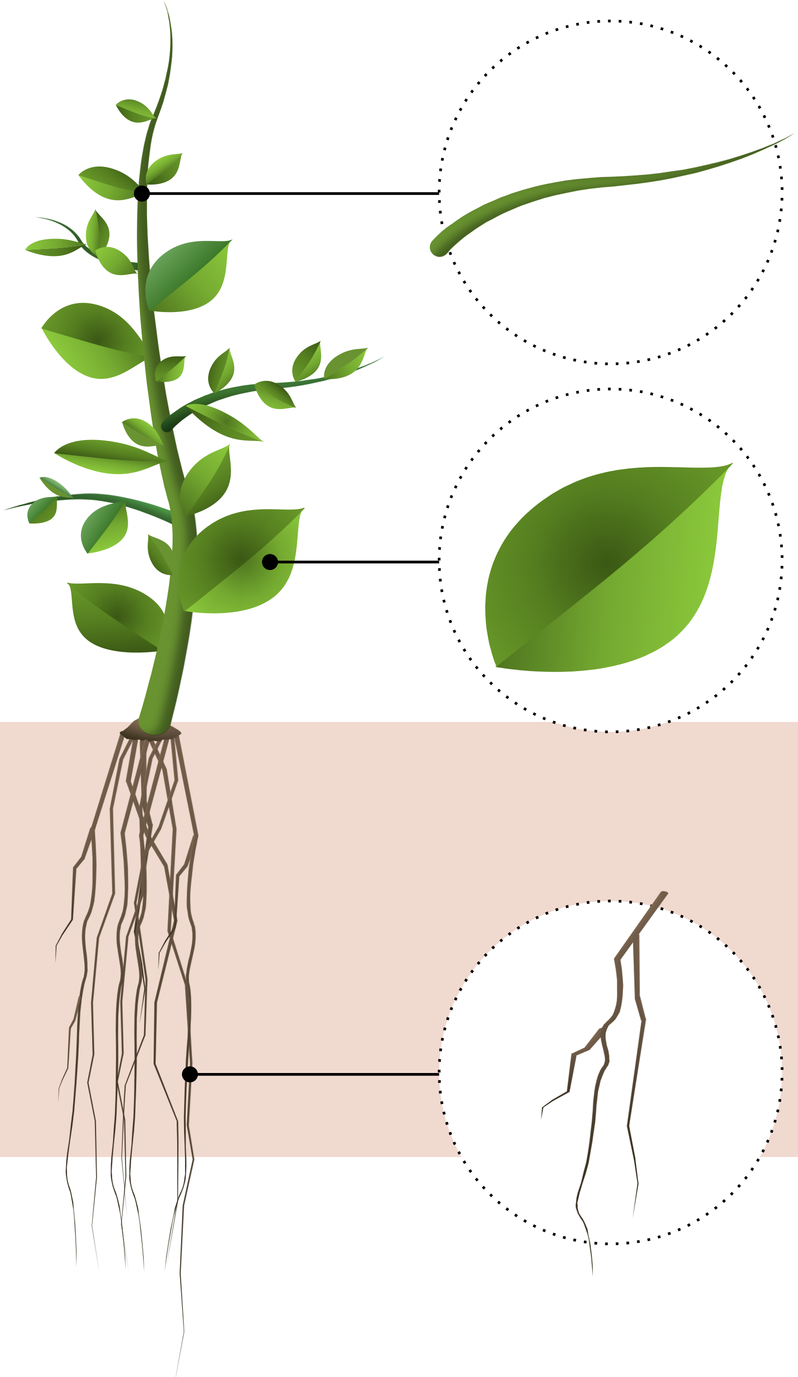
3

Decision Tree

FDP ANN & ML 2023

Dr. Uday Pratap Singh
Associate Professor
PIET, Jaipur



Decision

Leaf Node

Choice

Branch Node

Variable that Best Splits Data

Root Node

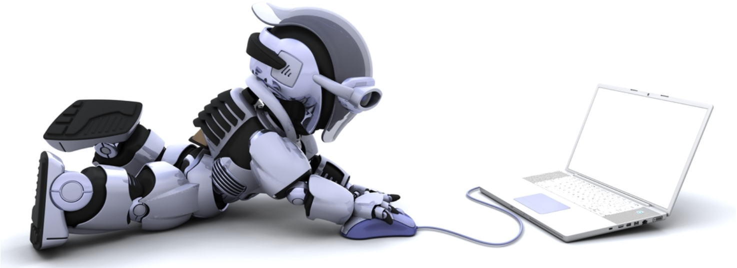
Agenda for Today Session

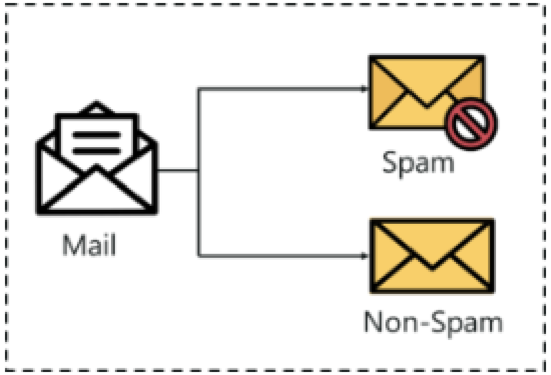
- What is Classification?
- Types of Classification
- Classification Use Case
- What is Decision Tree?
- Decision Tree Terminology
- Visualizing a Decision Tree
- Writing a Decision Tree Classifier from Scratch in Python using CART Algorithm

What is Classification

What is Classification?

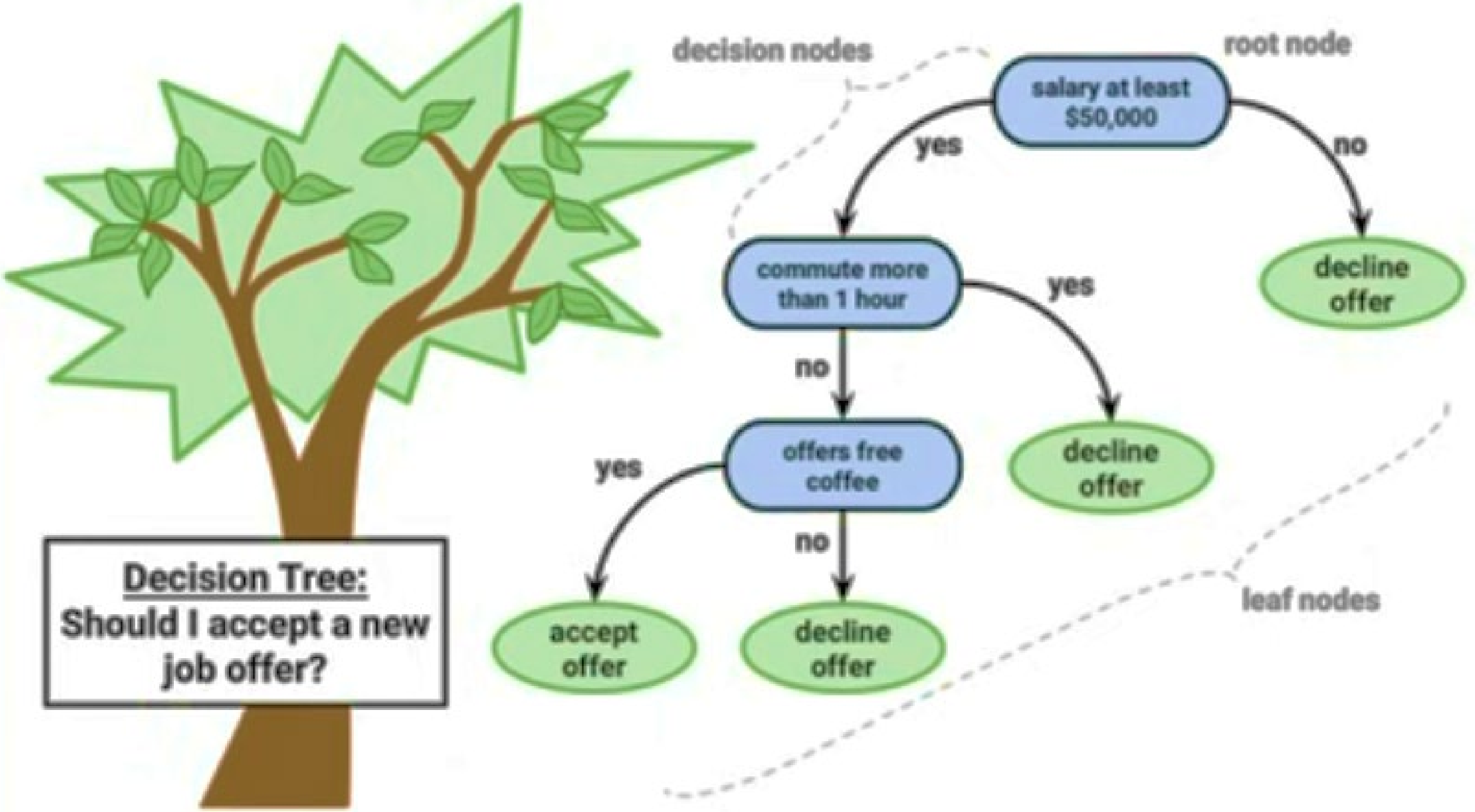
"Classification is *a supervised machine learning process of categorizing a given set of input data into classes based on one or more variables.*"

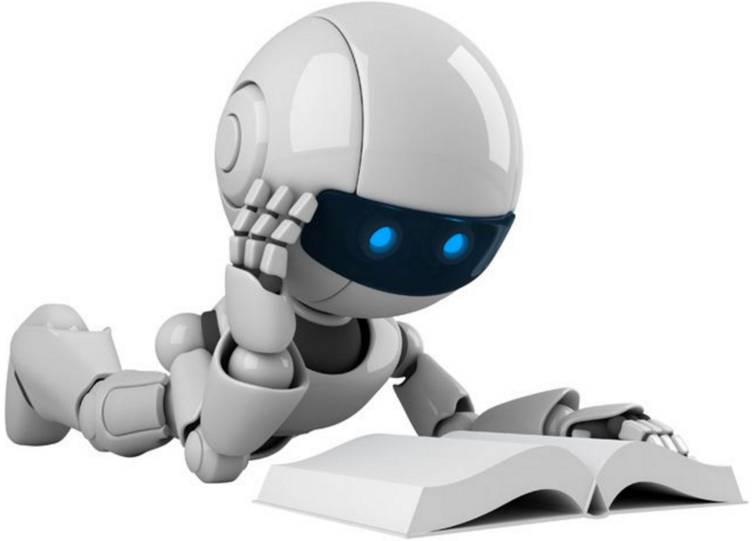




What is Decision Tree?

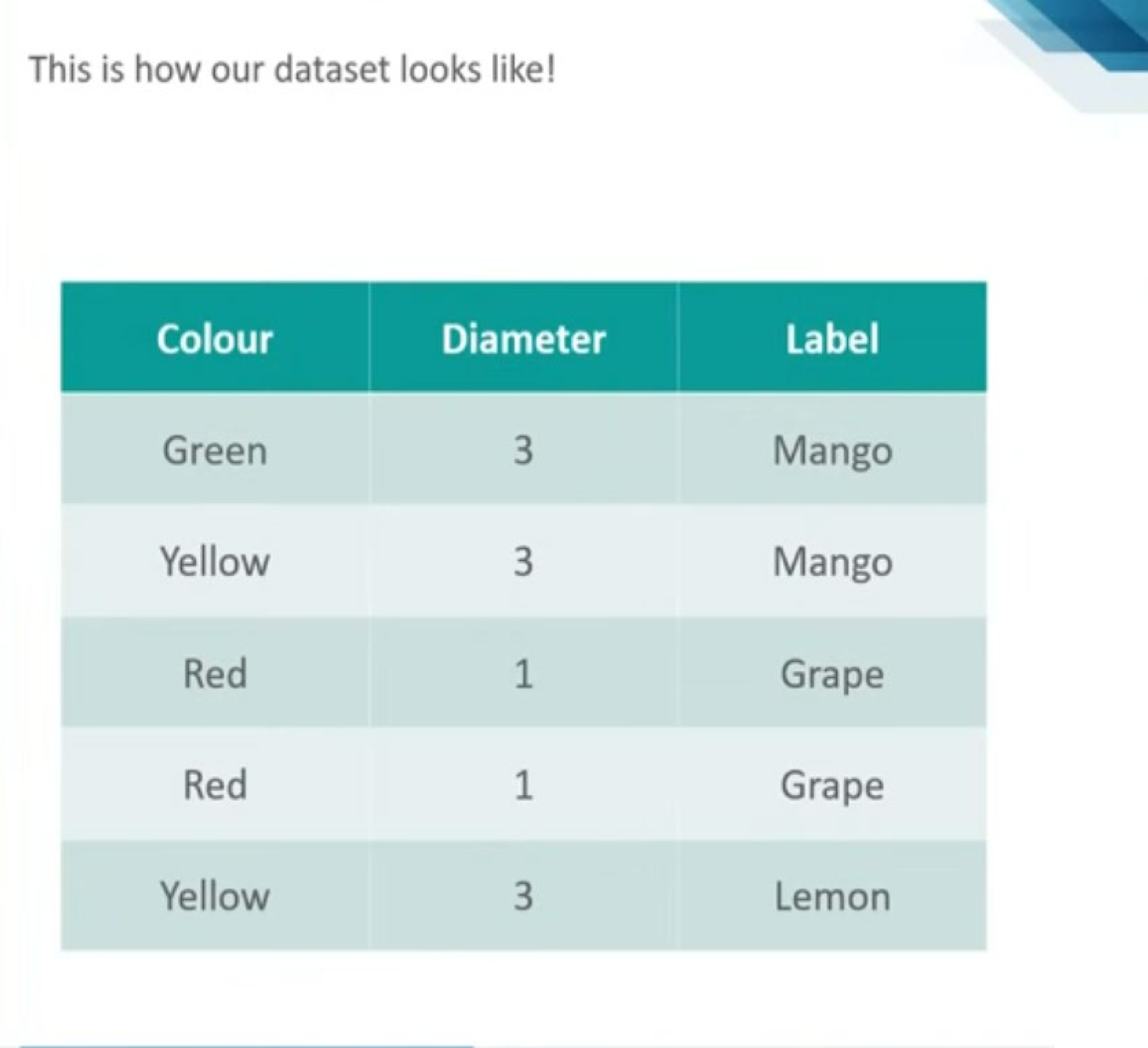
"A **decision tree** is a graphical representation of all the possible solutions to a decision based on certain conditions"





Understanding Decision Tree

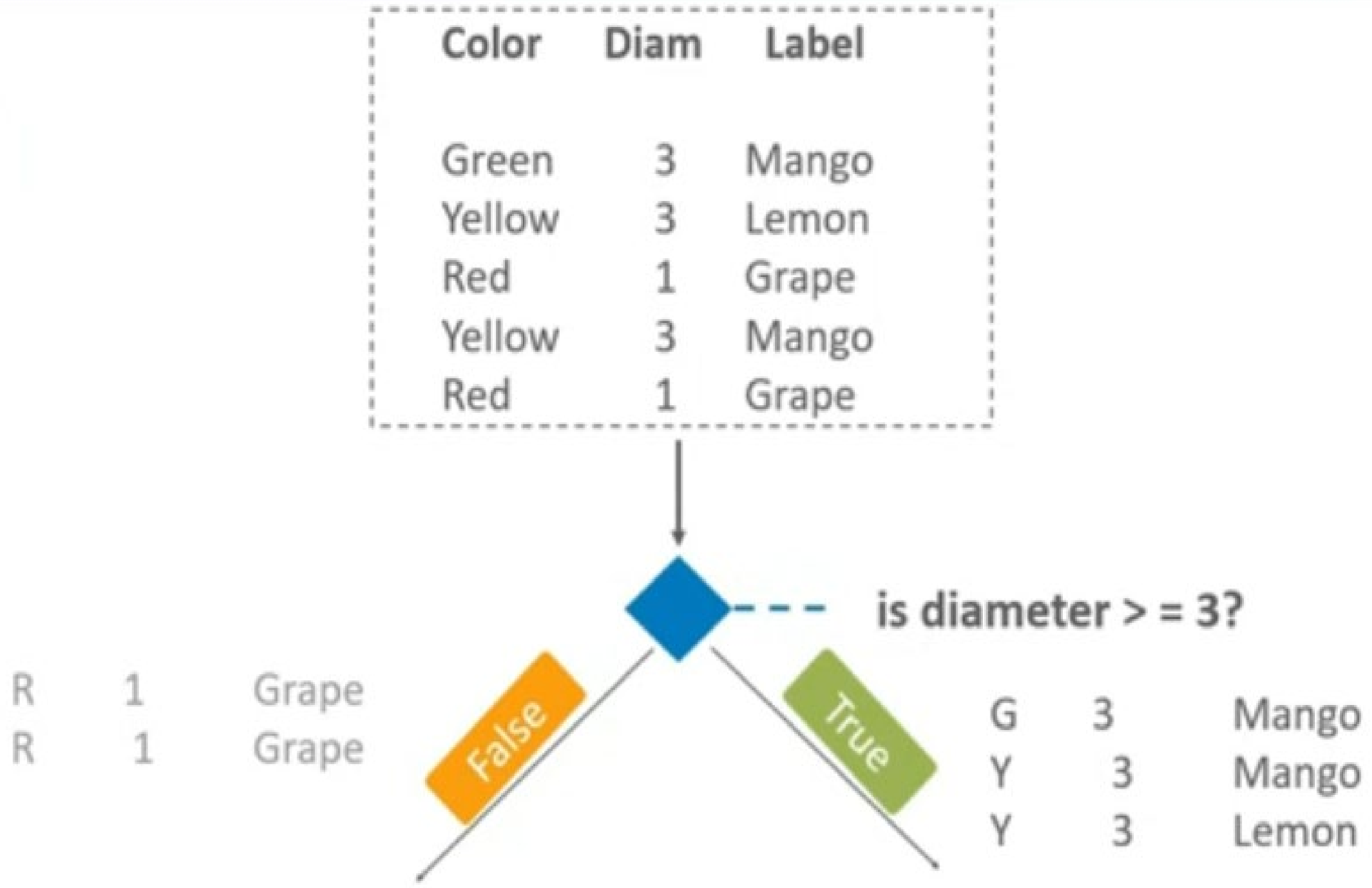
Data Set



This is how our dataset looks like!

Colour	Diameter	Label
Green	3	Mango
Yellow	3	Mango
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

Decision Tree



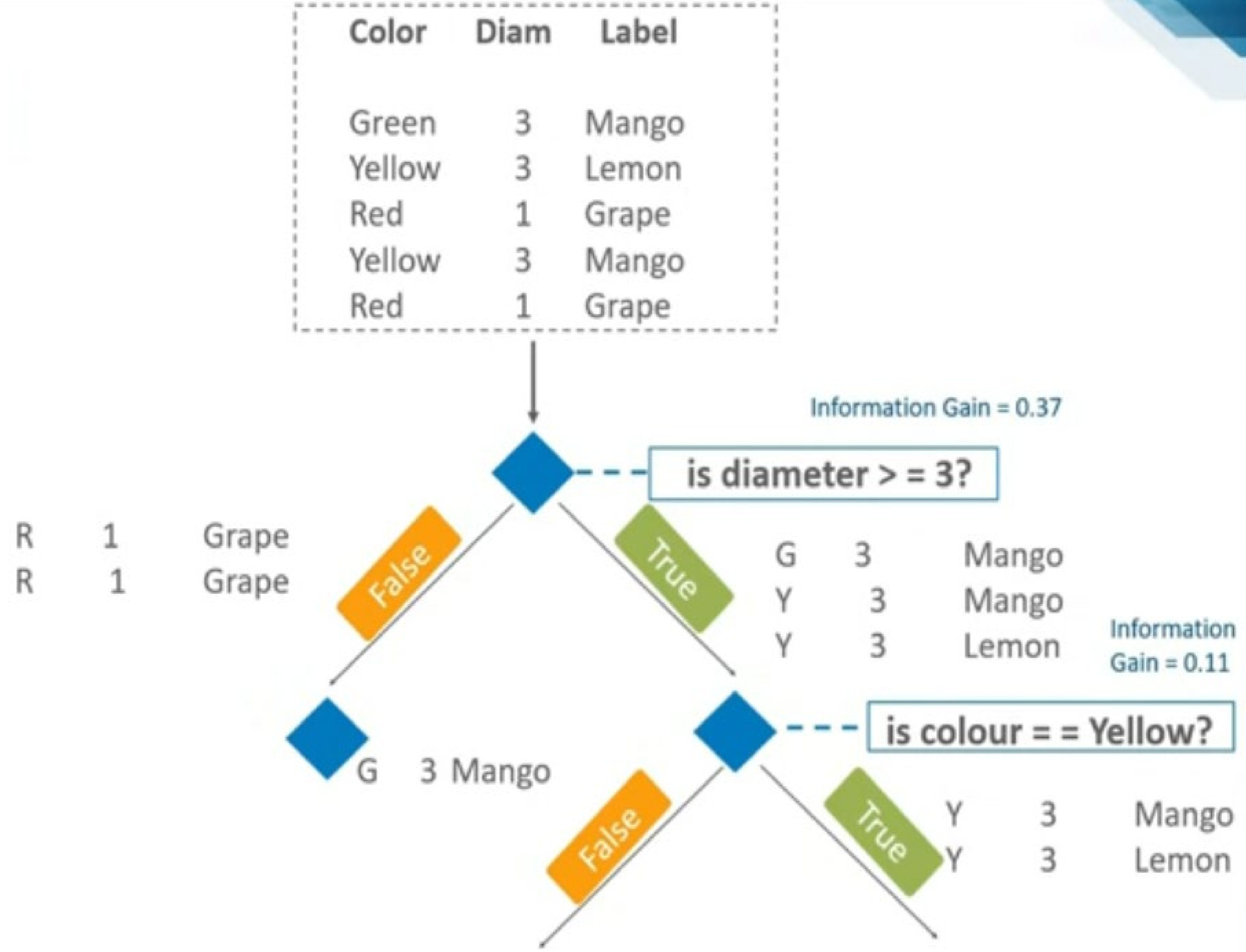
Color	Diam	Label
Green	3	Mango
Yellow	3	Lemon
Red	1	Grape
Yellow	3	Mango
Red	1	Grape

is diameter > = 3?

R	1	Grape
R	1	Grape

G	3	Mango
Y	3	Mango
Y	3	Lemon

Decision Tree



Color	Diam	Label
Green	3	Mango
Yellow	3	Lemon
Red	1	Grape
Yellow	3	Mango
Red	1	Grape

Information Gain = 0.37

is diameter > = 3?

R	1	Grape
R	1	Grape

G	3	Mango
Y	3	Mango
Y	3	Lemon

Information Gain = 0.11

is colour = = Yellow?

G 3 Mango

Y	3	Mango
Y	3	Lemon

Decision Tree Terminology

Pruning

Opposite of Splitting, basically removing unwanted branches from the tree

Branch/SubTree

Formed by splitting the tree/node

Splitting

Splitting is dividing the root node/sub node into different parts on the basis of some condition.

Root Node

It represents the entire population or sample and this further gets divided into two or more homogenous sets.

Leaf Node

Node cannot be further segregated into further nodes

How Does A Tree Decide Where To Split?

Gini Index

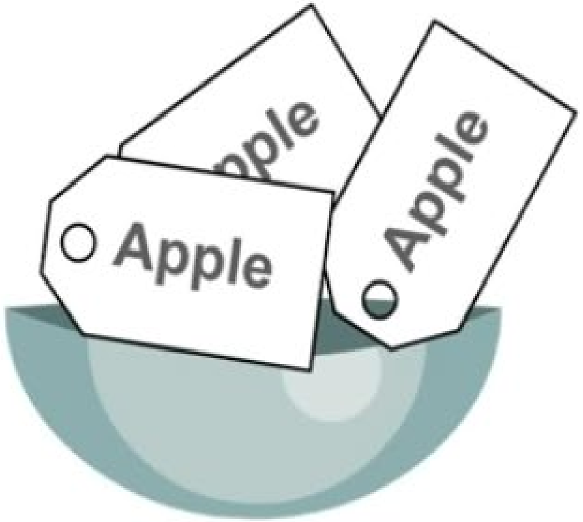
The measure of impurity (or purity) used in building decision tree in CART is Gini Index

$$Gini = 1 - \sum_{i=1}^{j} P(i)^2$$

Information Gain

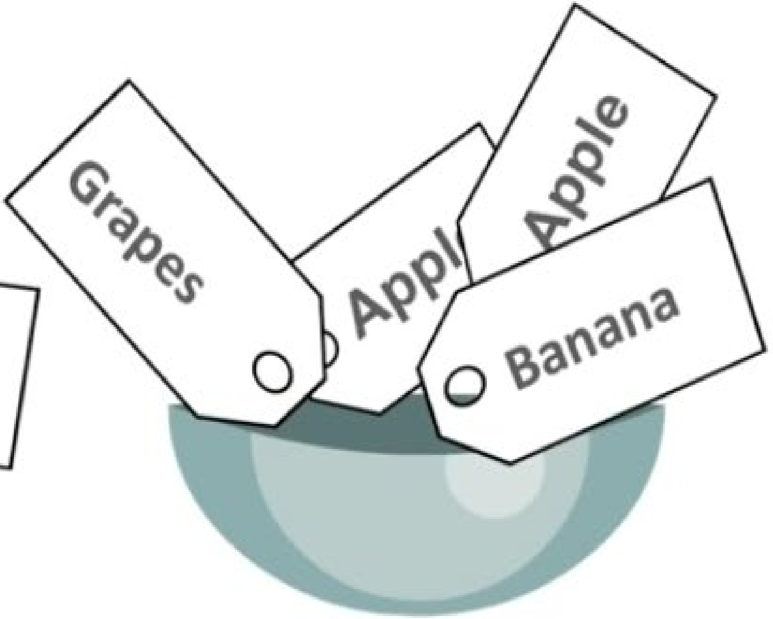
The information gain is the decrease in entropy after a dataset is split on the basis of an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain

Let’s First Understand What is Impurity



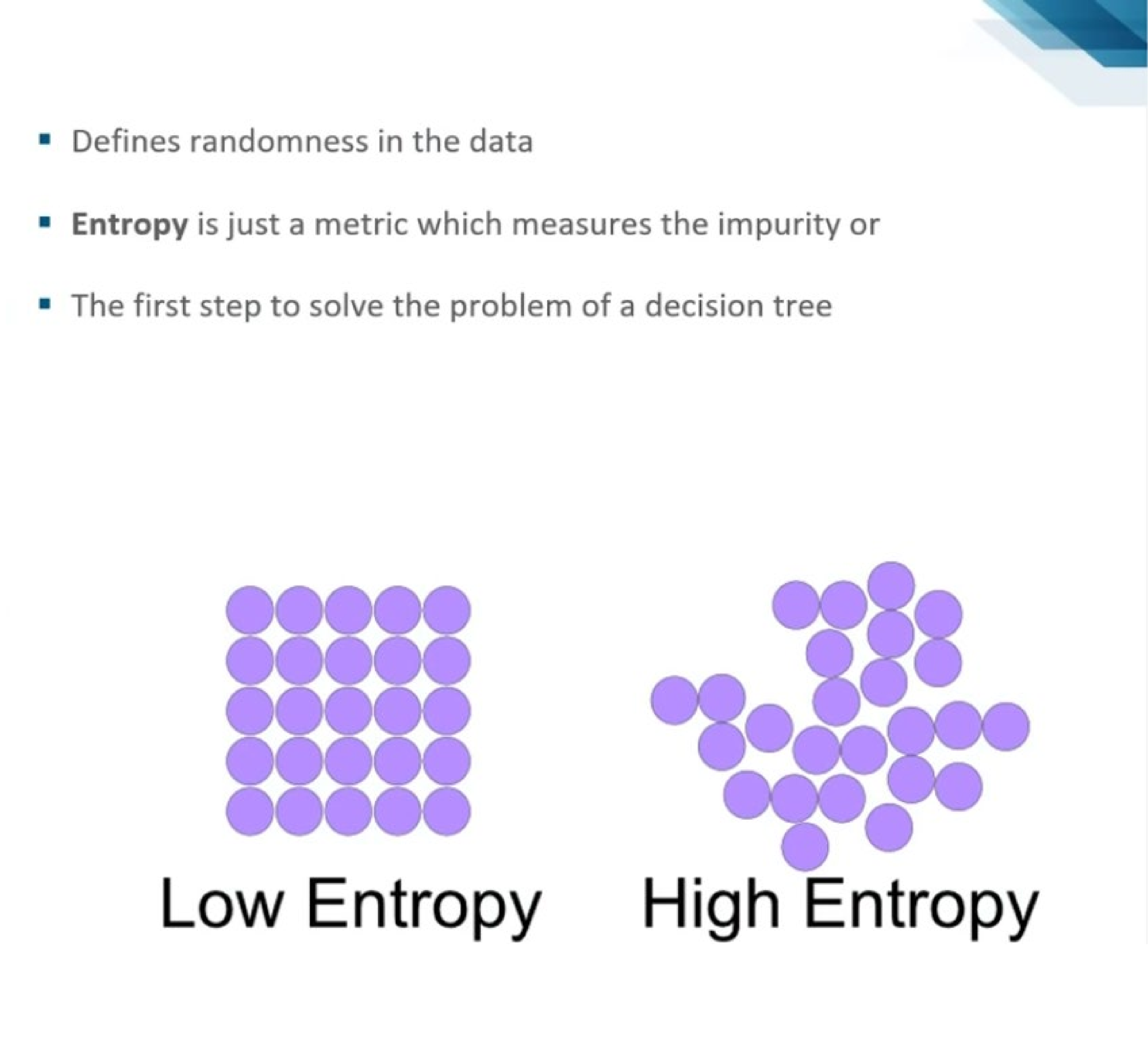
Impurity = 0

Let’s First Understand What is Impurity



Impurity $\neq 0$

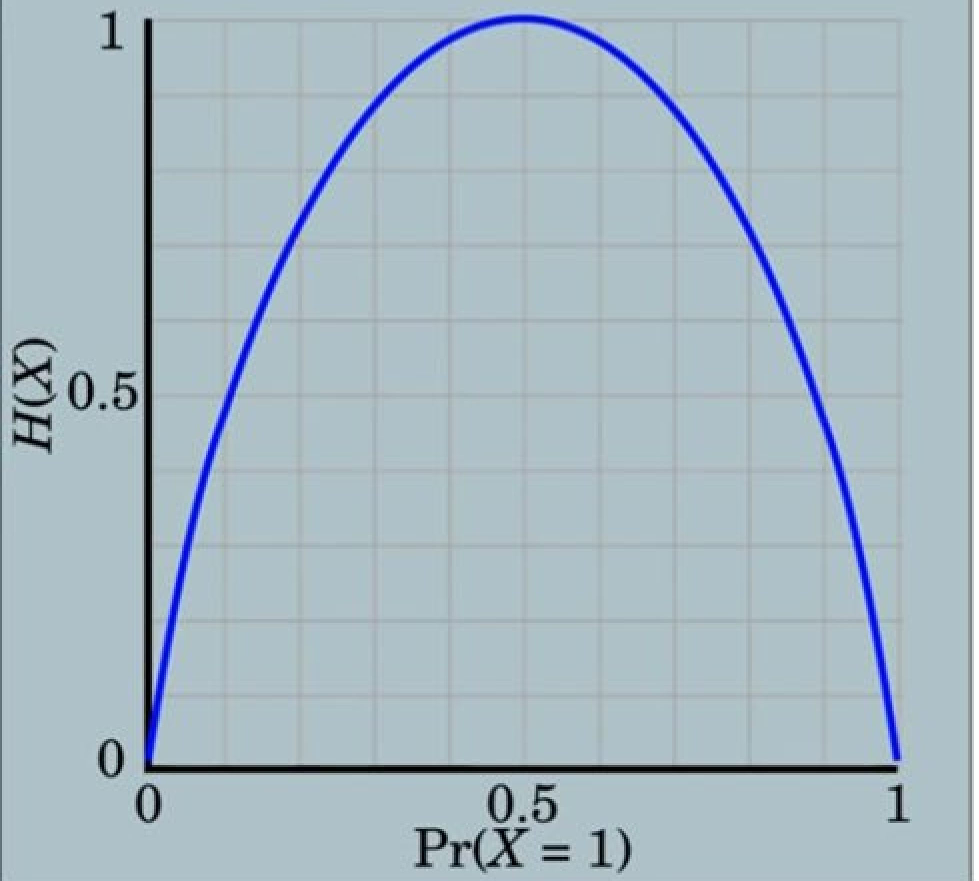
What is Entropy?



- Defines randomness in the data
- **Entropy** is just a metric which measures the impurity or
- The first step to solve the problem of a decision tree

Low Entropy
High Entropy

What is Entropy?



Entropy(s) =- P(yes) $\log_2 \mathrm{P(yes)}$ – P(no) $\log_2 \mathrm{P(no)}$

Where,

- S is the total sample space,
- P(yes) is probability of yes

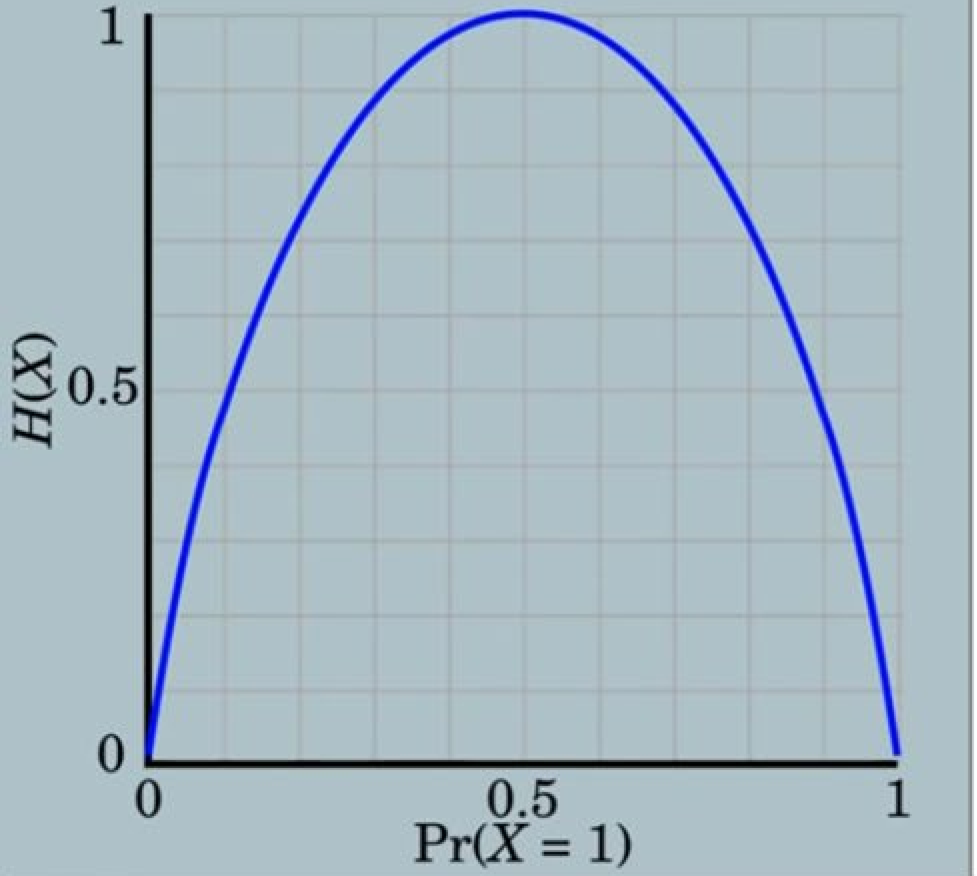
If number of *yes* = number of *no ie P(S) = 0.5*

⇒ *Entropy(s) = 1*

If it contains all yes or all no ie P(S) = 1 or 0

⇒ *Entropy(s) = 0*

What is Entropy?



$E(S) = -P(Yes)\log_2 P(Yes)$

When $P(Yes) = P(No) = 0.5$ ie YES + NO = Total Sample(S)

$E(S) = 0.5\ \log_2 0.5 - 0.5\ \log_2 0.5$

$E(S) = 0.5(\log_2 0.5 - \log_2 0.5)$

$E(S) = 1$

What is Information Gain?

- **Measures the reduction in entropy**
- **Decides which attribute should be selected as the decision node**

If S is our total collection,

Information Gain = Entropy(S) – [(Weighted Avg) x Entropy(each feature)]

Step 1: Compute the entropy for the Data set

Out of 14 instances we have 9 YES and 5 NO

So we have the formula,

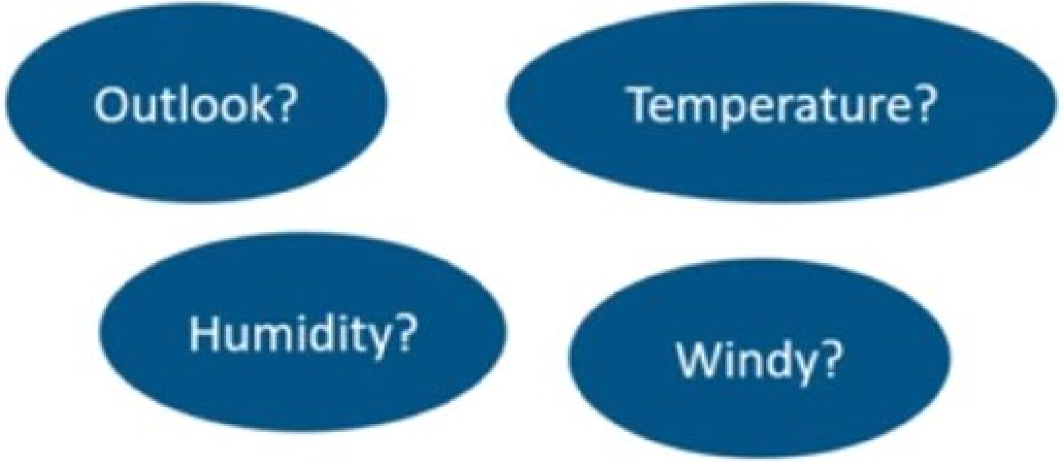
$E(S) = -P(Yes) \log_2 P(Yes) - P(No) \log_2 P(No)$

$E(S) = -(9/14)^* \log_2 9/14 - (5/14)^* \log_2 5/14$

$E(S) = 0.41+0.53 = 0.94$

	outlook	temp.	humidity	windy	play
D1	sunny	hot	high	false	no
D2	sunny	hot	high	true	no
D3	overcast	hot	high	false	yes
D4	rainy	mild	high	false	yes
D5	rainy	cool	normal	false	yes
D6	rainy	cool	normal	true	no
D7	overcast	cool	normal	true	yes
D8	sunny	mild	high	false	no
D9	sunny	cool	normal	false	yes
D10	rainy	mild	normal	false	yes
D11	sunny	mild	normal	true	yes
D12	overcast	mild	high	true	yes
D13	overcast	hot	normal	false	yes
D14	rainy	mild	high	true	no

Which Node To Select As Root Node?



Outlook?

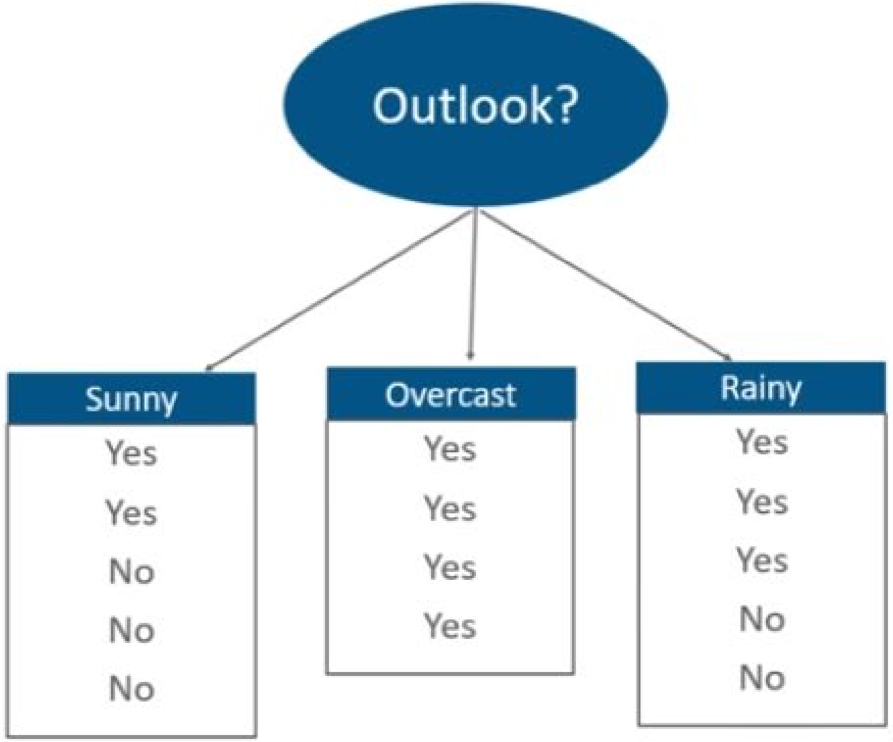
Temperature?

Humidity?

Windy?

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Which Node To Select As Root Node: Outlook



Outlook?

Sunny
Yes
Yes
No
No
No

Overcast
Yes
Yes
Yes
Yes

Rainy
Yes
Yes
Yes
No
No

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Which Node To Select As Root Node: Outlook

E(Outlook = Sunny) $= -2/5 \log_2 2/5 - 3/5 \log_2 3/5 = 0.971$

E(Outlook = Overcast) $= -1 \log_2 1 - 0 \log_2 0 = 0$

E(Outlook = rainy *)* $= -3/5 \log_2 3/5 - 2/5 \log_2 2/5 = 0.971$

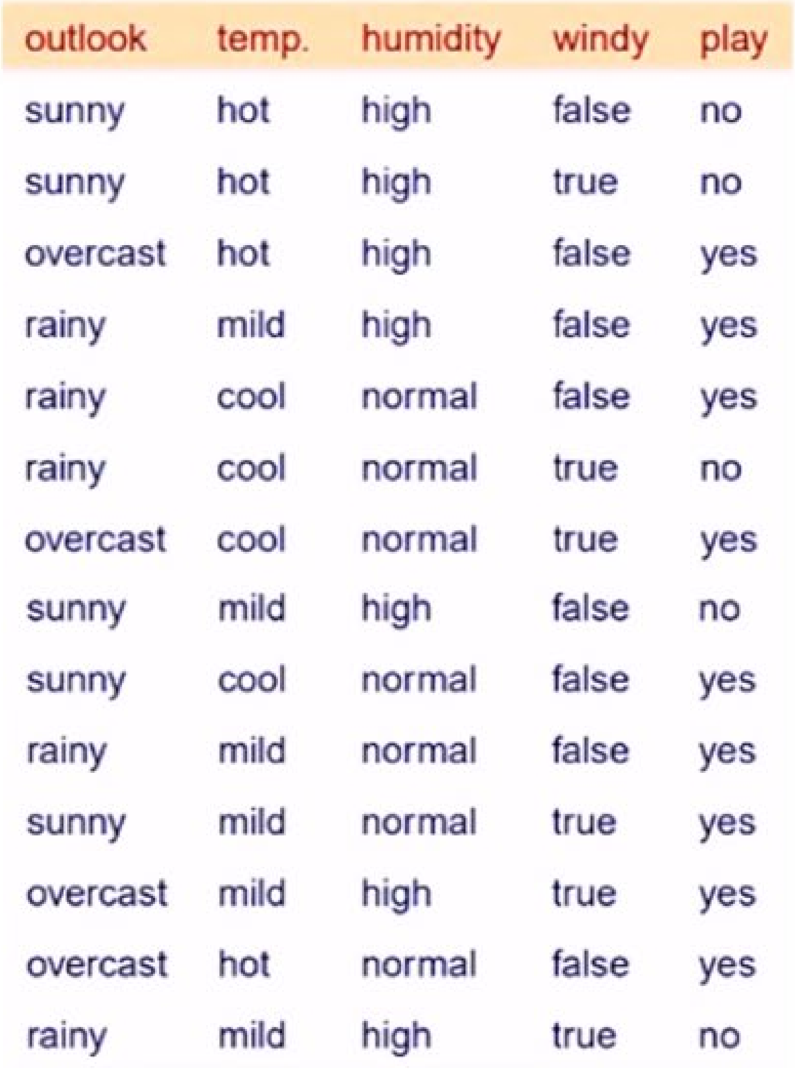
Information from outlook,

I(Outlook) = 5/14 x 0.971 + 4/14 x 0 + 5/14 x 0.971 = 0.693

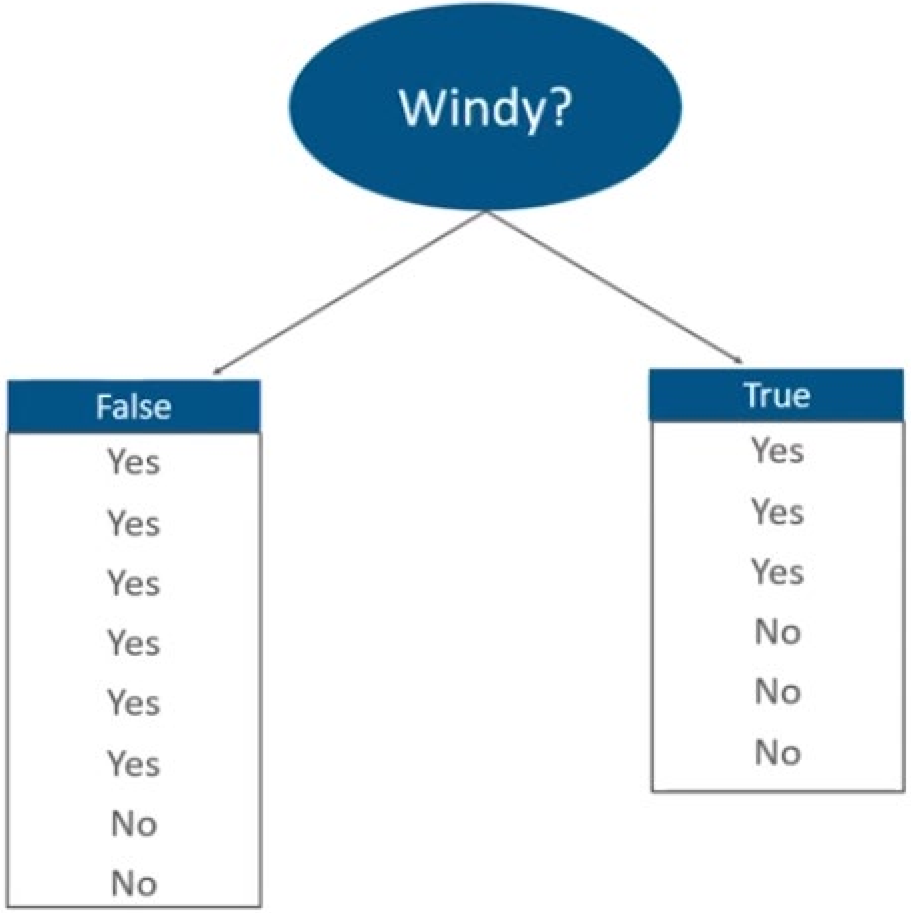
Information gained from outlook,

Gain(Outlook) = E(S) – I(Outlook)

0.94 – 0.693 = 0.247



outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no



Windy?

False
Yes
Yes
Yes
Yes
Yes
Yes
No
No

True
Yes
Yes
Yes
No
No
No

E(Windy = True) = 1

E(Windy = False) = 0.811

Information from windy,

I(Windy) = 8/14 x 0.811+ 6/14 x 1 = 0.892

Information gained from outlook,

Gain(Windy) = E(S) – I(Windy)

0.94 – 0.892 = 0.048

Which Node To Select As Root Node

Outlook:
Info 0.693
Gain: 0.940-0.693 $\boxed{0.247}$

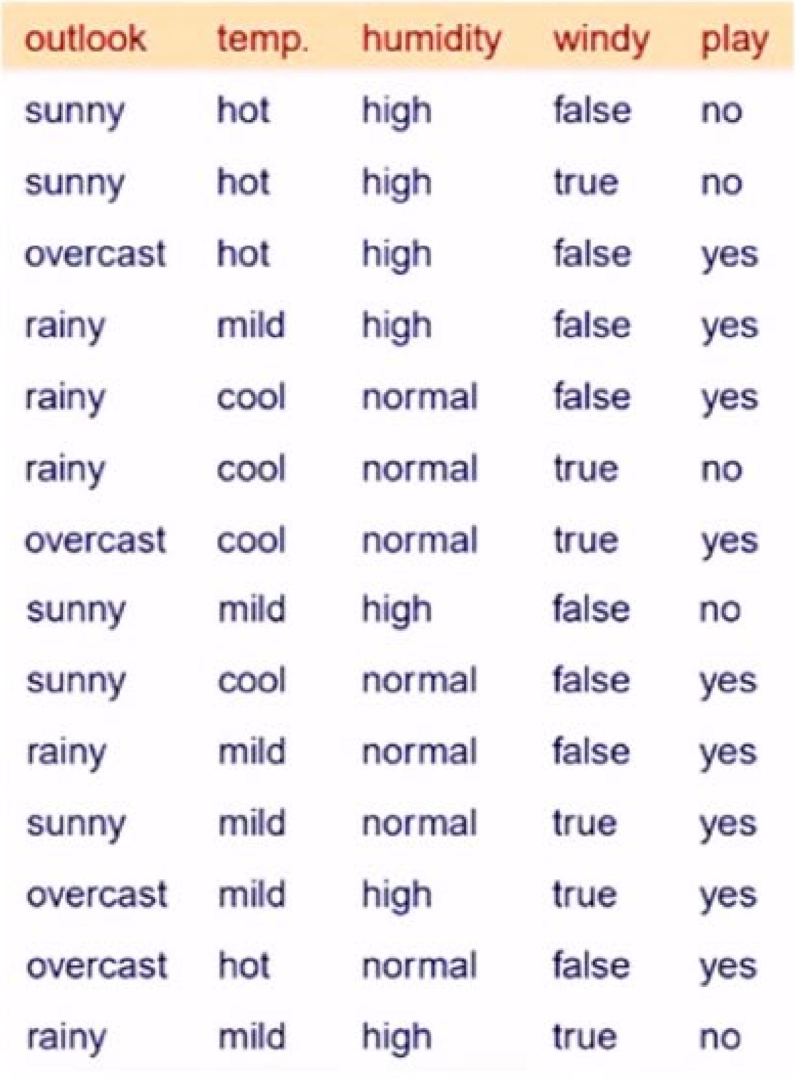
Temperature:
Info 0.911
Gain: 0.940-0.911 0.029

Humidity:
Info 0.788
Gain: 0.940-0.788 0.152

Windy:
Info 0.892
Gain: 0.940-0.982 0.048

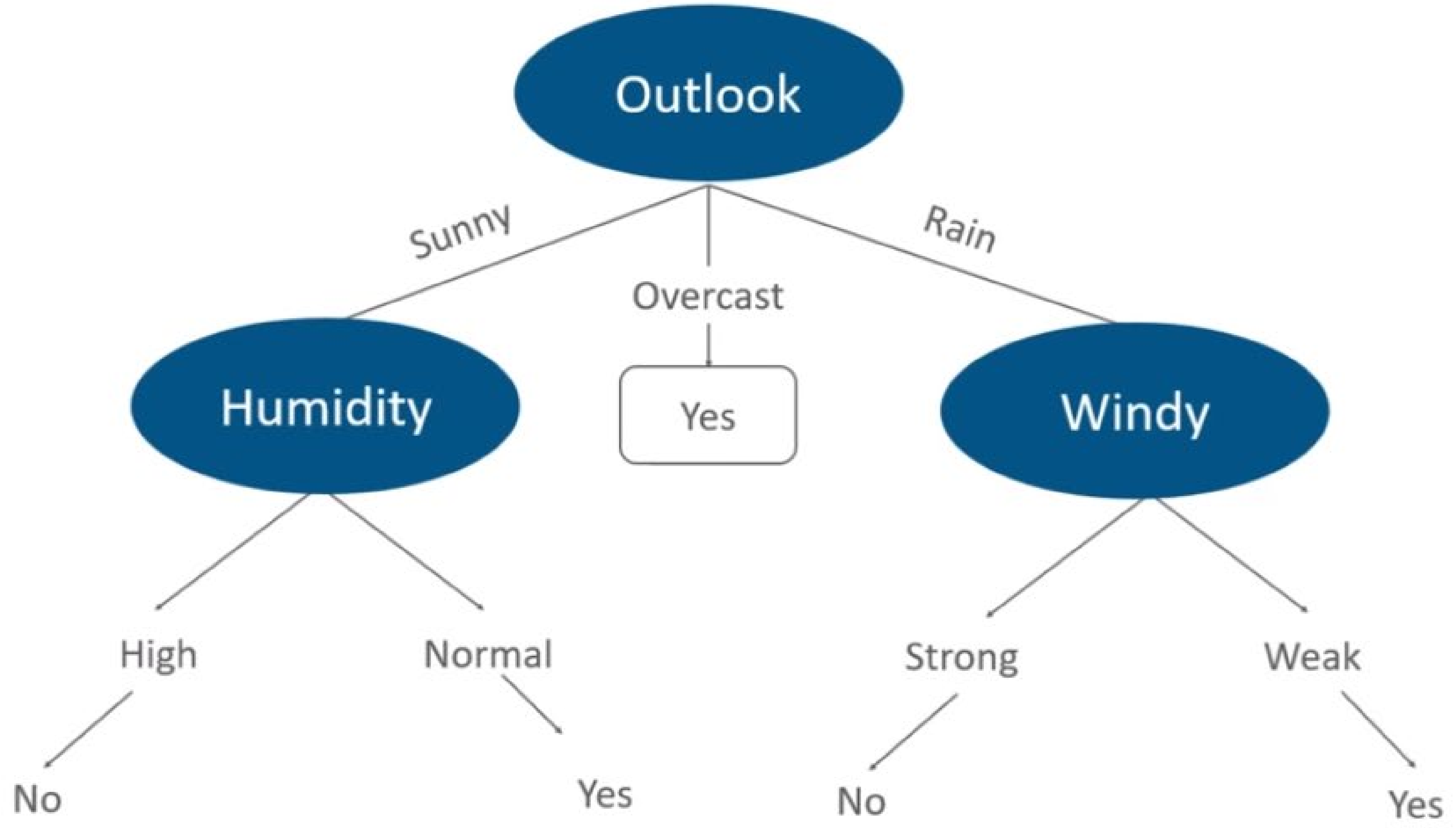
Since Max gain = 0.247,

Outlook is our ROOT Node



outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

This Is How Your Complete Tree Will Look Like

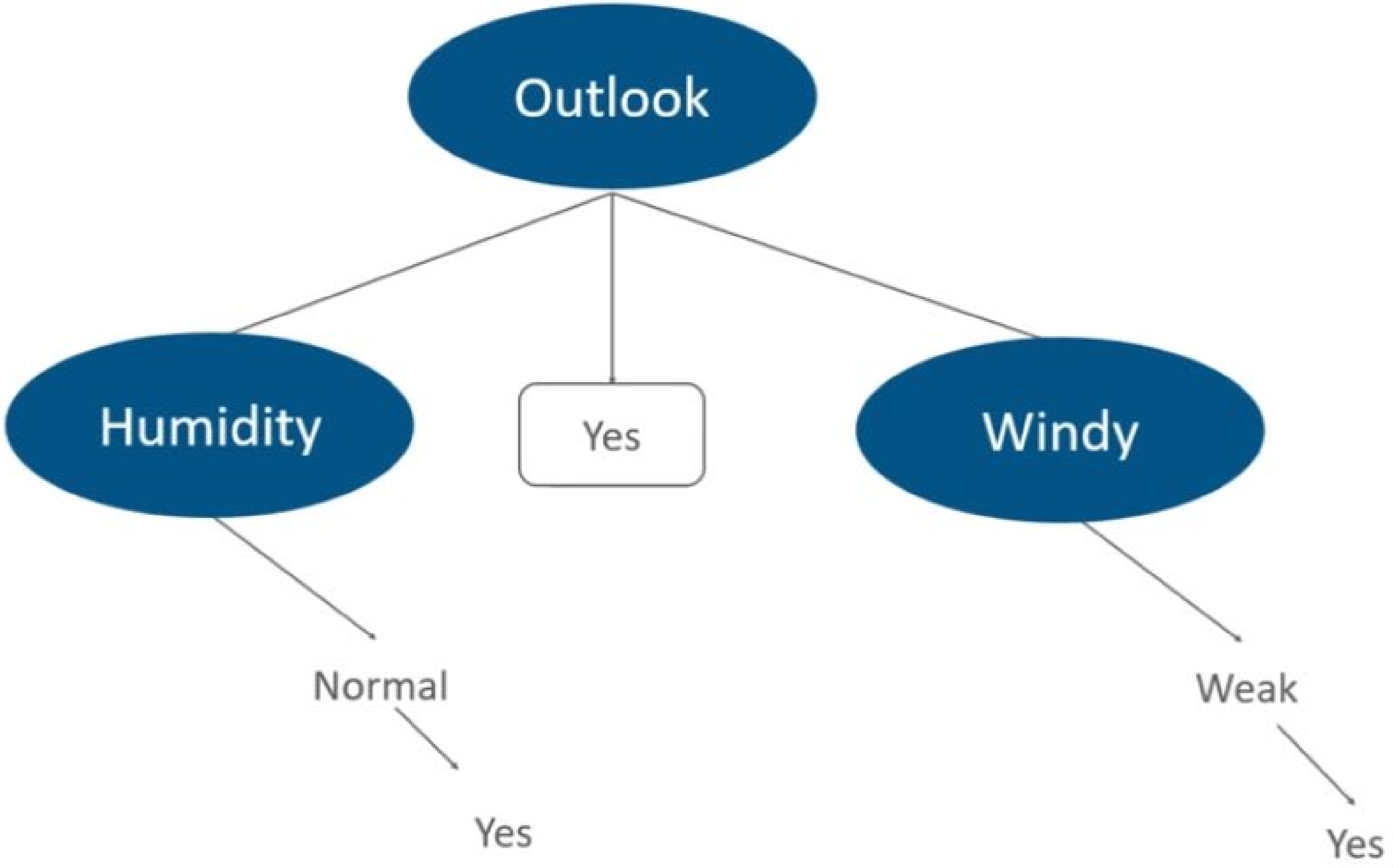


What Should I do to play? Pruning

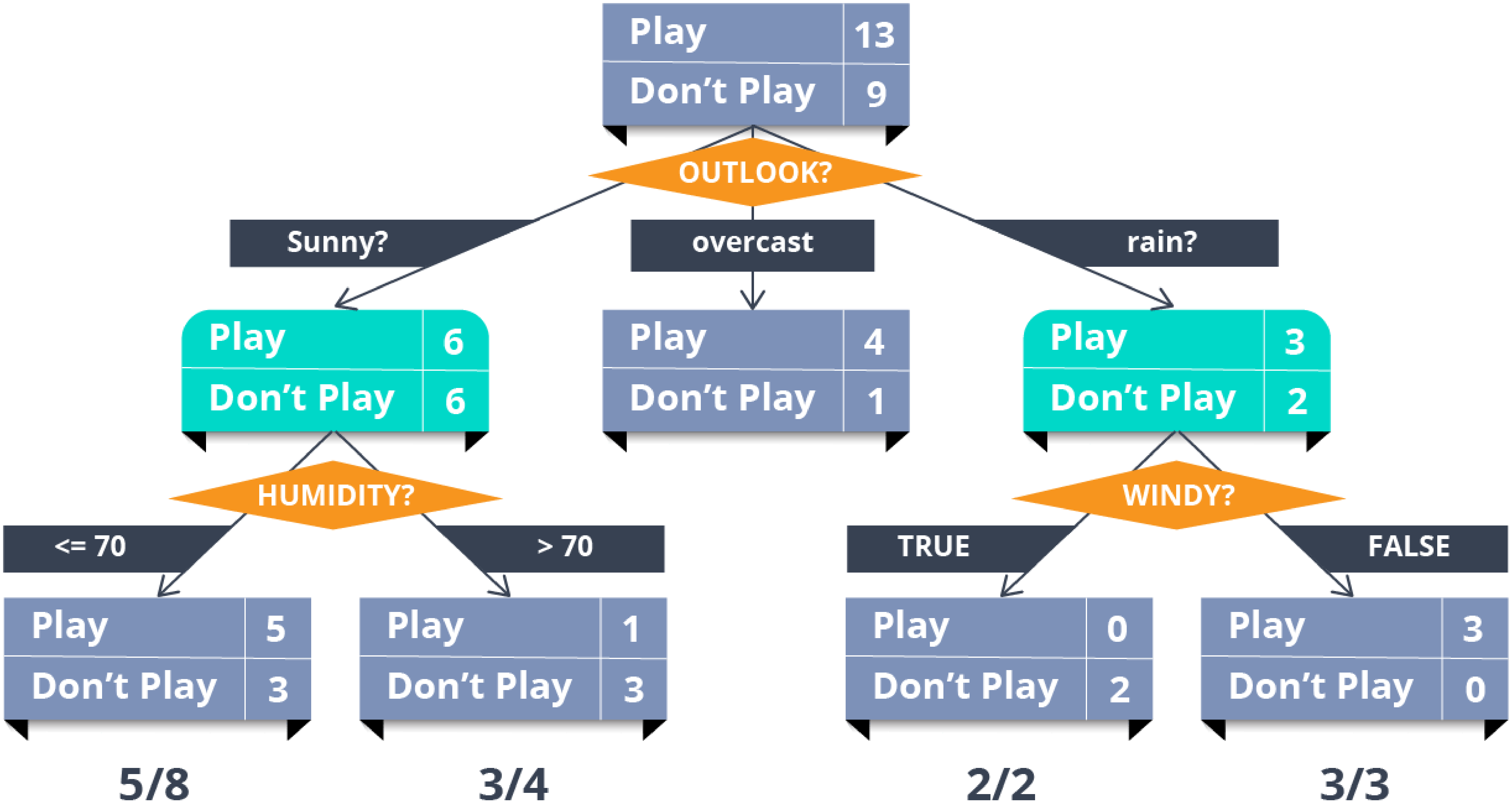
What is Pruning?

Pruning is to cutting down the nodes to get optimal solution.

Pruning: Reducing The Complexity



Dependent variable: PLAY


Play 13
Don't Play 9
OUTLOOK?
Sunny?
overcast
rain?
Play 6
Don't Play 6
Play 4
Don't Play 1
Play 3
Don't Play 2
HUMIDITY?
<= 70
> 70
WINDY?
TRUE
FALSE
Play 5
Don't Play 3
Play 1
Don't Play 3
Play 0
Don't Play 2
Play 3
Don't Play 0
5/8
3/4
2/2
3/3