Decision Tree Terminology

1 Introduction

This document provides a glossary of key terms related to the decision tree algorithm, extracted from the provided text. Each term is defined concisely to clarify its role in the context of decision trees.

2 Terminology

- **Decision Tree**: A supervised machine learning model that recursively splits data into regions based on feature values to make predictions for classification or regression tasks.
- **Root Node**: The topmost node in a decision tree, containing all samples before any splits are applied.
- **Child Node**: A node created by splitting a parent node based on a feature and threshold. Child nodes are further divided into left and right nodes.
- **Leaf Node**: A terminal node at the bottom of the tree that contains the final prediction, typically the most common class label (classification) or average value (regression).
- **Feature**: An input variable used to make decisions in the tree (e.g., ”Is it raining?” or ”Time available”).
- **Split**: The process of dividing a node’s samples into two or more child nodes based on a feature and a threshold.
- **Best Split**: The feature and threshold combination that maximizes information gain or minimizes impurity for a given node.
- **Threshold**: A specific value of a feature used to split a node (e.g., ”Time > 10 minutes”).
- **Entropy**: A measure of uncertainty in a node, calculated as:

$$H = -\sum_{k} p_k \log_2(p_k)$$

  where p_k is the proportion of samples belonging to class k. Lower entropy indicates purer nodes.

- **Information Gain**: The reduction in entropy achieved by splitting a node, calculated as:
$$\text{IG} = H_{\text{parent}} - \sum_i \frac{N_i}{N} H_i$$
where H_{parent} is the parent node's entropy, N_i is the number of samples in child node i, N is the total number of samples, and H_i is the entropy of child node i.
- **Greedy Search**: An algorithm that evaluates all possible features and thresholds at a node to select the split that maximizes information gain.
- **Class Label**: The predicted output for a sample (e.g., "Yes" or "No" for walking or taking the bus).
- **Overfitting**: When a decision tree grows too complex, capturing noise in the training data rather than general patterns, leading to poor performance on new data.
- **Stopping Criteria**: Rules to halt tree growth, such as:
  - **Maximum Depth**: The maximum number of levels in the tree (e.g., stop at depth 5).
  - **Minimum Samples Split**: The minimum number of samples required to split a node (e.g., 5 samples).
  - **Pure Node**: A node with samples of only one class, requiring no further splits.
- **Training Phase**: The process of building the decision tree by recursively selecting the best splits and storing features and thresholds.
- **Testing Phase**: The process of predicting a class label for a new sample by traversing the tree from the root to a leaf node, applying stored splits.