Problem Statement

0.1 Task 1

Modify the standard Information Gain (IG) splitting criterion by introducing a penalty term to discourage splits on attributes with high cardinality. The modified metric, denoted as $IG'(S, A)$, is defined as:

$$IG'(S, A) = IG(S, A) - \lambda \cdot \frac{v - 1}{|S|}$$

where:

- $IG(S, A)$ is the standard information gain of attribute A on dataset S,
- v is the number of distinct values of attribute A,
- $|S|$ is the number of instances in dataset S,
- λ is a tunable hyperparameter that controls the strength of the penalty.

0.2 Task 2

Your program should report the number of nodes that were forced to become leaf nodes due to the depth limit, even though they had sufficient data to be split further.

In other words, count how many nodes could have continued growing (i.e., were not pure and had more than one sample), but were not allowed to split because they reached the specified maximum depth.