Gini impurity is a measure used in decision trees to determine how often a randomly chosen element from the set would be incorrectly classified if it were randomly labeled according to the distribution of labels in the subset. It is used to evaluate the quality of a split; the lower the Gini impurity, the better the split.

The Gini impurity for a node is calculated as:

$$\text{Gini Impurity} = 1 - \sum_{i=1}^{n} p_i^2$$

where p_i is the probability of an element being classified into a particular class. In a binary classification, the formula simplifies to:

$$\text{Gini Impurity} = 2 \times p \times (1 - p)$$

Example

Suppose we have a dataset of customers where each customer either buys or doesn't buy a product. We're using a feature (e.g., **age**) to create a split in the dataset.

Age	Buys Product (Yes/No)
<30	Yes
<30	No
<30	Yes
30-60	Yes
30-60	No
>60	No
>60	No
>60	Yes

Let's consider a split on **age** and calculate the Gini impurity for each subset.

Step 1: Calculate Gini Impurity for Each Split

1. **Age < 30**:
   - 2 Yes, 1 No
   - $p_{\text{Yes}} = 2/3, p_{\text{No}} = 1/3$
   - Gini impurity = $1 - (2/3)^2 - (1/3)^2 = 1 - 4/9 - 1/9 = 4/9$

2. **Age 30-60**:
    - 1 Yes, 1 No
    - $p_{\text{Yes}} = 1/2$, $p_{\text{No}} = 1/2$
    - Gini impurity = $1 - (1/2)^2 - (1/2)^2 = 1 - 1/4 - 1/4 = 1/2$
3. **Age > 60**:
    - 1 Yes, 2 No
    - $p_{\text{Yes}} = 1/3$, $p_{\text{No}} = 2/3$
    - Gini impurity = $1 - (1/3)^2 - (2/3)^2 = 1 - 1/9 - 4/9 = 4/9$

Step 2: Calculate the Weighted Gini Impurity for the Split

If the split results in subsets of size 3 (Age < 30), 2 (Age 30-60), and 3 (Age > 60), we can calculate the weighted Gini impurity of the split as follows:

$$\text{Weighted Gini Impurity} = \frac{3}{8} \times \frac{4}{9} + \frac{2}{8} \times \frac{1}{2} + \frac{3}{8} \times \frac{4}{9}$$

The split with the lowest Gini impurity is considered optimal because it means the classes within each subset are more "pure" or homogeneous, thus improving the model's classification performance.

Intuition

- **Pure nodes** have a Gini impurity of 0 (i.e., all samples belong to one class).
- Higher Gini impurity means more mixed classes and less clarity in the classification, prompting further splitting to achieve better classification clarity.