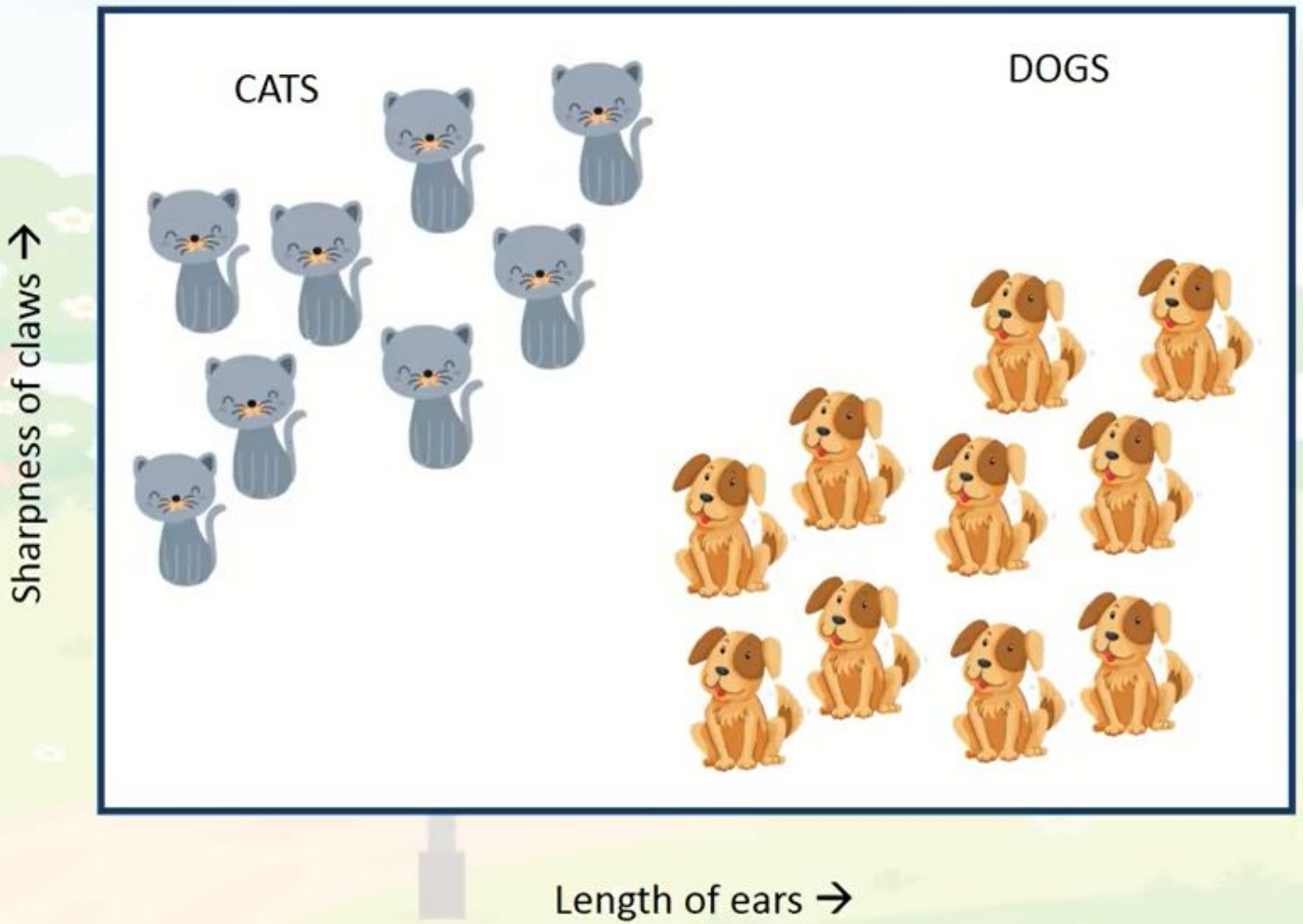
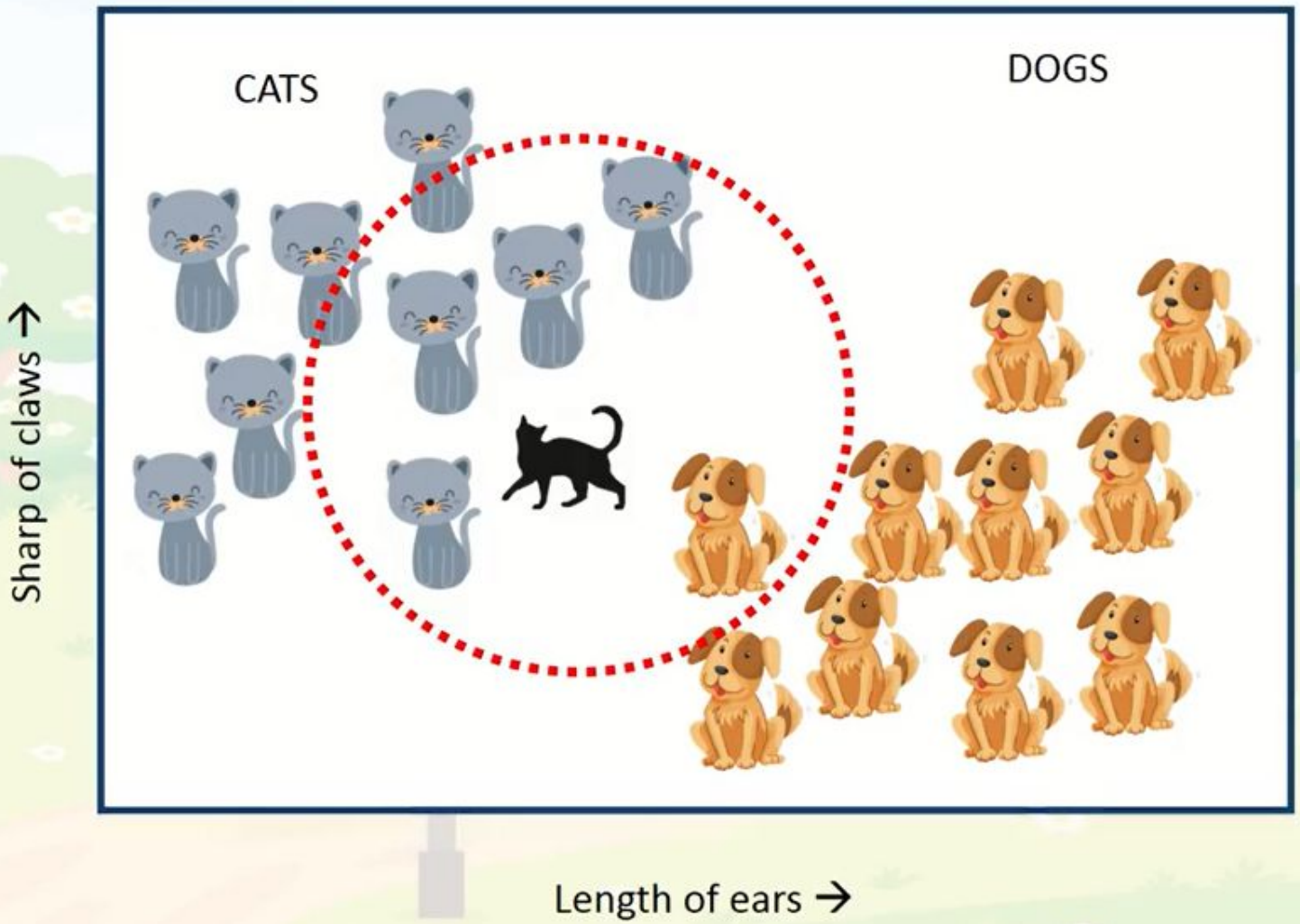
MACHINE LEARNING

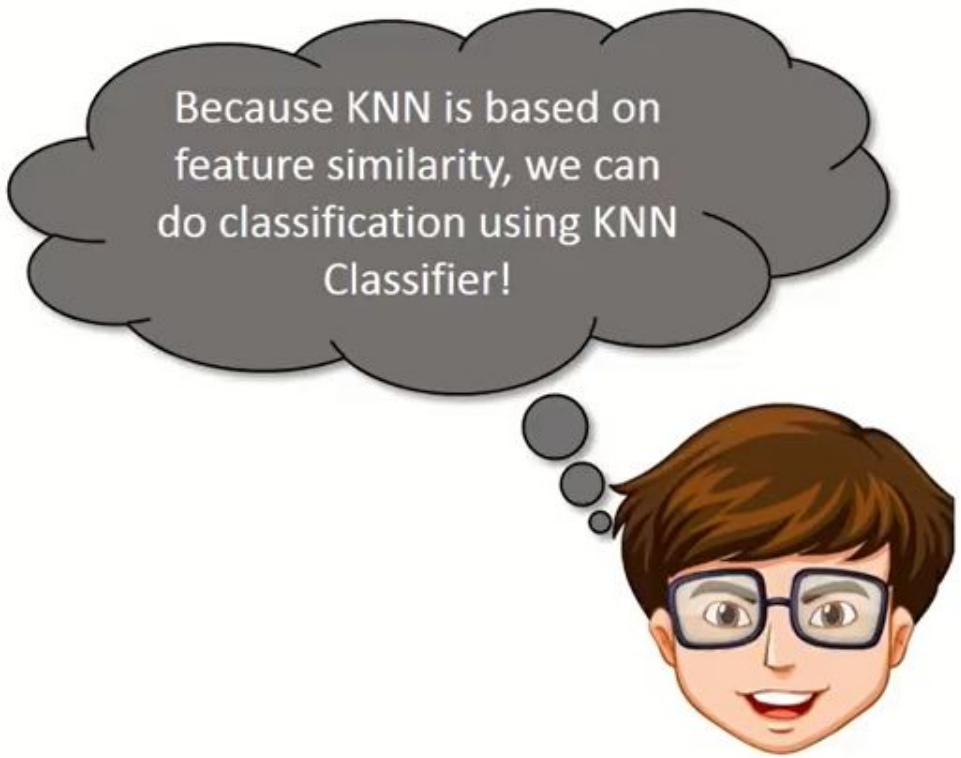
KNN

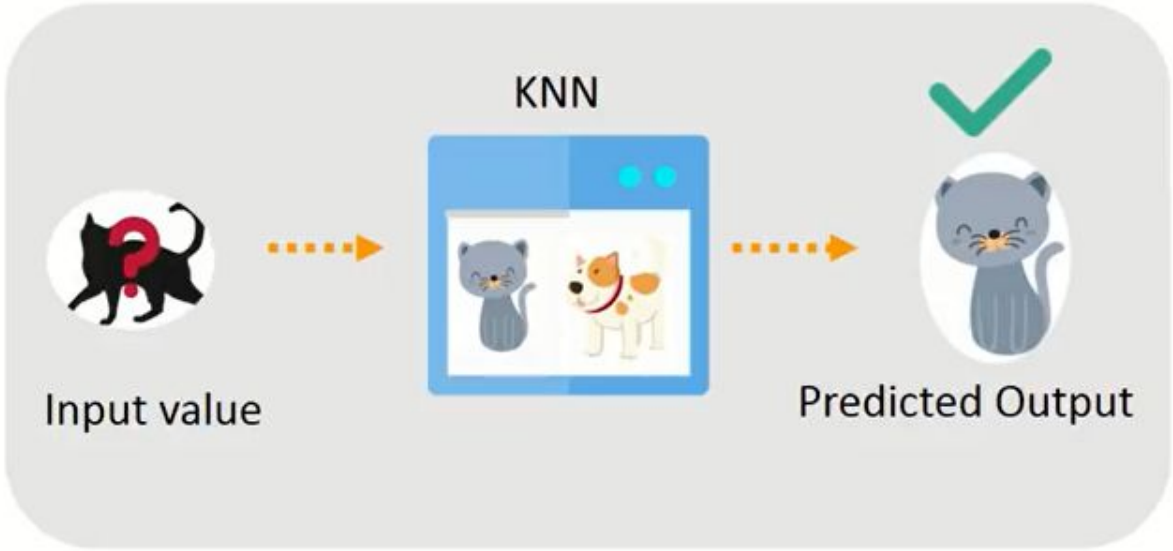


No dear, you can differentiate between a cat and a dog based on their characteristics


CATS
DOGS
Sharp of claws →
Length of ears →

Why KNN?





What is KNN Algorithm?

KNN – K Nearest Neighbors, is one of the simplest **Supervised** Machine Learning algorithm mostly used for

Classification

It classifies a data point based on how its neighbors are classified

Next video

simplilearn

How does KNN Algorithm work?

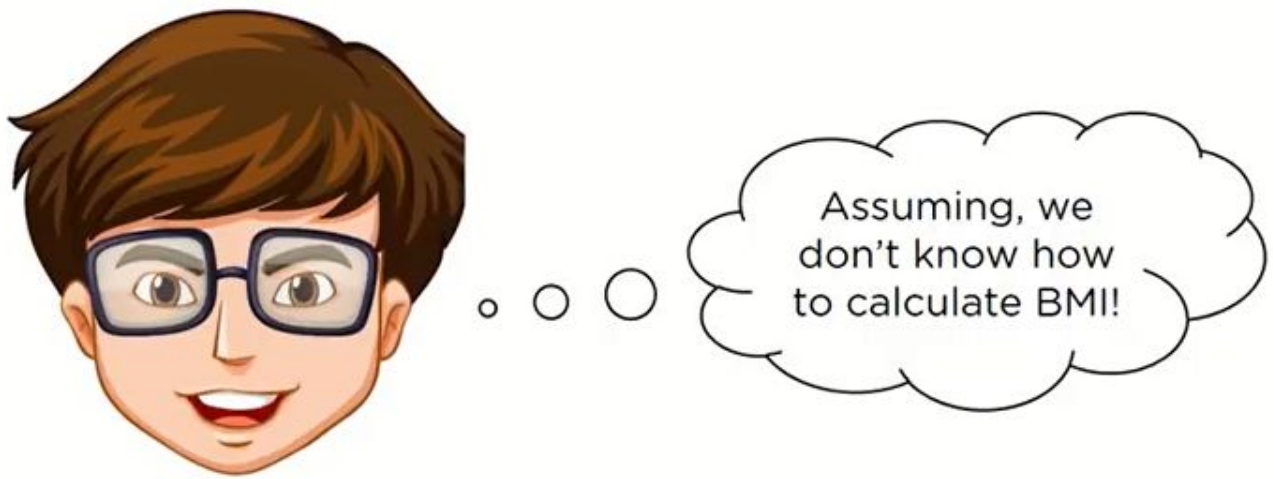
Consider a dataset having two variables: height (cm) & weight (kg) and each point is classified as Normal or Underweight

Weight(x2)	Height(y2)	Class
51	167	Underweight
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Underweight
58	169	Normal
57	173	Normal
55	170	Normal

How does KNN Algorithm work?

On the basis of the given data we have to classify the below set as Normal or Underweight using KNN

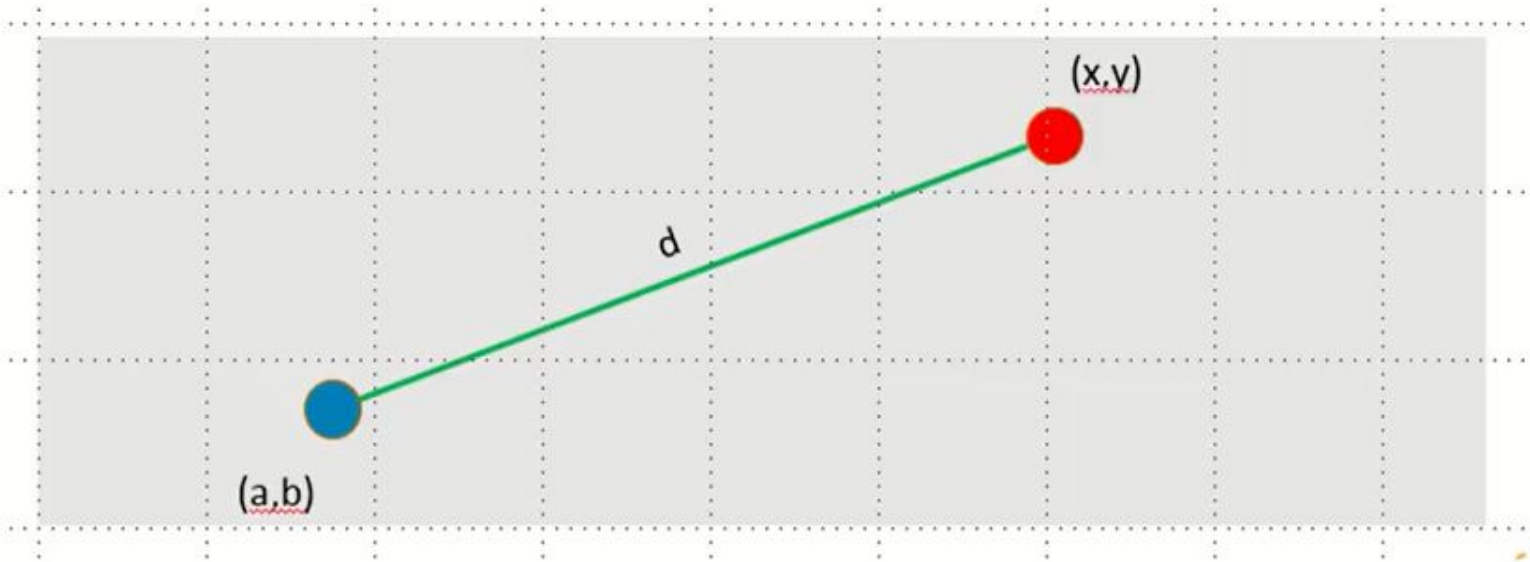
57 kg	170 cm	?



How does KNN Algorithm work?

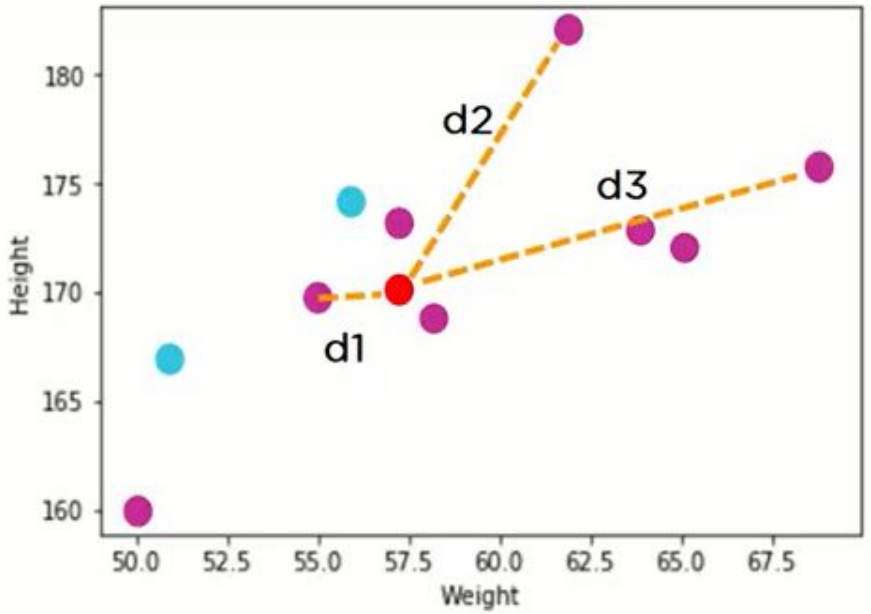
According to the **Euclidean distance** formula, the **distance** between two points in the plane with coordinates (x, y) and (a, b) is given by:

$$dist(d) = \sqrt{(x - a)^2 + (y - b)^2}$$



How does KNN Algorithm work?

Let's calculate it to understand clearly:



dist(**d1**)= √(170-167)² + (57-51)² ~= 6.7

dist(**d2**)= √(170-182)² + (57-62)² ~= 13

dist(**d3**)= √(170-176)² + (57-69)² ~= 13.4

● Unknown data point

How does KNN Algorithm work?

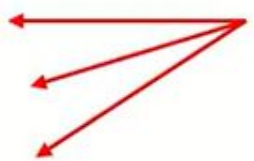
Hence, we have calculated the Euclidean distance of unknown data point from all the points as shown:

Where (x1, y1) = (57, 170) whose class we have to classify

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

How does KNN Algorithm work?

Class	Euclidean Distance
Underweight	6.7
Normal	13
Normal	13.4
Normal	7.6
Normal	8.2
Underweight	4.1
Normal	1.4
Normal	3
Normal	2



k = 3

So, majority neighbors are pointing towards '*Normal*'

Hence, as per KNN algorithm the class of (57, 170) should be 'Normal'

Movies & TV

KNN Algorithms Supermarket.xlsx - Excel — Dimitri Patarroyo

File | Home | Developer | Insert | Page Layout | View | Formulas | Data | Review | Power Pivot | Tell me what you want to do | Share

Cut | Copy | Format Painter | Paste | Calibri | 11 | Wrap Text | General | Merge & Center | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | AutoSum | Fill | Clear | Sort & Filter | Find & Select

Clipboard | Font | Alignment | Number | Styles | Cells | Editing

E6

The Big Elephant - Supermarket

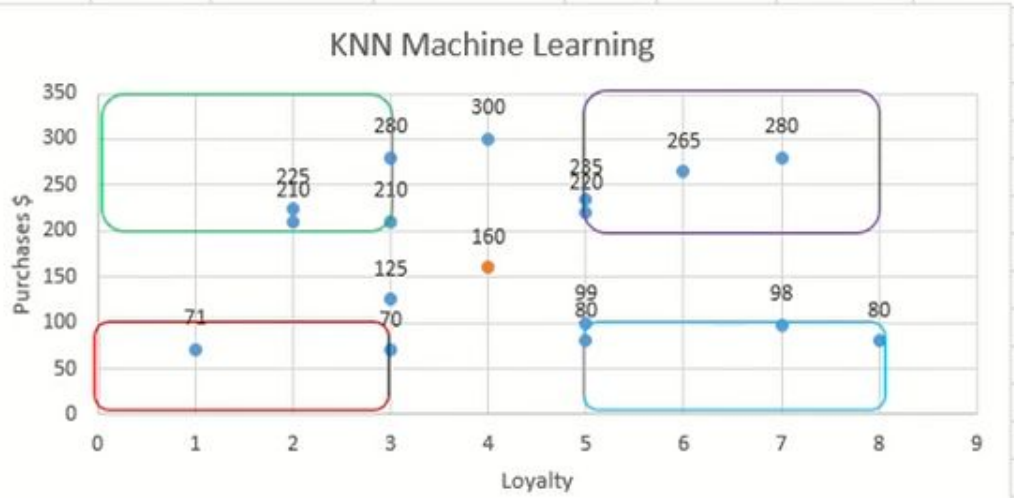
The Big Elephant Measures Loyalty by number of times a customer buy per week in avegrage

LIST OF NEW CUSTOMERS TO BE CLASSIFIED:

Customer ID	Loyalty	Purchase ($)	Customer
10001	3	280	Straightforward
10002	5	220	Prime
10003	8	80	WindowShopper
10004	2	225	Straightforward
10005	5	80	WindowShopper
10006	6	265	Prime
10007	1	71	Basic
10008	5	235	Prime
10009	7	98	WindowShopper
10010	4	300	Not defined
10011	2	210	Straightforward
10012	3	70	Basic
10013	5	99	WindowShopper
10014	3	210	Straightforward
10015	3	125	Not defined
10016	7	280	Prime

Identify this customer:

Loyalty	4
Purchases	160
Customer	??



Loyalty	Purchase $	Type of Customer
0-3	1-100	Basic
0-3	>200	Straightforward
>5	1-100	Windowshopper
>5	>200	Prime

Rank	Distance	Label	Cust
1	120.0041666	Straightforward	

k	Label
1	Straightforward
2	#N/A
3	#N/A
4	#N/A

All Customers | BI Customers | Goals and Strategies

Ready

KNN Algorithms Supermarket.xlsx - Excel — Dimitri Patarroyo

File | Home | Developer | Insert | Page Layout | View | Formulas | Data | Review | Power Pivot | Tell me what you want to do | Share

Cut | Copy | Format Painter | Paste | Clipboard
Calibri | 11 | Font
Wrap Text | Merge & Center | Alignment
General | Number
Conditional Formatting | Format as Table | Cell Styles | Styles
Insert | Delete | Format | Cells
AutoSum | Fill | Clear | Sort & Filter | Find & Select | Editing

N2 =SQRT((B6-$H$5)^2+(C6-$H$6)^2)

The Big Elephant - Supermarket

The Big Elephant Measures Loyalty by number of times a customer buys per week in avegrage

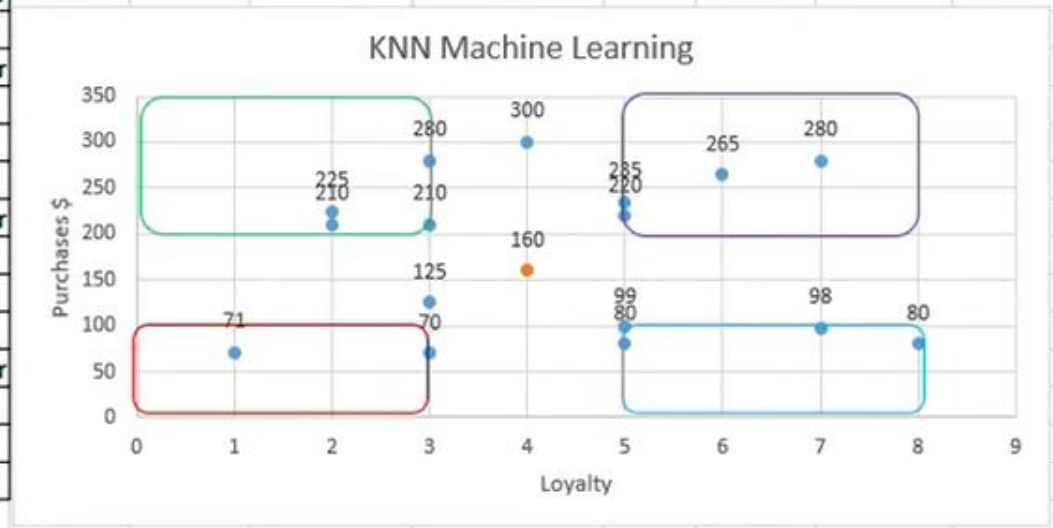
LIST OF NEW CUSTOMERS TO BE CLASSIFIED:

Customer ID	Loyalty	Purchase ($)	Customer
10001	3	280	Straightforward
10002	5	220	Prime
10003	8	80	WindowShopper
10004	2	225	Straightforward
10005	5	80	WindowShopper
10006	6	265	Prime
10007	1	71	Basic
10008	5	235	Prime
10009	7	98	WindowShopper
10010	4	300	Not defined
10011	2	210	Straightforward
10012	3	70	Basic
10013	5	99	WindowShopper
10014	3	210	Straightforward
10015	3	125	Not defined
10016	7	280	Prime

Identify this customer:

Loyalty	4
Purchases	160
Customer	??

Loyalty	Purchase $	Type of Customer
0-3	1-100	Basic
0-3	>200	Straightforward
>5	1-100	Windowshopper
>5	>200	Prime



Rank	Distance	Label	Cust
1	120.0041666	Straightforward	

k	Label
1	Straightforward
2	#N/A
3	#N/A
4	#N/A

All Customers | BI Customers | Goals and Strategies

ENG 07:52 م ٢٠١٩/١٢/٠٩

KNN Algorithms Supermarket.xlsx - Excel — Dimitri Patarroyo

File | Home | Developer | Insert | Page Layout | View | Formulas | Data | Review | Power Pivot | Tell me what you want to do | Share

Cut | Copy | Format Painter | Paste | Clipboard
Calibri | 11 | Font
Wrap Text | Merge & Center | Alignment
General | Number
Conditional Formatting | Format as Table | Cell Styles | Styles
Insert | Delete | Format | Cells
AutoSum | Fill | Clear | Sort & Filter | Find & Select | Editing

N2 =SQRT((B6-$H$5)^2+(C6-$H$6)^2)

The Big Elephant - Supermarket

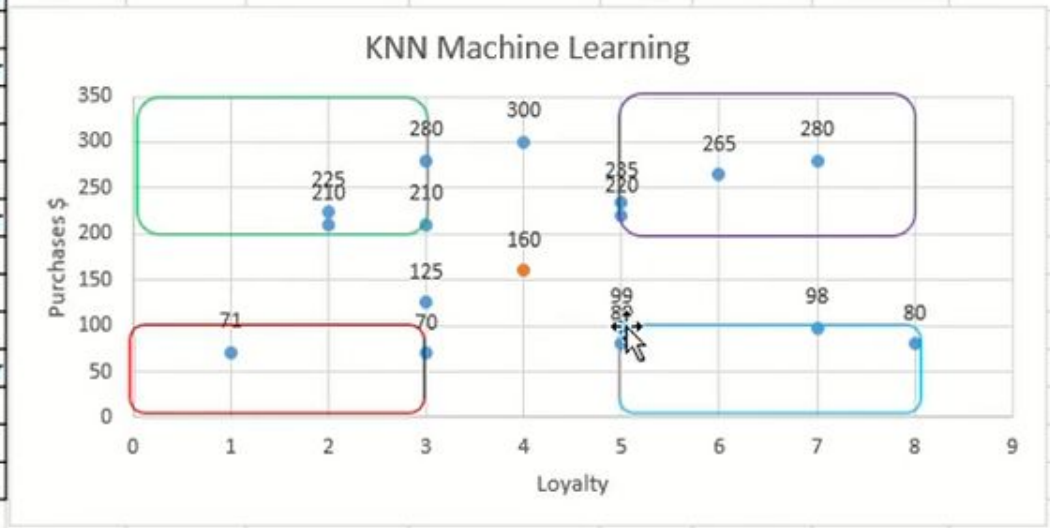
The Big Elephant Measures Loyalty by number of times a customer buys per week in avegrage

LIST OF NEW CUSTOMERS TO BE CLASSIFIED:

Customer ID	Loyalty	Purchase ($)	Customer
10001	3	280	Straightforward
10002	5	220	Prime
10003	8	80	WindowShopper
10004	2	225	Straightforward
10005	5	80	WindowShopper
10006	6	265	Prime
10007	1	71	Basic
10008	5	235	Prime
10009	7	98	WindowShopper
10010	4	300	Not defined
10011	2	210	Straightforward
10012	3	70	Basic
10013	5	99	WindowShopper
10014	3	210	Straightforward
10015	3	125	Not defined
10016	7	280	Prime

Identify this customer:

Loyalty	4
Purchases	160
Customer	??



Loyalty	Purchase $	Type of Customer
0-3	1-100	Basic
0-3	>200	Straightforward
>5	1-100	Windowshopper
>5	>200	Prime

Rank	Distance	Label	Cust
14	120.0041666	Straightforward	
	60.00833275		
	80.09993758		
	65.03076195		
	80.00624976		
	105.0190459		
	89.05054744		
	75.00666637		
	62.07253821		
	140		
	50.03998401		
	90.00555538		
	61.00819617		
	50.009999		
	35.0142828		
	120.0374941		

k	Label
1	#N/A
2	#N/A
3	#N/A
4	#N/A

All Customers | BI Customers | Goals and Strategies

Ready — Average: 80.15085988 Count: 16 Sum: 1282.413758 — 100%

KNN Algorithms Supermarket.xlsx - Excel — Dimitri Patarroyo

File | Home | Developer | Insert | Page Layout | View | Formulas | Data | Review | Power Pivot | Tell me what you want to do | Share

Cut | Copy | Format Painter | Paste | Clipboard
Calibri | 11 | Font
Wrap Text | Merge & Center | Alignment
General | Number
Conditional Formatting | Format as Table | Cell Styles | Styles
Insert | Delete | Format | Cells
AutoSum | Fill | Clear | Sort & Filter | Find & Select | Editing

N13 =SQRT((B17-$H$5)^2+(C17-$H$6)^2)

The Big Elephant - Supermarket

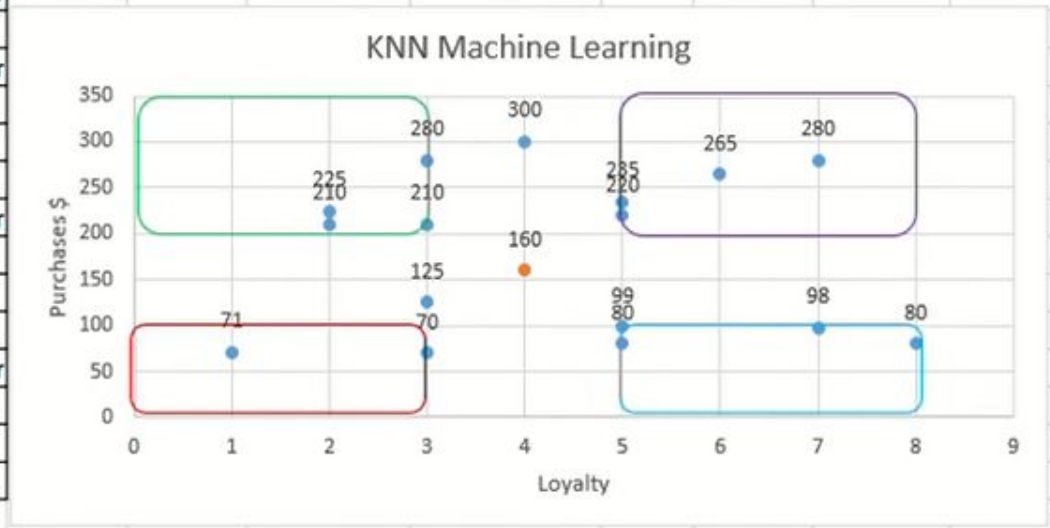
The Big Elephant Measures Loyalty by number of times a customer buys per week in avegrage

LIST OF NEW CUSTOMERS TO BE CLASSIFIED:

Customer ID	Loyalty	Purchase ($)	Customer
10001	3	280	Straightforward
10002	5	220	Prime
10003	8	80	WindowShopper
10004	2	225	Straightforward
10005	5	80	WindowShopper
10006	6	265	Prime
10007	1	71	Basic
10008	5	235	Prime
10009	7	98	WindowShopper
10010	4	300	Not defined
10011	2	210	Straightforward
10012	3	70	Basic
10013	5	99	WindowShopper
10014	3	210	Straightforward
10015	3	125	Not defined
10016	7	280	Prime

Identify this customer:

Loyalty	4
Purchases	160
Customer	??



Loyalty	Purchase $	Type of Customer
0-3	1-100	Basic
0-3	>200	Straightforward
>5	1-100	Windowshopper
>5	>200	Prime

Rank	Distance	Label	Cust
14	120.0041666	Straightforward	
4	60.00833275		
10	80.09993758		
7	65.03076195		
9	80.00624976		
13	105.0190459		
11	89.05054744		
8	75.00666637		
6	62.07253821		
16	140		
3	50.03998401		
12	90.00555538		
5	61.00819617		
2	50.009999		
1	35.0142828		
15	120.0374941		

k	el
1	0
2	0
3	0
4	0

All Customers | BI Customers | Goals and Strategies

Ready

KNN Algorithms Supermarket.xlsx - Excel — Dimitri Patarroyo

File | Home | Developer | Insert | Page Layout | View | Formulas | Data | Review | Power Pivot | Tell me what you want to do | Share

Cut | Copy | Format Painter | Paste | Calibri | 11 | Wrap Text | Merge & Center | General | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | AutoSum | Fill | Clear | Sort & Filter | Find & Select

Clipboard | Font | Alignment | Number | Styles | Cells | Editing

M12 | =RANK(N12,$N$2:$N$17,1)

The Big Elephant - Supermarket

res Loyalty by number of times a customer buys per week in avegrage

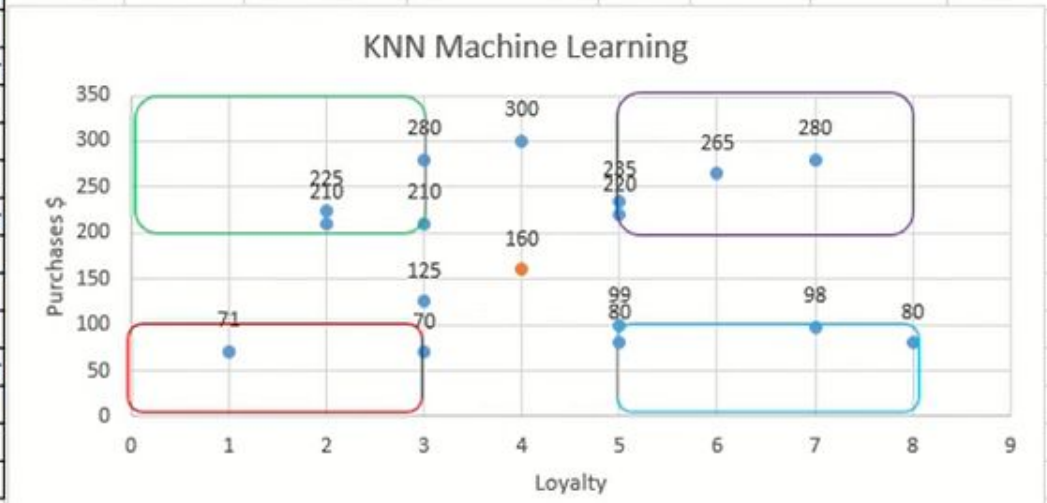
S TO BE CLASSIFIED:

Purchase ($)	Customer
280	Straightforward
220	Prime
80	WindowShopper
225	Straightforward
80	WindowShopper
265	Prime
71	Basic
235	Prime
98	WindowShopper
300	Not defined
210	Straightforward
70	Basic
99	WindowShopper
210	Straightforward
125	Not defined
280	Prime

Identify this customer:

Loyalty	4
Purchases	160
Customer	??

Loyalty	Purchase $	Type of Customer
0-3	1-100	Basic
0-3	>200	Straightforward
>5	1-100	Windowshopper
>5	>200	Prime



Rank	Distance	Label	Customer ID
14	120.0041666	Straightforward	10001
4	60.00833275	Prime	10002
10	80.09993758	WindowShopper	10003
7	65.03076195	Straightforward	10004
9	80.00624976	WindowShopper	10005
13	105.0190459	Prime	10006
11	89.05054744	Basic	10007
8	75.00666637	Prime	10008
6	62.07253821	WindowShopper	10009
16	140	Not defined	10010
3	50.03998401	Straightforward	10011
12	90.00555538	Basic	10012
5	61.00819617	WindowShopper	10013
2	50.009999	Straightforward	10014
1	35.0142828	Not defined	10015
15	120.0374941	Prime	10016

k	Label
1	Not defined
2	Straightforward
3	Straightforward
4	Prime

All Customers | BI Customers | Goals and Strategies

Ready | Average: 26.51999201 | Count: 3 | Sum: 53.03998401 | 100%

Movies & TV

KNN Algorithms Supermarket.xlsx - Excel — Dimitri Patarroyo

File | Home | Developer | Insert | Page Layout | View | Formulas | Data | Review | Power Pivot | Tell me what you want to do | Share

Cut | Copy | Format Painter | Paste | Calibri | 11 | Wrap Text | General | Merge & Center | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | AutoSum | Fill | Clear | Sort & Filter | Find & Select

Clipboard | Font | Alignment | Number | Styles | Cells | Editing

A20 | 10015

The Big Elephant - Supermarket

The Big Elephant Measures Loyalty by number of times a customer buys per week in avegrage

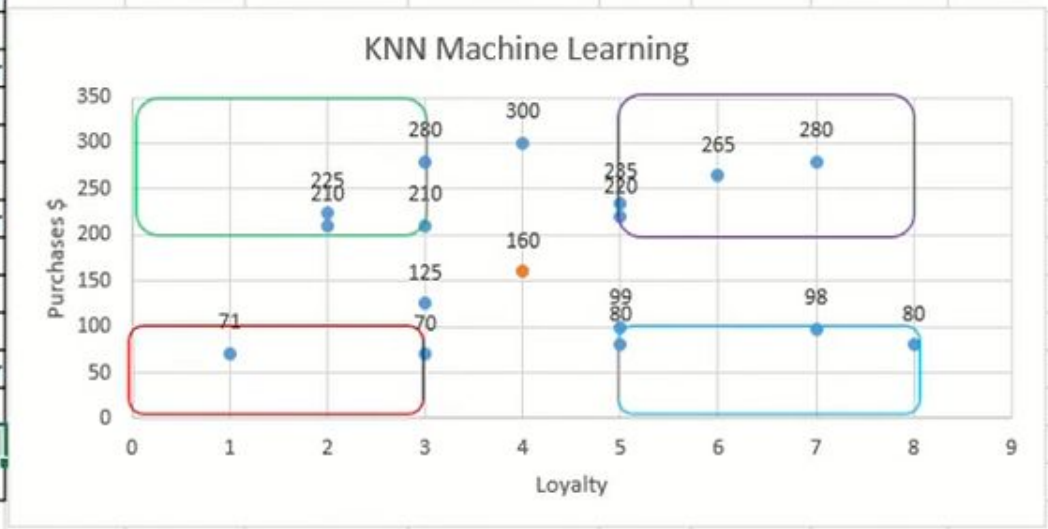
LIST OF NEW CUSTOMERS TO BE CLASSIFIED:

Customer ID	Loyalty	Purchase ($)	Customer
10001	3	280	Straightforward
10002	5	220	Prime
10003	8	80	WindowShopper
10004	2	225	Straightforward
10005	5	80	WindowShopper
10006	6	265	Prime
10007	1	71	Basic
10008	5	235	Prime
10009	7	98	WindowShopper
10010	4	300	Not defined
10011	2	210	Straightforward
10012	3	70	Basic
10013	5	99	WindowShopper
10014	3	210	Straightforward
10015	3	125	Not defined
10016	7	280	Prime

Identify this customer:

Loyalty	4
Purchases	160
Customer	??

Loyalty	Purchase $	Type of Customer
0-3	1-100	Basic
0-3	>200	Straightforward
>5	1-100	Windowshopper
>5	>200	Prime



Rank	Distance	Label
14	120.0041666	Straightforward
4	60.00833275	Prime
10	80.09993758	WindowShopper
7	65.03076195	Straightforward
9	80.00624976	WindowShopper
13	105.0190459	Prime
11	89.05054744	Basic
8	75.00666637	Prime
6	62.07253821	WindowShopper
16	140	Not defined
3	50.03998401	Straightforward
12	90.00555538	Basic
5	61.00819617	WindowShopper
2	50.009999	Straightforward
1	35.0142828	Not defined
15	120.0374941	Prime

k	Label
1	Not defined
2	Straightforward
3	Straightforward
4	Prime

All Customers | BI Customers | Goals and Strategies

Ready | Average: 3381 | Count: 4 | Sum: 10143 | 100%

KNN Algorithms Supermarket.xlsx - Excel — Dimitri Patarroyo

File | Home | Developer | Insert | Page Layout | View | Formulas | Data | Review | Power Pivot | Tell me what you want to do

Cut | Copy | Format Painter | Paste | Clipboard | Calibri | 11 | Font | Wrap Text | Merge & Center | Alignment | General | Number | Conditional Formatting | Format as Table | Cell Styles | Styles | Insert | Delete | Format | Cells | AutoSum | Fill | Clear | Sort & Filter | Find & Select | Editing

3R x 1C | =VLOOKUP(M19,$M$2:$O$17,3,0)

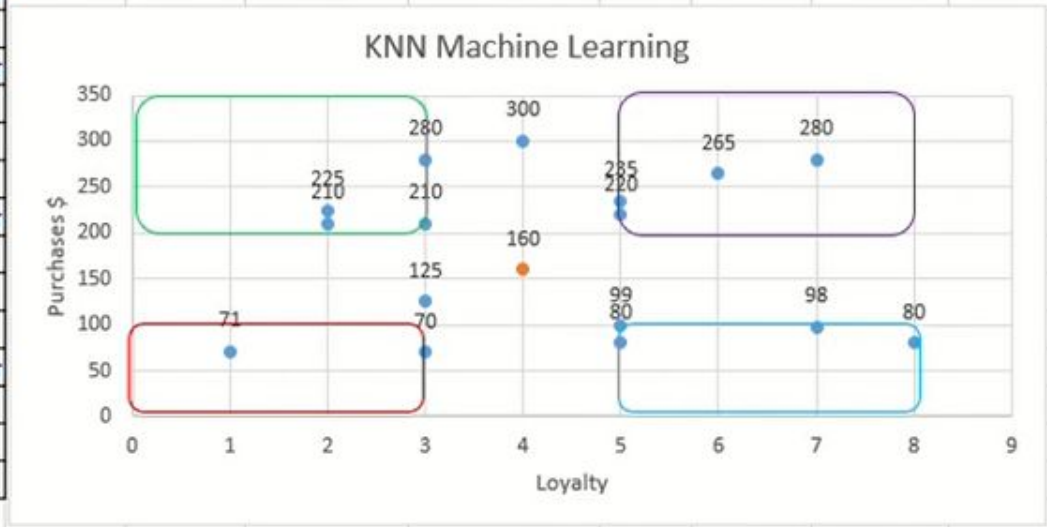
CUSTOMERS TO BE CLASSIFIED:

Loyalty	Purchase ($)	Customer
3	280	Straightforward
5	220	Prime
8	80	WindowShopper
2	225	Straightforward
5	80	WindowShopper
6	265	Prime
1	71	Basic
5	235	Prime
7	98	WindowShopper
4	300	Not defined
2	210	Straightforward
3	70	Basic
5	99	WindowShopper
3	210	Straightforward
3	125	Not defined
7	280	Prime

Identify this customer:

Loyalty	4
Purchases	160
Customer	??

>5	1-100	Windowshopper
>5	>200	Prime



10	80.09993758	WindowShopper	10003
7	65.03076195	Straightforward	10004
9	80.00624976	WindowShopper	10005
13	105.0190459	Prime	10006
11	89.05054744	Basic	10007
8	75.00666637	Prime	10008
6	62.07253821	WindowShopper	10009
16	140	Not defined	10010
3	50.03998401	Straightforward	10011
12	90.00555538	Basic	10012
5	61.00819617	WindowShopper	10013
2	50.009999	Straightforward	10014
1	35.0142828	Not defined	10015
15	120.0374941	Prime	10016

k	Label	
1	Not defined	
2	Straightforward	k=3
3	Straightforward	

All Customers | BI Customers | Goals and Strategies

Ready | Count: 3

KNN Algorithms Supermarket.xlsx - Excel — Dimitri Patarroyo

File | Home | Developer | Insert | Page Layout | View | Formulas | Data | Review | Power Pivot | Tell me what you want to do | Share

Cut | Copy | Format Painter | Paste | Calibri | 11 | Wrap Text | Merge & Center | General | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | AutoSum | Fill | Clear | Sort & Filter | Find & Select

Clipboard | Font | Alignment | Number | Styles | Cells | Editing

H8

The Big Elephant - Supermarket

ant Measures Loyalty by number of times a customer buys per week in avegrage

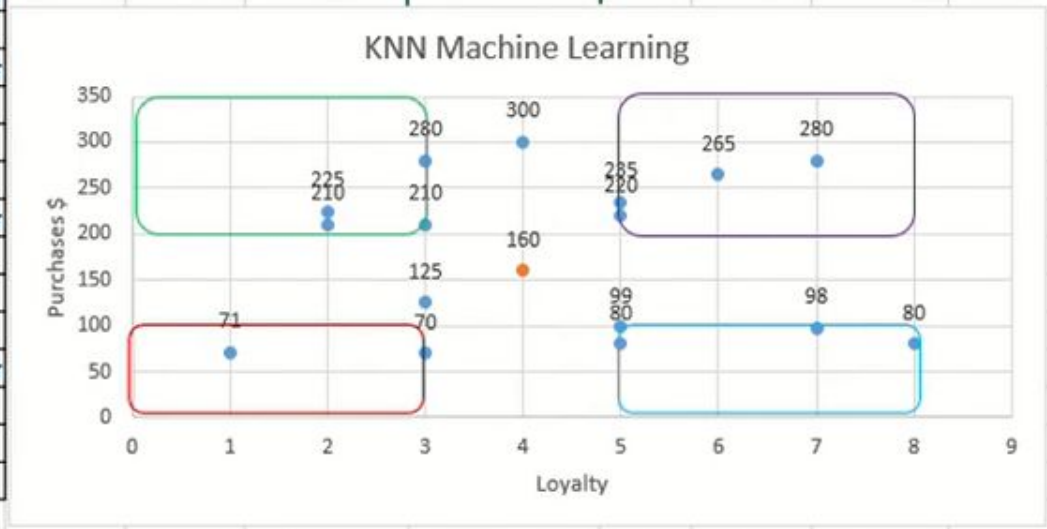
CUSTOMERS TO BE CLASSIFIED:

Loyalty	Purchase ($)	Customer
3	280	Straightforward
5	220	Prime
8	80	WindowShopper
2	225	Straightforward
5	80	WindowShopper
6	265	Prime
1	71	Basic
5	235	Prime
7	98	WindowShopper
4	300	Not defined
2	210	Straightforward
3	70	Basic
5	99	WindowShopper
3	210	Straightforward
3	125	Not defined
7	280	Prime

Identify this customer:

Loyalty	4
Purchases	160
Customer	Straightforward

Loyalty	Purchase $	Type of Customer
0-3	1-100	Basic
0-3	>200	Straightforward
>5	1-100	Windowshopper
>5	>200	Prime



Rank	Distance	Label	Customer ID
14	120.0041666	Straightforward	10001
4	60.00833275	Prime	10002
10	80.09993758	WindowShopper	10003
7	65.03076195	Straightforward	10004
9	80.00624976	WindowShopper	10005
13	105.0190459	Prime	10006
11	89.05054744	Basic	10007
8	75.00666637	Prime	10008
6	62.07253821	WindowShopper	10009
16	140	Not defined	10010
3	50.03998401	Straightforward	10011
12	90.00555538	Basic	10012
5	61.00819617	WindowShopper	10013
2	50.009999	Straightforward	10014
1	35.0142828	Not defined	10015
15	120.0374941	Prime	10016

k	Label	
1	Not defined	
2	Straightforward	k=5
3	Straightforward	

All Customers | BI Customers | Goals and Strategies

Ready

Advantages of K-nearest neighbors

A. Knn is simple to implement.
B. Knn executes quickly for small training data sets.
C. performance the same as performance of the Bayes Classifier.
D. Don’t need any prior knowledge about the structure of data in the training set.
E. No retraining is required if the new training pattern is added to the existing training set.

Limitation to K-nearest neighbors algorithm

A. When the training set is large, it may take a lot of space.
B. For every test data, the distance should be computed between test data and all the training data. Thus a lot of time may be needed for the testing.

Knn Algorithm Pseudocode:

1. Calculate “$d(x, x_i)$” i =1, 2, ..., **n**; where **d** denotes the Euclidean distance between the points.
2. Arrange the calculated **n** Euclidean distances in non-decreasing order.
3. Let **k** be a +ve integer, take the first **k** distances from this sorted list.
4. Find those **k**-points corresponding to these **k**-distances.
5. Let $\mathbf{k}_i$ denotes the number of points belonging to the i^{th} class among **k** points i.e. $k \geq 0$
6. If $k_i > k_j \ \forall \ i \neq j$ then put x in class i.

RANDOM FOREST

Application of Random Forest

.it creates a forest using multiple decision trees randomly

Remote Sensing

Used in ETM devices to acquire images of the earth's surface.

Accuracy is higher and training time is less

Object Detection

Multiclass object detection is done using Random Forest algorithms

Provides better detection in complicated environments

Kinect

Random Forest is used in a game console called Kinect

Tracks body movements and recreates it in the game

Enhanced Thematic Mapper

User performs a step

Kinect registers the movement

Marks the user based on accuracy

Training set to identify body parts

User performs a step

Kinect registers the movement

Marks the user based on accuracy

Training set to identify body parts

Random forest classifier learns

Identifies the body parts while dancing

Score game avatar based on accuracy

Why Random Forest?

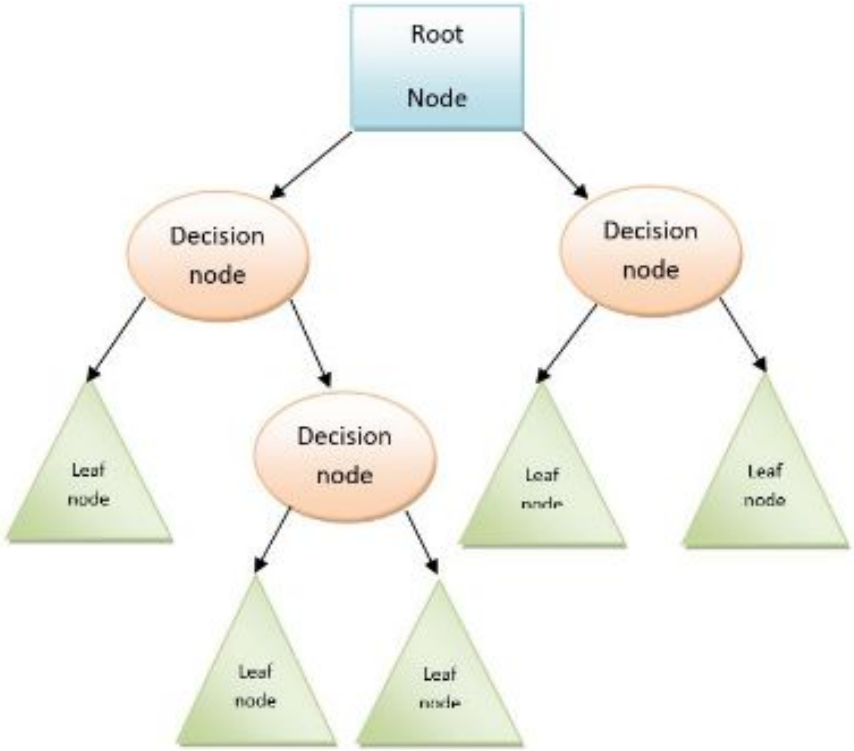
when we train it with a lot of data, When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set. Then the model does not categorize the data correctly, because of too much of details and noise.

Use of multiple trees reduce the risk of overfitting

Training time is less

?What is a Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a test or a condition on an attribute, each branch represents an outcome of the test and each leaf/terminal node holds a class label. It is considered to be a **non-parametric** method which means that it makes **no assumptions about the space** distribution and the classifier structure.



Terminologies related to Decision Tree

- **Root node:** the top most tree node which divides into two homogeneous sets.
- **Decision node:** a sub-node which further splits into other two sub-nodes.
- **Terminal/Leaf node:** the lowermost nodes or the nodes with no children that represents a class label (decision taken after computing all attributes)
- **Splitting**: dividing a node into two or more nodes. The splitting technique results in fully grown trees until the criteria of a class attribute are met. But a fully grown tree is likely to over-fit the data which leads to poor accuracy on unseen observations. This is when Pruning comes into the picture.
- **Pruning**: Process of reducing the size of the tree by removing the nodes which play a minimal role in classifying an instance without reducing the predictive accuracy as measured by a cross-validation set.

- it cannot capture the underlying trend of the data. Underfitting **destroys the accuracy of** our machine learning model.
- our model or the algorithm does not fit the data well enough.
- It usually happens when we have less data to build an accurate model and also when we try to build a linear model with a non-linear data.

Overfitting

when we train it with a lot of data.

When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set.

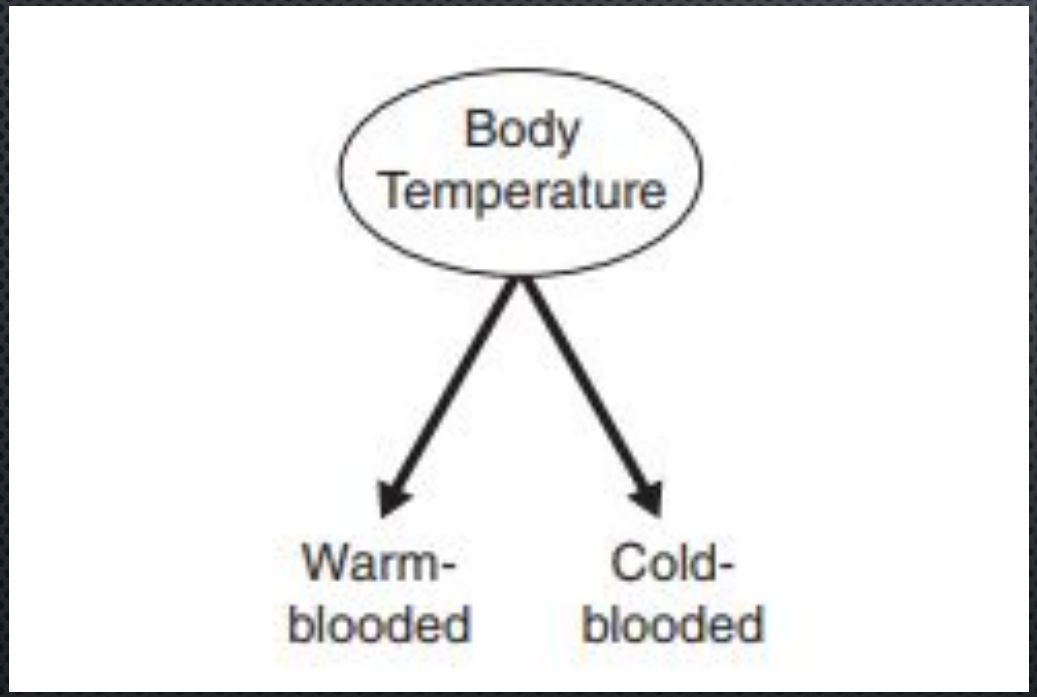
The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models.

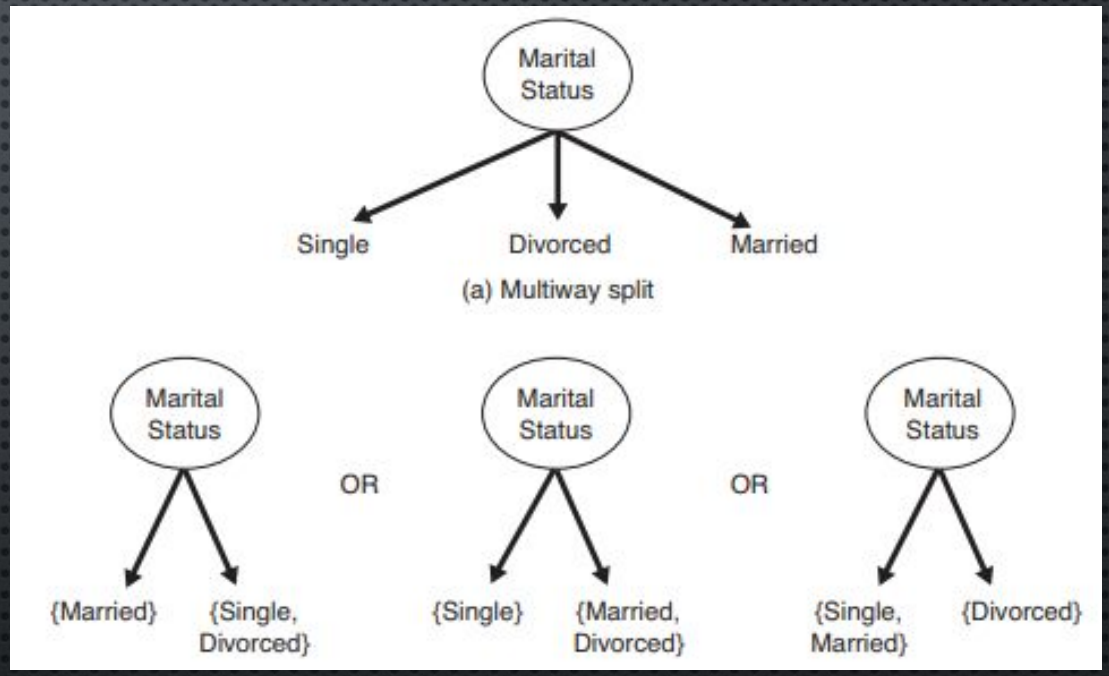
Pseudo-code for Decision Tree Algorithm

1. Select the most powerful attribute as the root node.
2. Split the training set into sub-nodes such that each sub-node has identical attribute values.
3. Repeat Step 1 and 2 until you meet the criteria of the class attribute.
4. Remove unwanted nodes if you have a fully grown tree such that it doesn’t affect the prediction accuracy.

Methods of Splitting Attribute Test Condition

Binary attributes: a test condition that generates•
.only two potential outcomes

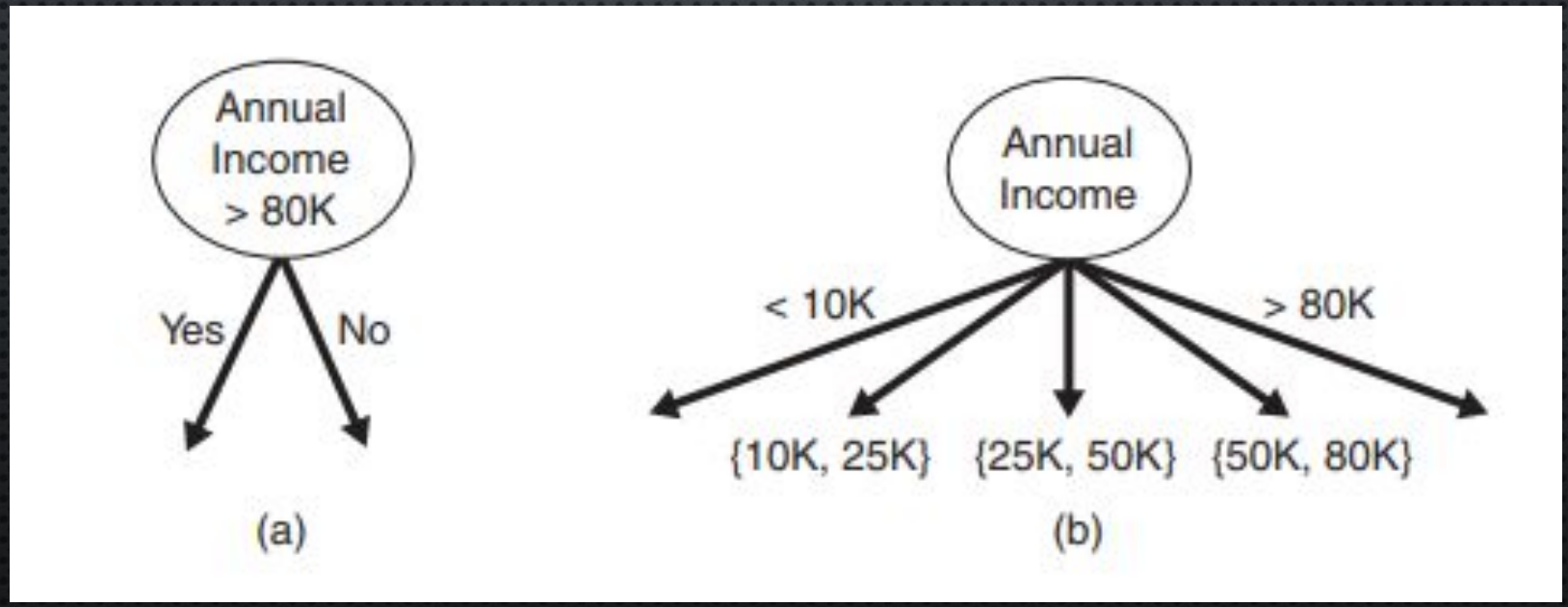




Nominal attributes: It can have multiple outcomes depending upon the number of distinct values that the corresponding attribute can hold.

Ordinal Attributes: it can have binary or multiple potential outcomes.

Continuous attributes: test condition can have either binary outcomes or different ranges of outcomes.



No overfitting

Use of multiple trees reduce the risk of overfitting

Training time is less

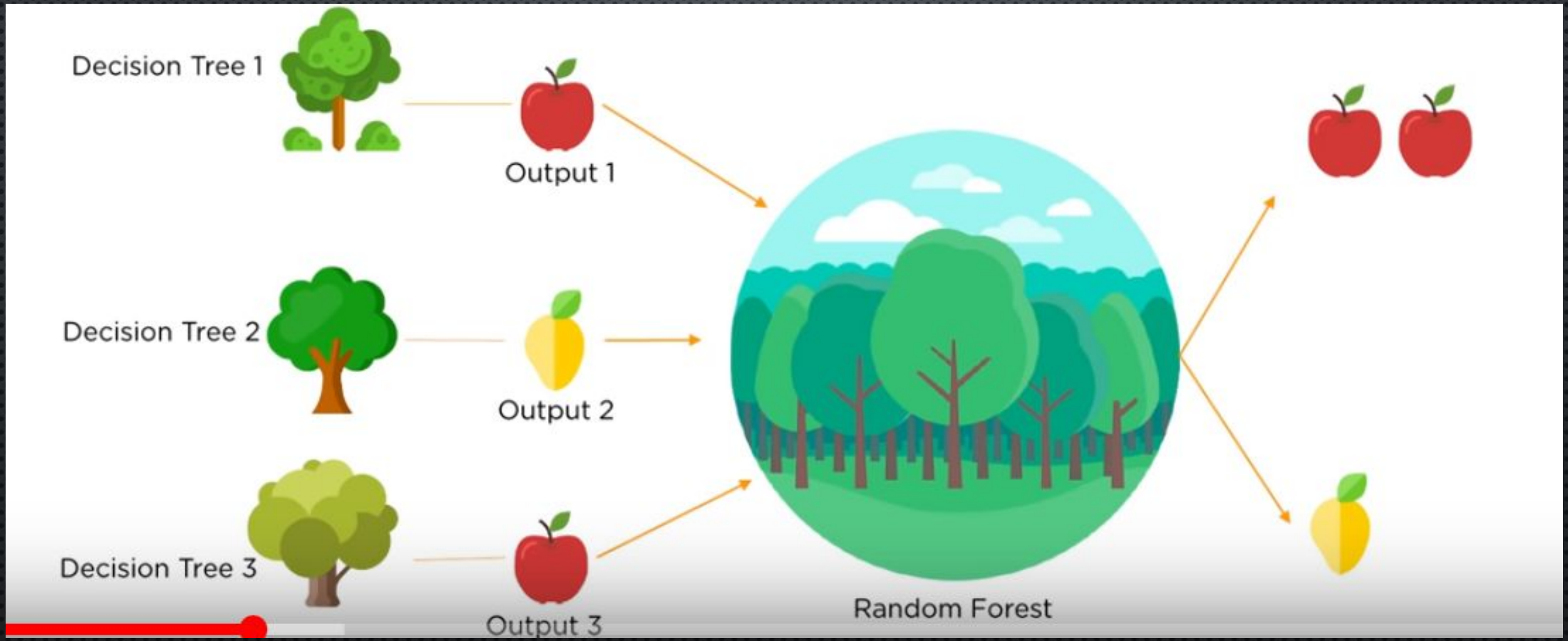
High accuracy

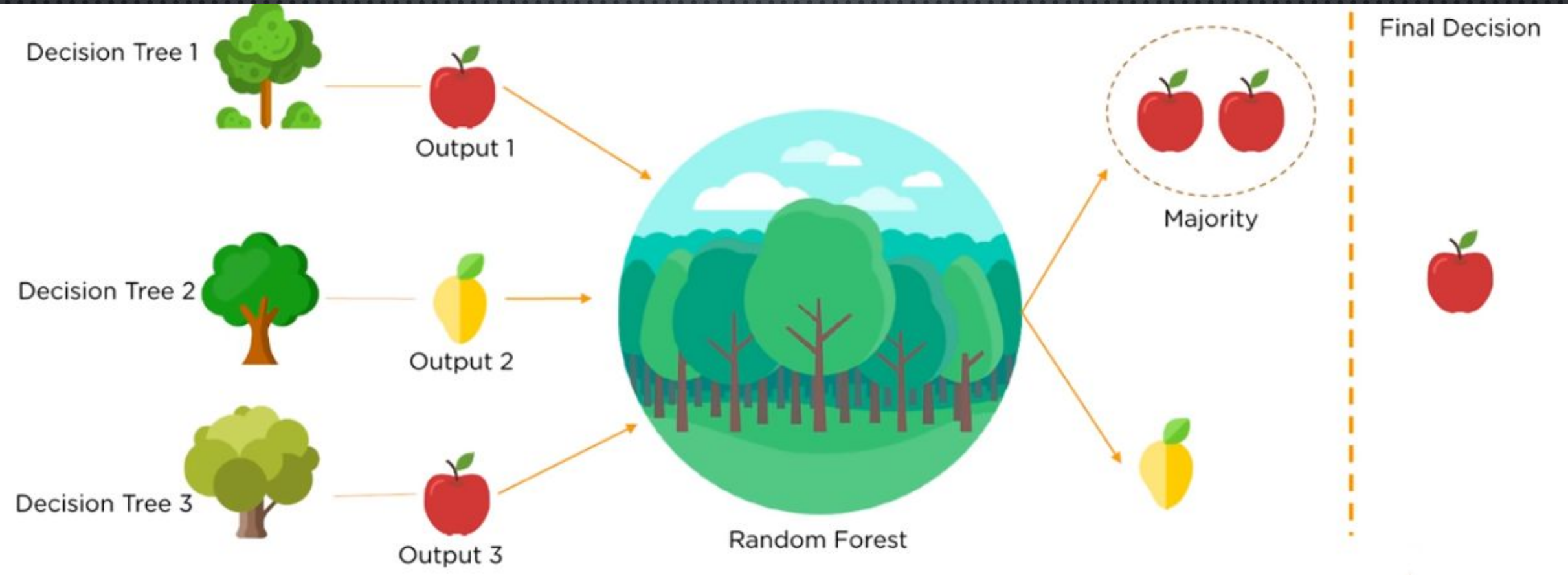
Runs efficiently on large database

For large data, it produces highly accurate predictions

Estimates missing data

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


Decision Tree 1
Output 1
Decision Tree 2
Output 2
Decision Tree 3
Output 3
Random Forest


Decision Tree 1
Output 1
Decision Tree 2
Output 2
Decision Tree 3
Output 3
Random Forest
Majority
Final Decision


Random Forest and Decision Tree
simplilearn
SUBSCRIBE

Decision Tree- Important Terms

Entropy

Entropy is the measure of randomness or unpredictability in the dataset

Information gain

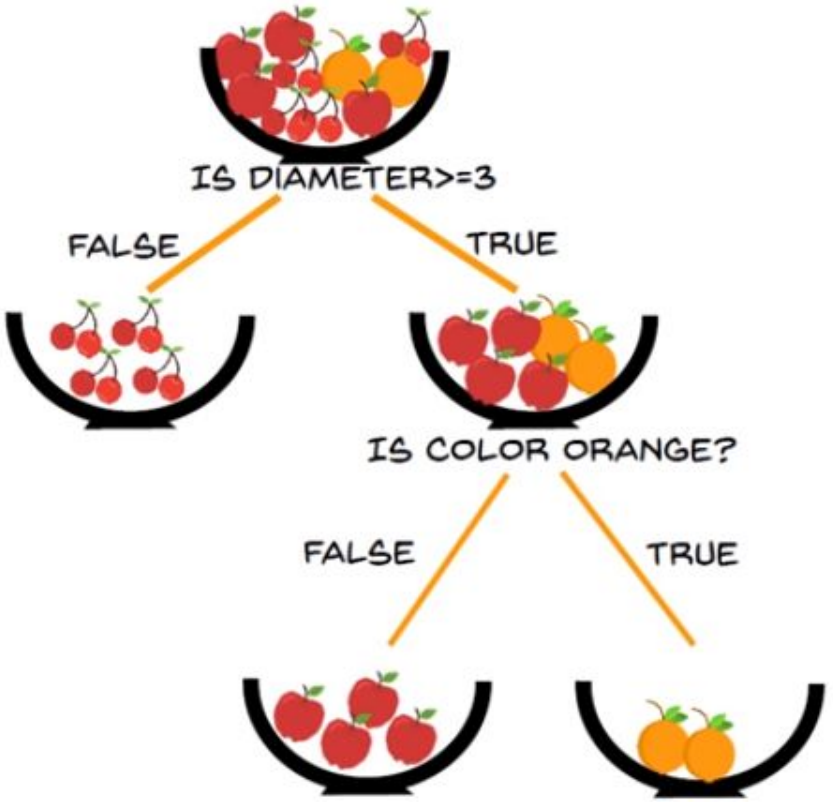
Leaf Node

Decision Node

Root Node

Decision Tree

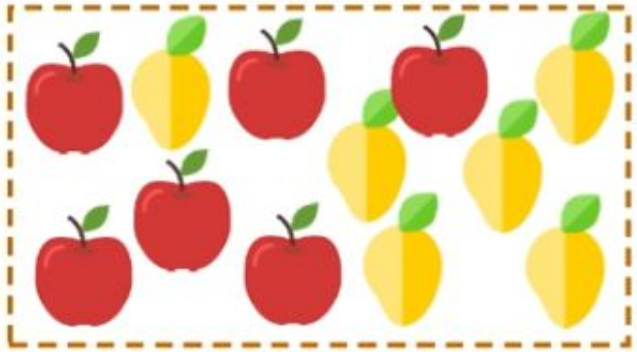
Decision Tree is a tree shaped diagram used to determine a course of action. Each branch of the tree represents a possible decision, occurrence or reaction



Decision Tree - Important Terms

Entropy

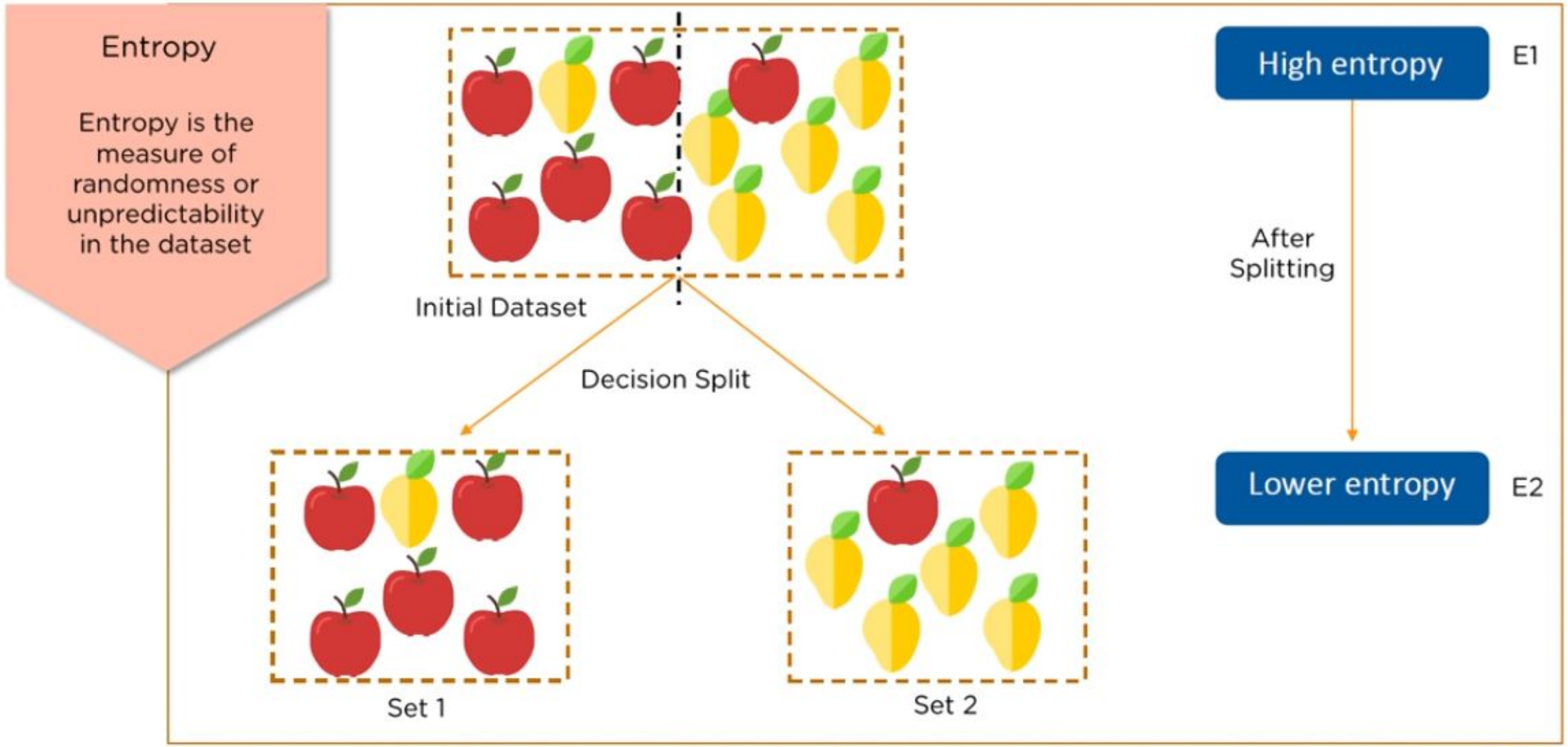
Entropy is the measure of randomness or unpredictability in the dataset



High entropy

E1

Decision Tree - Important Terms



Decision Tree - Important Terms

Entropy

Information gain

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

Leaf Node

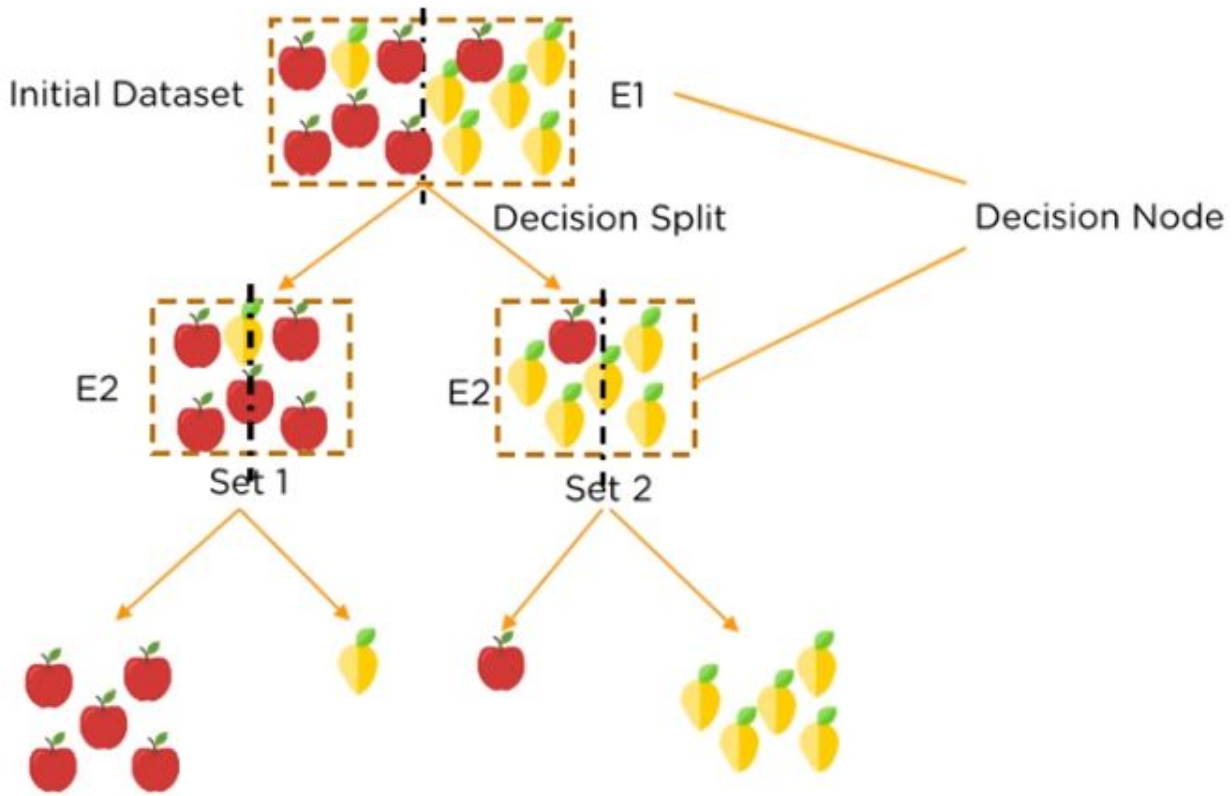
Decision Node

Root Node

Decision Tree - Important Terms

Decision Node

Decision node has two or more branches



How does a Decision Tree work?

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

How does a Decision Tree work?

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 does a Decision Tree work?

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

How does a Decision Tree work?

CONDITIONS

COLOR== PURPLE?

DIAMETER=3

COLOR== YELLOW?

COLOR== RED?

DIAMETER=1

TRAINING DATASET

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

How does a Decision Tree work?

CONDITIONS

COLOR== PURPLE?

DIAMETER=3

COLOR== YELLOW?

COLOR== RED?

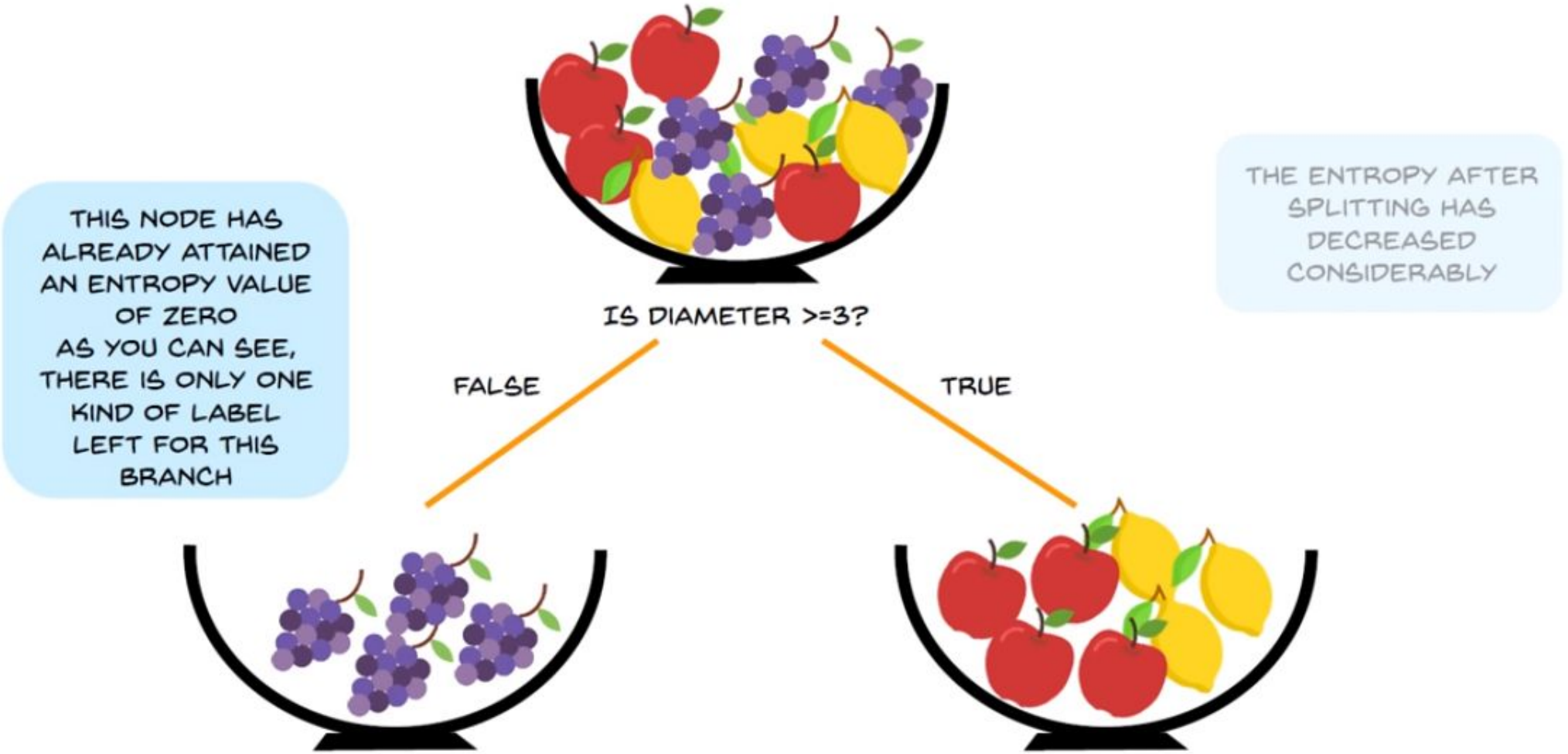
DIAMETER=1

LET'S SAY THIS CONDITION GIVES US THE MAXIMUM GAIN

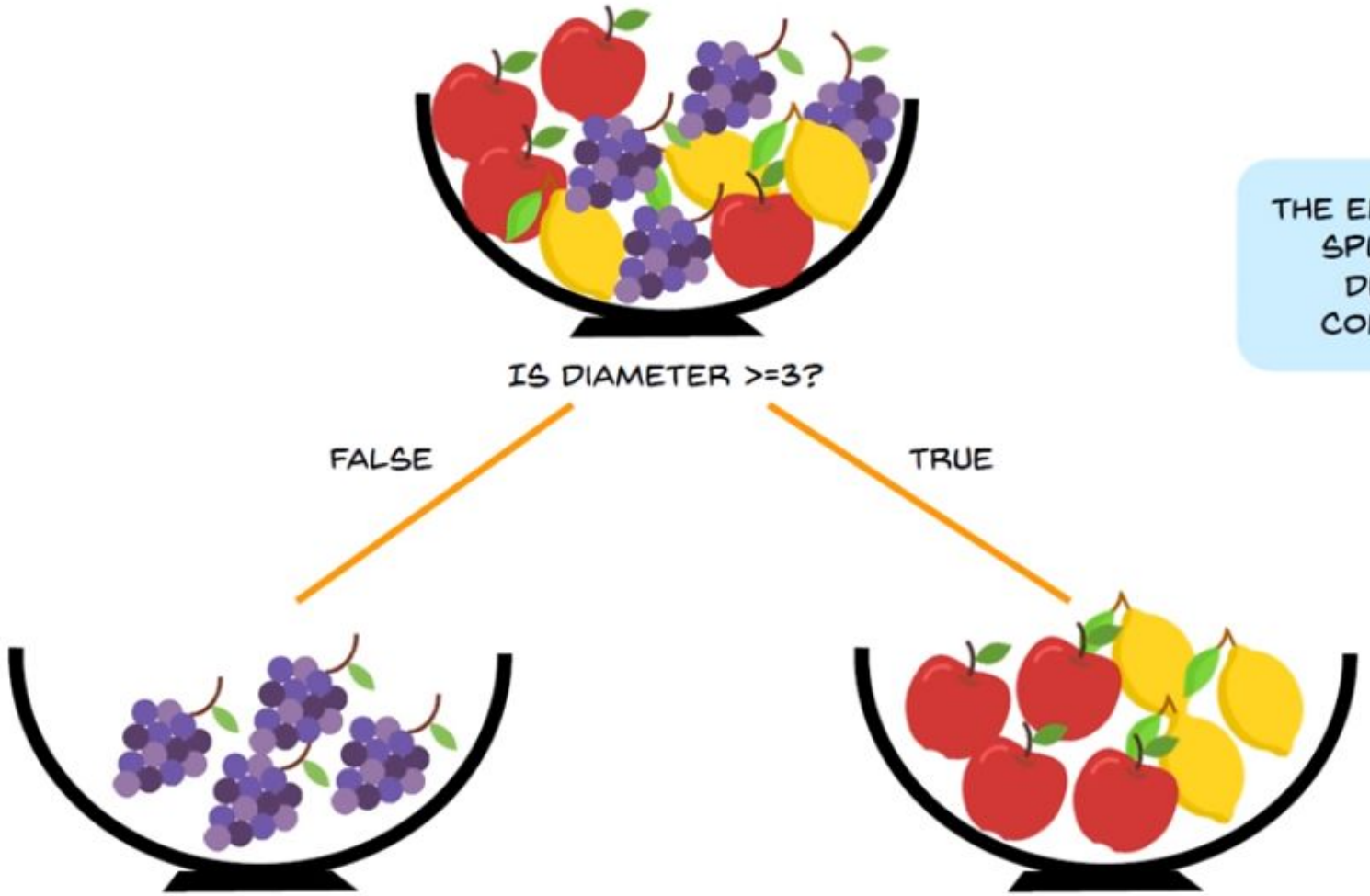
TRAINING DATASET

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

How does a Decision Tree work?

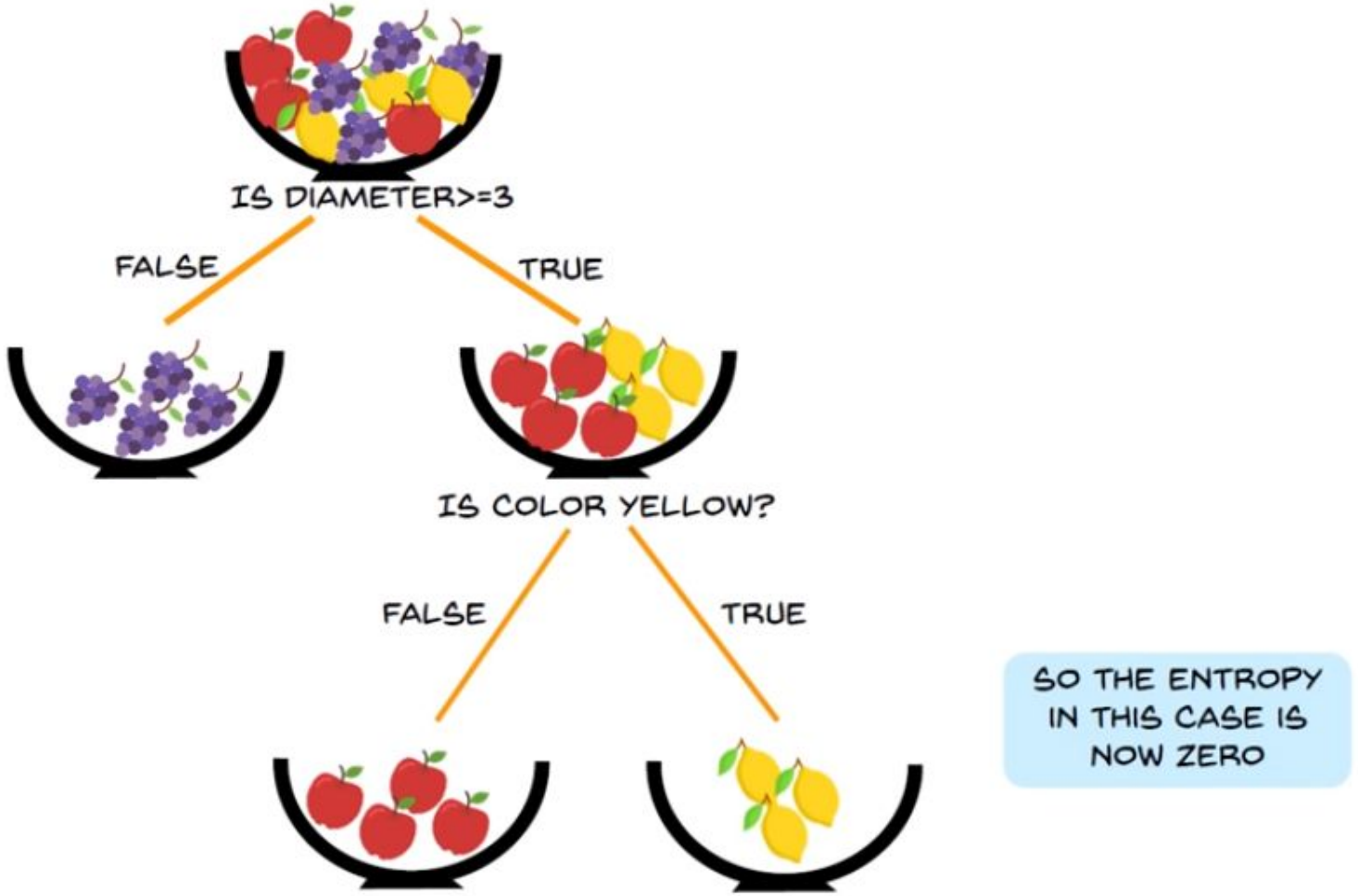


How does a Decision Tree work?

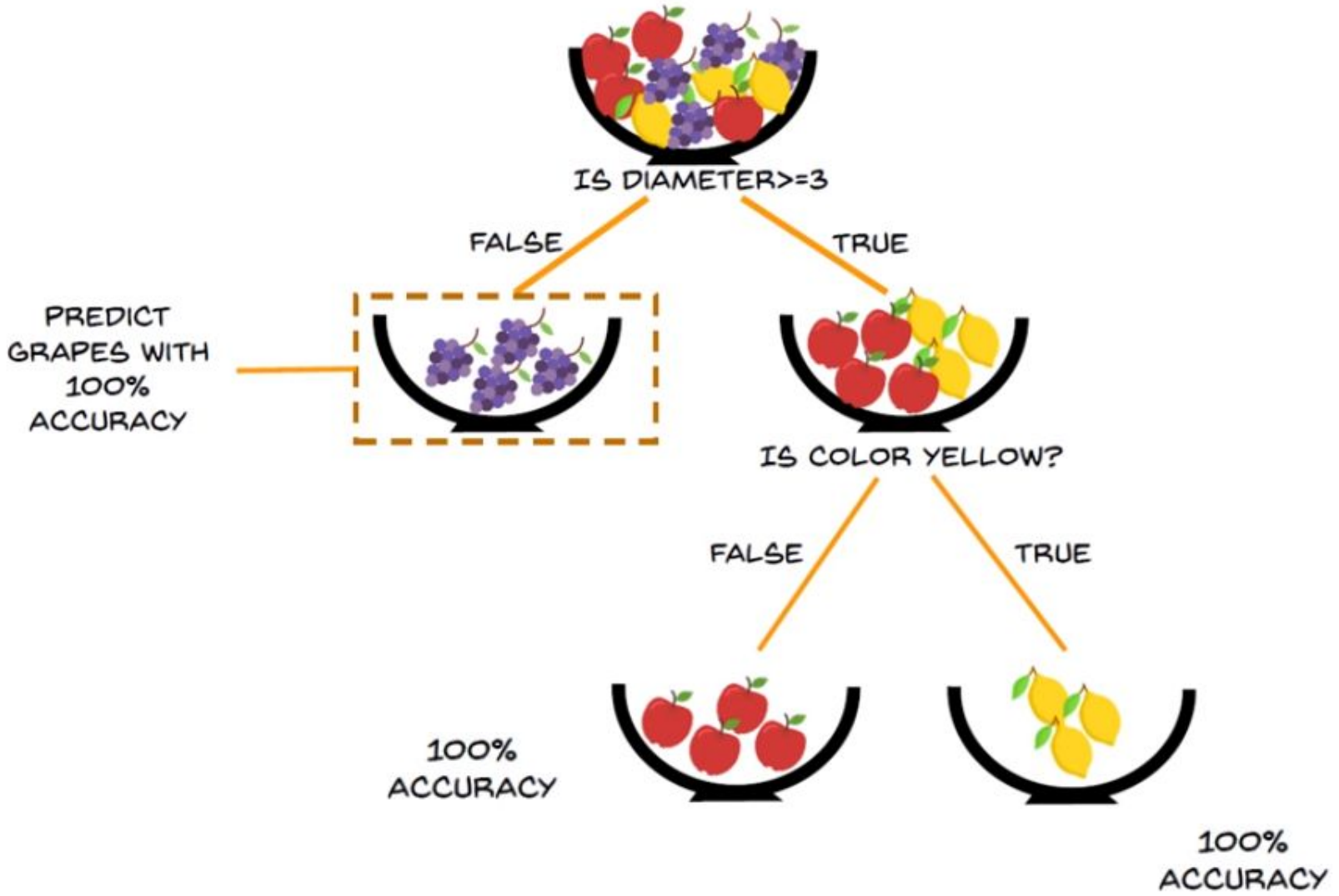


THE ENTROPY AFTER SPLITTING HAS DECREASED CONSIDERABLY

How does a Decision Tree work?

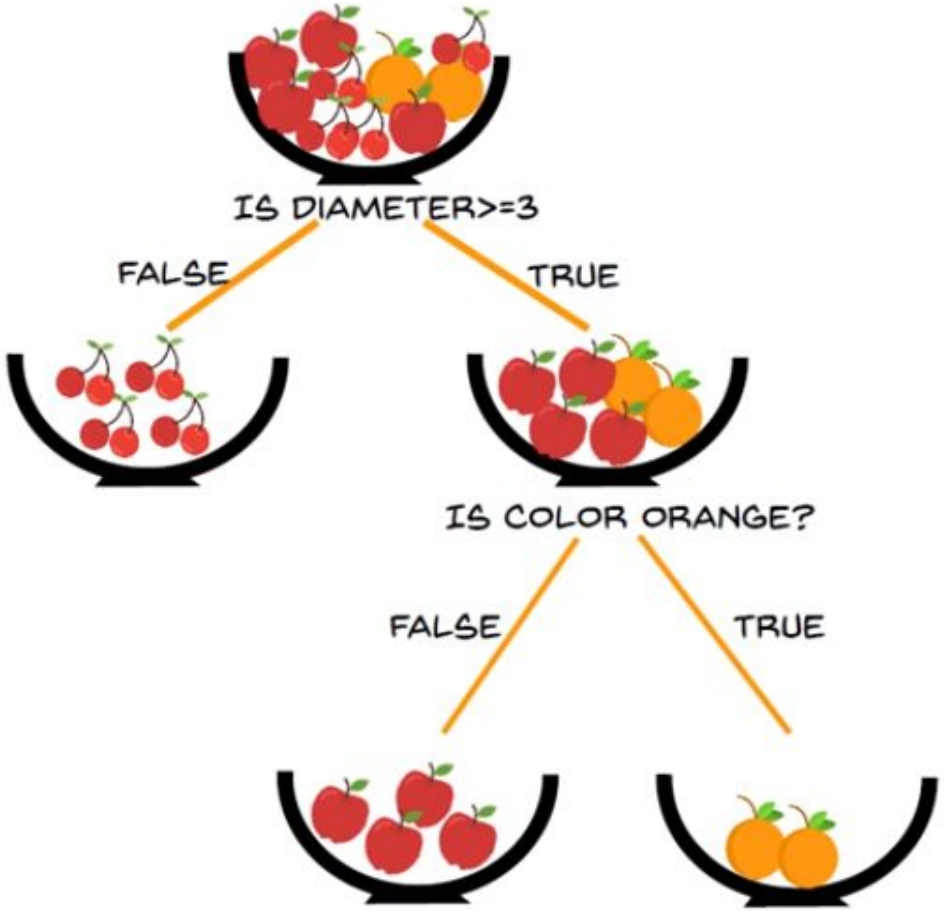


How does a Decision Tree work?



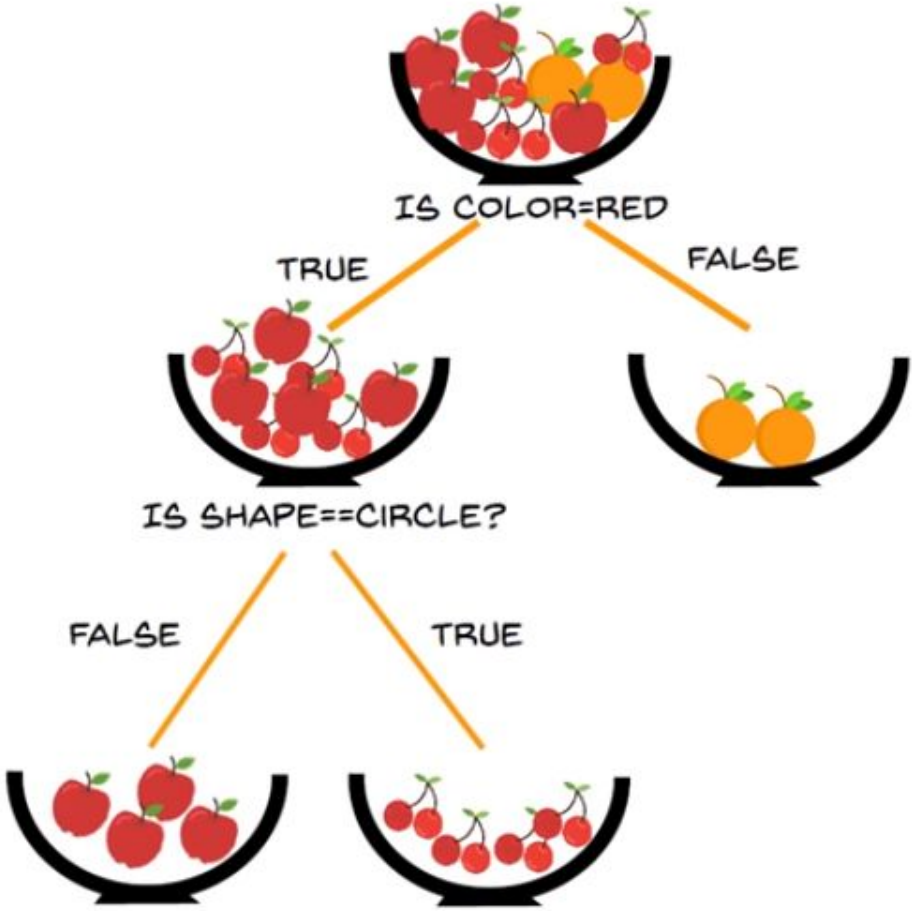
How does a Random Forest work?

LET THIS BE TREE 1



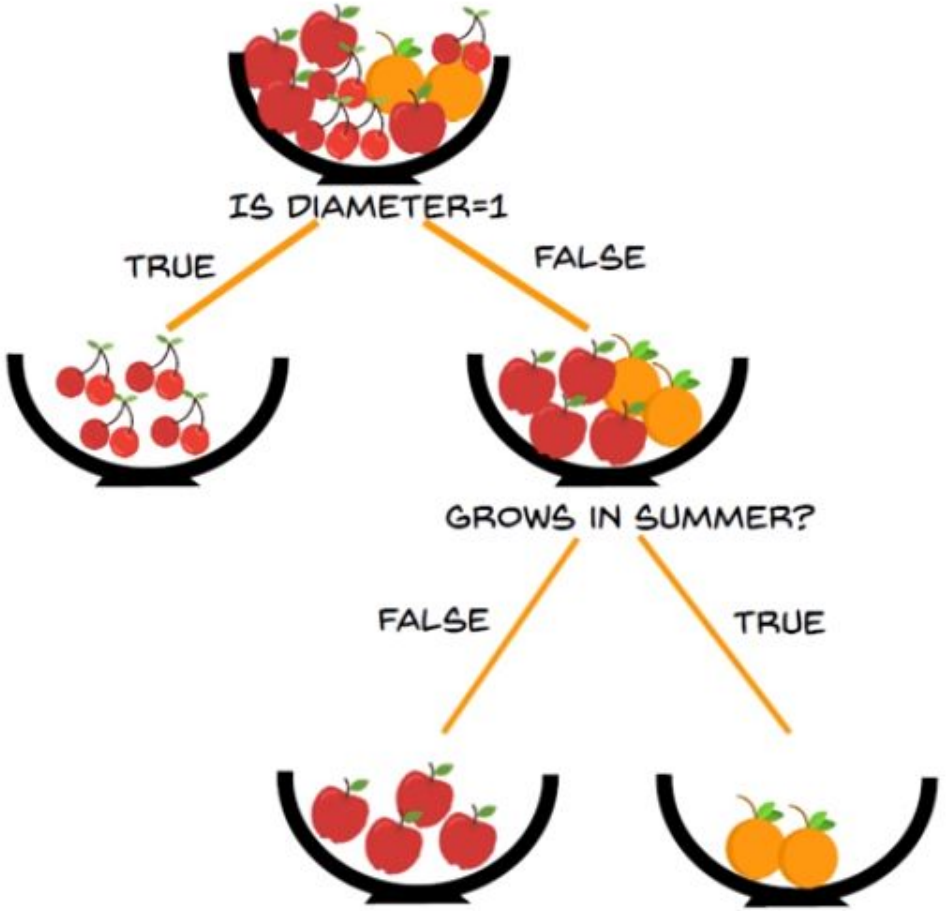
How does a Random Forest work?

LET THIS BE TREE 2

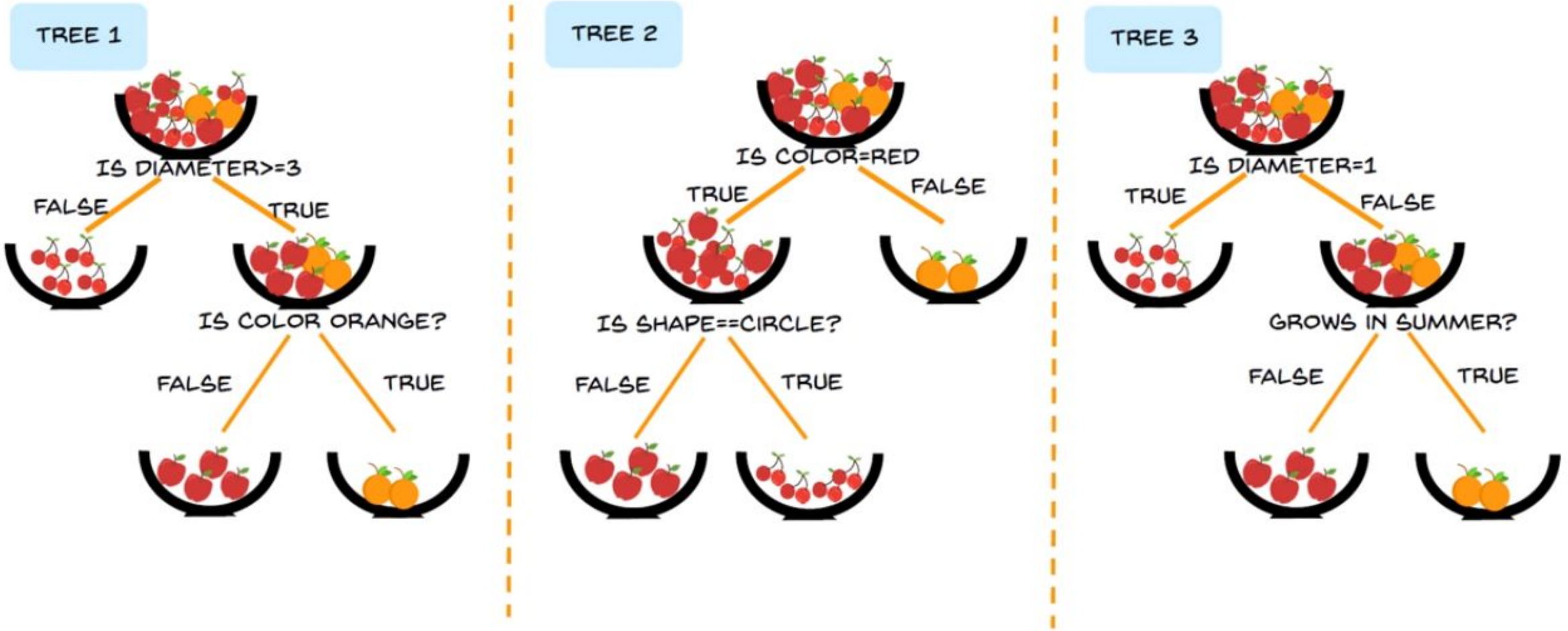


How does a Random Forest work?

LET THIS BE TREE 3



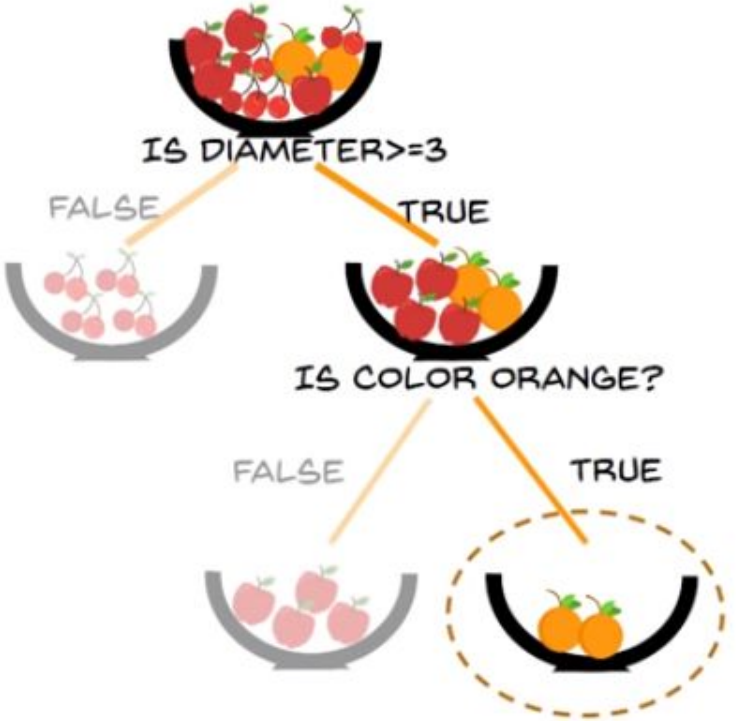
How does a Random Forest work?



How does a Random Forest work?

TREE 1 CLASSIFIES IT AS AN ORANGE

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



How does a Random Forest work?

Properties of Random Forest Algorithm

- ✔ Random forest is a predictive modeling algorithm
- ✔ The random forest can be used for both classification and regression tasks.
- ✔ It works well with default hyper-parameters.
- ✔ It can be used to rank the importance of variables in a regression or classification problem.
- ✔ The correlation between any two trees in the forest. Increasing the correlation increases the forest error rate.
- ✔ A tree with a low error rate is a strong classifier. Increasing the strength of the individual trees decreases the forest error rate.
- ✔ It runs efficiently on large datasets.