# Decision Tree

NILOY KUMAR MONDAL

July 2025

## Contents

# 1 Introduction

This document explores the effect of maximum tree depth on decision tree performance using different selection criteria.

# 2 Iris Dataset

## 2.1 Accuracy vs. Max Tree Depth

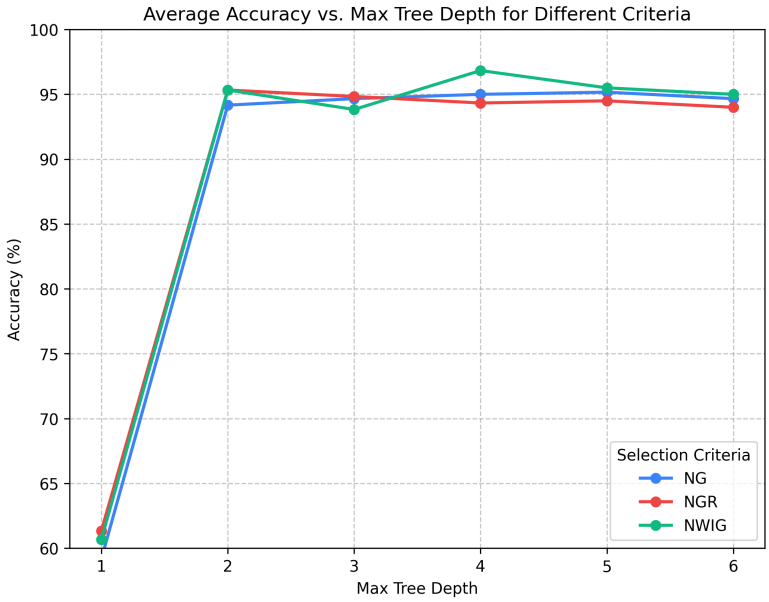


Figure 1: Average Accuracy vs. Max Tree Depth for Different Criteria (NG, NGR, NWIG)

## 2.2 Tree Size vs. Max Tree Depth

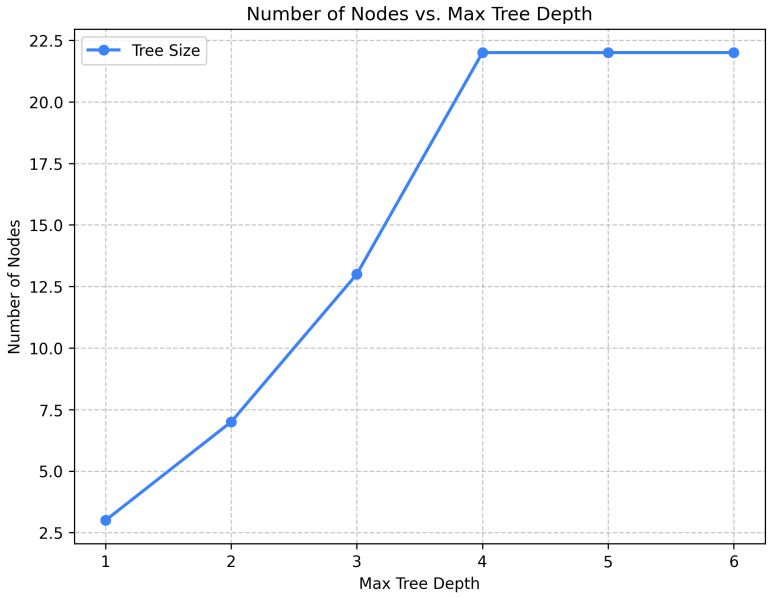


Figure 2: Number of Nodes vs. Max Tree Depth

## 2.3 Observations and Analysis

- **Accuracy Trends**:
  - At **Max Depth = 1**, all criteria (NG, NGR, NWIG) achieve low accuracy ( 61–62%), indicating underfitting.
  - A significant increase in accuracy is observed at **Depth = 2** ( 94–95%), as the tree becomes expressive enough to capture key splits.
  - From **Depth 3 to 6**, accuracy remains high and fairly stable for all criteria.
  - **NWIG** peaks around **Depth 4** with accuracy close to **97%**, slightly outperforming the others.
- **Tree Size Analysis**:
  - The number of nodes grows quickly from Depth 1 to 4.
  - After **Depth 4**, the number of nodes saturates at 22 — further increasing depth doesn't increase complexity.
  - This suggests stopping conditions or pruning effects limiting tree growth, helping to avoid overfitting.

- **Criterion Performance Comparison**:

| Criterion | Strengths | Weaknesses |
|---|---|---|
| NG (Information Gain) | High accuracy and stable across depths. | May prefer attributes with more distinct values, possibly leading to overfitting. |
| NGR (Information Gain Ratio) | Balances information gain with attribute diversity; good generalization. | Slightly underperforms at higher depths. |
| NWIG (Normalized Weighted IG) | Achieves best accuracy (peak at depth 4); consistently strong. | May be computationally more intensive. |

Table 1: Comparison of Selection Criteria

- **Summary and Trade-offs**:
  - Depths **2 to 4** provide an optimal trade-off between tree complexity and accuracy.
  - No significant signs of overfitting beyond depth 4 due to limited tree growth.
  - **NWIG** is the best-performing criterion overall.
  - Increasing tree depth beyond 4 yields **diminishing returns**.

## 2.4 Training Time Analysis

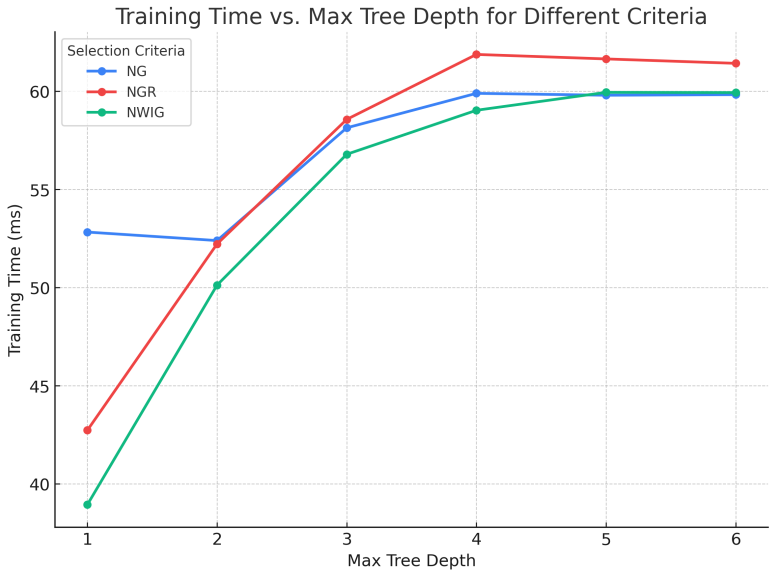


Figure 3: Training Time vs. Max Tree Depth for Different Criteria

- **General Trends:**
  - Training time increases with tree depth for all criteria.
  - Growth slows and stabilizes after depth 4.
- **Criterion-wise Analysis:**
  - **NG (Information Gain)**:
    - Starts with higher time at depth 1.
    - Stabilizes around 60 ms beyond depth 3.
  - **NGR (Information Gain Ratio)**:
    - Fastest rise in training time.
    - Peaks at depth 4 with the highest training time among all.
  - **NWIG (Normalized Weighted IG)**:
    - Most efficient at lower depths.
    - Increases gradually and catches up to NG by depth 5.
- **Insights:**
  - **NGR** may require more computation, especially with deeper trees.
  - **NWIG** is the most training-efficient at shallow depths.

– All criteria converge to similar training times ( 60 ms) after depth 4.

| Criterion | Training Time Behavior | Implication |
|---|---|---|
| NG (Information Gain) | High at depth 1, then stable from depth 3 onward. | Reliable and consistent training time. |
| NGR (Information Gain Ratio) | Steep rise; consistently highest time. | Might be costlier in real-time systems. |
| NWIG (Normalized Weighted IG) | Lowest at start, gradually increases. | Best suited when fast training is a priority. |

Table 2: Training Time Behavior by Criterion

# 3 Adult Dataset

## 3.1 Accuracy vs. Max Tree Depth

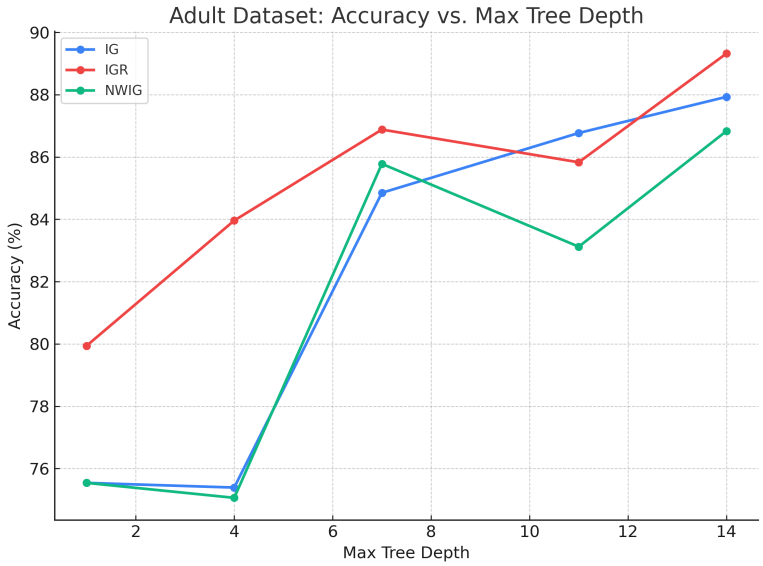


Figure 4: Adult Dataset: Accuracy vs. Max Tree Depth for Different Criteria

## 3.2 Tree Size vs. Max Tree Depth

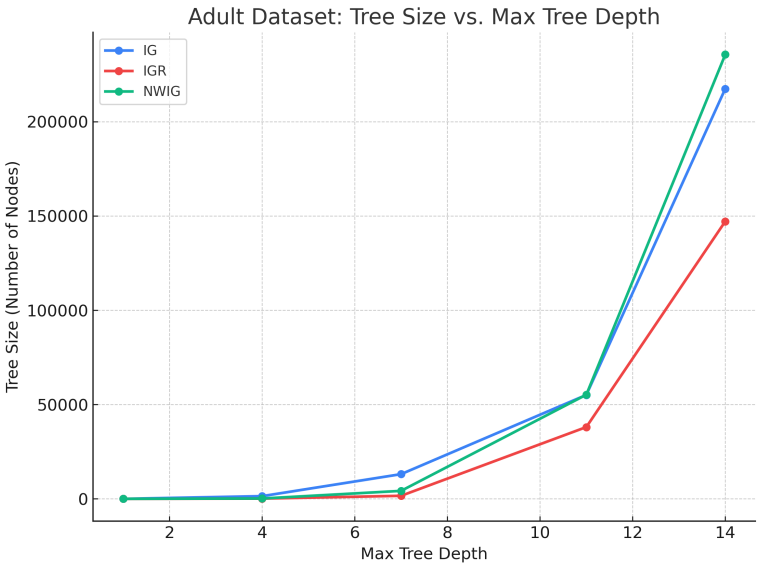


Figure 5: Adult Dataset: Tree Size vs. Max Tree Depth

### 3.3 Training Time vs. Max Tree Depth

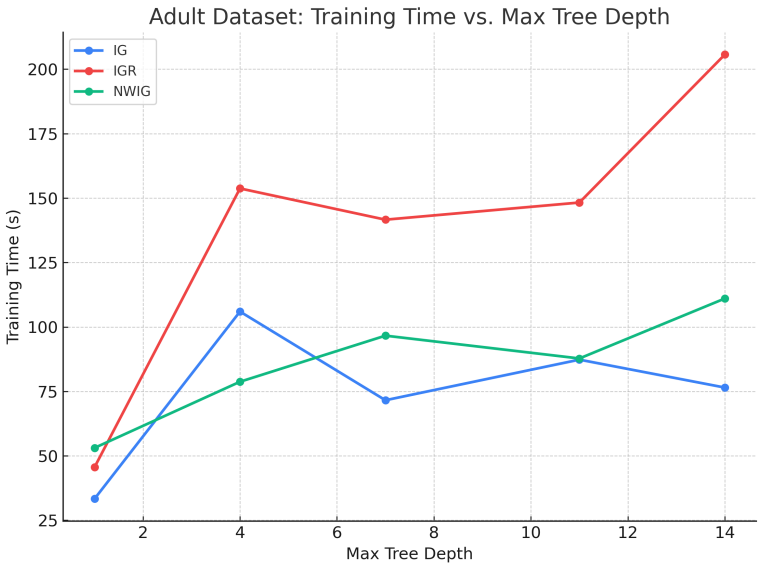


Figure 6: Adult Dataset: Training Time vs. Max Tree Depth

### 3.4 Observations and Analysis

- **Accuracy Trends**:
  - All criteria show improved accuracy as depth increases.
  - **IGR (Information Gain Ratio)** achieves the highest accuracy, peaking at **89.32%** at depth 14.
  - **IG** and **NWIG** follow with slightly lower accuracy but show consistent improvement.
- **Tree Size Analysis**:
  - Tree size increases drastically with depth, especially for **IG** and **NWIG**.
  - At depth 14, **NWIG** has the largest tree (**235,560 nodes**), while **IGR** has significantly smaller trees.
  - **IGR** appears to reduce overfitting by controlling tree size effectively.
- **Training Time**:
  - **IGR** consistently requires the highest training time, especially at deeper levels.

- **NWIG** maintains a balance between training time and accuracy.
- **IG** has moderate training time but can lead to excessively large trees.

| Criterion | Strengths | Weaknesses |
|---|---|---|
| IG (Information Gain) | Improves accuracy with depth; moderate training time. | Leads to very large trees, potentially overfitting. |
| IGR (Information Gain Ratio) | Highest accuracy; controls tree size effectively. | Highest training time at almost all depths. |
| NWIG (Normalized Weighted IG) | Balanced performance in terms of size, time, and accuracy. | Tree size can be large at deep depths. |

Table 3: Comparison of Criteria on Adult Dataset

- **Summary and Trade-offs**:
  - **IGR** is the best-performing criterion in terms of accuracy and generalization.
  - **NWIG** offers good balance between size and performance.
  - **IG** might be less optimal due to overfitting risk from deep, large trees.