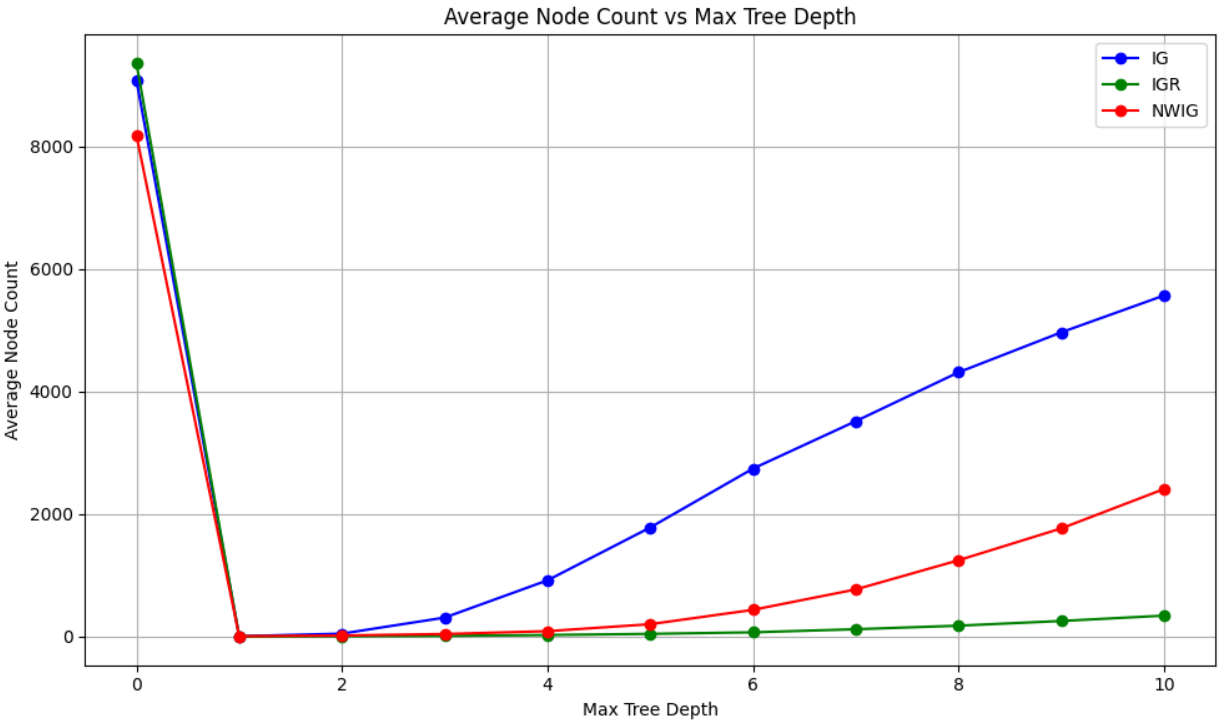
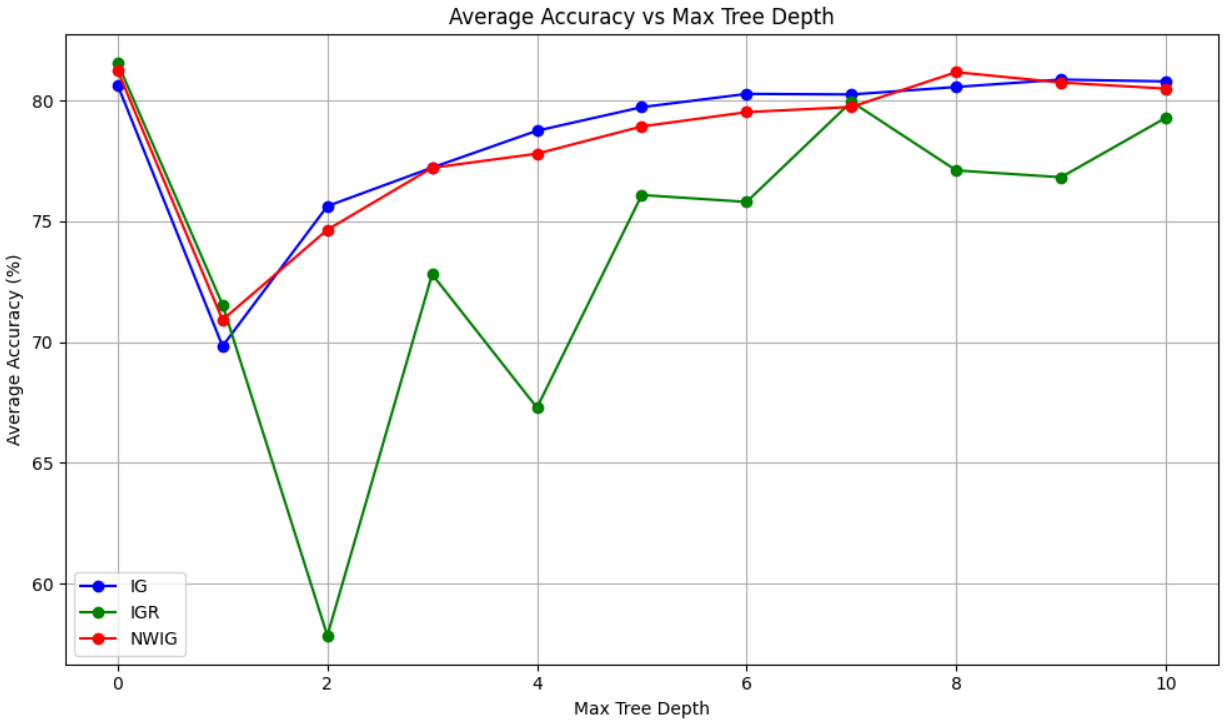
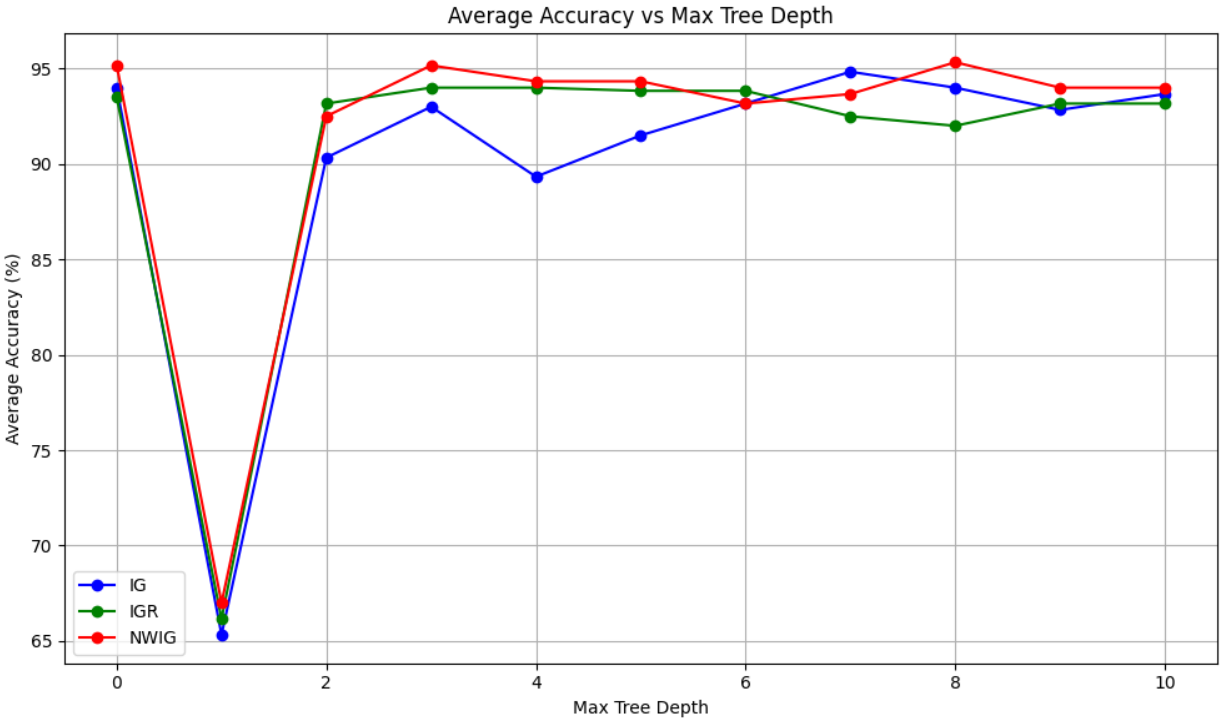
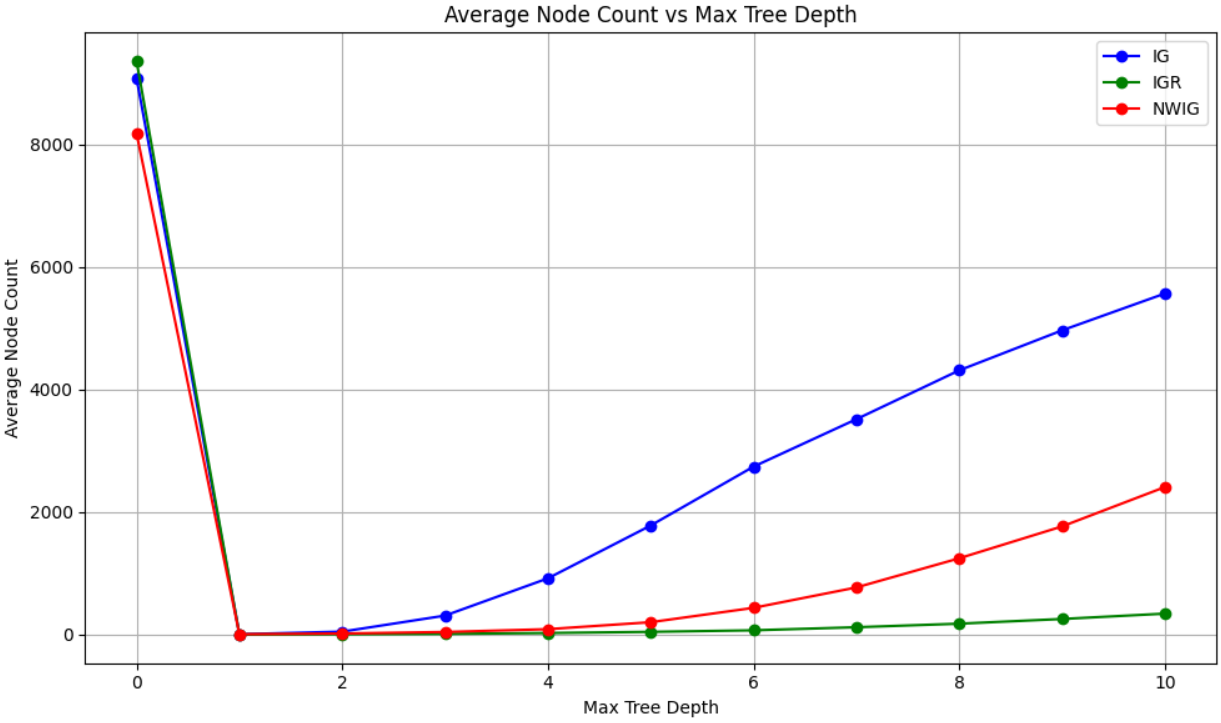




ADULT Dataset


Average Accuracy vs Max Tree Depth
Average Accuracy (%)
Max Tree Depth
IG
IGR
NWIG
95
90
85
80
75
70
65
0
2
4
6
8
10


Average Node Count vs Max Tree Depth
Average Node Count
Max Tree Depth
IG
IGR
NWIG
8000
6000
4000
2000
0
0
2
4
6
8
10

IRIS DATA SET

Decision Tree: Accuracy vs Depth - Observations & Analysis

This analysis compares how different attribute selection criteria (IG, IGR, NWIG) affect the decision tree's accuracy and size (node count) as the maximum tree depth increases. The datasets used are Adult and Iris.

1. Performance Across Depths (Adult Dataset)

Information Gain (IG)

- Performs **consistently well** starting from depth 5 onwards.
- Accuracy peaks around depth **9** (80.86%) with over **4900 nodes**, indicating some **overfitting**.
- At depth 0 (root only), it still gives **~80.6%**, due to majority class dominance.

Information Gain Ratio (IGR)

- Very **low accuracy** at shallow depths (e.g., 57% at depth 2).
- Starts performing better after depth 5.
- **Least overfitting** (e.g., at depth 10, it uses only 343 nodes vs 5500+ in IG).
- Good trade-off between size and accuracy.

Normalized Weighted IG (NWIG)

- Shows a **gradual and stable improvement** in accuracy with depth.
- Performance is very close to IG but with **fewer nodes** (e.g., 2400 at depth 10).
- Appears more **efficient** than IG at deeper levels.

---

2. Performance Across Depths (Iris Dataset)

IG:

- Already **very accurate at shallow depths**: ~90% at depth 2.
- Peaks around **depth 7** with **94.8% accuracy**.
- Node count stays low due to smaller dataset.

IGR:

- Slightly lower accuracy than IG and NWIG.
- Still above **92-94%** from depth 4 onward.
- Node count remains compact (under 19).

NWIG:

- Most **consistent high accuracy** among all.
- **Best accuracy** (95.3%) at depth 8.
- Works great with very small trees too.

Summary & Key Observations

- **IG** is powerful but **prone to overfitting** on large datasets (e.g., Adult). Performs best with deep trees but creates large trees.
- **IGR** tends to underfit at shallow depths but keeps the tree small. Best for **compact and interpretable** trees.
- **NWIG** offers a **balanced trade-off** between accuracy and node count, performing **consistently well** across depths and datasets.

Final Thoughts

- **Pruning** (limiting depth) is essential to avoid overfitting, especially for IG.
- **NWIG** is recommended for scalable performance.
- Choice of criterion can depend on:
    - Dataset size
    - Desired model simplicity vs accuracy
    - Available computation resources