Conducted by:

# **Tree Based Methods**

Ritvij Bharat Private Limited (RBPL)

# Tree Based Methods

#### Three main methods:

- Decision Trees
- Random Forests
- Boosted Trees

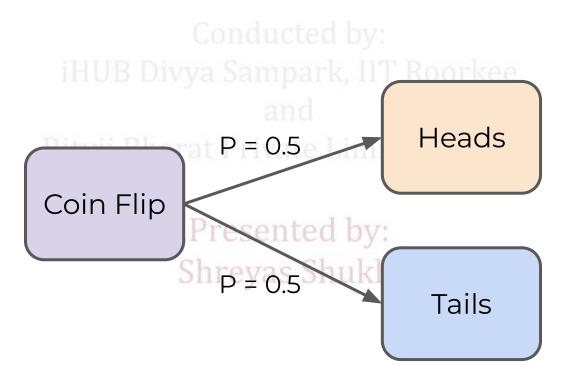
Conducted by:

# **Decision Trees**

Theory and Intuition

- While the use of basic decision trees for modeling choices and outcomes have been around for a very long time, statistical decision trees are a more recent development.
- Note the difference here!

The general term "decision tree" can refer to a flowchart mapping out outcomes.



Decision Tree Learning refers to the statistical modeling that uses a form of decision trees, where node splits are decided based on an information metric.

Ritvij Bharat Private Limited (RBPL)

Decision trees methods is basically the ability to split data based on information from features.

We need a mathematical definition of information and the ability to measure it.

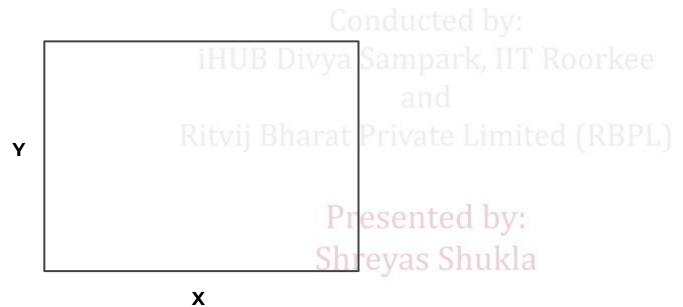
The ability to measure and define information will become more important as we learn the mathematics of tree based methods.

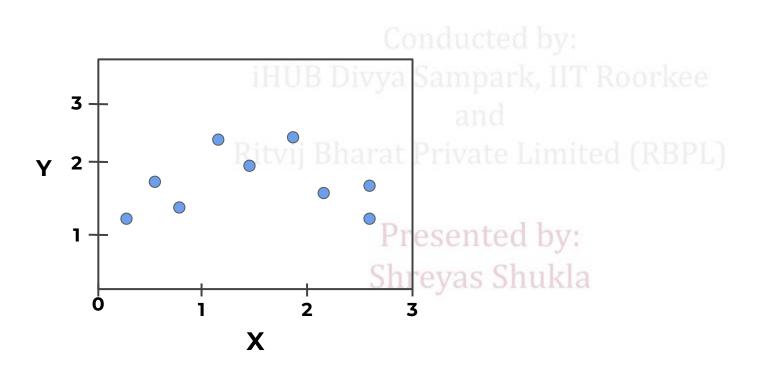
Let's talk about the development of decision trees.

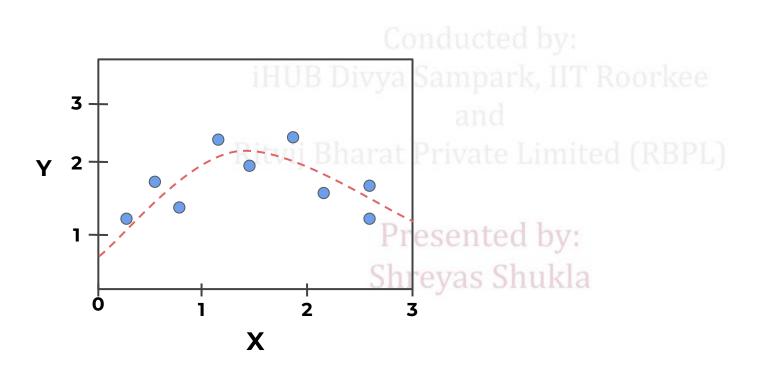
1963: First publication of regression tree algorithm by Morgan and Sonquist

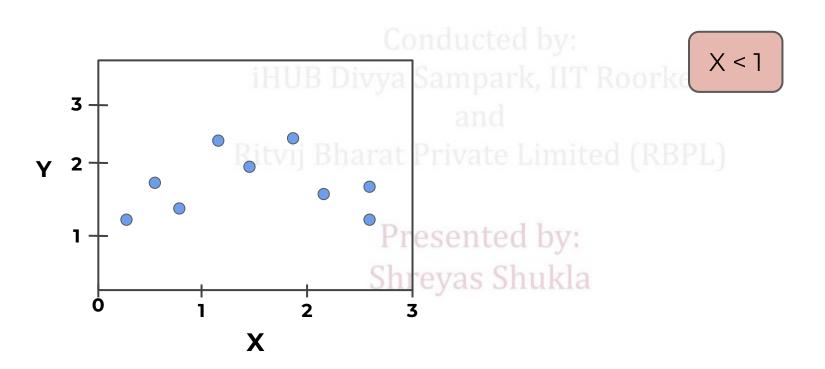
1963: Morgan and Sonquist created piecewise-constant model with splits.

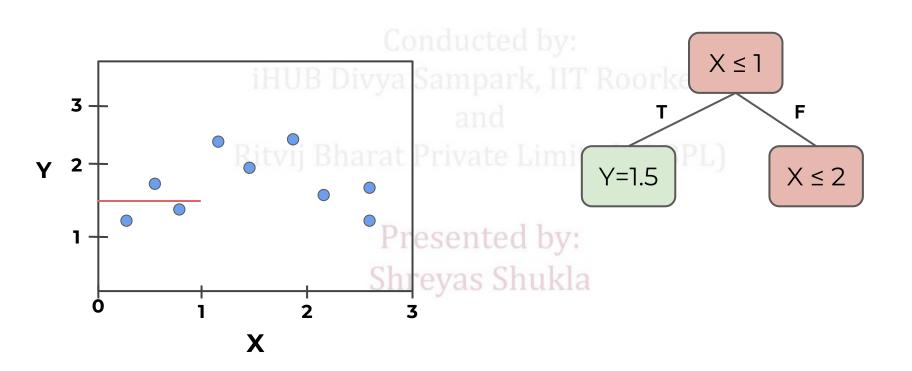
### 1963: Piecewise-constant regression tree

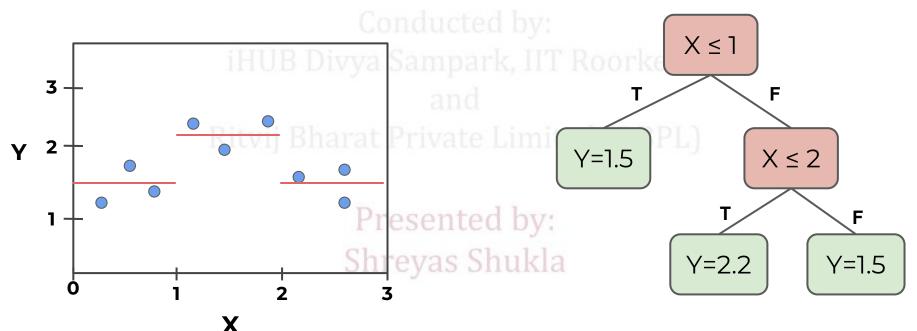


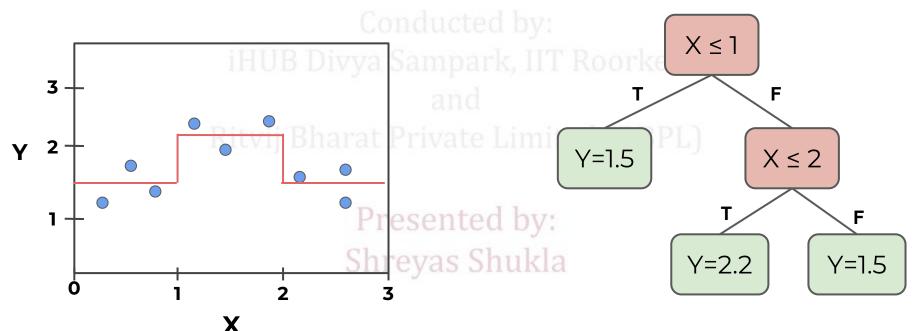












### **Node impurity**

Conducted by: iHUB Divya Sampark, IIT Roorkee and

Rituii Rharat Drivata Limitad (DRDI)

$$\phi(t) = \sum_{i \in t} (y_i - \bar{y})^2$$

Shreyas Shukla

Conducted by:

# **Decision Trees**

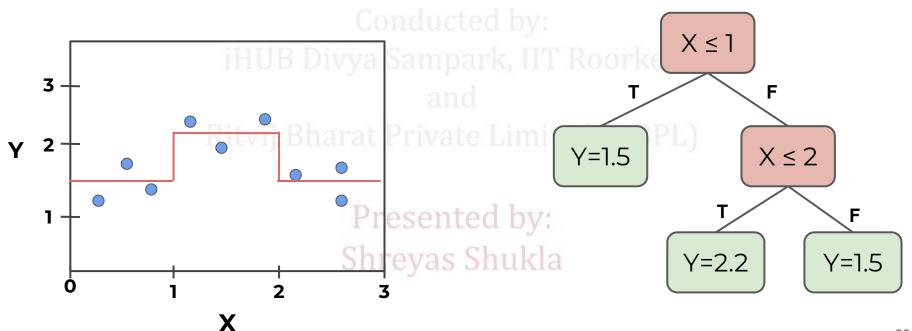
Decision Tree Basics

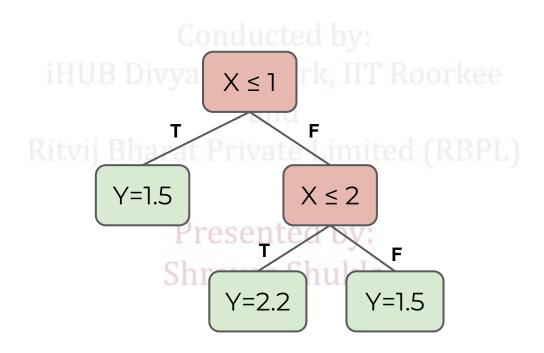
Let us understand some terminology about the decision tree components.

iHUB Divya Sampark, IIT Roorkee and

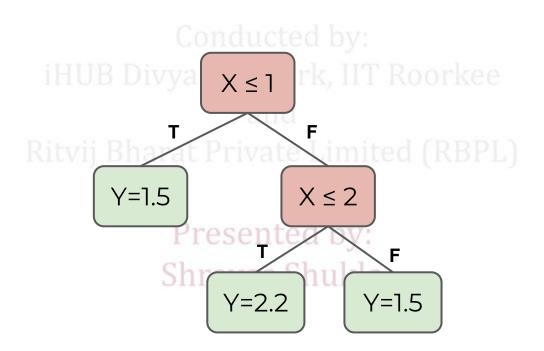
Ritvij Bharat Private Limited (RBPL)

#### Recall our simple regression tree:

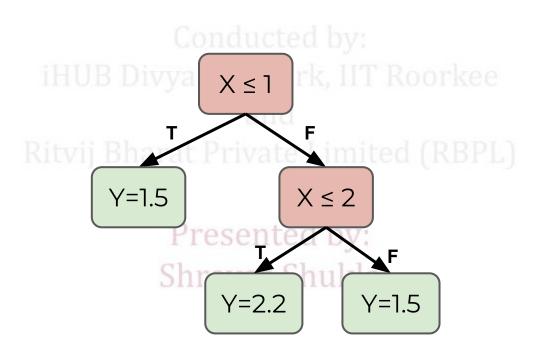




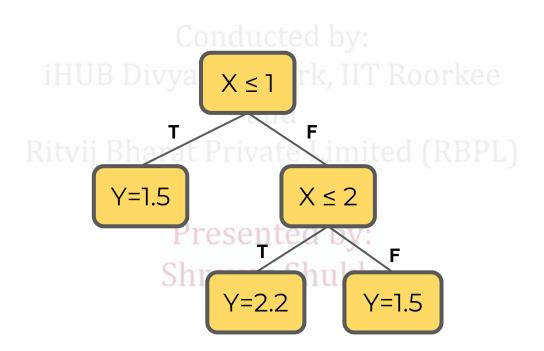
# Splitting



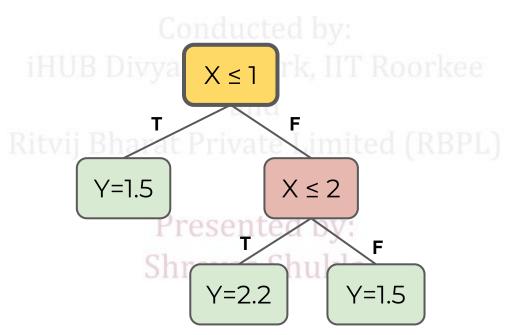
# Splitting



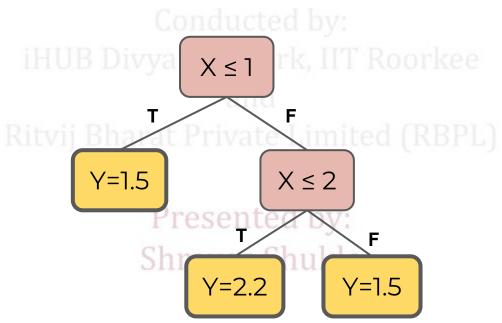
#### Nodes:



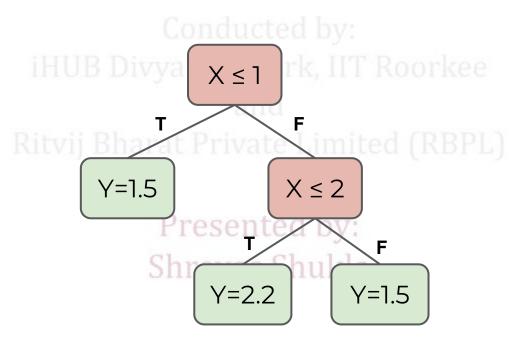
#### Root Node:



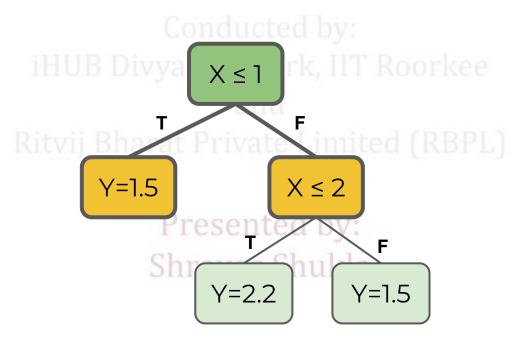
## Leaf (Terminal) Nodes:



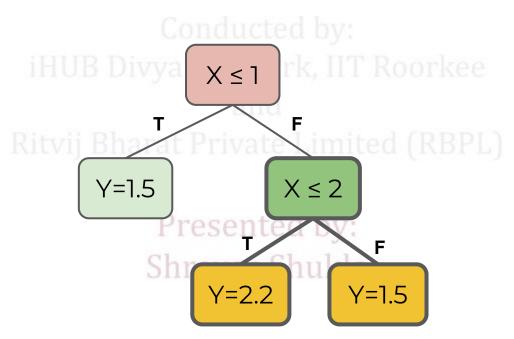
#### Parent and Children Nodes:



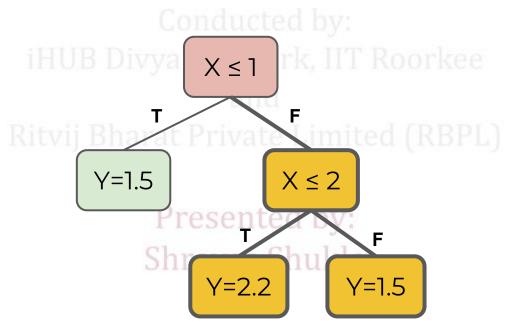
#### Parent and Children Nodes:



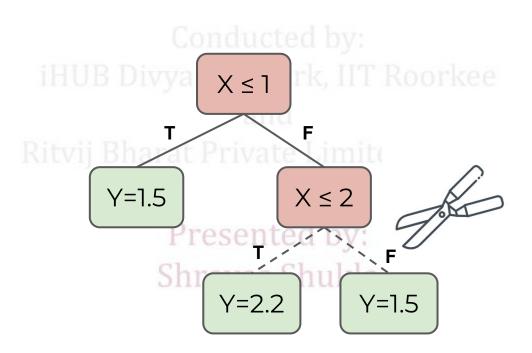
#### Parent and Children Nodes:



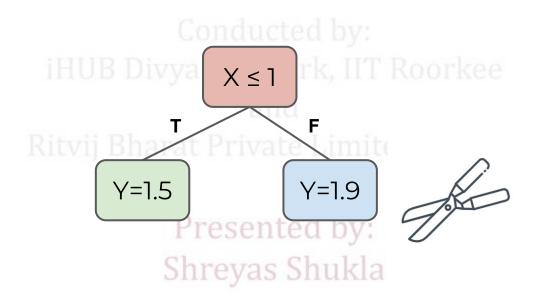
### Tree Branches (Sub Trees):



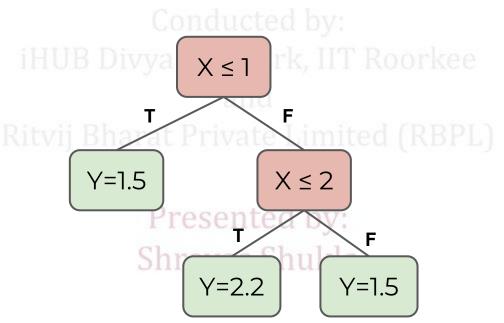
# Pruning:



# Pruning:



## Let's begin constructing a tree!



Conducted by:

# **Decision Trees**

Gini Impurity

# An Introduction to Wachine Learning with Python Programming Gini Impurity

A mathematical measurement of how "pure" the information in a data set is.

We can think of this as a measurement of class uniformity.

### Gini Impurity for Classification:

• For a set of classes **C** for a given dataset **Q**:

and

Ritvij Bharat Pr
$$G(Q) = \sum_{c \in C} p_c (1-p_c)$$
Pres $C \in C$ Shreyas Snukia

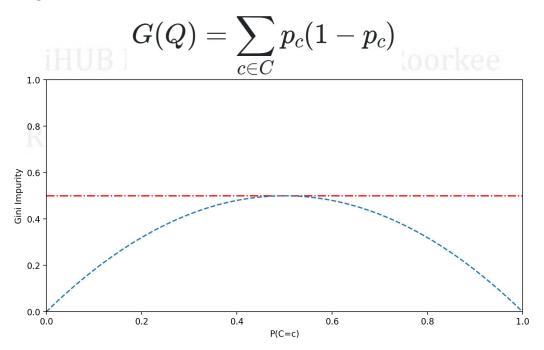
#### Gini Impurity for Classification:

For a set of classes C for a given dataset Q, p<sub>c</sub> is probability of class c.

$$p_c = rac{1}{N_Q} \sum_{x \in Q} \mathbb{1}(y_{class} = c) egin{array}{c} G(Q) = \sum_{c \in C} p_c (1 - p_c) \end{array}$$

Shreyas Shukia

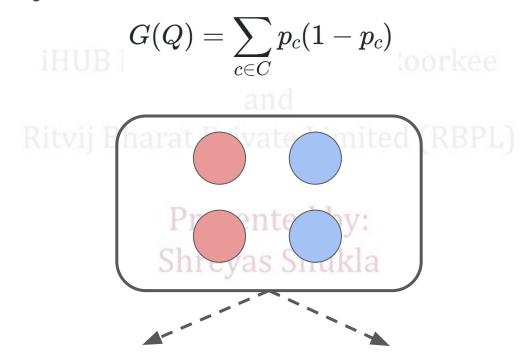
#### Gini Impurity for Classification:

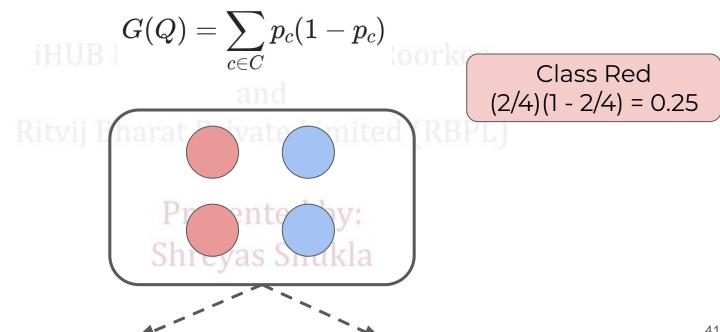


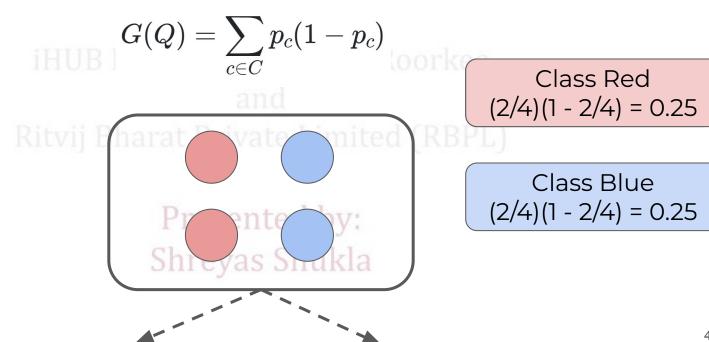
#### Gini Impurity for Classification:

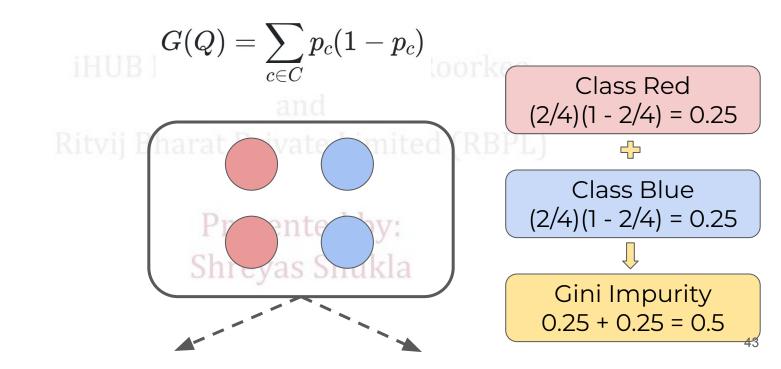
$$G(Q) = \sum_{c \in C} p_c (1 - p_c)$$
 loorkee and Ritvij Pharat Private Limited (RBPL) Presented by: Shreyas Shukla

#### Gini Impurity for Classification:

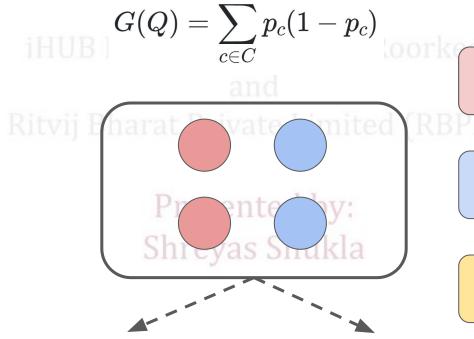








### "Maximum" Impurity Possible



Class Red (2/4)(1 - 2/4) = 0.25

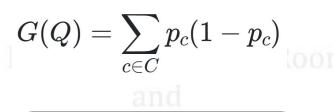


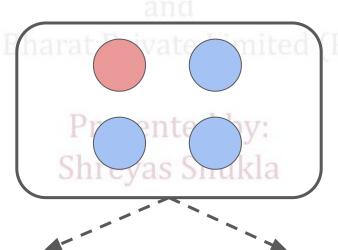
Class Blue (2/4)(1 - 2/4) = 0.25



Gini Impurity 0.25 + 0.25 = 0.5

### Data is more "pure" (less impurity)





Class Red (1/4)(1 - 1/4) = 0.1875

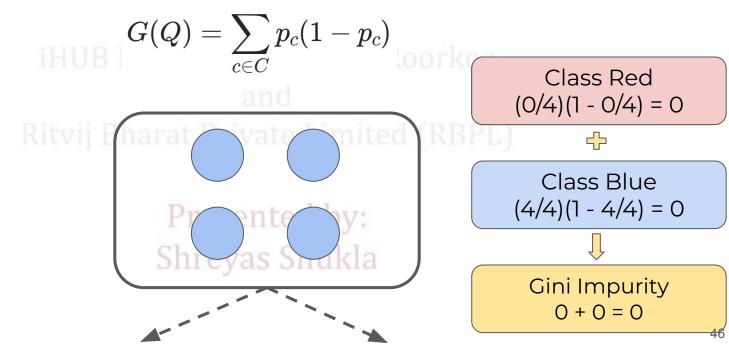


Class Blue (3/4)(1 - 3/4) = 0.1875



Gini Impurity 0.1875+0.1875 = 0.375

### Data is completely "pure" (no impurity)



An Introduction to Machine Learning with Python Programming

If the goal of a decision tree is to separate out classes, we can use gini impurity to decide on data split values.

We want to minimize the gini impurity at leaf nodes.

Minimized impurity at leaf nodes means we are separating classes effectively

Shreyas Shukla

Conducted by:

## **Decision Trees**

Gini Impurity in Trees

Presented by: Shreyas Shukla

For constructing a tree, we have to decide what feature will be root node.

Use gini impurity to compare the information contained within features for the training data.

Presented by: Shreyas Shukla

#### Gini Impurity for Classification:

For a set of classes C for a given dataset Q, p<sub>c</sub> is probability of class c.

$$p_c = egin{array}{c} rac{1}{N_Q} \sum_{x \in Q} \mathbb{1}(y_{class} = c) & G(Q) = \sum_{c \in C} p_c (1 - p_c)$$

# Create a decision tree to predict spam.

X - URL Link	Y-Spam	acted by:
Yes	HUB Dyesya San	ipark, IIT Roorkee
Yes	Yes	and
No Ri	vij Bhnoat Priv	rate Limited (RBPL)
No	No	
No	Yes Prese	nted by:
No	NShreya	as Shukla
Yes	No	

## Only one X feature to use for a node.

X - URL Link	Y-Spam	acted by:
Yes	HUB Dyesya San	npark, IIT Roorl URL
Yes	Yes	and
No	vij Bhnoat Priv	rate Limited (RBPL)
No	No	
No	Yes Prese	nted by:
No	NShreya	as Shukla
Yes	No	

## Predict if email is spam if it contains a URL:

X - URL Link	Y-Spam
Yes	HUB Dyesya San
Yes	Yes
No Ri	vij Bhnoat Priv
No	No
No	Yes Prese
No	мShreya
Yes	No

X - URL Link	Y-Spam	acted by:
Yes	Yes	npark, IIT Roorl URL
Yes	Yes	ınd T
No Ri	vij Bh <sub>No</sub> at Pri	
No	No	Spam Yes: 2
No	Yes Prese	nted (No: 1
No	NShreya	as Shukla
Yes	No	

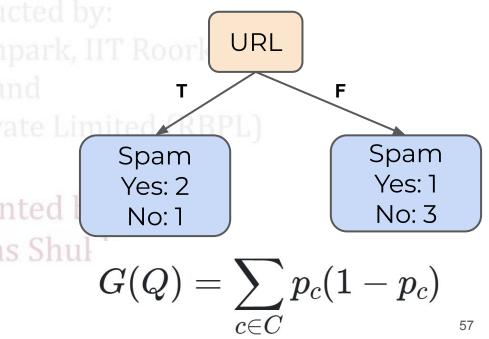
X - URL Link	Y-Spam
Yes	HUB Dyesya San
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No

## Predict if email is spam if it contains a URL:

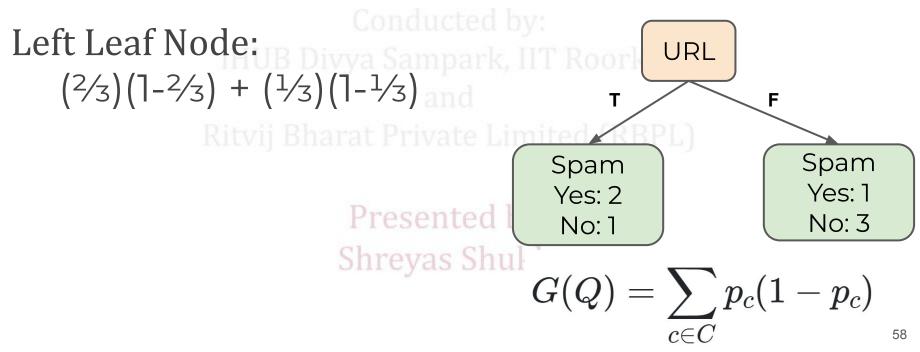
X - URL Link	Y-Spam
Yes	HUB Dyesya Sar
Yes	Yes
No R	itvij Bh <sub>No</sub> at Pri
No	No
No	Yes Prese
No	Nohrey
Yes	No

### Recall the gini impurity formula: The Python Programming

X - URL Link	Y-Spam	
Yes	HUB Dyesya San	
Yes	Yes	
No R	vij Bh <sub>No</sub> at Pri	
No	No	
No	YesPrese	
No	NShreya	
Yes	No	



#### Treat Yes Spam and No Spam as **c** classes:



#### Treat Yes Spam and No Spam as **c** classes:

Left Leaf Node:

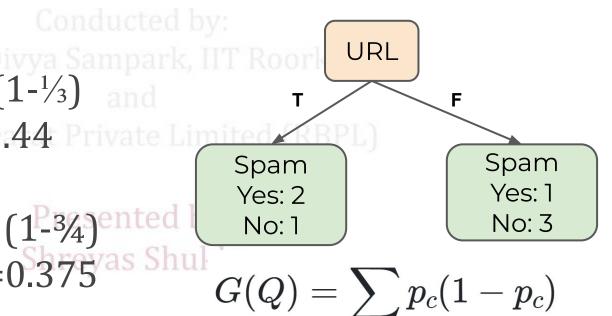
$$(\frac{2}{3})(1-\frac{2}{3}) + (\frac{1}{3})(1-\frac{1}{3})$$

Left Leaf Gini=0.44 Private Li

Right Leaf Node:

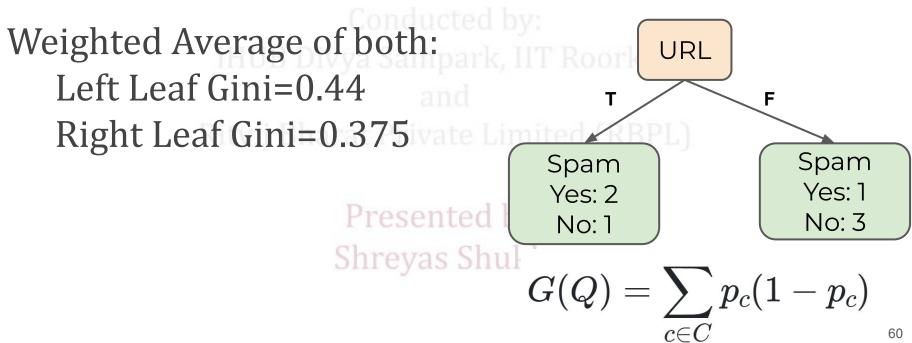
$$(\frac{1}{4})(1-\frac{1}{4}) + (\frac{3}{4})(1-\frac{3}{4})^{\text{ented}}$$

Right Leaf Gini=0.375



59

Calculate gini impurity of URL feature.



Total Emails: 
$$(2+1)+(1+3)=7$$
Left Leaf Gini=0.44
Right Leaf Gini=0.375
Left Emails: 3
Right Emails: 4

Spam
Yes: 2
No: 1
No: 3
Shreyas Shul
 $G(Q) = \sum_{c \in C} p_c(1-p_c)$ 

Total Emails: 
$$(2+1) + (1+3) = 7$$
  
Left Leaf Gini=0.44

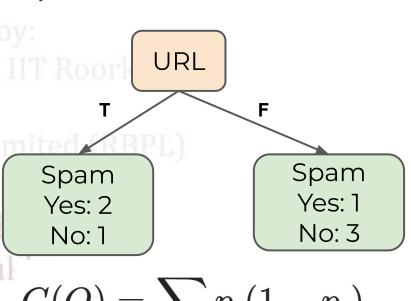
Right Leaf Gini=0.375

Left Emails: 3

Right Emails: 4

(3/7)\*0.44 + (4/7)\*0.375

Gini Impurity: 0.403eyas Shul



$$G(Q) = \sum_{c \in C} p_c (1-p_c)$$

#### More issues to consider:

- Multiple Features and ucted by:
- Continuous Features
- Multi-categorical Features

We use the gini impurity to each of these issues to solve for best root nodes and best split parameters for leaves.

Conducted by:

## **Decision Trees**

Gini Impurity Part Two

Presented by: Shreyas Shukla

#### Let's explore:

- Continuous numeric features
- Multi-categorical features (N>2)
- Choosing a root node feature

Presented by: Shreyas Shukla

# Imagine a continuous feature. Calculate the feature gini impurity:

X - Words in Email	Y-Spam	ipark, IIT Roorkee
10	Yes	and
40 RI	vij Bhnoat Priv	rate Limited (RBPL
20	Yes	
50	No Prese	nted by:
30	n&hreya	as Shukla

#### Sort data:

Conducted by:

X - Words in Email	Y-Spam	ipark, IIT Roorkee
10	Yes	and
40 Ri	vij Bhnoat Priv	rate Limited (RBPL
20	Yes	
50	No Prese	nted by:
30	NShreya	as Shukla

#### Calculate potential split values for node

	\		
X - Words in Email	Y-Spam	ipark, IIT Roc	Words ≤ N
10	Yes	and	
20 Ri	Yes	rate Limited (	
30	No		
40	No Prese	nted by:	
50	N&hrey?	as Shukla	

#### Use averages between rows as values:

Conducted by:			VA/a ada a NI	
X - Wo	ords in Email	Y-Spam	npark, IIT Roc	Words ≤ N
15	10	Yes	and	
	20	Yes	rate Limited (	
25	30	No		
35	40	No Prese	nted by:	
45	50	NShrey	as Shukla	

#### Perform all the potential split:

Conducted by:				\
X - Wo	ords in Email	Y-Spam	npark, IIT Roc	Words ≤ 15
15	10	Yes	and	
	20 RI	Yes	vate Limited (	
25	30	No		
35	40	No Prese	nted by:	
45	50	NShrey	as Shukla	

### Calculate gini impurity for each split:

	Words ≤ 15				
X - Words in Email		Y-Spam	ipark, IIT Roc	\ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \	
15	10	Yes	and		
	20	VIJ BOYes at Prin	rate Limited (		
	30	No			
	40	No Prese	ented by:		
	50	NShreya	as Shukla		

#### Calculate gini impurity for each split:

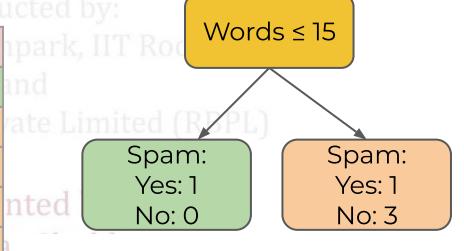
	\		
X - Words in Email	Y-Spam	npark, IIT Roo	Words ≤ 15
15 10	Yes	and	
20	Yes	vate Limited (	
30	No		
40	No	nted by:	
50	No	as Shukla	

#### Calculate gini impurity for each split:

	Con	ducted by:	uala (IC)
X - Words in Email	Y-Spam	ipark, IIT Roc	rds ≤ 15
15 10	Yes	and	
20	Yes	rate Limited (RZPI	
30	No	Spam: Yes: 1	Spam: Yes: 1
40	No	nted No: 0	No: 3
50	No		
·		$G(Q) = \sum p_c(Q)$	$(1-p_c)$
		$c \in C$	

#### Calculate gini impurity for each split:

		COHU	u
X - Words in Email		Y-Spam	h
15	10	Yes	Í
	20	Yes	7
	30	No	
	40	No	
	50	No	2



$$G(Q) = (\%)(0+0) + (\%)((1/4)(1-1/4)+(3/4)(1-3/4)$$
  
= 0.3

#### Do it for all possible splits:

X - Wo	rds in Email	Y-Spam	ipark, IIT Roorkee
15	10	Yes	Gini=0.3
	20 Ri	VI Boyes at Prin	Gini=0.3
25	30	No	
35	40	No Prese	Gini=0.26
45	50	мShreya	Gini=0.4

### Choose lowest impurity split value

X - Wo	ords in Email	Y-Spam		
	10	Yes	and	
25	20 Ri	VIJ BOYes at PTII	→ Gini=0	
25	30	No	GIIII-0	
	40	No Prese	nted by:	
	50	мShreya	as Shukla	

#### Choose this as split value for node

	//ords < 25			
X - Words in Email	Y-Spam	Tipark, IIT Ro Words ≤ 25		\$ 25
10	Yes	and		
20 Ri	Yes	vate Limited (	RPL	
30	No	Spar Yes:	n:	Spam: Yes: 0
40	No Prese	nted Yes: No:		Yes: 0 No: 3
50	NShrey			
G(Q) = 0				

## Multicategorical feature

### Calculate gini impurity for all combinations:

X - Sender	Y-Spam	ipark, IIT Roorkee
Abe	Yes	and
Bob	Yes Pri	vate Limited (RBPL
Claire	No	
Abe	No Prese	nted by:
Bob	NShrey	as Shukla

#### Calculate gini impurity for all combinations:

Conducted by:					
X - Sender	Y-Spam	npark, IIT F Sender == Abe			
Abe	Yes	and			
Bob	Yes	rate Limited (RZPL)			
Claire	No	Spam: Spam: Yes: 1			
Abe	No	nted No: 1			
Bob	No	as Shukla			

#### Calculate gini impurity for all combinations

	Cona	
X - Sender	Y-Spam	Sender == Abe
Abe	Yes	and
Bob	Ri Vij Bi Yes	vata Limitad (PRPI)
Claire	No	Sender == Bob
Abe	No Prese	nte
Bob	Nohrey	Sender == Claire
		Serider Claire

### Calculate gini impurity for all combinations

X - Sender	Y-Spam	Sender == Abe	Sender == Abe or Bob
Abe	Yes	and	
Bob	Yes	rate Limited (PRP)	Sender ==
Claire	No	Sender == Bob	Claire or Bob
Abe	No Prese	nte	
Bob	Nohreya	Sender == Claire	Sender ==
		Scrider Claire	Abe or Claire
Choose lowest impuri	ty split combination	on	81

Now we can split any type of feature.

# How does the decision tree decide on the root node of a multi-feature dataset?

Calculate the gini impurity values of each feature and choose the lowest impurity value to split on first.

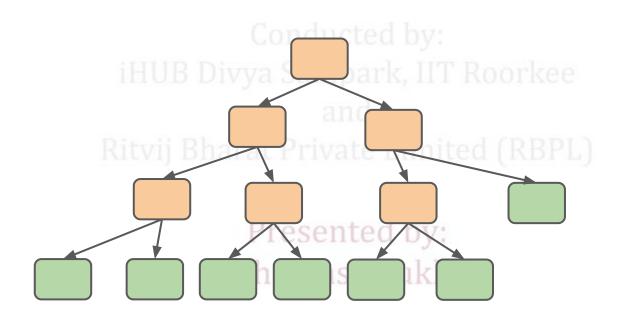
By choosing the feature with the lowest resulting gini impurity in its leaf nodes, we are choosing the feature that best splits the data into "pure" classes.

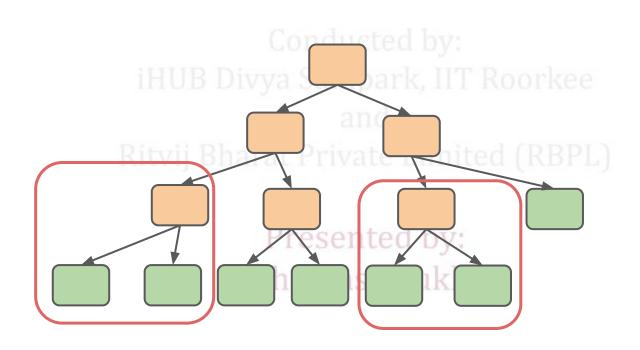
By using gini impurity as a measurement of the effectiveness of a node split, we can perform automatic feature selection by mandating an impurity threshold for an additional feature based split to occur.

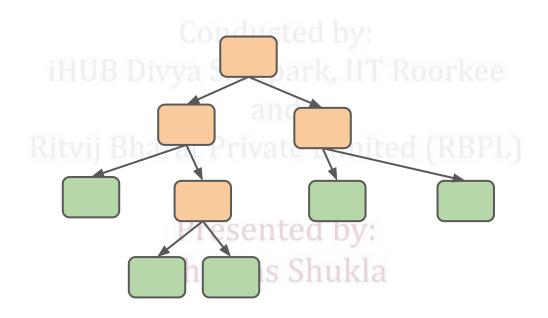
Presented by: Shreyas Shukla

# A large overfitted tree.

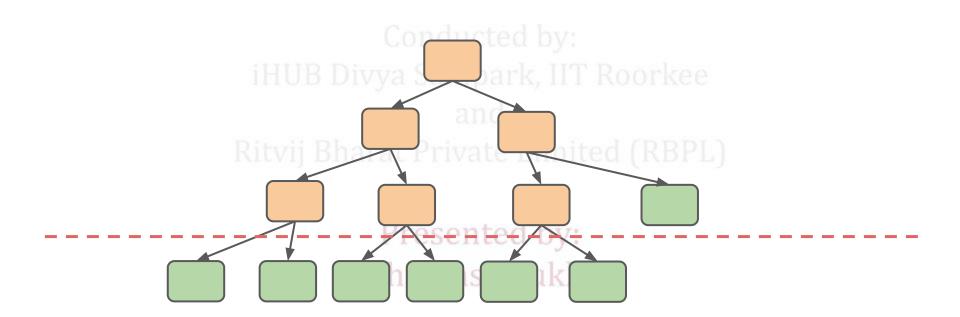
### Add minimum gini impurity decrease

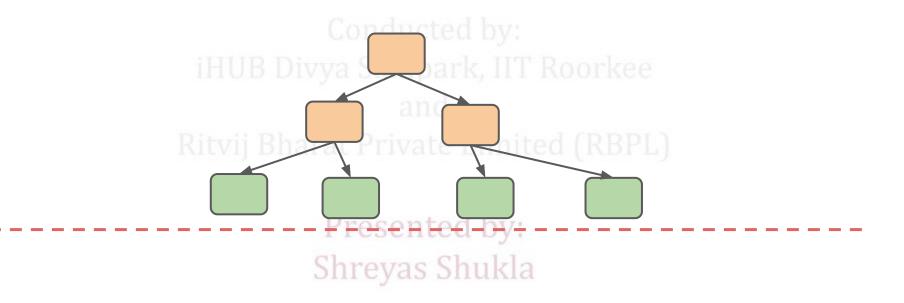






### We can also mandate a max depth





Conducted by: iHUB Divya Sampark, IIT Roorkee and

Let's code!!

Presented by: Shreyas Shukla