

Decision Tree Worksheet: Calculating Gini Impurity**Dataset:**

ID	Feature A	Feature B	Feature C	Target
1	Yes	No	Yes	Yes
2	No	Yes	No	No
3	Yes	No	No	Yes
4	No	Yes	Yes	No
5	Yes	Yes	No	Yes
6	No	No	Yes	No
7	Yes	No	Yes	Yes
8	No	Yes	No	No
9	Yes	No	No	No
10	No	No	Yes	No

Step 1: Compute Initial Gini Impurity for the Target Variable

- Count how many 'Yes' and 'No' values are in the Target column.
- Compute the initial Gini Impurity using the formula:

$$Gini = 1 - (p_{Yes}^2 + p_{No}^2)$$

Feature A:

Number of rows with Feature A = Yes:

- Number of rows Target = Yes:
- Number of rows Target = No:
- Calculate Gini score:

$$Gini_A(Yes) = 1 - \sum p^2$$

Number of rows with Feature A = No:

- Number of rows Target = Yes:
- Number of rows Target = No:
- Calculate Gini score:

$$Gini_A(No) = 1 - \sum p^2$$

Weighted Gini Score:

$$Gini_A = \frac{\text{Number of Yes}}{\text{Total Rows}} * Gini_A(Yes) + \frac{\text{Number of No}}{\text{Total Rows}} * Gini_A(No)$$

Feature B:

Number of rows with Feature B = Yes:

- Number of rows Target = Yes:
- Number of rows Target = No:
- Calculate Gini score:

$$Gini_B(Yes) = 1 - \sum p^2$$

Number of rows with Feature B = No:

- Number of rows Target = Yes:
- Number of rows Target = No:
- Calculate Gini score:

$$Gini_B(No) = 1 - \sum p^2$$

Weighted Gini Score:

$$Gini\ B = \frac{Number\ of\ Yes}{Total\ Rows} * Gini_B(Yes) + \frac{Number\ of\ No}{Total\ Rows} * Gini_B(No)$$

After Branching at A:

ID	Feature A	Feature B	Feature C	Target
1	Yes	No	Yes	Yes
3	Yes	No	No	Yes
5	Yes	Yes	No	Yes
7	Yes	No	Yes	Yes
9	Yes	No	No	No

For B: Yes = 1 (1/0) , No = 4 (3/1)

For C: Yes = 2 (2/0), No = 3 (2/1)

- Calculate Gini Score:

$$Gini_B(Yes) = 1 - \sum p^2$$

$$Gini_B(No) = 1 - \sum p^2$$