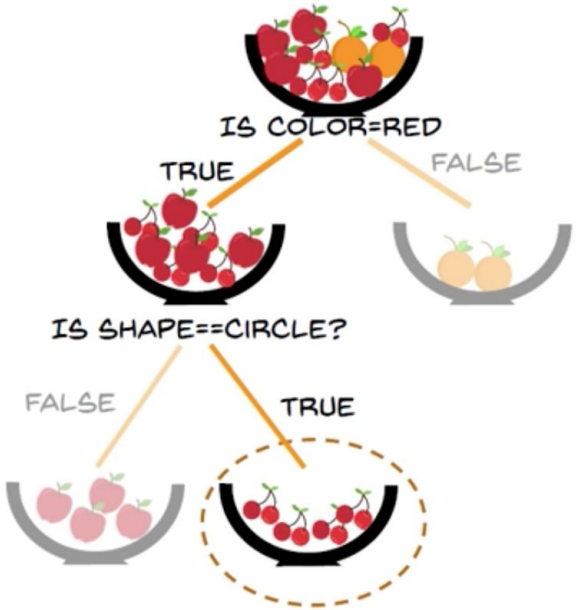
TREE 1 CLASSIFIES
IT AS AN ORANGE

DIAMETER = 3
COLOUR = ORANGE
GROWS IN SUMMER = YES
SHAPE = CIRCLE



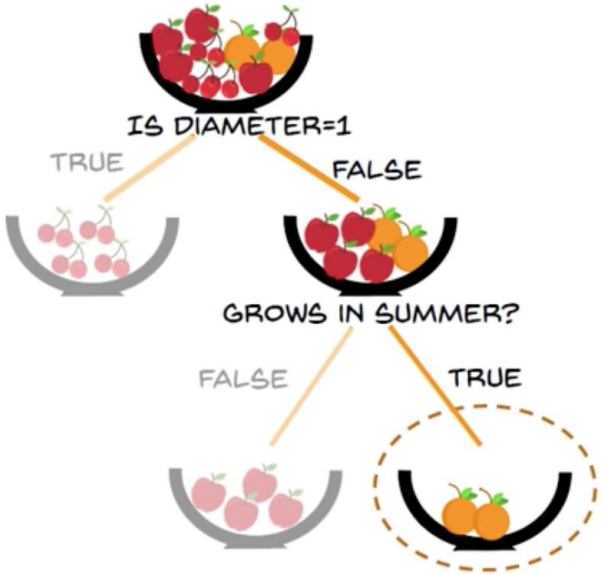
TREE 2 CLASSIFIES
IT AS CHERRIES

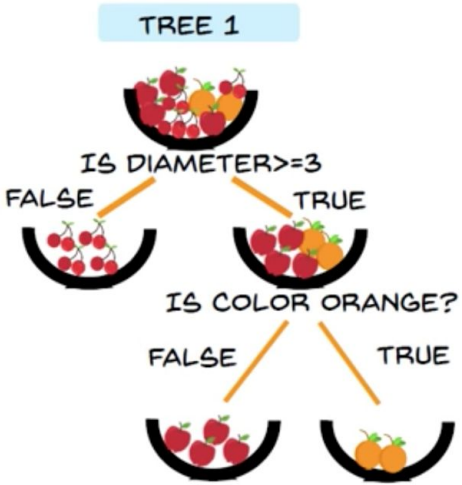
DIAMETER = 3
COLOUR = ORANGE
GROWS IN SUMMER = YES
SHAPE = CIRCLE

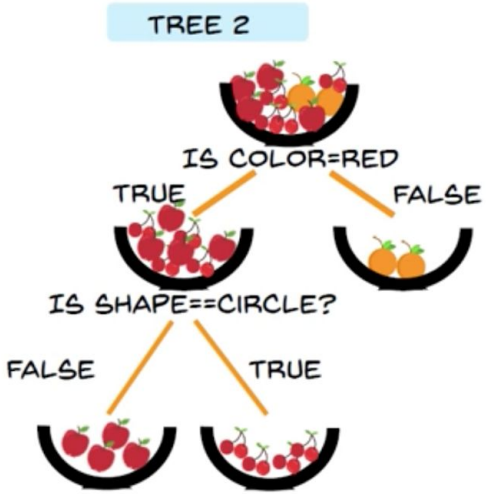


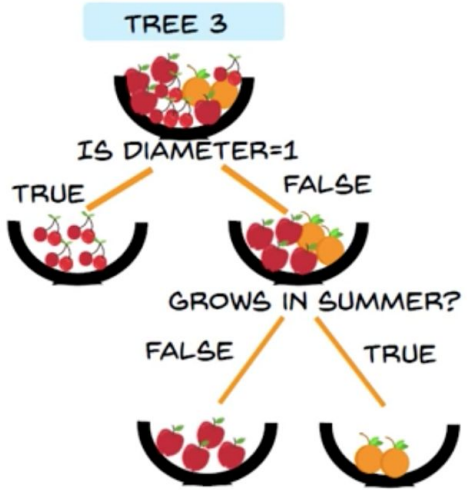
TREE 3 CLASSIFIES IT AS ORANGE

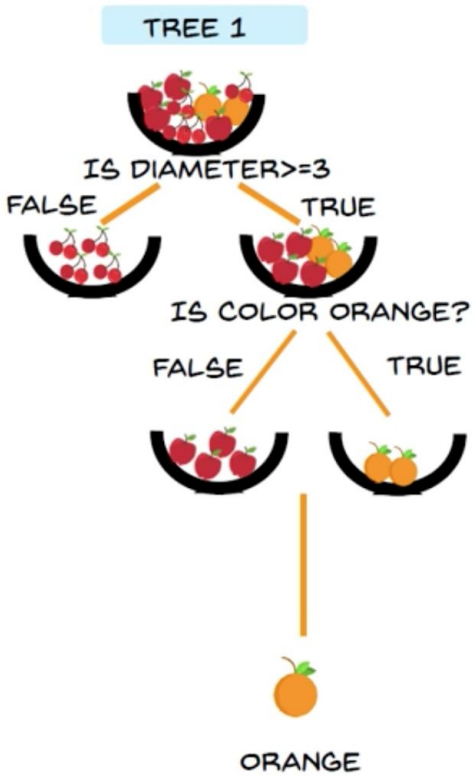
DIAMETER = 3
COLOUR = ORANGE
GROWS IN SUMMER = YES
SHAPE = CIRCLE

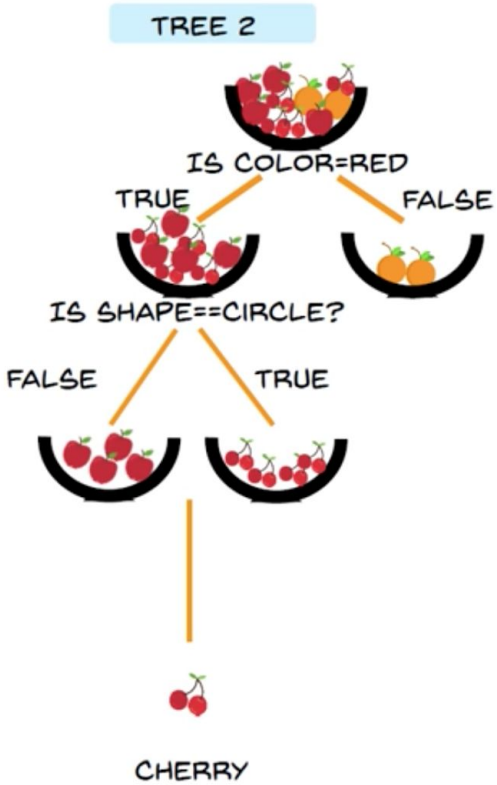


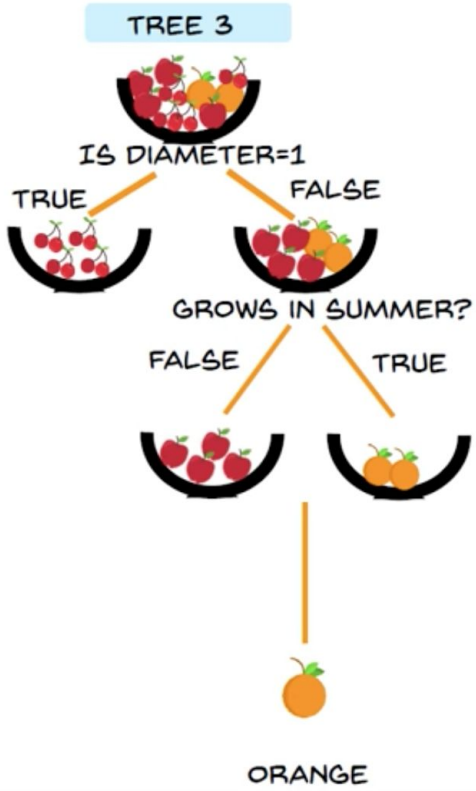

TREE 1
IS DIAMETER>=3
FALSE
TRUE
IS COLOR ORANGE?
FALSE
TRUE


TREE 2
IS COLOR=RED
TRUE
FALSE
IS SHAPE==CIRCLE?
FALSE
TRUE


TREE 3
IS DIAMETER=1
TRUE
FALSE
GROWS IN SUMMER?
FALSE
TRUE


TREE 1
IS DIAMETER>=3
FALSE
TRUE
IS COLOR ORANGE?
FALSE
TRUE
ORANGE


TREE 2
IS COLOR=RED
TRUE
FALSE
IS SHAPE==CIRCLE?
FALSE
TRUE
CHERRY


TREE 3
IS DIAMETER=1
TRUE
FALSE
GROWS IN SUMMER?
FALSE
TRUE
ORANGE

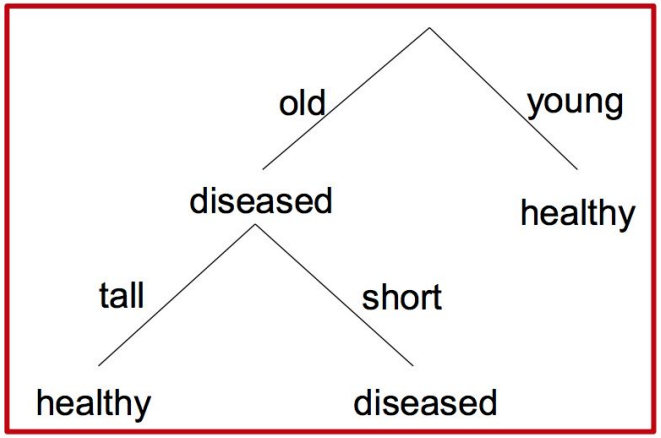
ORANGE

MAJORITY
VOTED

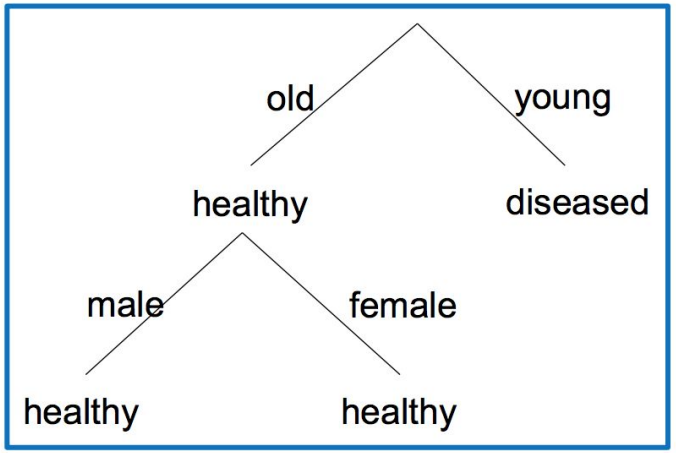
CHERRY

SO THE FRUIT IS CLASSIFIED AS AN ORANGE

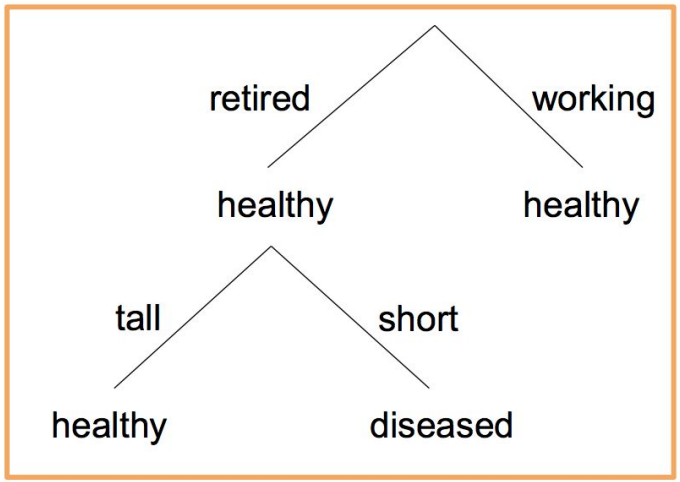
Tree 1



Tree 2



Tree 3



New sample:

old, retired, male, short

Tree predictions:

diseased, healthy, diseased

Majority rule:

diseased

Points to remember

- For classification a good default is: **m = sqrt(p)**
- For regression a good default is: **m = p/3**

Estimated Performance

- For each bootstrap sample taken from the training data, there will be samples left behind that were not included. These samples are called **Out-Of-Bag samples** or OOB.
- The performance of each model on its left out samples when averaged can provide an estimated accuracy of the bagged models. This estimated performance is often called the OOB estimate of performance.
- These performance measures are reliable test error estimate and correlate well with cross validation estimates.

Row	Feature1	Feature2	Label
1	5	3	A
2	7	8	B
3	6	4	A
4	9	6	B
5	2	7	A
6	4	3	B
7	1	5	A
8	3	6	B
9	8	7	A
10	7	2	B

Building the First Tree (Tree 1)

Bootstrap Sample: For Tree 1, we randomly select 10 rows **with replacement**. Let's say we get:

```
1, 2, 2, 4, 5, 6, 6, 7, 8, 10
```

Notice that rows 2 and 6 appear twice, and some rows (e.g., 3 and 9) are not selected.

OOB Samples for Tree 1: The rows that were **not chosen** in the bootstrap sample are:

```
Rows 3 and 9
```

Validation for Tree 1: After training Tree 1 on its bootstrap sample, we use it to predict the labels for rows 3 and 9. The accuracy on these rows is recorded as the **OOB accuracy** for Tree 1.

Repeating for Multiple Trees

Let’s build two more trees to see how the OOB samples change:

- **Tree 2**
  - **Bootstrap Sample**: `1, 1, 3, 4, 5, 5, 6, 7, 8, 10`
  - **OOB Samples**: Rows `2` and `9`
- **Tree 3**
  - **Bootstrap Sample**: `1, 3, 3, 4, 6, 7, 7, 8, 9, 10`
  - **OOB Samples**: Rows `2` and `5`

This variability in OOB samples helps ensure that every row in the dataset is left out at least a few times across all trees, allowing multiple trees to provide OOB predictions for most rows.

Calculating the OOB Error

Once all trees are built, each row will have a collection of **OOB predictions** from the trees where it was not part of the bootstrap sample.

1. **Collect OOB Predictions**: For each row in the original dataset, collect predictions from all trees where that row was an OOB sample.
2. **Majority Vote**: For classification problems, take the **majority vote** of these OOB predictions for each row to determine the final predicted label.
3. **Compare with Actual Label**: Calculate the percentage of rows where the OOB predictions **do not match the actual label**—this gives the **OOB error**.

For instance, if rows 3 and 9 have majority OOB predictions matching their actual labels, but row 2 has a mismatch, then:

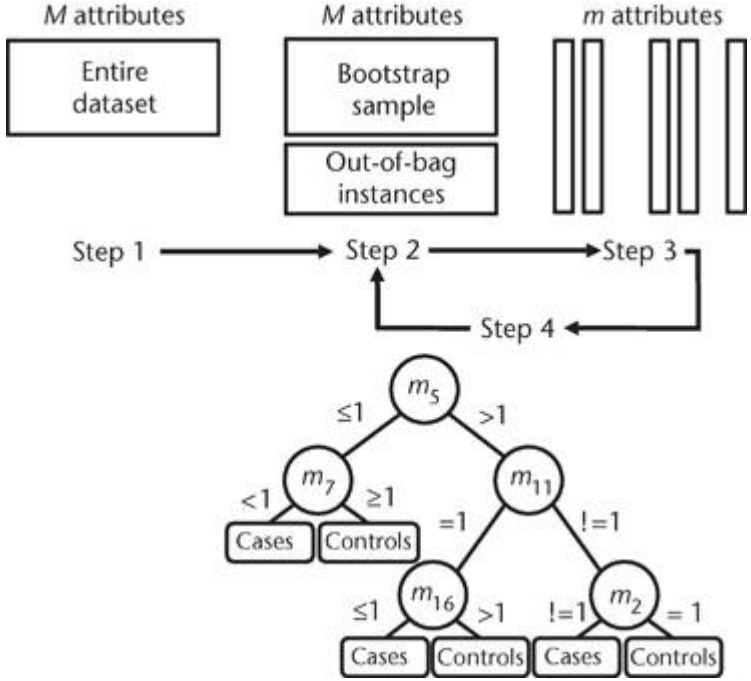
$$\text{OOB Error} = \frac{\text{Number of Incorrect OOB Predictions}}{\text{Total Rows}} = \frac{1}{10} = 10\%$$

Benefits of Using OOB Samples

The **OOB error** serves as an accurate estimate of how the Random Forest will perform on new data, making it a valuable metric:

- **Efficiency**: It doesn’t require a separate test set or cross-validation.
- **Reliability**: It provides an unbiased estimate of error as it is based on predictions for unseen samples.

This is why OOB samples are a fundamental component of Random Forests, ensuring reliable and efficient performance evaluation directly during the training phase.


M attributes
Entire dataset
M attributes
Bootstrap sample
Out-of-bag instances
m attributes
Step 1
Step 2
Step 3
Step 4
m_5
≤1
>1
m_7
m_{11}
<1
≥1
=1
!=1
Cases
Controls
m_{16}
m_2
≤1
>1
!=1
= 1
Cases
Controls
Cases
Controls

Random forest cons

It is one of the most accurate learning algorithms available.

For many data sets, it produces a highly accurate classifier.

It runs efficiently on large databases.

Conclusion

- Random forests are an effective tool in prediction.
- Forests give results competitive with boosting and adaptive bagging, yet do not progressively change the training set.
- Random inputs and random features produce good results in classification- less so in regression.
- For larger data sets, we can gain accuracy by combining random features with boosting.