

Decison Tree

Gini Index Formula

We use:

$$Gini = 1 - \sum_{i=1}^k p_i^2$$

Where p_i is the proportion of class i in the node.

A split's total Gini is:

$$Gini_{\text{split}} = \frac{n_{\text{left}}}{n} Gini_{\text{left}} + \frac{n_{\text{right}}}{n} Gini_{\text{right}}$$

Simplified Python Code

```
import numpy as np
from collections import Counter

# --- Gini impurity function ---
def gini(y):
    counts = Counter(y)
    total = len(y)
    impurity = 1.0 - sum((count / total) ** 2 for count in counts.values())
    return impurity

# --- Weighted Gini for a split ---
def gini_split(y_left, y_right):
    total = len(y_left) + len(y_right)
    gini_left = gini(y_left)
    gini_right = gini(y_right)
    return (len(y_left) / total) * gini_left + (len(y_right) / total) * gini_right

# --- Try all splits on all features ---
def find_best_split(X, y):
    best_gini = float('inf')
    best_feature = None
    best_threshold = None

    for feature in range(X.shape[1]):
        thresholds = np.unique(X[:, feature])
        for thresh in thresholds:
            left_idx = X[:, feature] <= thresh
            right_idx = X[:, feature] > thresh
            y_left, y_right = y[left_idx], y[right_idx]

            if len(y_left) == 0 or len(y_right) == 0:
                continue

            gini_score = gini_split(y_left, y_right)
            print(f"Feature {feature}, Threshold {thresh}, Gini: {gini_score:.4f}")

            if gini_score < best_gini:
                best_gini = gini_score
                best_feature = feature
                best_threshold = thresh
```

```
print(f"\nBest Split → Feature {best_feature}, Threshold {best_threshold}, Gini: {b
return best_feature, best_threshold
```



Example Dataset

```
# Simple 2D dataset
X = np.array([
    [2.7], [1.0], [3.1], [1.8], [3.0], [2.5], [0.5]
])
y = np.array([0, 0, 1, 0, 1, 1, 0]) # binary classes

find_best_split(X, y)
```



Output (Illustrative)

```
Feature 0, Threshold 0.5, Gini: 0.4898
Feature 0, Threshold 1.0, Gini: 0.4898
Feature 0, Threshold 1.8, Gini: 0.4444
Feature 0, Threshold 2.5, Gini: 0.3750
Feature 0, Threshold 2.7, Gini: 0.4444
Feature 0, Threshold 3.0, Gini: 0.4898
Feature 0, Threshold 3.1, Gini: 0.4898
```

```
Best Split → Feature 0, Threshold 2.5, Gini: 0.3750
```