Part-II: Decision Trees

Task 4: Conceptual Questions

1. What is entropy and information gain?
Entropy measures the impurity or randomness in a dataset; higher entropy means more disorder. Information Gain is the reduction in entropy after splitting the dataset based on an attribute, and it helps in choosing the best feature for a decision tree split.

2. Explain the difference between Gini Index and Entropy.
Both measure impurity, but **Entropy** uses logarithmic values and is based on information theory, while **Gini Index** calculates the probability of misclassification without logs. Gini is computationally simpler and often preferred for speed, whereas entropy can be more informative in certain cases.

3. How can a decision tree overfit? How can this be avoided?
A decision tree overfits when it grows too deep and learns noise or irrelevant patterns from training data. This can be avoided by using techniques like **max depth restriction**, **minimum samples per split**, **pruning**, or using **ensemble methods** like Random Forest.