

Task 1 – Calculate Entropy and Information Gain

Step 1: Calculate the entropy of the whole dataset S.

There are 5 'Yes' and 3 'No' out of 8 records.

So:

$$p(\text{Yes}) = 5/8 = 0.625$$

$$p(\text{No}) = 3/8 = 0.375$$

Entropy formula:

$$H(S) = -[p(\text{Yes}) * \log_2(p(\text{Yes})) + p(\text{No}) * \log_2(p(\text{No}))]$$

$$H(S) = -[0.625 * \log_2(0.625) + 0.375 * \log_2(0.375)] \approx 0.954 \text{ bits}$$

Step 2: Choose an attribute (e.g., Age) and compute the entropy after the split.

Age has three values: Young, Middle, Old.

- Young: records 1, 2, 8 → (Yes = 1, No = 2)
- Middle: records 3, 7 → (Yes = 2, No = 0)
- Old: records 4, 5, 6 → (Yes = 2, No = 1)

$$H(\text{Young}) = -[(1/3)*\log_2(1/3) + (2/3)*\log_2(2/3)] \approx 0.918 \text{ bits}$$

$$H(\text{Middle}) = -[(2/2)*\log_2(2/2)] = 0 \text{ (pure set)} \approx 0.000 \text{ bits}$$

$$H(\text{Old}) = -[(2/3)*\log_2(2/3) + (1/3)*\log_2(1/3)] \approx 0.918 \text{ bits}$$

Step 3: Compute the weighted average entropy after splitting on Age:

$$H(S | \text{Age}) = (3/8)*H(\text{Young}) + (2/8)*H(\text{Middle}) + (3/8)*H(\text{Old})$$

$$H(S | \text{Age}) \approx (3/8)*0.918 + (2/8)*0.000 + (3/8)*0.918$$

$$\approx 0.689 \text{ bits}$$

Step 4: Information Gain for Age:

$$\text{IG}(S, \text{Age}) = H(S) - H(S | \text{Age})$$

$$\text{IG}(S, \text{Age}) \approx 0.954 - 0.689 = 0.266 \text{ bits}$$

You would repeat this process for the other attributes (Income, Student, etc.) and choose the attribute with the highest Information Gain as the root.

Task 2 – Calculate Gini Impurity and Compare Splits

Gini impurity for a node S with k classes is defined as:

$$\text{Gini}(S) = 1 - \sum p_i^2 \text{ (sum over all classes i)}$$

For the whole dataset S:

$$\text{Gini}(S) = 1 - [(0.625)^2 + (0.375)^2] \approx 0.469$$

$$\text{Gini(Young)} = 1 - [(1/3)^2 + (2/3)^2] \approx 0.444$$

$$\text{Gini(Middle)} = 1 - [(2/2)^2 + 0^2] = 0 \approx 0.000$$

$$\text{Gini(Old)} = 1 - [(2/3)^2 + (1/3)^2] \approx 0.444$$

Weighted Gini after splitting on Age:

$$\text{Gini}(S | \text{Age}) \approx (3/8)*0.444 + (2/8)*0.000 + (3/8)*0.444$$

$$\approx 0.333$$

You would compute $\text{Gini}(S | \text{Income})$, $\text{Gini}(S | \text{Student})$, etc., the same way. The best split is the attribute with the smallest weighted Gini after the split. You can then compare which attribute is chosen by Entropy vs. Gini.

Task 3 – Construct the Decision Tree

Once you know which attribute has the highest Information Gain (or lowest Gini), you can build the tree:

1. Root node: attribute with highest IG (e.g., Age).
2. For each branch (Young, Middle, Old), look at the subset of records:
 - If all records in a subset are the same class → make a leaf node with that class.
 - If not pure → repeat the Entropy/IG or Gini calculations using only that subset, and choose the next best attribute to split on.

For example, in our case:

- Middle subset was pure (all 'Yes') → leaf node: Yes.
- Young and Old subsets were not pure → you would test Income or Student on those subsets to continue growing the tree until all leaves are pure or you reach a stopping condition.

To apply this to your assignment, replace the counts (numbers of Yes/No) in each step with the counts from your own dataset, then follow exactly the same formulas.

PORTFOLIO:

In this coursework I studied decision tree learning in depth and tried to understand not just the formulas but the logic behind every split. At the beginning I only knew that a decision tree asks a series of questions and ends in a prediction. Through the tasks I learned exactly how those questions are chosen. First, I calculated the overall entropy of the dataset and then, for each attribute, I worked out the entropy of every subset and the weighted average. From this I computed the information gain and identified which attribute should be used at the root of the tree.

Next, I repeated a similar process using Gini impurity. I calculated the Gini value for the whole dataset and for each split, then compared which attribute produced the lowest weighted Gini. This allowed me to compare entropy/information gain and Gini as criteria for choosing splits. Finally, using these results, I constructed the decision tree step by step, showing which attribute appears at each level and explaining why certain branches become pure leaves.

The most challenging part was managing the fractions and logarithms without making mistakes, so I wrote every step clearly and checked the calculations carefully. By the end, I could explain every node in my tree and justify each decision using entropy, information gain and Gini. This portfolio contains all my reasoning, so my understanding is clear without needing a viva.