The Gini impurity index Intuition

Measuring the diversity in a dataset

Which one is more diverse?

Which one is more diverse?

Which one is more diverse?

Gini = 0.42

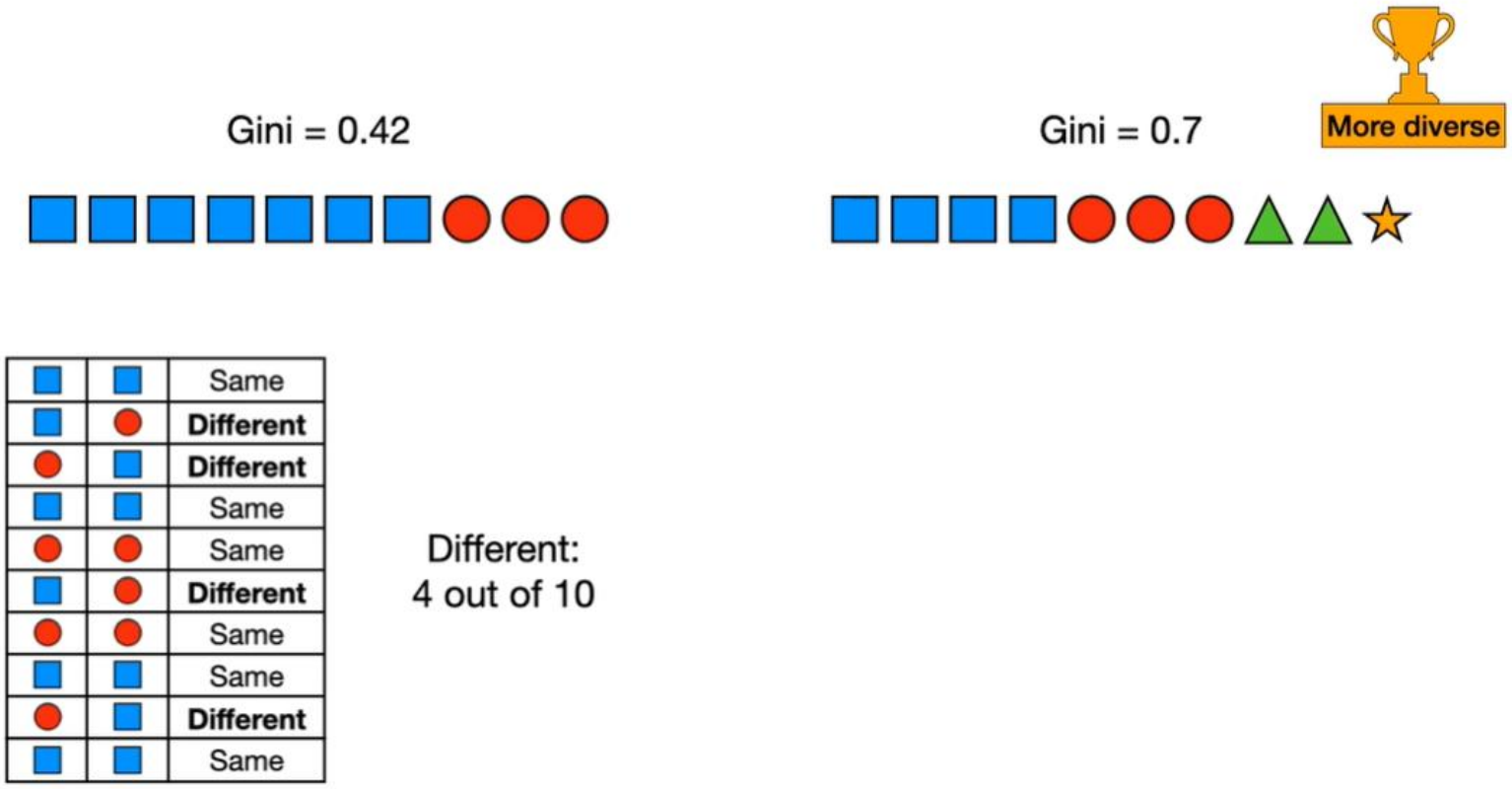
Gini = 0.7

How to calculate Gini index?

- Randomly pick any two elements from the dataset and determine if they are similar or different.

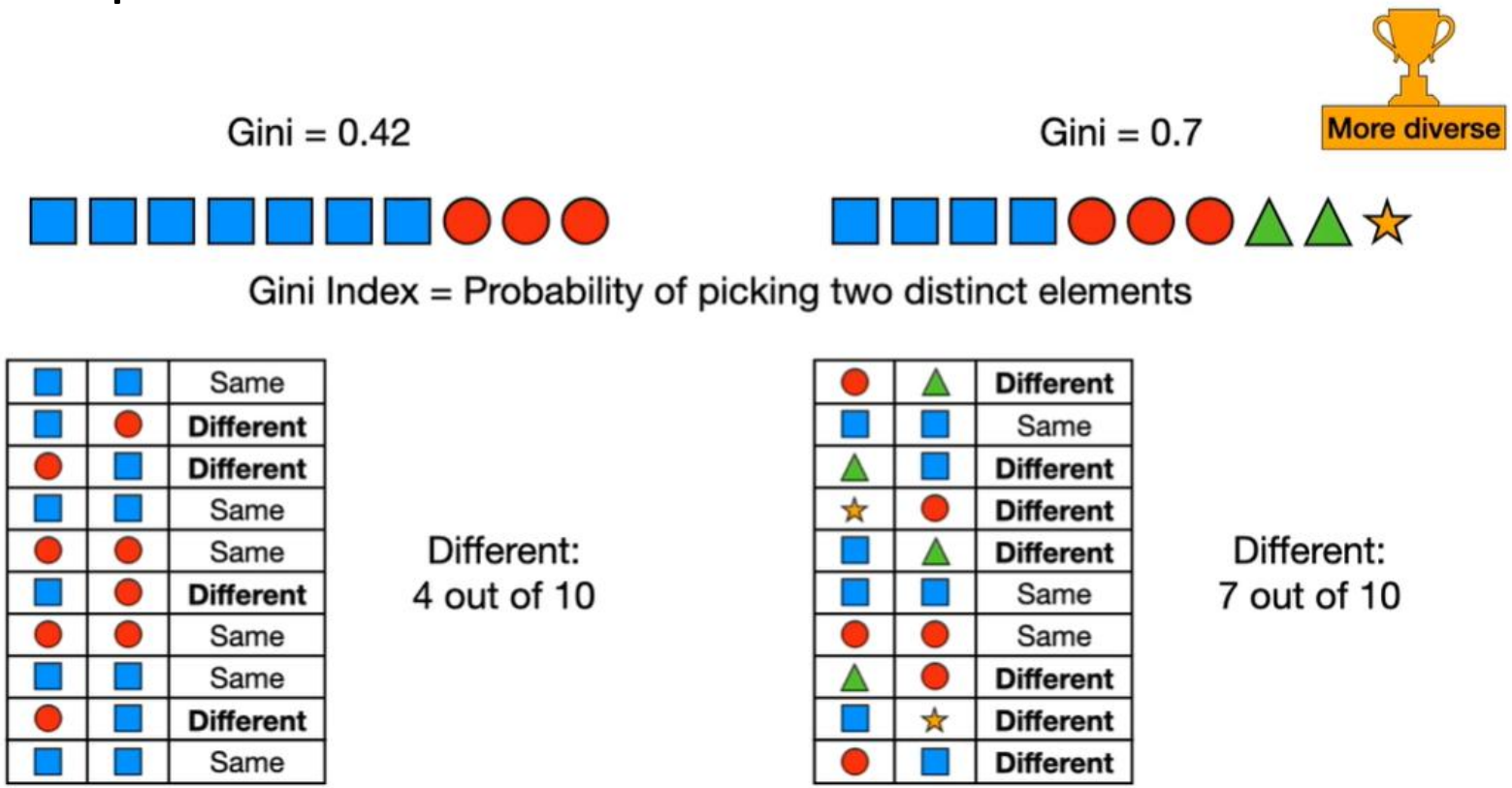
How to calculate Gini index?

- Repeat the process a number of times.

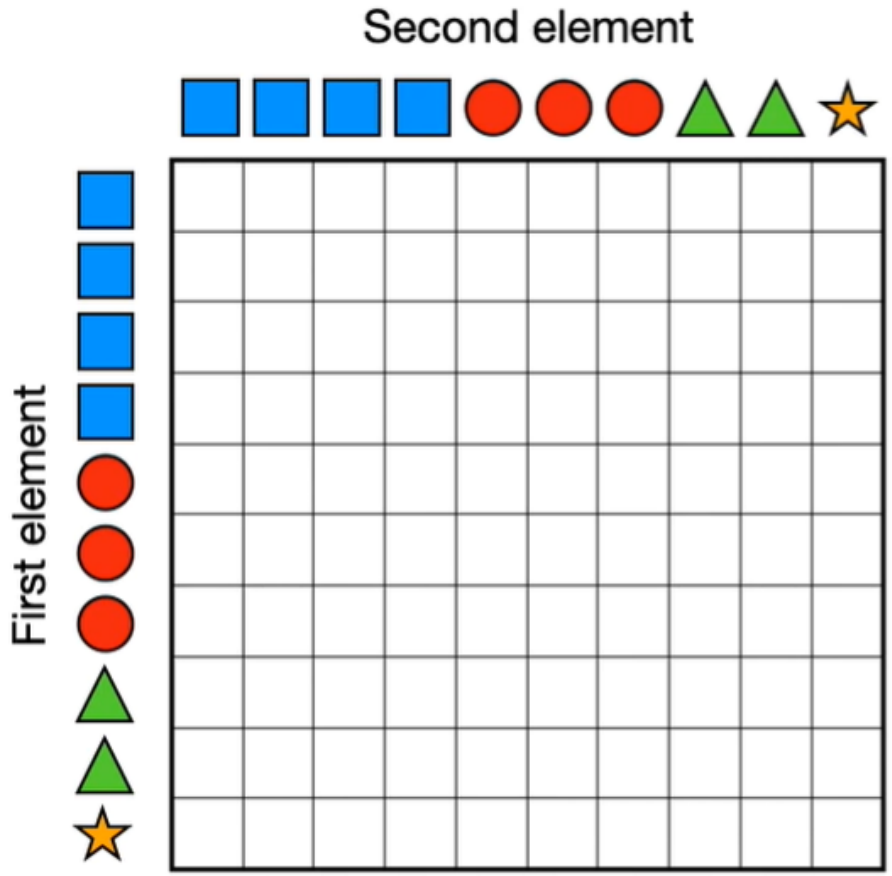


How to calculate Gini index?

- Do the same process for the second dataset.

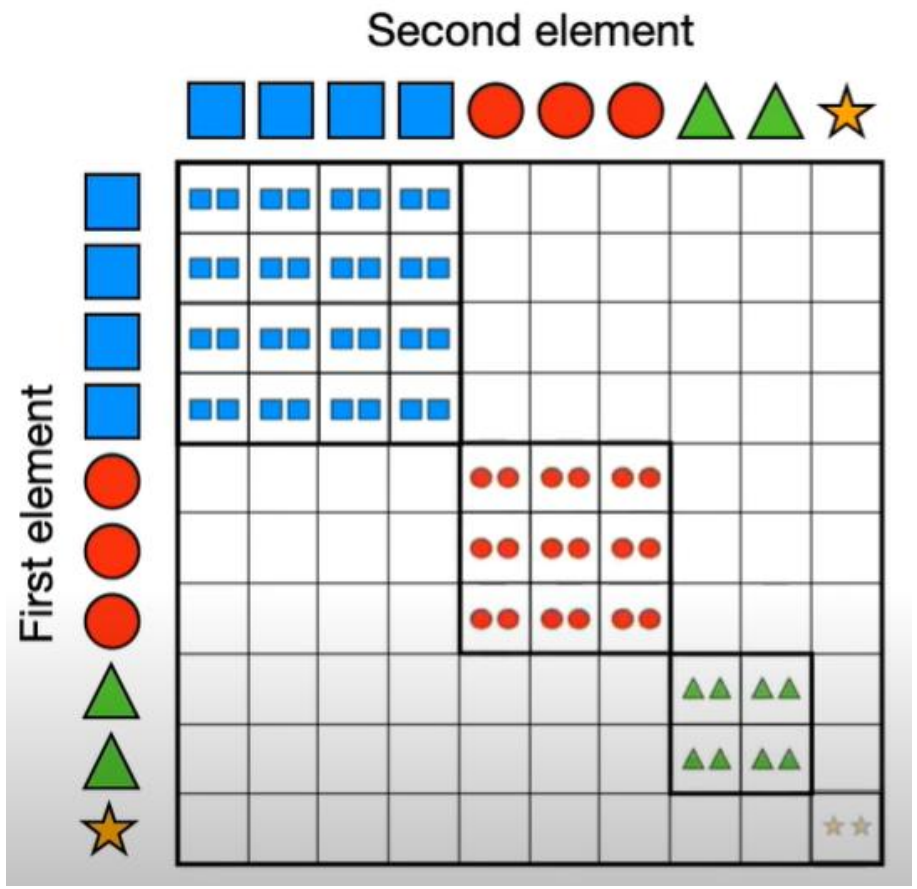


How to calculate Gini index?



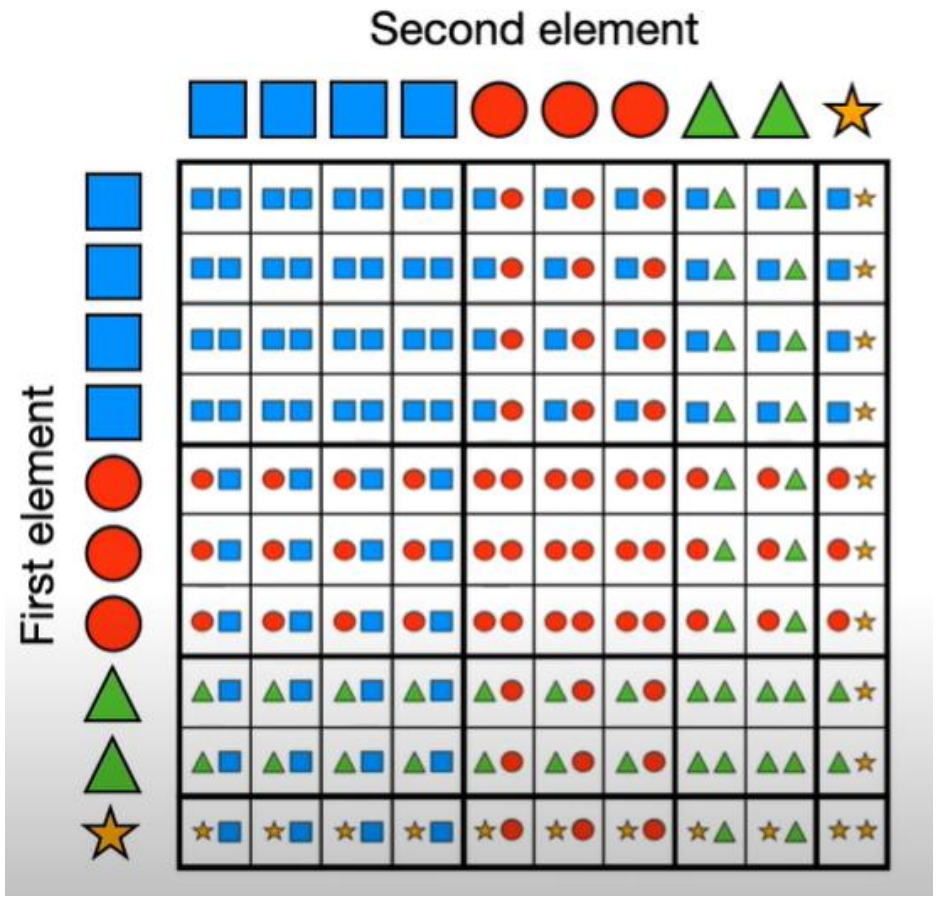
How to calculate Gini index?

- Let's start with the second dataset.
- Enumerate all the possible combinations.



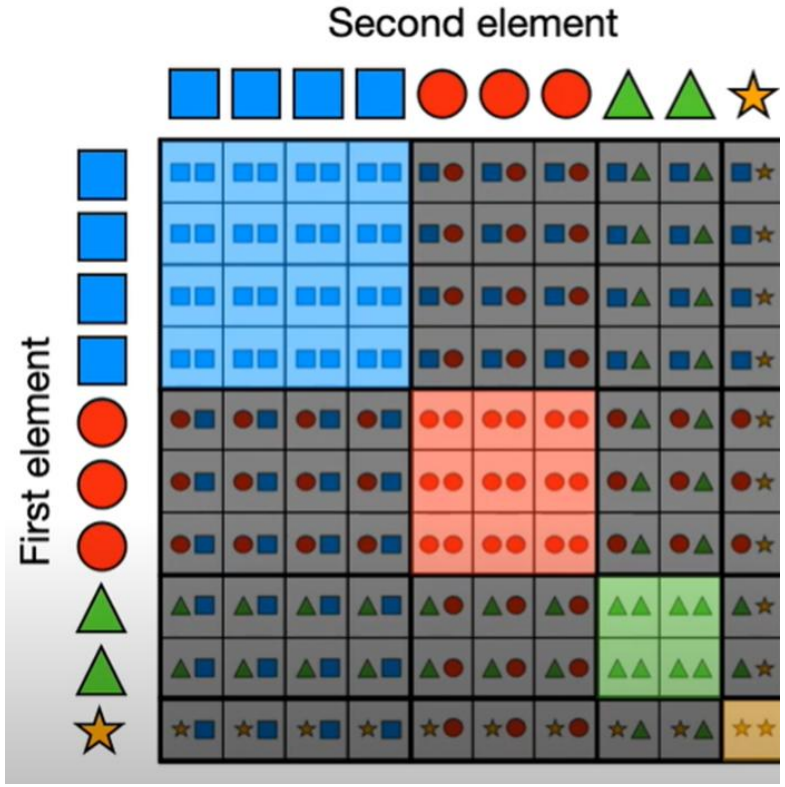
How to calculate Gini index?

- Let's start with the second dataset.
- Enumerate all the possible combinations.



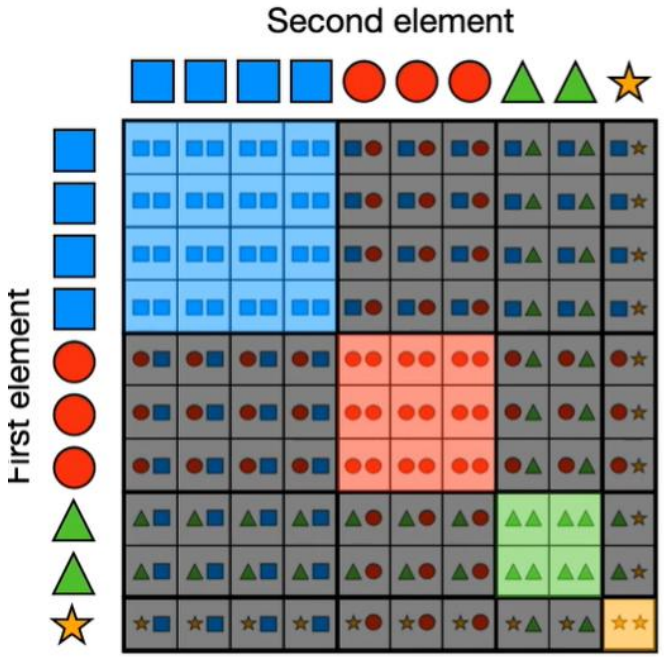
How to calculate Gini index?

- We need to calculate the percentage of the black area (When both elements are different).



How to calculate Gini index?

- We need to calculate the percentage of the black area (When both elements are different).

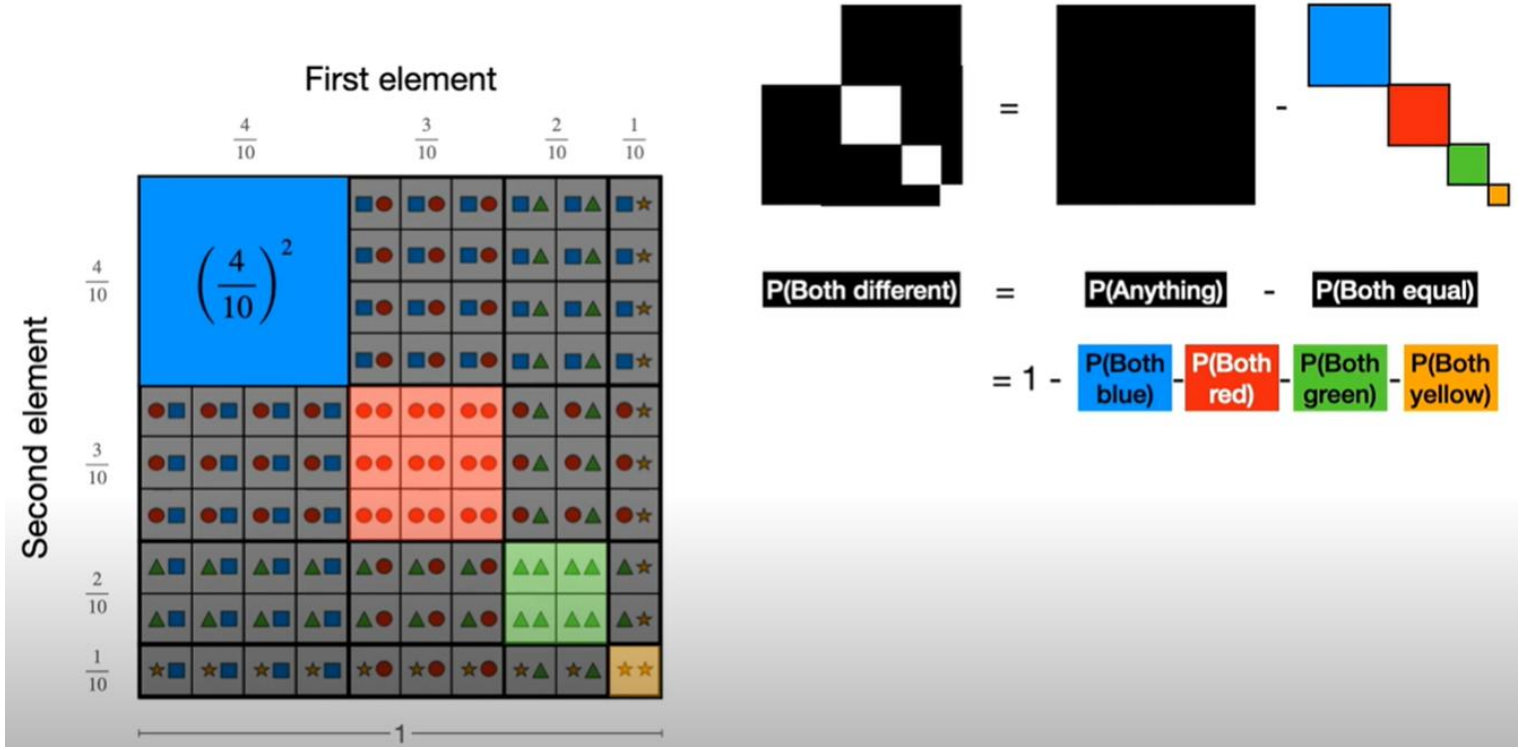


P(Both different) = P(Anything) - P(Both equal)

= 1 - P(Both blue) - P(Both red) - P(Both green) - P(Both yellow)

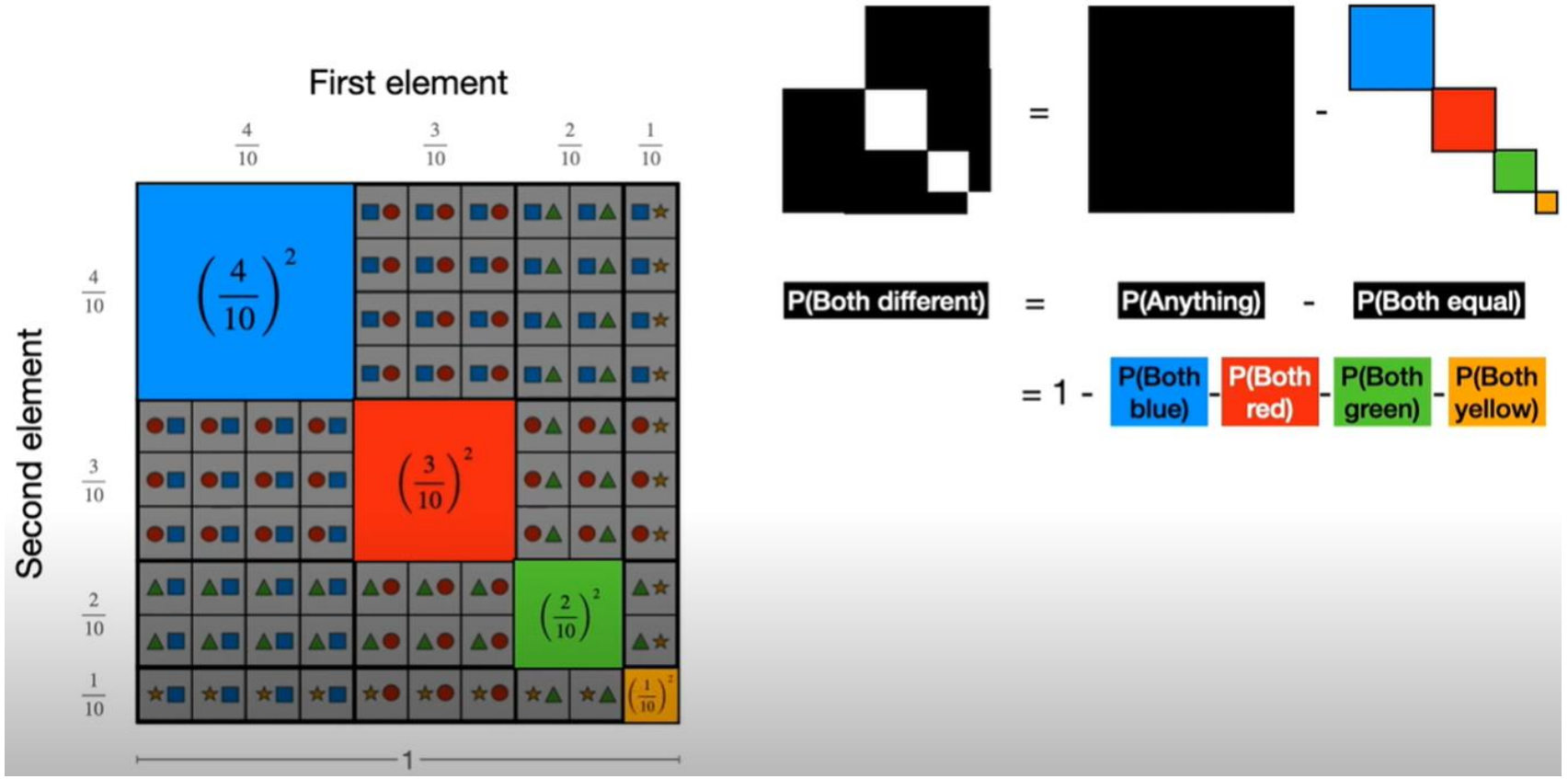
How to calculate Gini index?

- We need to calculate the percentage of the black area (When both elements are different).



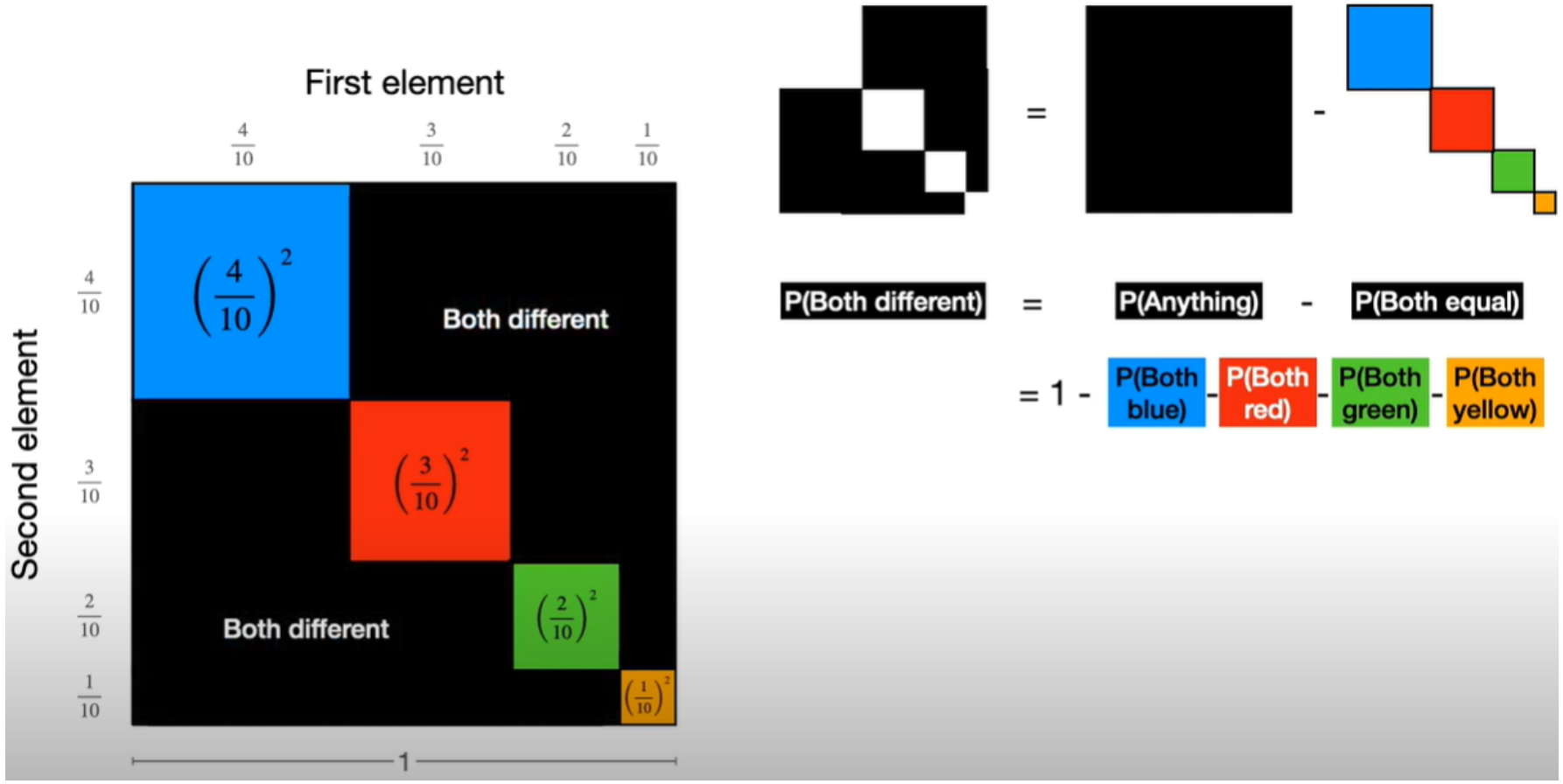
How to calculate Gini index?

- We need to calculate the black area.



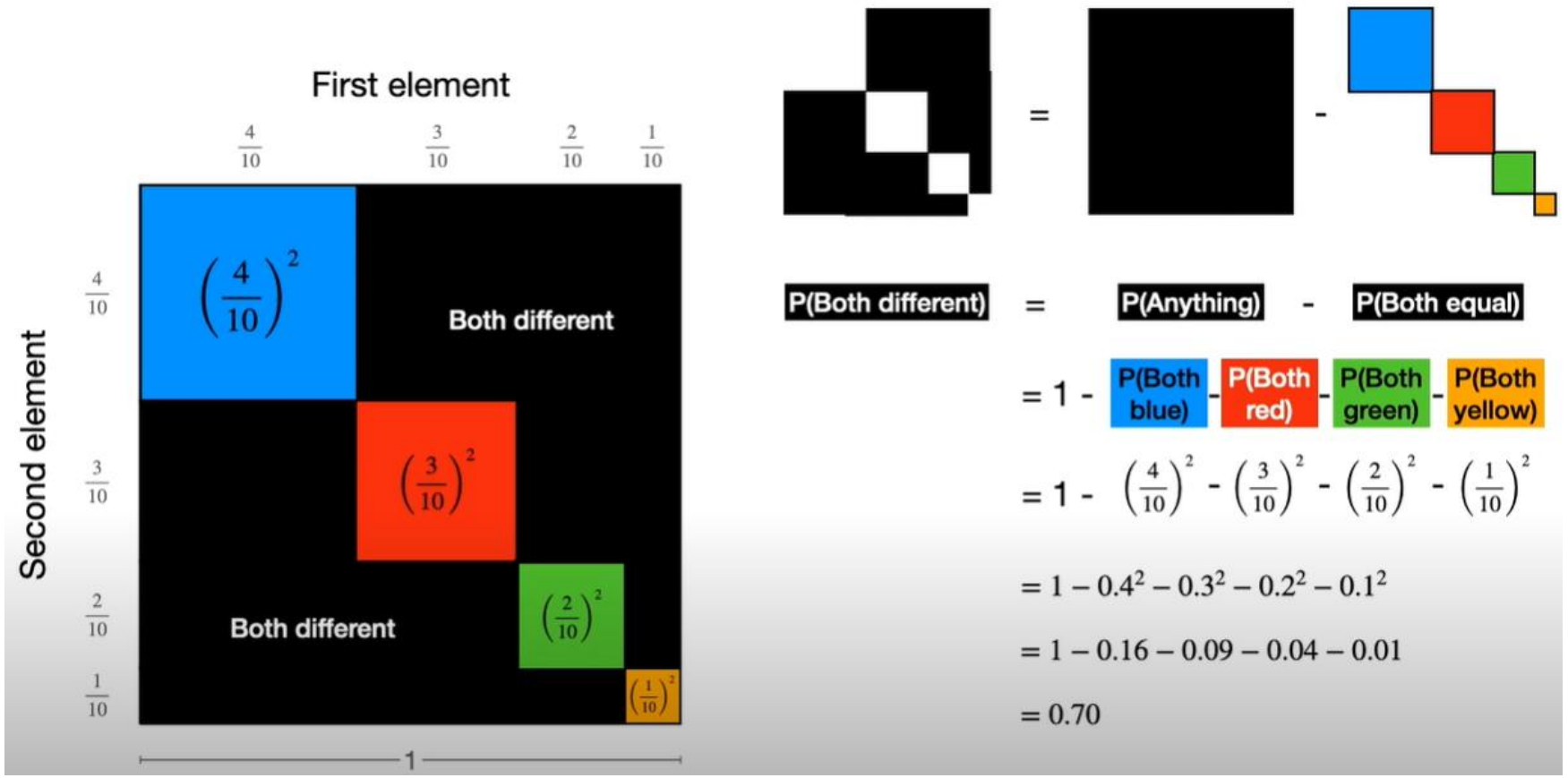
How to calculate Gini index?

- We need to calculate the black area



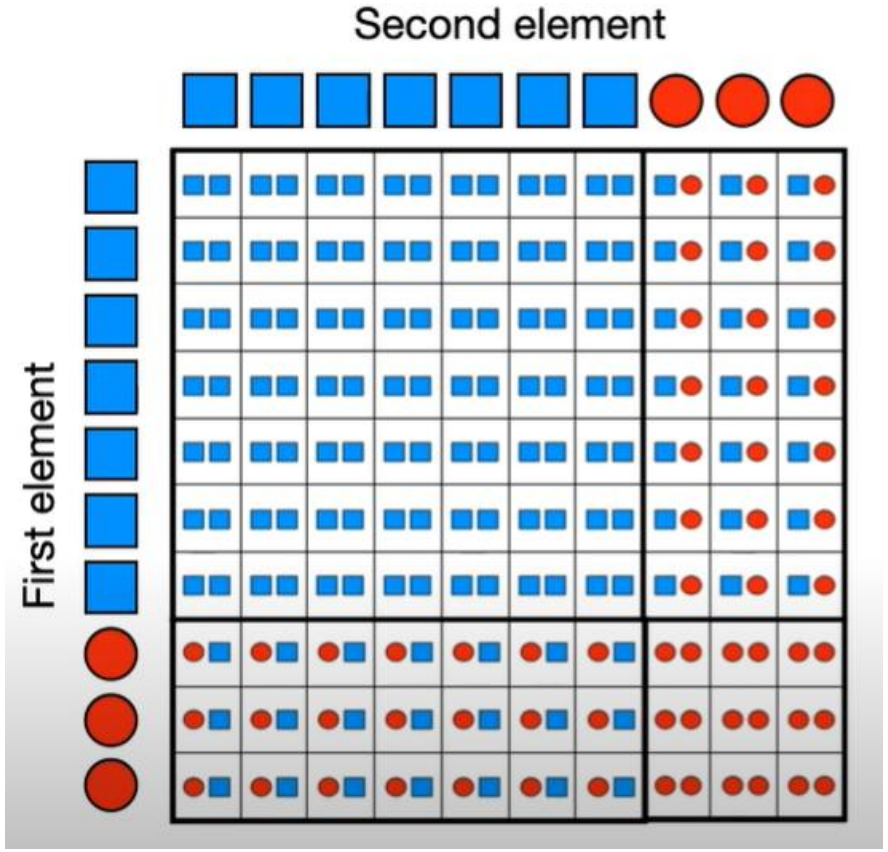
How to calculate Gini index?

- We need to calculate the black area



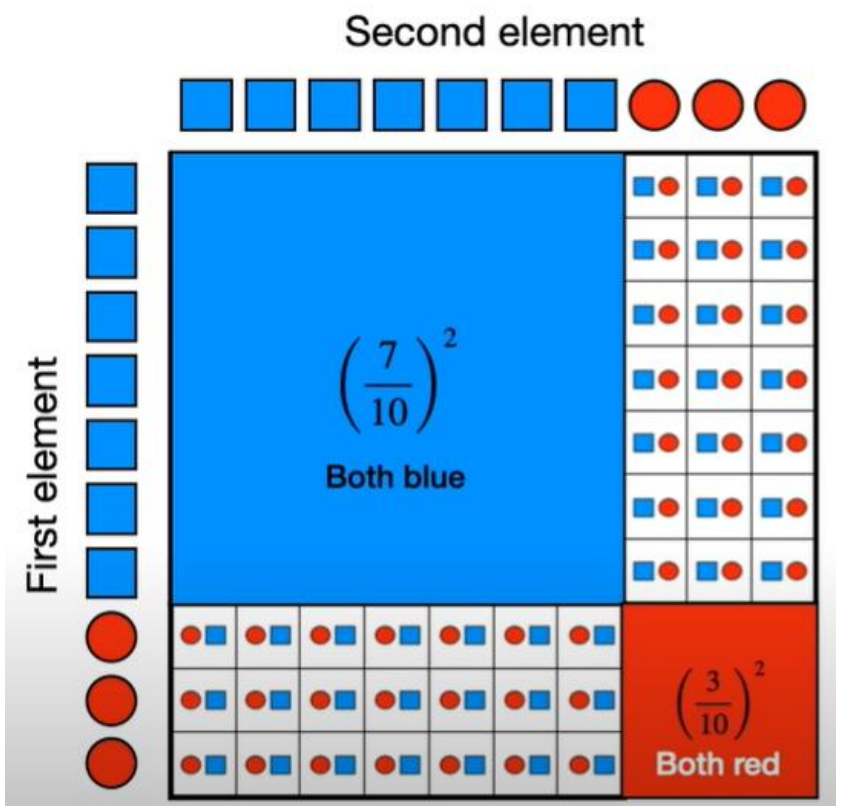
How to calculate Gini index?

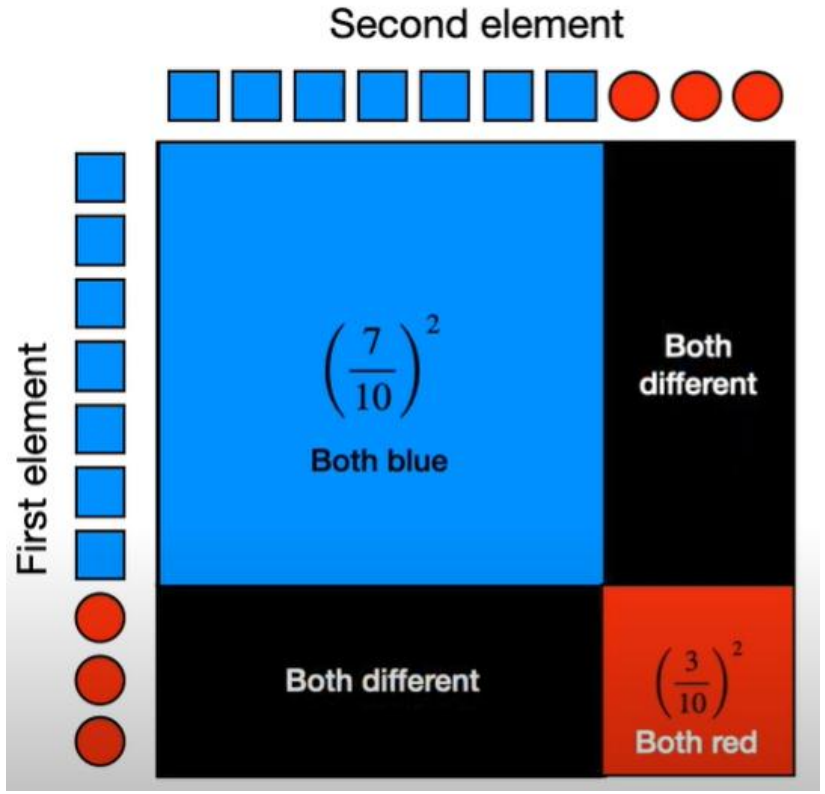
- Let's calculate the Gini index for the first dataset.



How to calculate Gini index?

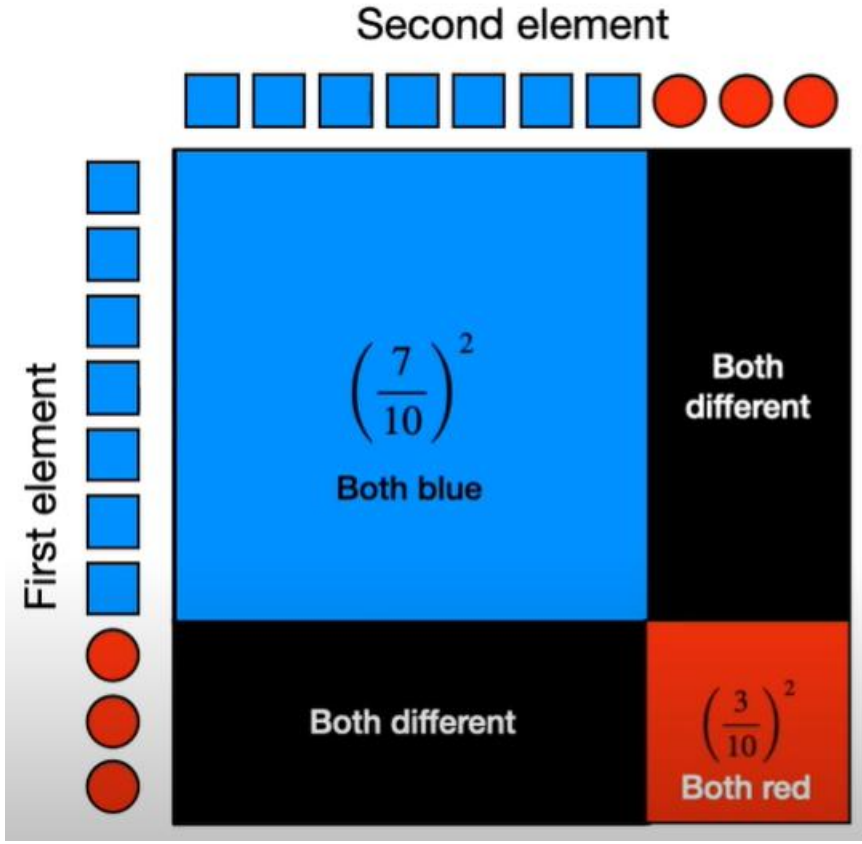
- Let's calculate the Gini index for the first dataset.





How to calculate Gini index?

- Let's calculate the Gini index for the first dataset.



P(Both different) = P(Anything) - P(Both equal)

= 1 - P(Both blue) - P(Both red)

$= 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2$

$= 1 - 0.7^2 - 0.3^2$

$= 1 - 0.49 - 0.09$

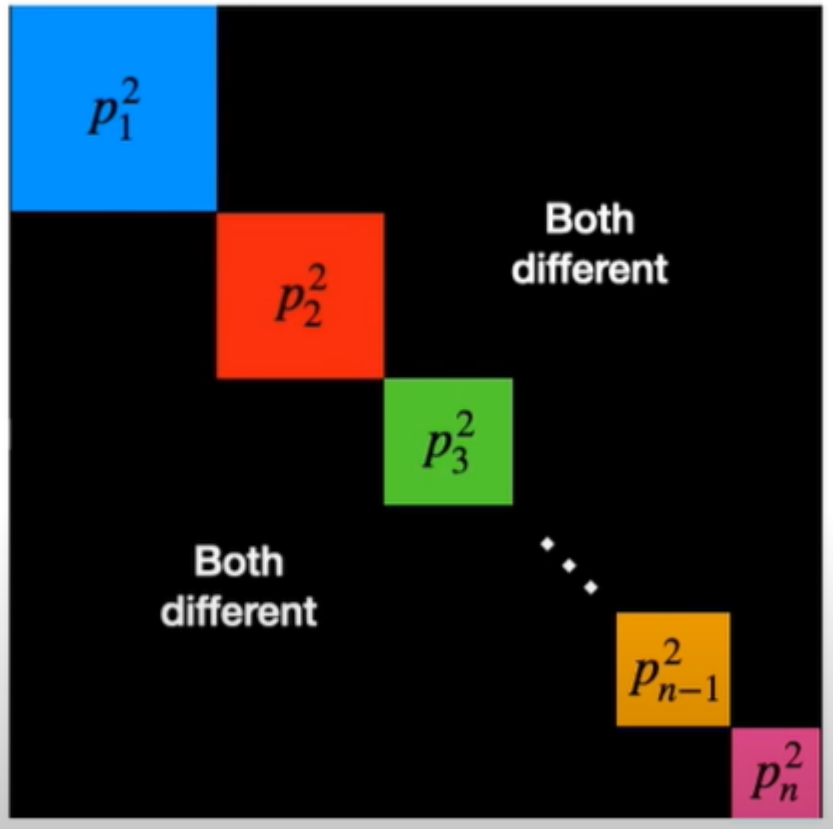
$= 0.42$

The General Formula

n classes

Proportions: $p_1, p_2, \ldots, p_n$

Gini impurity index: $1 - p_1^2 - p_2^2 - \cdots - p_n^2$



The General Formula

n classes

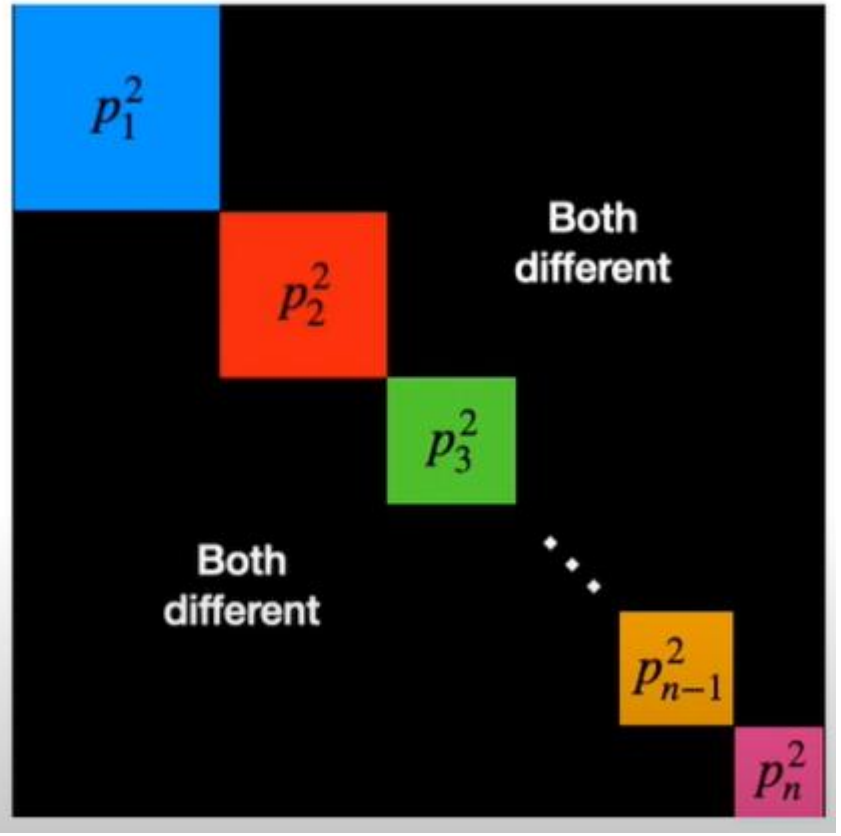
Proportions: $p_1, p_2, \ldots, p_n$

Gini impurity index: $1 - p_1^2 - p_2^2 - \cdots - p_n^2$

P(Both different)

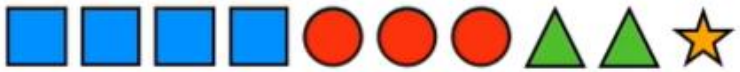
P(Anything)

P(Both equal)



Gini index examples, all similar or all different

Gini = 0.42



Gini = 0.7

Gini = 0

$1 - 1^2 = 0$

Gini = ?

$1 - 0.1^2 - 0.1^2 - \cdots - 0.1^2 = 0.9$

10 times

Note: The Gini index can not be one, because its always 1 – the sum of squares, unless the dataset is infinite.

References

- https://www.youtube.com/watch?v=u4IxOk2ijSs