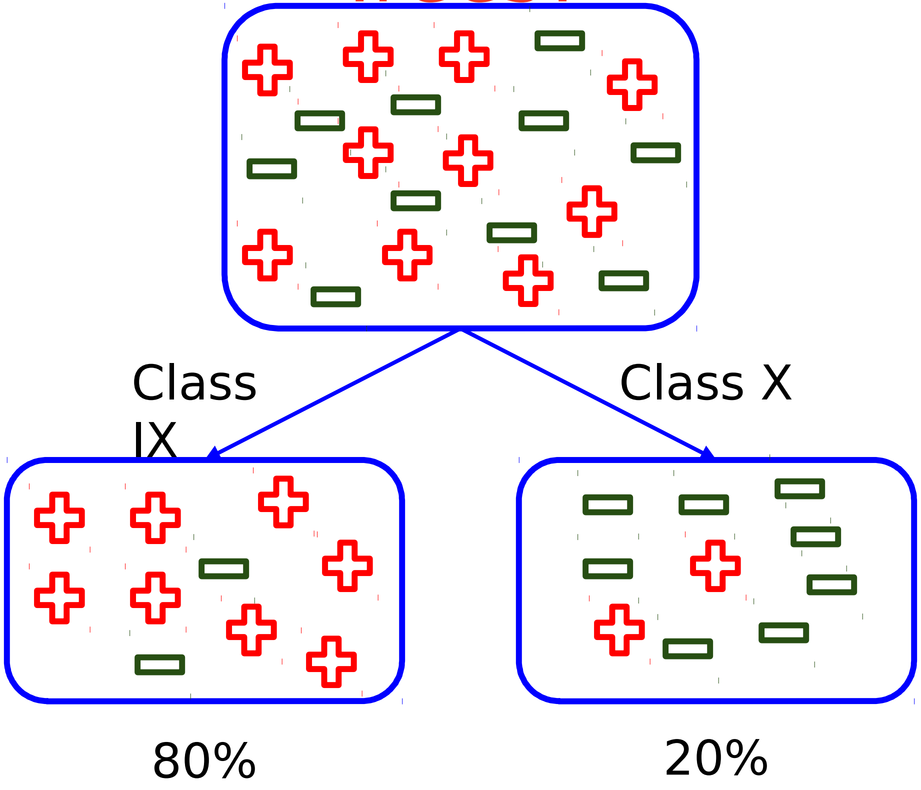
How to select best split point in Decision Trees?

Analytics Vidhya
Learn everything about analytics

How to select best split point in Decision Trees?



Split on Class

Analytics Vidhya
Learn everything about analytics

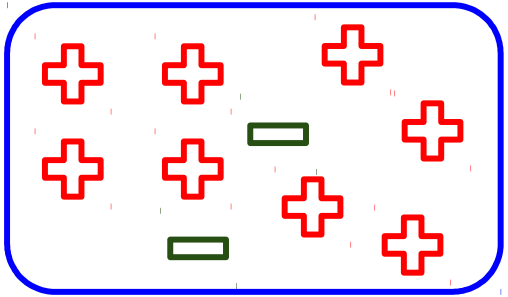
How to select best split point in Decision Trees?

- Decision tree splits the nodes on all available variables
- Selects the split which results in most homogeneous sub-nodes

Gini Impurity

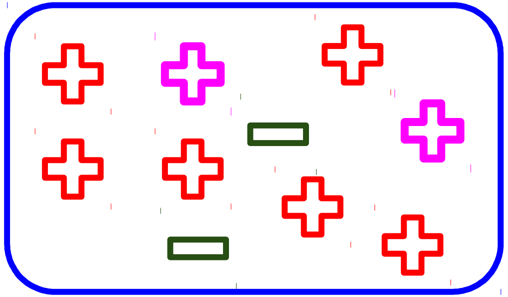
Gini Impurity = 1 - Gini

Gini

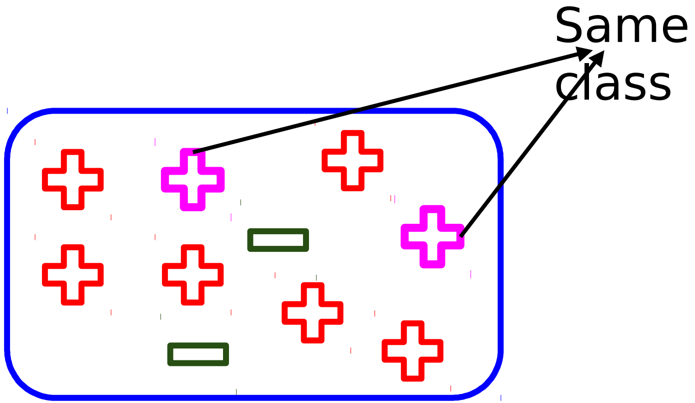


Analytics Vidhya
Learn everything about analytics

Gini



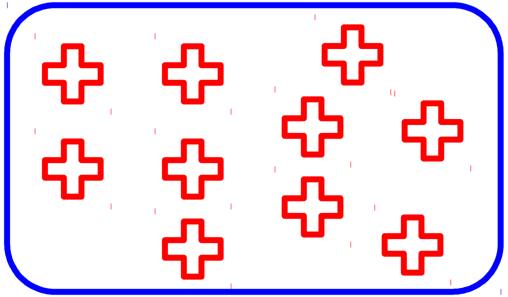
Gini



If we select two items from a population at random, they must be of same class

Analytics Vidhya
Learn everything about analytics

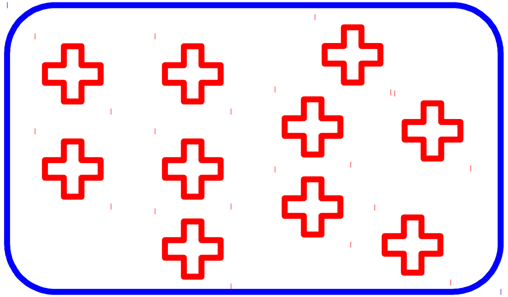
Gini



Probability that randomly picked points belong to same class?

Analytics Vidhya
Learn everything about analytics

Gini



Probability = 1

Properties of Gini Impurity

- Node split is decided based on the gini impurity

$$Gini\ Impurity = 1 - Gini$$

- Lower the gini impurity, higher the homogeneity of nodes
- Works only with categorical targets
- Only performs binary splits

Steps to calculate Gini Impurity for a split

- Calculate the gini impurity for sub-nodes :

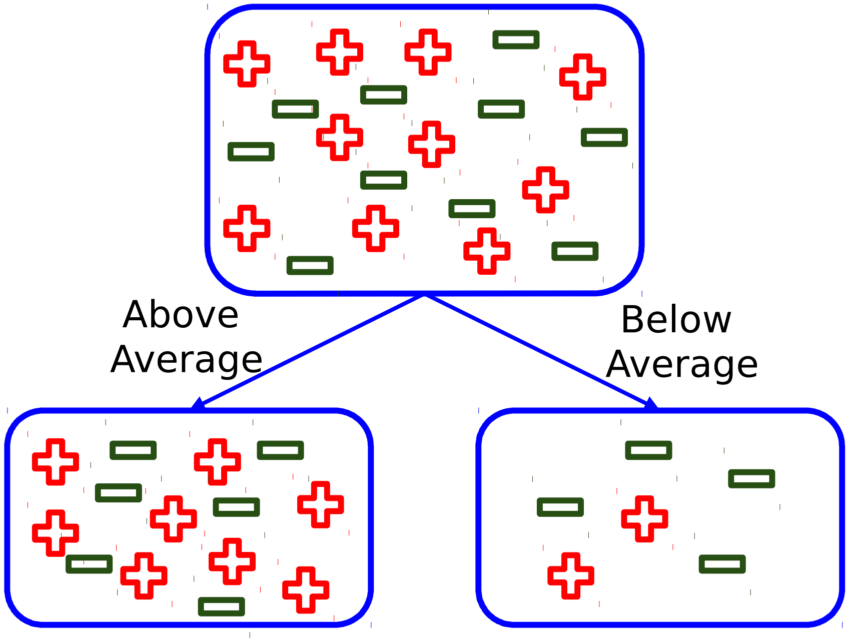
$$\textit{Gini Impurity} = 1 - \textit{Gini}$$

- Gini = Sum of square of probabilities for each class/category

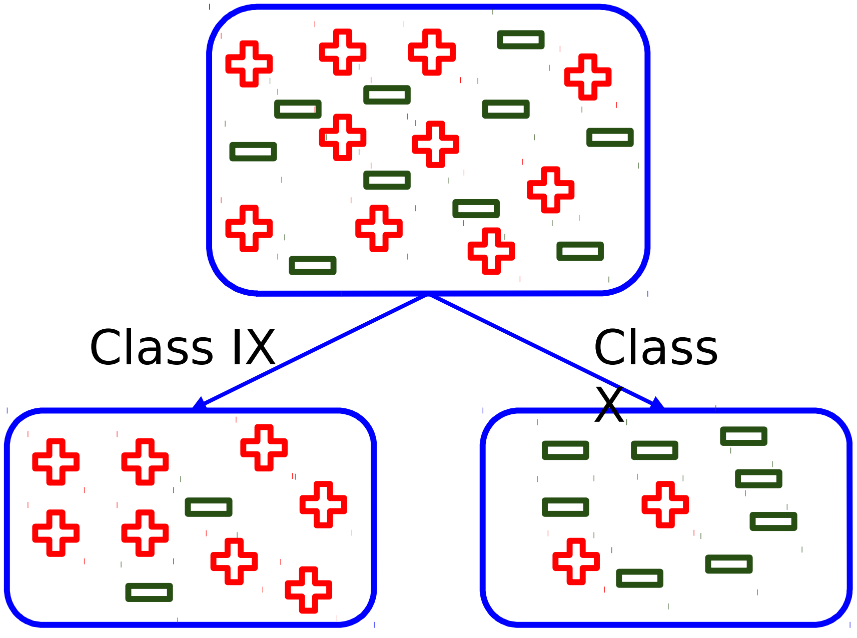
$$\textit{Gini} = (p_1^2 + p_2^2 + p_3^2 + \ldots + p_n^2)$$

- To calculate the gini impurity for split, take weighted gini impurity of both sub-nodes of that split

Steps to calculate Gini Impurity for a split



Split on Performance in Class



Split on Class

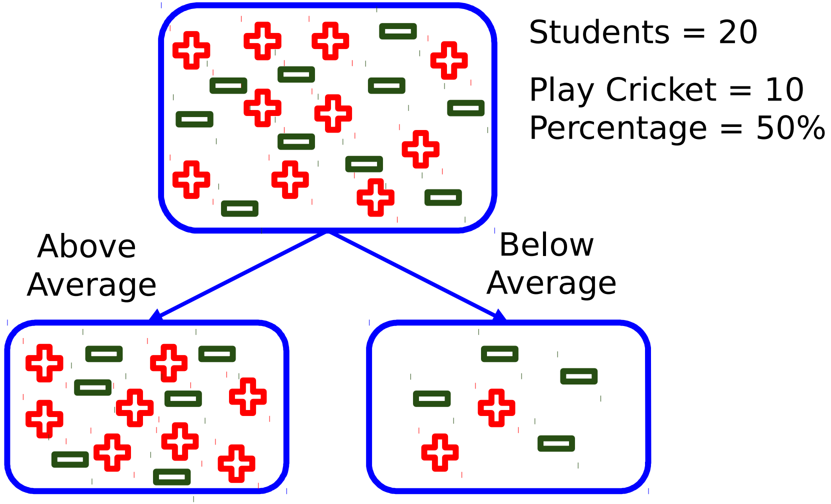
Analytics Vidhya
Learn everything about analytics

Steps to calculate Gini Impurity for a split

Split on Performance in Class

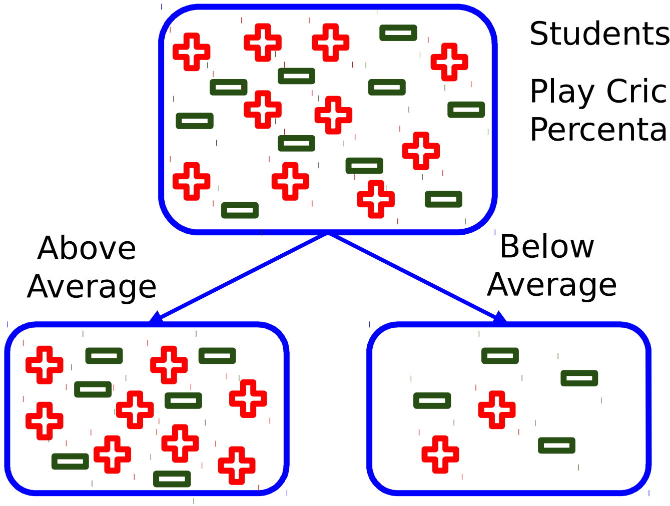
Steps to calculate Gini Impurity for a split

Split on Performance in Class



Steps to calculate Gini Impurity for a split

Split on Performance in Class



Above Average

Below Average

Students = 20

Play Cricket = 10
Percentage = 50%

Students = 14
Play Cricket = 8
Do not play = 6
Prob. play = 0.57
Prob. Not play = 0.43

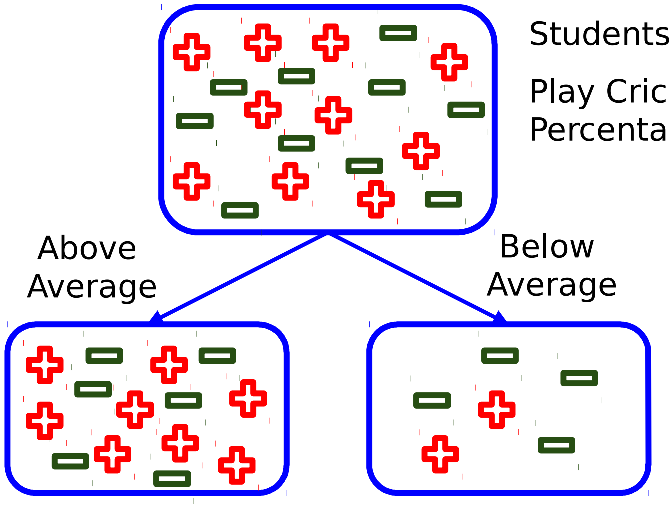
Analytics Vidhya
Learn everything about analytics

Steps to calculate Gini Impurity for a split

Split on Performance in Class

- Gini Impurity: sub-node Above Average:
  1 - [(0.57)*(0.57) + (0.43)*(0.43)] = 0.49

- Gini Impurity: sub-node Below Average:
  1 - [(0.33)*(0.33) + (0.67)*(0.67)] = 0.44



Students = 20

Play Cricket = 10
Percentage = 50%

Above Average

Below Average

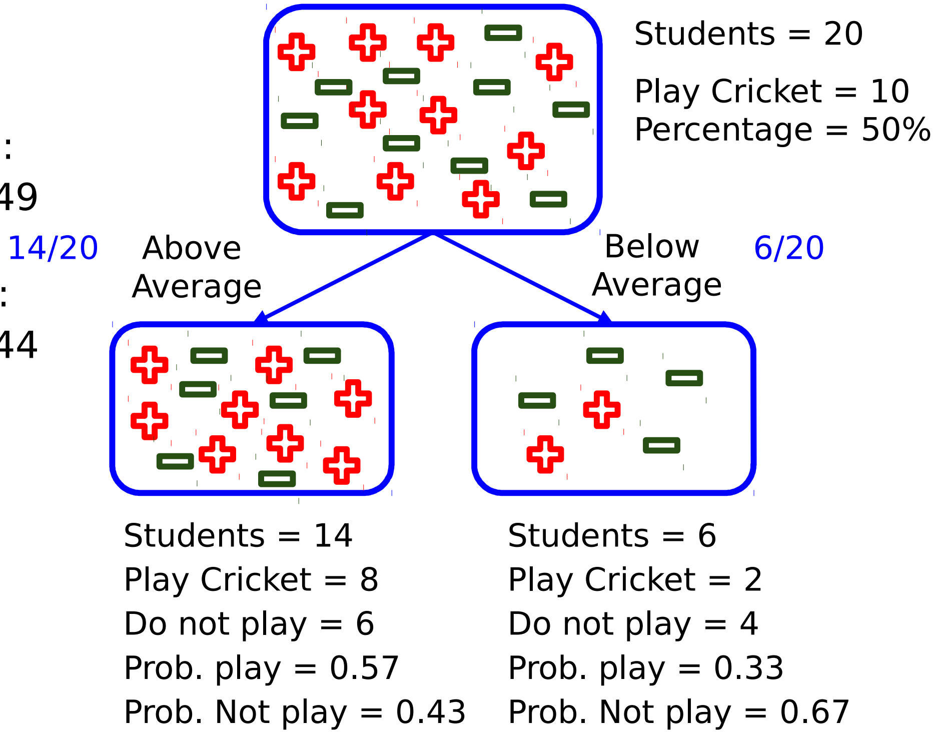
Students = 14
Play Cricket = 8
Do not play = 6
Prob. play = 0.57
Prob. Not play = 0.43

Students = 6
Play Cricket = 2
Do not play = 4
Prob. play = 0.33
Prob. Not play = 0.67

Steps to calculate Gini Impurity for a split

Split on Performance in Class

- Gini Impurity: sub-node Above Average:
  1 - [(0.57)*(0.57) + (0.43)*(0.43)] = 0.49
- Gini Impurity: sub-node Below Average:
  1 - [(0.33)*(0.33) + (0.67)*(0.67)] = 0.44



Students = 20

Play Cricket = 10
Percentage = 50%

14/20 Above Average

Below Average 6/20

Students = 14
Play Cricket = 8
Do not play = 6
Prob. play = 0.57
Prob. Not play = 0.43

Students = 6
Play Cricket = 2
Do not play = 4
Prob. play = 0.33
Prob. Not play = 0.67

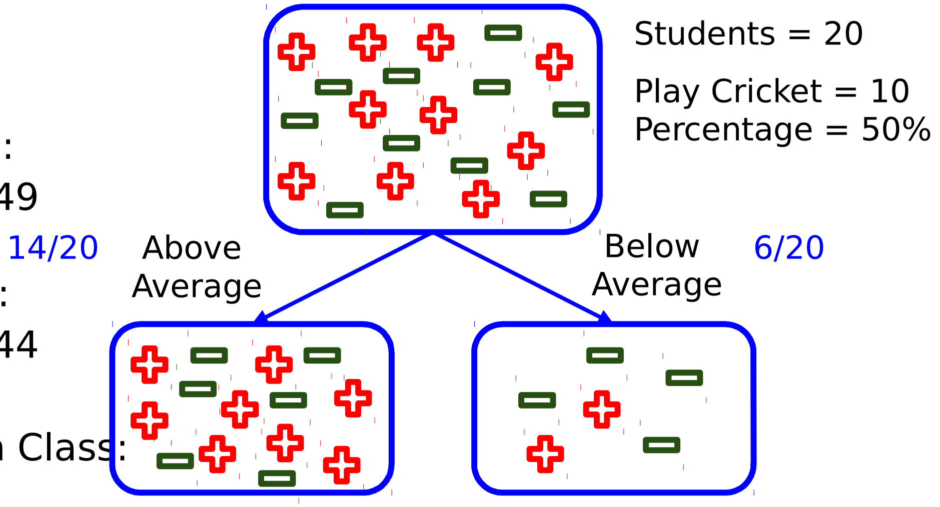
Steps to calculate Gini Impurity for a split

Split on Performance in Class

- Gini Impurity: sub-node Above Average:
  1 - [(0.57)*(0.57) + (0.43)*(0.43)] = 0.49

- Gini Impurity: sub-node Below Average:
  1 - [(0.33)*(0.33) + (0.67)*(0.67)] = 0.44

- Weighted Gini Impurity: Performance in Class:
  (14/20)*0.49 + (6/20)*0.44 = 0.475



Students = 20

Play Cricket = 10
Percentage = 50%

Students = 14
Play Cricket = 8
Do not play = 6
Prob. play = 0.57
Prob. Not play = 0.43

Students = 6
Play Cricket = 2
Do not play = 4
Prob. play = 0.33
Prob. Not play = 0.67

Analytics Vidhya
Learn everything about analytics

Steps to calculate Gini Impurity for a split

Split on Class

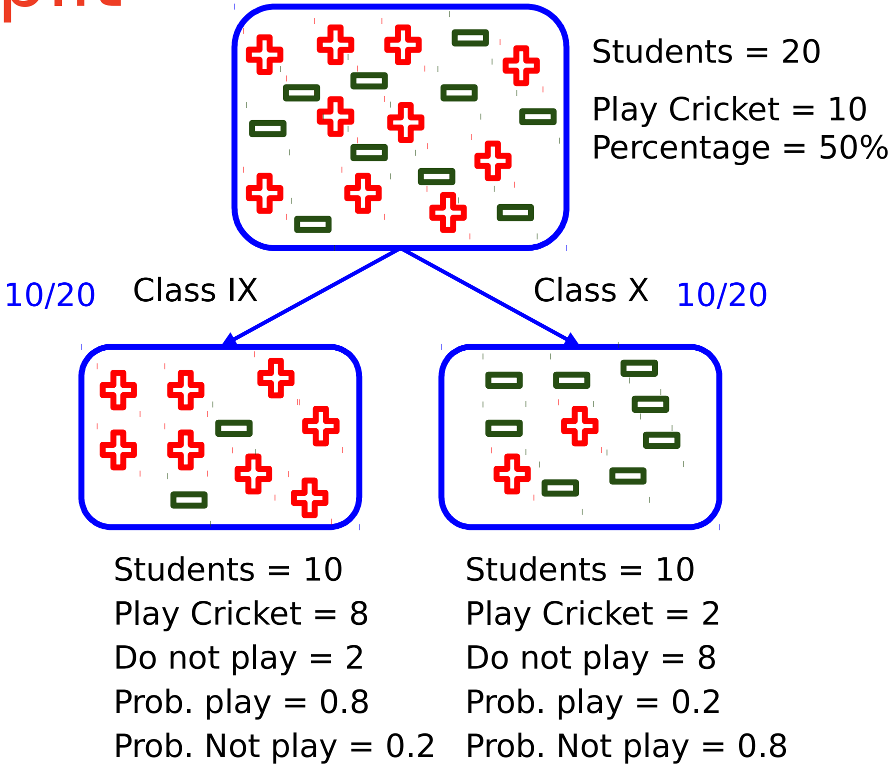
Steps to calculate Gini Impurity for a split

Split on Class

- Gini Impurity: sub-node Class IX:
  1 - [(0.8)*(0.8) + (0.2)*(0.2)] = 0.32

- Gini Impurity: sub-node Class X:
  1 - [(0.2)*(0.2) + (0.8)*(0.8)] = 0.32

- Weighted Gini Impurity: Class:
  (10/20)*0.32 + (10/20)*0.32 = 0.32



Students = 20

Play Cricket = 10
Percentage = 50%

Students = 10
Play Cricket = 8
Do not play = 2
Prob. play = 0.8
Prob. Not play = 0.2

Students = 10
Play Cricket = 2
Do not play = 8
Prob. play = 0.2
Prob. Not play = 0.8

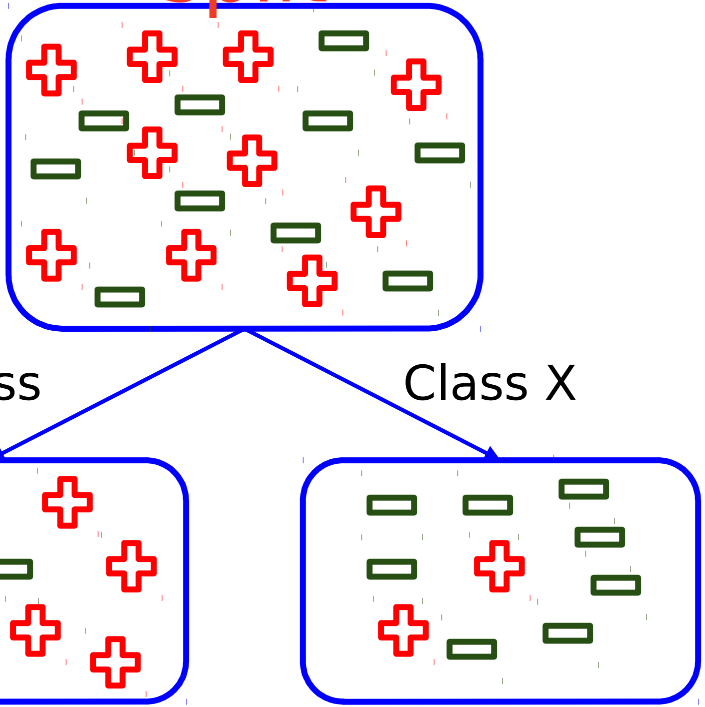
Steps to calculate Gini Impurity for a split

Split	**Weighted Gini Impurity**
Performance in Class	0.475
Class	0.32

Steps to calculate Gini Impurity for a split

Split	Weighted Gini Impurity
Performance in Class	0.475
Class	0.32

Steps to calculate Gini Impurity for a split



Split on Class

Thank You!

Analytics Vidhya
Learn everything about analytics