Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Decision Trees

Gini Impurity

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.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Kiwij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Gini Impurity for Classification:

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

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

Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Bitvij Bharat Private Limited

Gini Impurity for Classification:

- For a set of classes **C** for a given dataset **Q**, $\mathbf{p_c}$ is probability of class **c**.

$$p_c = \frac{1}{N_Q} \sum_{x \in Q} \mathbb{1}(y_{class} = c) \qquad G(Q) = \sum_{c \in C} p_c(1 - p_c)$$

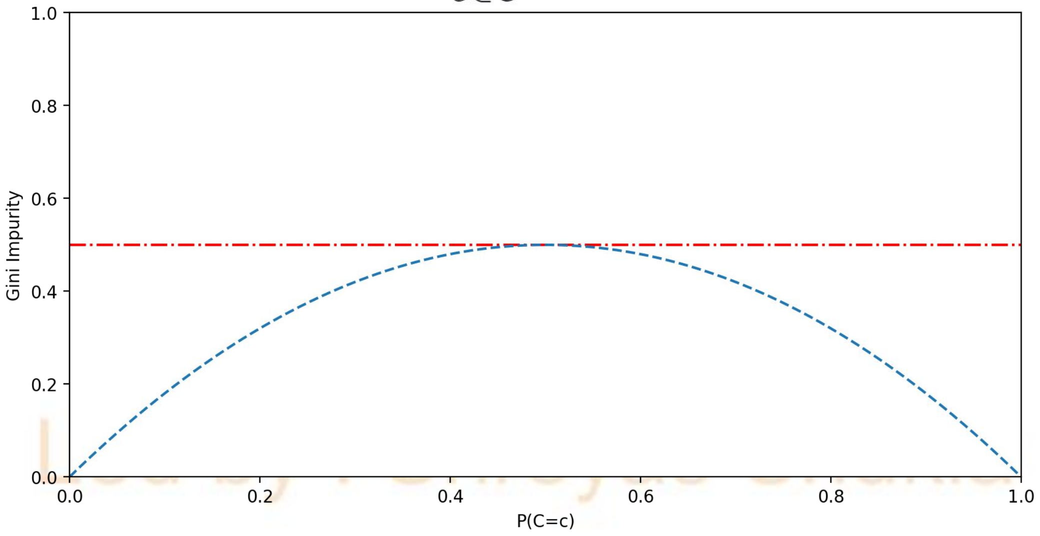
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Gini Impurity for Classification:

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

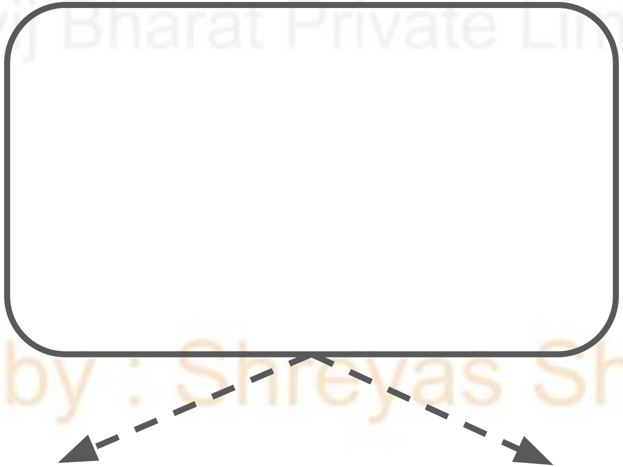
iHUB D Roorkee



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB D Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Gini Impurity for Classification:

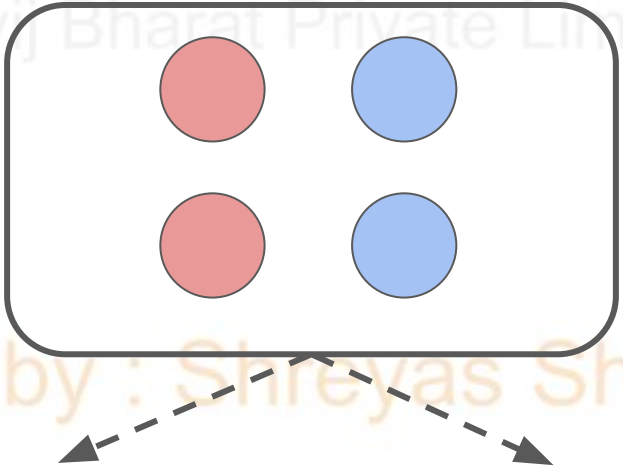
$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB D Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Gini Impurity for Classification:

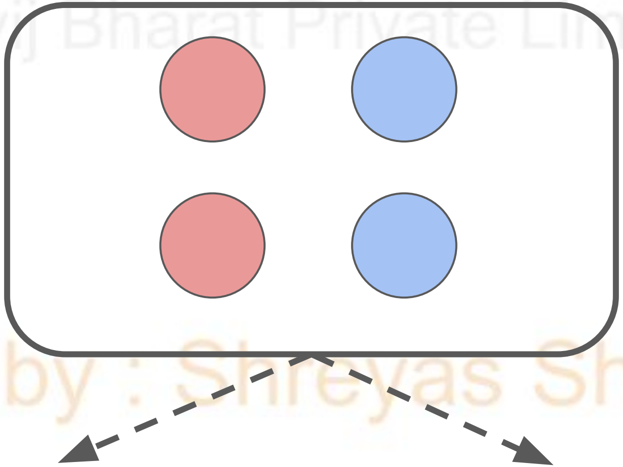
$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
Led by : Shreyas Shukla

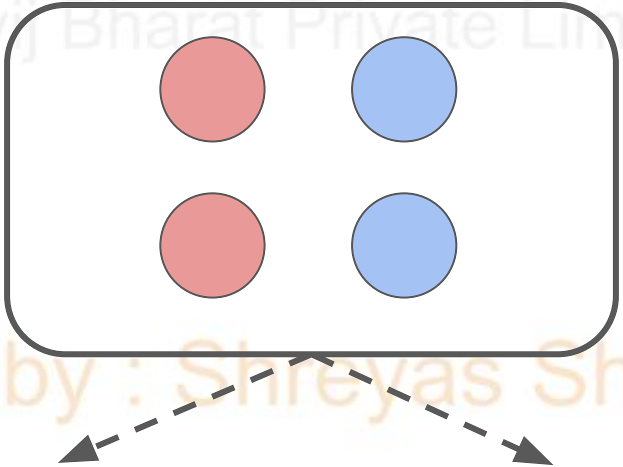
$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$

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



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB Divyasampark IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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

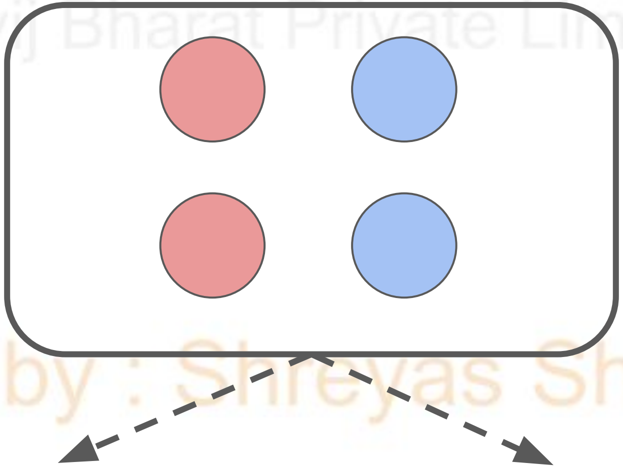


Class Red
(2/4)(1 - 2/4) = 0.25

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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB D... Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shuk...

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



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

+

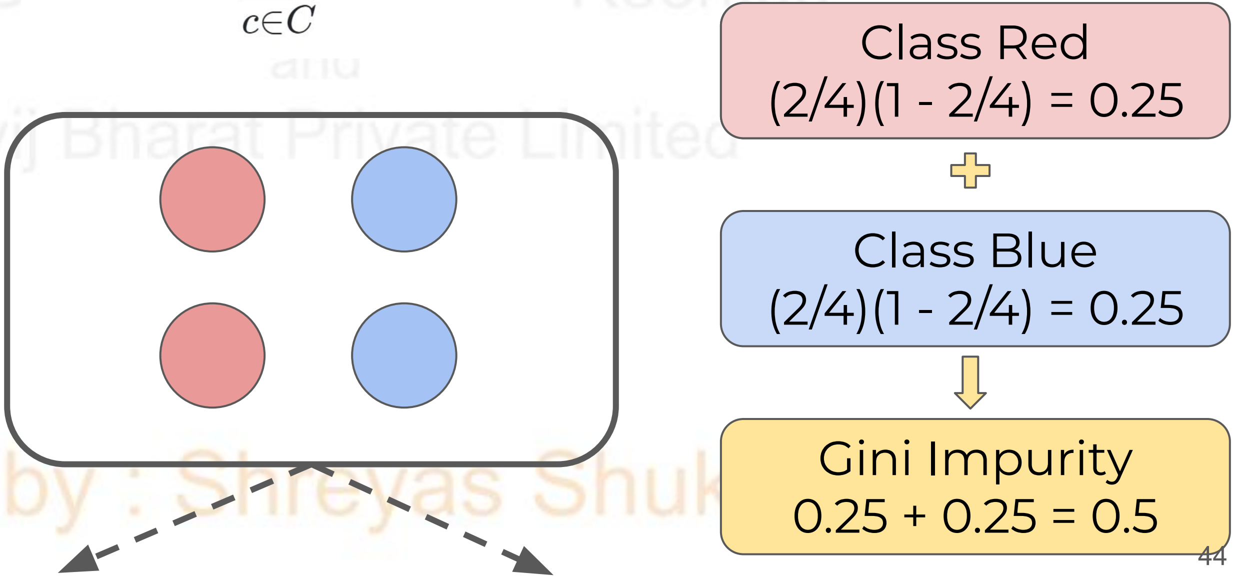
Class Blue
(2/4)(1 - 2/4) = 0.25

Gini Impurity
0.25 + 0.25 = 0.5

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB D... Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shuk...

"Maximum" Impurity Possible

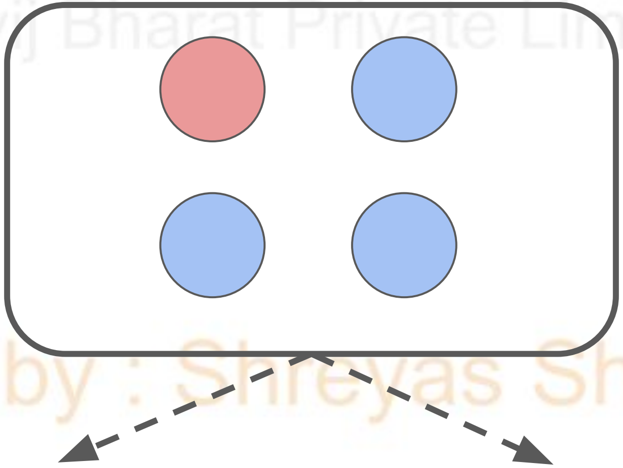
$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB D Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shuk

Data is more “pure” (less impurity)

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



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

+

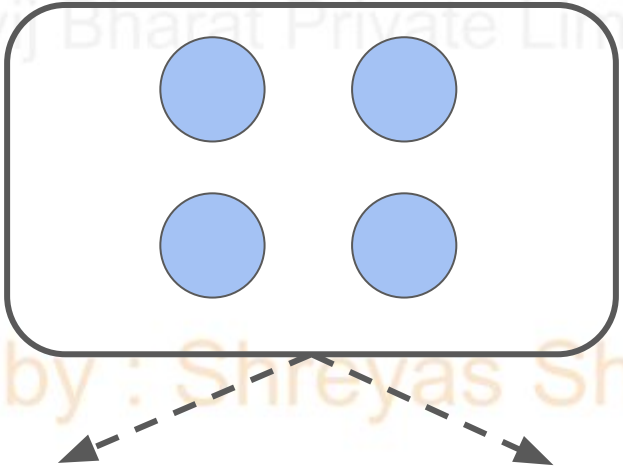
Class Blue (3/4)(1 - 3/4) =0.1875

Gini Impurity
0.1875+0.1875 = 0.375

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Data is completely “pure” (no impurity)

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



Class Red
(0/4)(1 - 0/4) = 0

+

Class Blue
(4/4)(1 - 4/4) = 0

Gini Impurity
0 + 0 = 0

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

Mastering Machine Learning with Python
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Decision Trees

Gini Impurity in Trees

Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ahvij Bharat Private Limited
Led 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.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Gini Impurity for Classification:

- For a set of classes **C** for a given dataset **Q**, $\mathbf{p_c}$ is probability of class **c**.

iHUB DivyaSampark, IIT Roorkee
and
Bharat Private Limited

$$p_c = \boxed{\frac{1}{N_Q} \sum_{x \in Q} \mathbb{1}(y_{class} = c)} \qquad G(Q) = \sum_{c \in C} \boxed{p_c}(1 - \boxed{p_c})$$

Led by : Shreyas Shukla

Create a decision tree to predict spam.

X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No

Mastering Machine Learning with Python
(27th Aug 2024 - 15th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Only one X feature to use for a node.

X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No

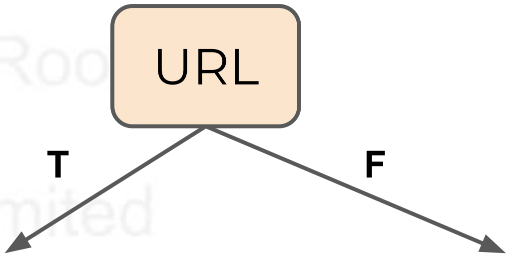
URL

iHUB DivyaSampark, IIT Roo
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Predict if email is spam if it contains a URL:

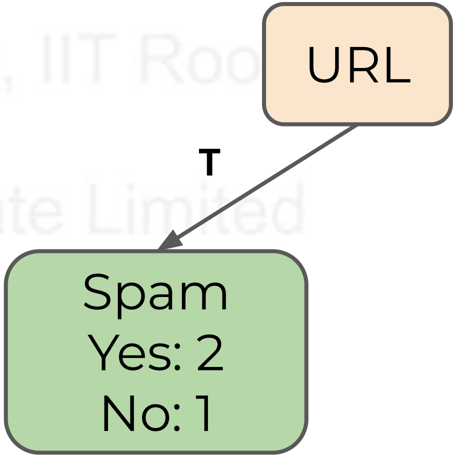
X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No



iHUB DivyaSampark, IIT Roo
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

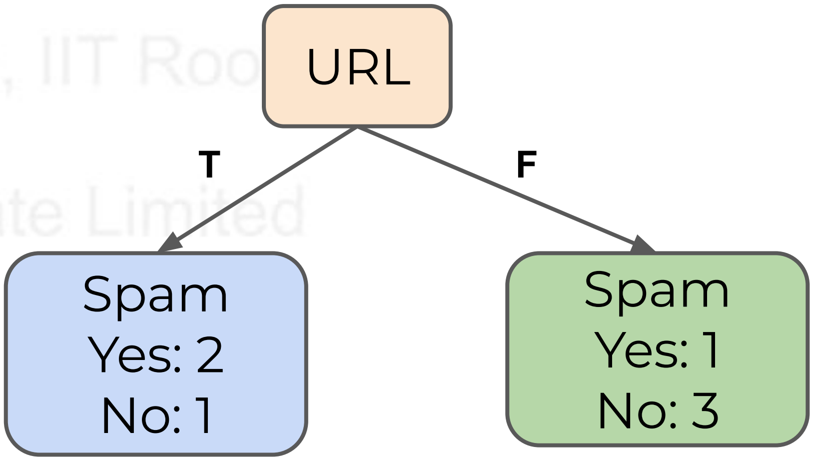
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

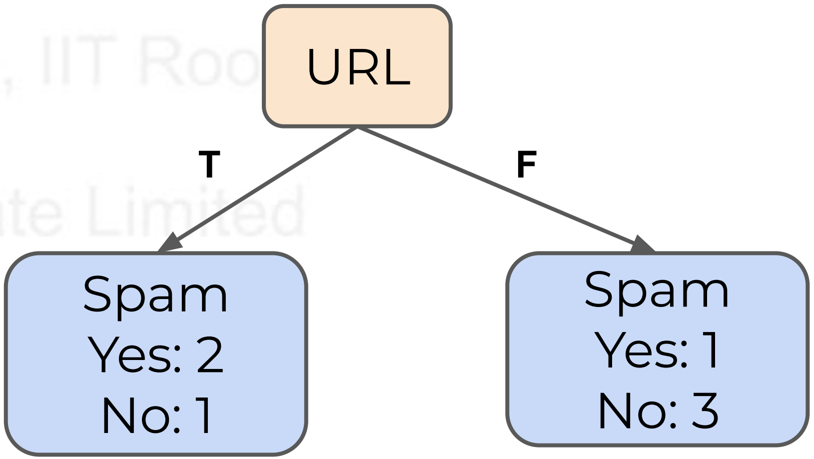
X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No



Led by : Shreyas Shukla

Predict if email is spam if it contains a URL:

X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No

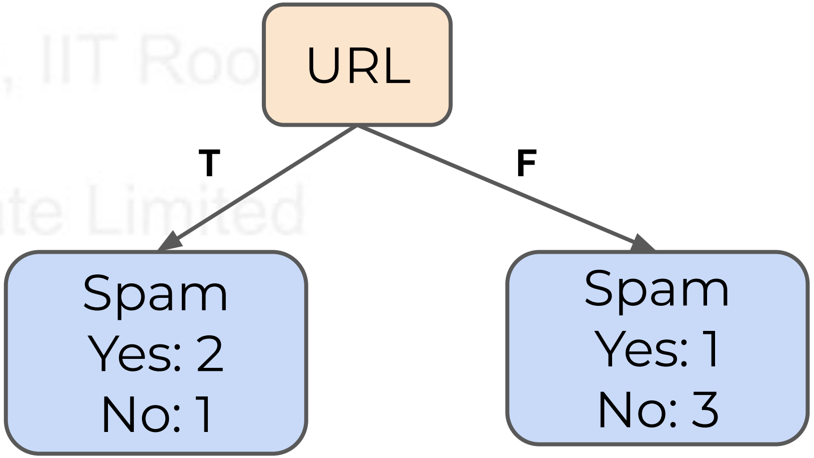


Mastering Machine Learning with Python (27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee and Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Recall the gini impurity formula:

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreya

X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No



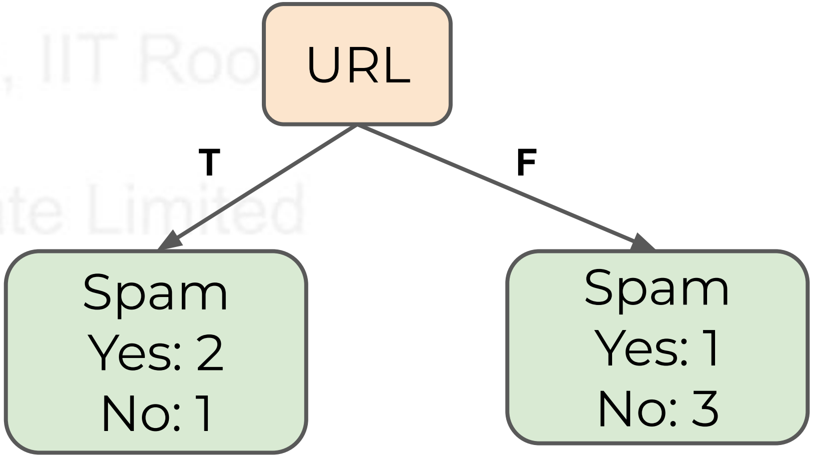
$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee and
Ritvij Bharat Private Limited
Led by : Shreya

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

Left Leaf Node:

$(2/3)(1-2/3) + (1/3)(1-1/3)$



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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

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

Left Leaf Node:

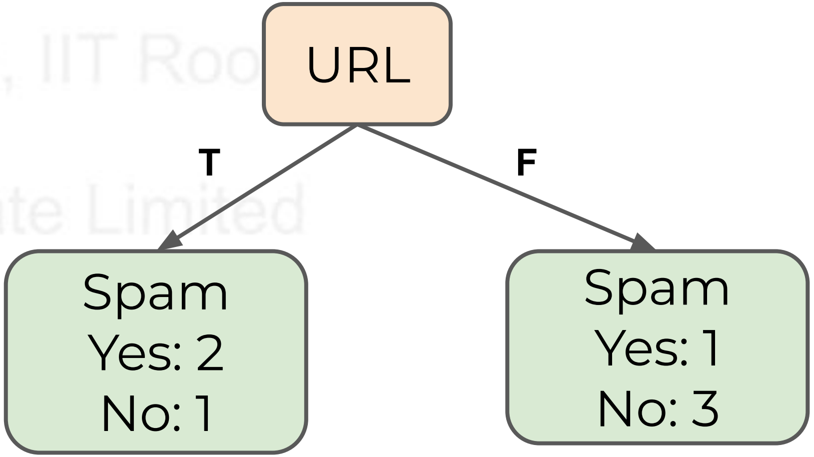
$(2/3)(1-2/3) + (1/3)(1-1/3)$

Left Leaf Gini=0.44

Right Leaf Node:

$(1/4)(1-1/4) + (3/4)(1-3/4)$

Right Leaf Gini=0.375



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

iHUB DivyaSampark, IIT Roorkee and Ritvij Bharat Private Limited
Led by : Shreya

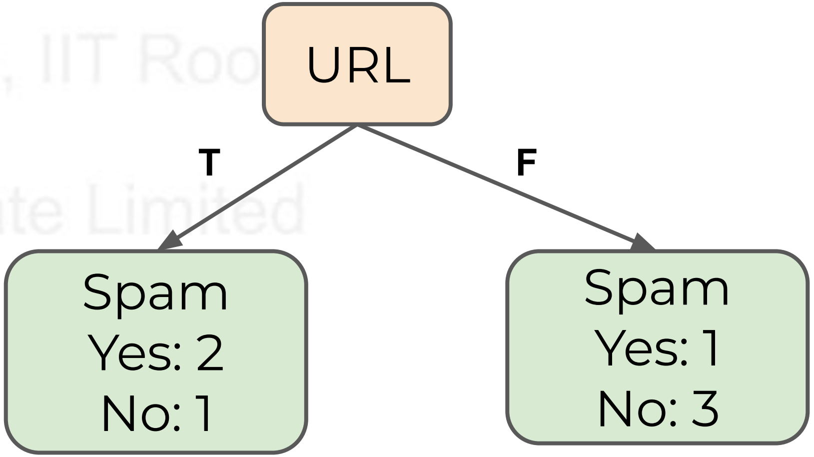
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Calculate gini impurity of URL feature.

Weighted Average of both:

Left Leaf Gini=0.44

Right Leaf Gini=0.375



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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreya

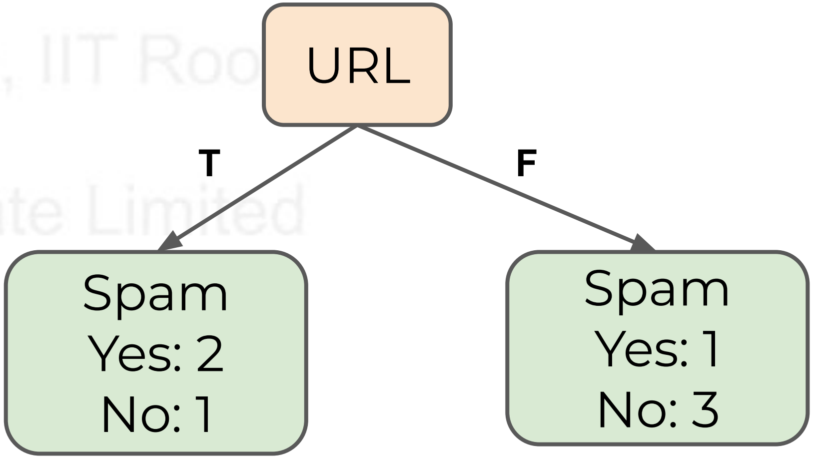
Total Emails: (2+1) + (1+3) = 7

Left Leaf Gini=0.44

Right Leaf Gini=0.375

Left Emails: 3

Right Emails: 4



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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roo
and
Ritvij Bharat Private Limited
Led by : Shreya

Total Emails: (2+1) + (1+3) = 7

Left Leaf Gini=0.44

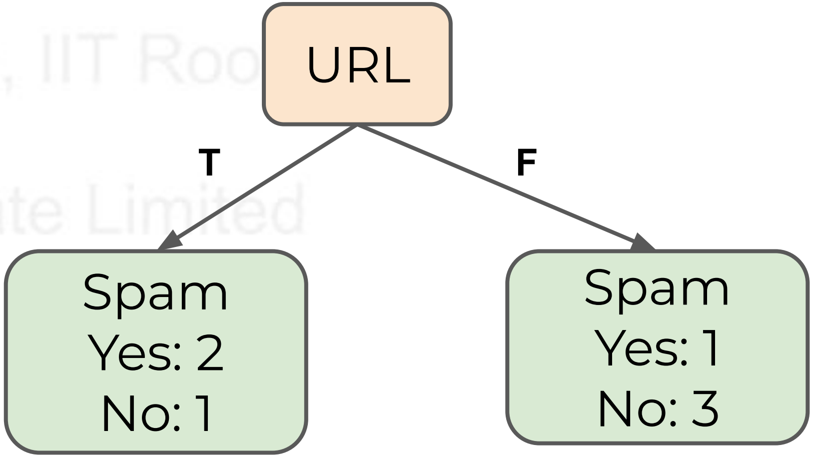
Right Leaf Gini=0.375

Left Emails: 3

Right Emails: 4

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

Gini Impurity: 0.403



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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

More issues to consider:

- Multiple Features
- 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.

Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Decision Trees

Gini Impurity Part Two

Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ikigai Bharat Private Limited
Led by : Shreyas Shukla

Let’s explore:

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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Imagine a continuous feature.
Calculate the feature gini impurity:

X - Words in Email	Y-Spam
10	Yes
40	No
20	Yes
50	No
30	No

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Sort data:

X - Words in Email	Y-Spam
10	Yes
40	No
20	Yes
50	No
30	No

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
IHUB Divya Sampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Calculate potential split values for node

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

Words ≤ N

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Use averages between rows as values:

Words ≤ N

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

15

25

35

45

...mpark, IIT Ro...
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
IHUB Divya Sampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Perform all the potential split:

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

15

25

35

45

Words ≤ 15

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Calculate gini impurity for each split:

Words ≤ 15

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

15

iHUB DivyaSampark, IIT Roorkee and Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Calculate gini impurity for each split:

Words ≤ 15

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

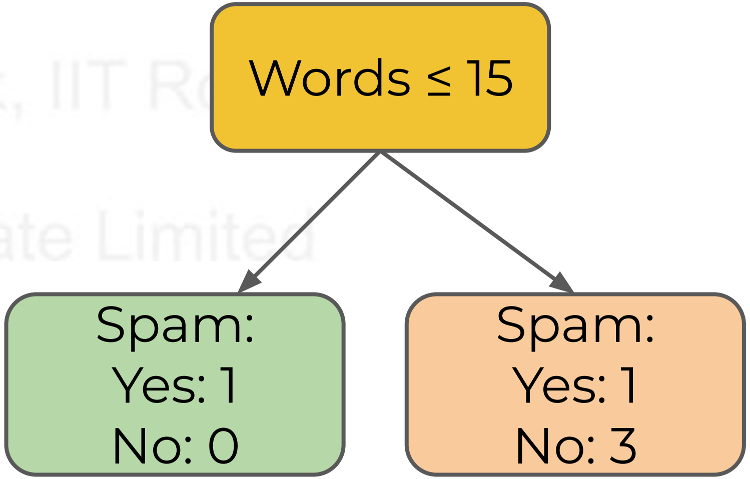
15

iHUB DivyaSampark, IIT Roorkee
and
Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Calculate gini impurity for each split:

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No



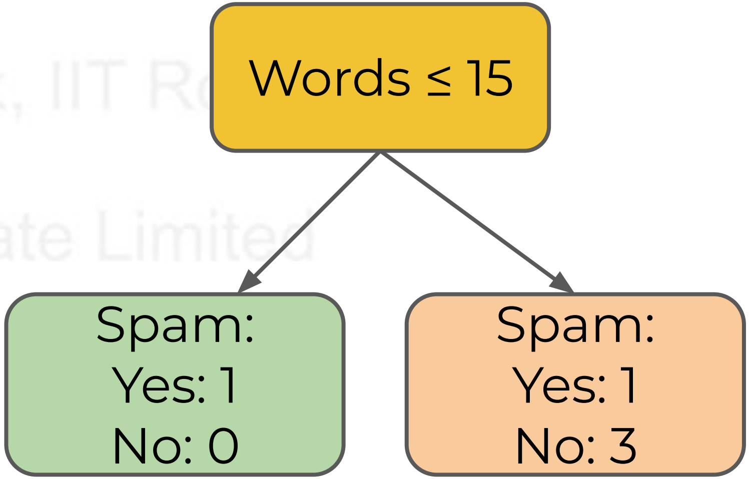
$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Calculate gini impurity for each split:

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

15



$$G(Q) = (\tfrac{1}{5})(0+0) + (\tfrac{4}{5})((\tfrac{1}{4})(1-\tfrac{1}{4})+(\tfrac{3}{4})(1-\tfrac{3}{4})$$
$$= 0.3$$

...park, IIT R...
and
Private Limited
Led by : Sh... Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Do it for all possible splits:

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

15 → Gini=0.3

25 → Gini=0

35 → Gini=0.26

45 → Gini=0.4

…park, IIT Roorkee
and
RiVi Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Choose lowest impurity split value

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

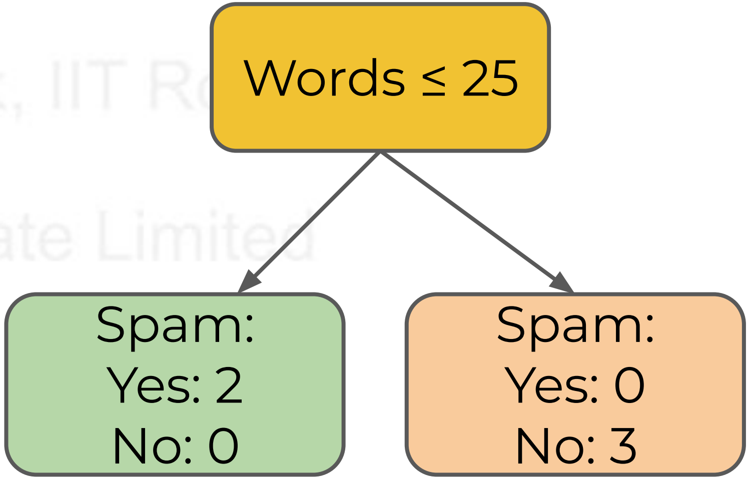
25 → Gini=0

Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Choose this as split value for node

X - Words in Email	Y-Spam
10	Yes
20	Yes
25	
30	No
40	No
50	No



$$G(Q) = 0$$

Multicategorical feature

Calculate gini impurity for all combinations:

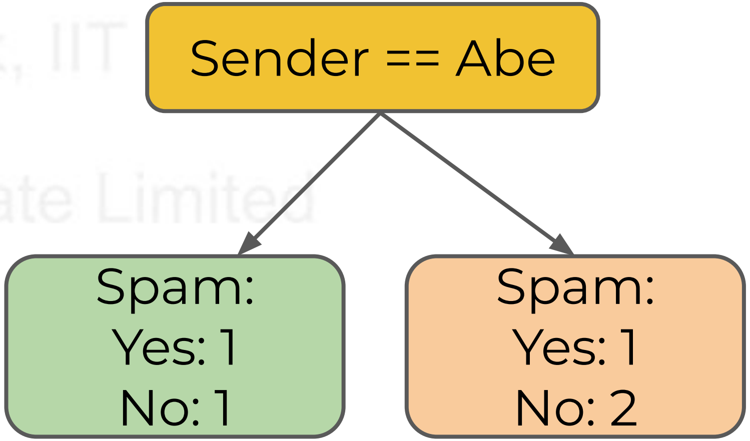
X - Sender	Y-Spam
Abe	Yes
Bob	Yes
Claire	No
Abe	No
Bob	No

Mastering Machine Learning with Python (27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee and Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Calculate gini impurity for all combinations:

X - Sender	**Y-Spam**
Abe	Yes
Bob	Yes
Claire	No
Abe	No
Bob	No

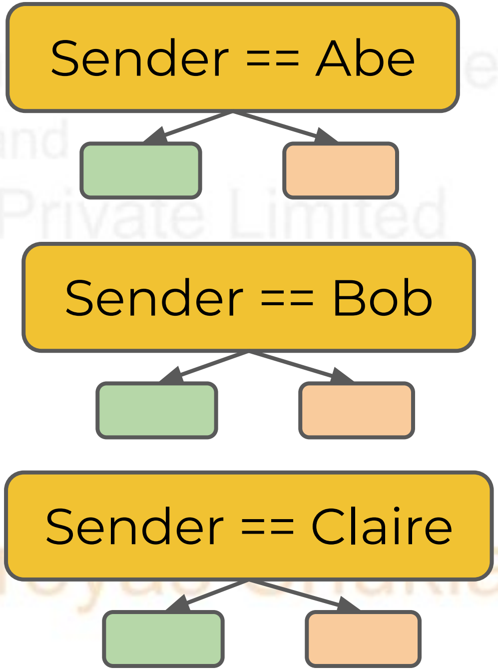


...park, IIT
...and
Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
Ritvij Bharat Private Limited

Calculate gini impurity for all combinations

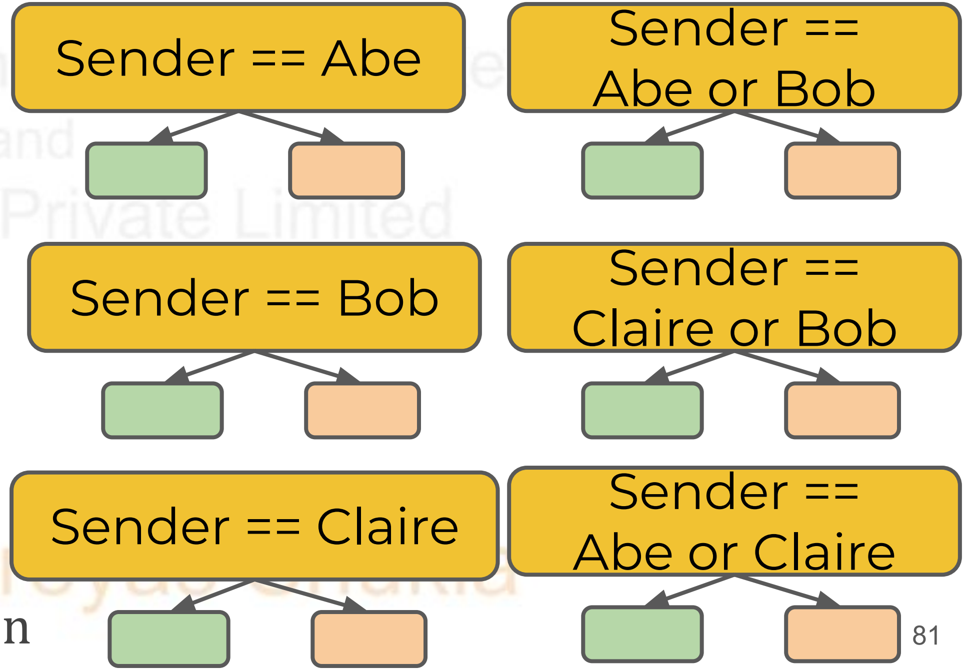
X - Sender	Y-Spam
Abe	Yes
Bob	Yes
Claire	No
Abe	No
Bob	No



Calculate gini impurity for all combinations

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
Ritvij Bharat Private Limited

X - Sender	Y-Spam
Abe	Yes
Bob	Yes
Claire	No
Abe	No
Bob	No



Choose lowest impurity split combination

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.

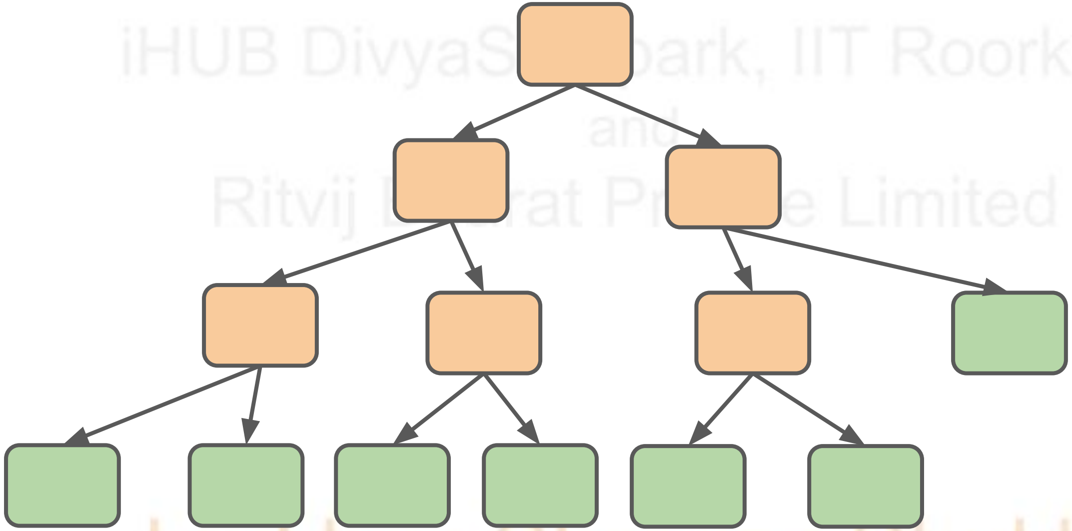
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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.

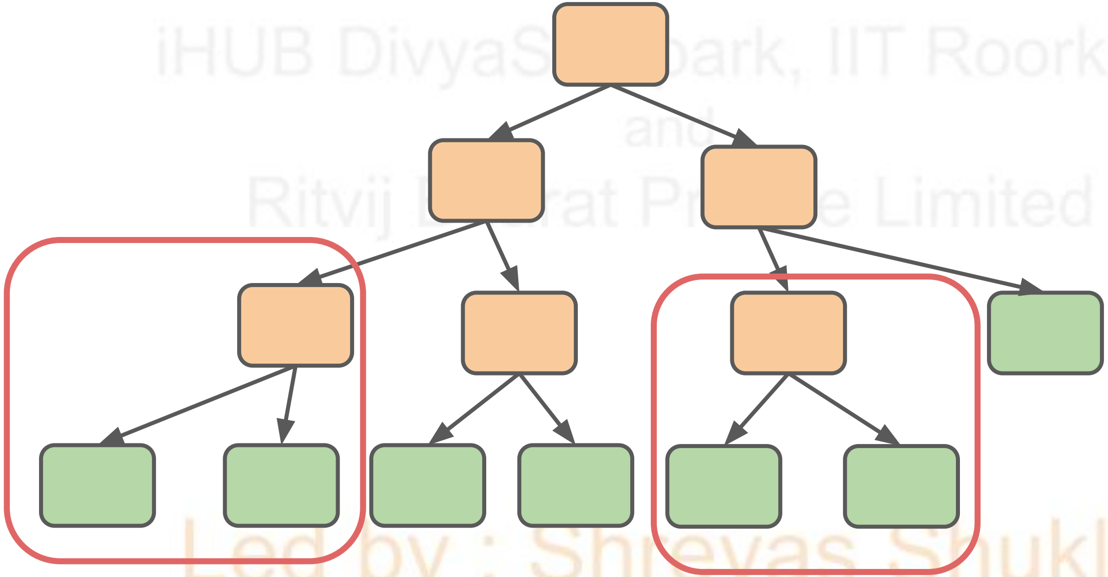
A large overfitted tree.

Add minimum gini impurity decrease



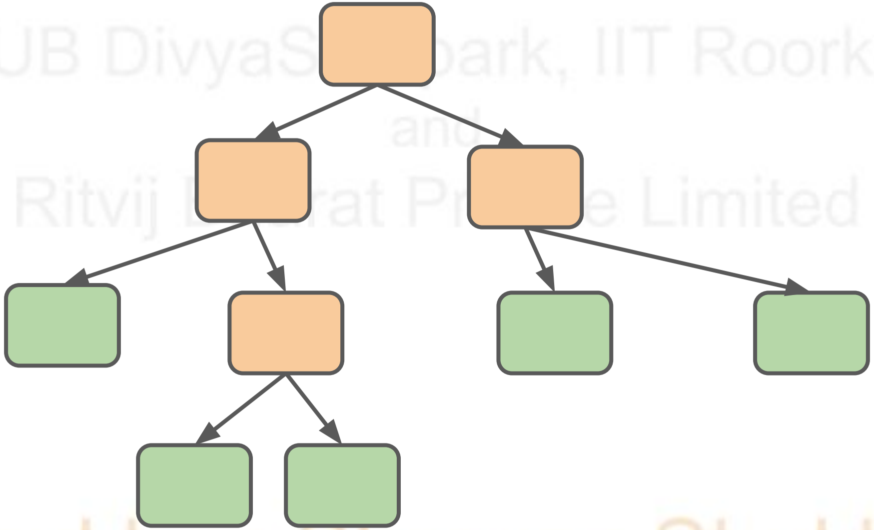
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)



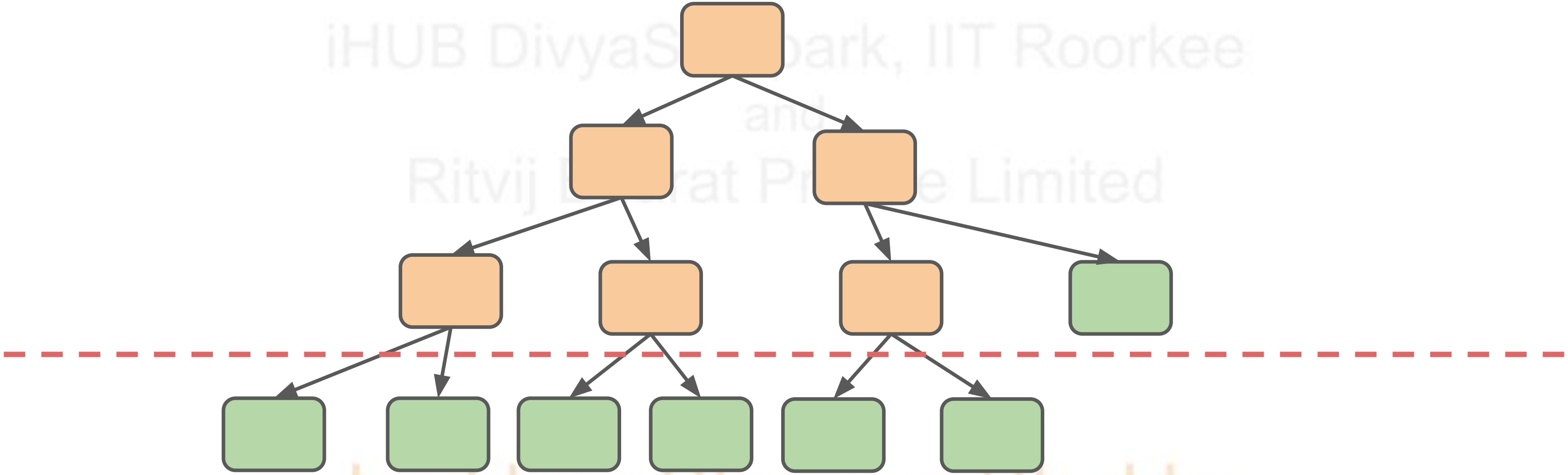
iHUB DivyaS... park, IIT Roorkee
and
Ritvij B... rat Pr... e Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

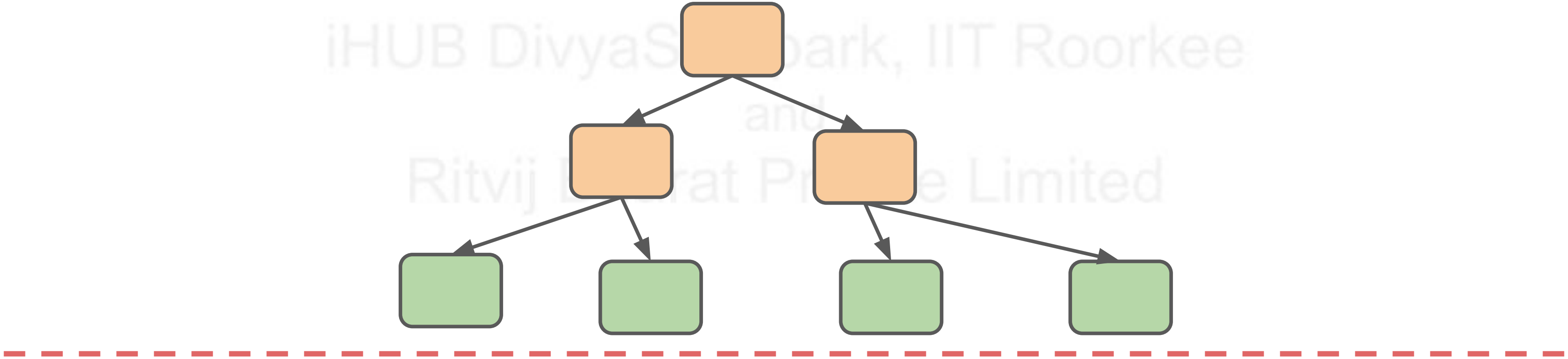


Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

We can also mandate a max depth



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Let’s code !!