Gini vs Entropy in Decision Trees

# 🔍 Overview

Both Gini Impurity and Entropy are metrics used to evaluate the quality of splits in Decision Trees. They measure how 'impure' or 'mixed' a node is. A pure node contains only one class. The goal is to minimize impurity at each split.

# 🟦 Gini Impurity

Formula:

Gini = 1 - ∑(p\_i)^2

- Measures the probability of misclassifying a sample  
- Ranges from 0 (pure) to (1 - 1/n) for n classes  
- Faster to compute (no logarithms involved)  
- Intuition: 'If I randomly assign a class, how often would I be wrong?'

# 🟥 Entropy (Information Gain)

Formula:

Entropy = - ∑(p\_i \* log₂(p\_i))

- Measures uncertainty or information needed to describe class labels  
- Ranges from 0 (pure) to log₂(n) for n classes  
- More computationally intensive (uses log)  
- Intuition: 'How much information (in bits) is needed to represent the class label?'

# 🔁 Comparison Summary

| Feature | Gini Impurity | Entropy (Information Gain) |  
|---------------------|--------------------------|------------------------------|  
| Formula | 1 - ∑ p\_i² | - ∑ p\_i log₂(p\_i) |  
| Measure Type | Probability of mistake | Information/uncertainty |  
| Binary Range | 0 to 0.5 | 0 to 1 |  
| Computation Speed | Faster | Slower (log involved) |  
| sklearn Default | ✅ Yes | ❌ No (must specify) |  
| Best for | Simplicity, speed | Theory, pure information |

# 🤔 When to Use Which?

- Use \*\*Gini\*\* when you want speed and simplicity (e.g., most real-world projects).  
- Use \*\*Entropy\*\* if you're focused on the theory or need the purest information-driven splits.  
- In practice, both produce very similar results.