CSE 318: Artificial Intelligence Sessional

Decision Trees

Abrar Zahin Raihan
Student ID: 2105009

July 7, 2025

Contents

1 Introduction

This assignment implements and compares three decision tree attribute selection criteria: Information Gain (IG), Information Gain Ratio (IGR), and Normalized Weighted Information Gain (NWIG). The goal is to analyze their performance across different datasets and maximum depth constraints.

1.1 Implemented Criteria

1.1.1 Information Gain (IG)

Information Gain measures entropy reduction after splitting on an attribute:

$$IG(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v) \tag{1}$$

Where entropy is:

$$Entropy(S) = - \sum_{c \in Classes} p(c) \log_2 p(c) \tag{2}$$

IG tends to favor attributes with many distinct values, potentially causing overfitting.

1.1.2 Information Gain Ratio (IGR)

IGR normalizes Information Gain by the intrinsic value to reduce bias toward multi-valued attributes:

$$GainRatio(S, A) = \frac{IG(S, A)}{IV(A)} \tag{3}$$

Where:

$$IV(A) = - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot \log_2 \left(\frac{|S_v|}{|S|} \right) \tag{4}$$

1.1.3 Normalized Weighted Information Gain (NWIG)

NWIG is a custom criterion that penalizes high-cardinality attributes and adjusts for dataset size:

$$NWIG(S, A) = \left(\frac{IG(S, A)}{\log_2(k+1)} \right) \cdot \left(1 - \frac{k-1}{|S|} \right) \tag{5}$$

Where k is the number of unique values for attribute A.

2 Experimental Setup

2.1 Datasets

Two datasets were used to evaluate the criteria:

Iris Dataset: 150 instances, 4 numerical features, 3 classes. This dataset has clear class boundaries and minimal noise.

Adult Dataset: 32,561 instances, 14 mixed-type features, 2 classes. This dataset has class imbalance, missing values, and complex feature interactions.

2.2 Methodology

For each criterion and maximum depth combination:

- Performed 20 random 80/20 train-test splits
- Calculated average accuracy and standard deviation
- Recorded tree complexity (number of nodes)
- Tested maximum depths from 1 to 5

Missing values were handled by replacing numerical unknowns with zero and categorical unknowns with empty strings.

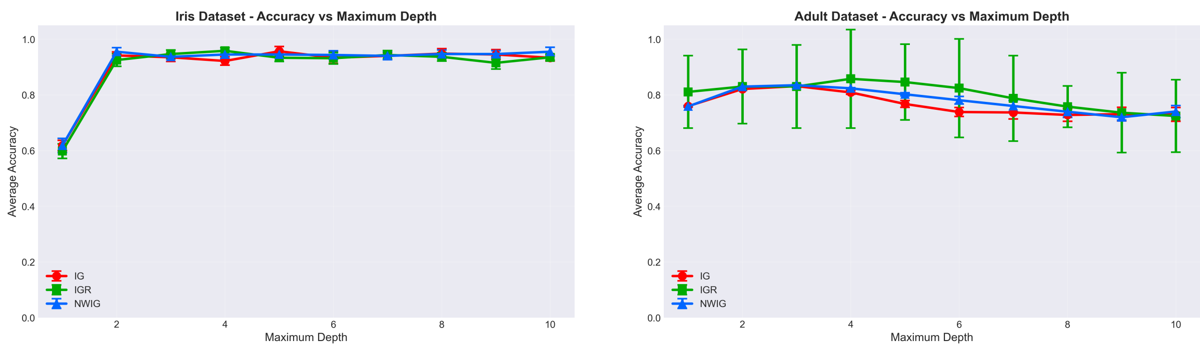
3 Results

3.1 Accuracy Performance

Table 1 shows the best performance for each criterion:

Table 1: Best Performance Summary by Criterion

Dataset	Criterion	Best Depth	Accuracy	Std Dev	Nodes
3*Iris	IG	5	0.9567	0.0392	16
	IGR	4	**0.9583**	0.0262	13
	NWIG	2	0.9550	**0.0329**	**5**
3*Adult	IG	3	0.8315	0.0093	1723
	IGR	4	**0.8577**	0.4033	475
	NWIG	3	0.8348	**0.0067**	**933**



(a) Iris Dataset (b) Adult Dataset

Figure 1: Accuracy vs Maximum Depth with 95% Confidence Intervals

IGR achieved the highest accuracy on both datasets (95.83% on Iris, 85.77% on Adult). However, IGR showed high variance on the Adult dataset (std dev: 0.4033).

3.2 Consistency Analysis

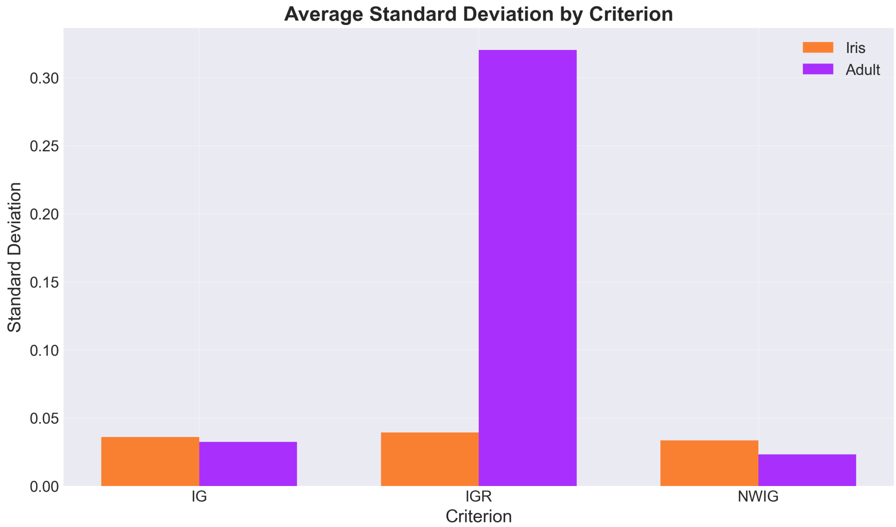
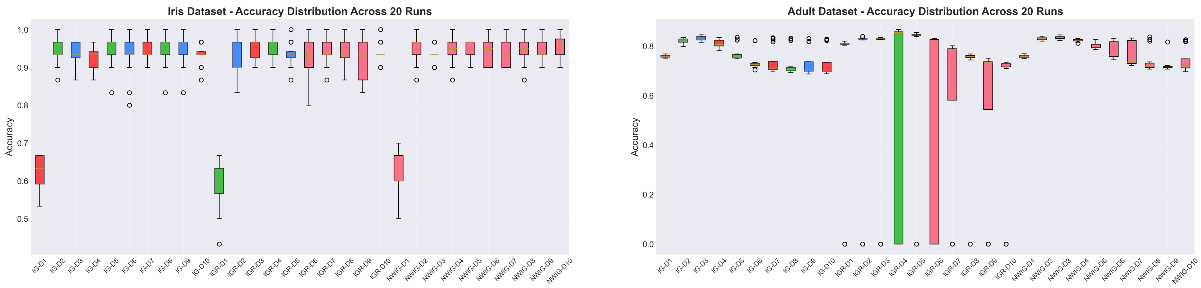


Figure 2: Average Standard Deviation by Criterion

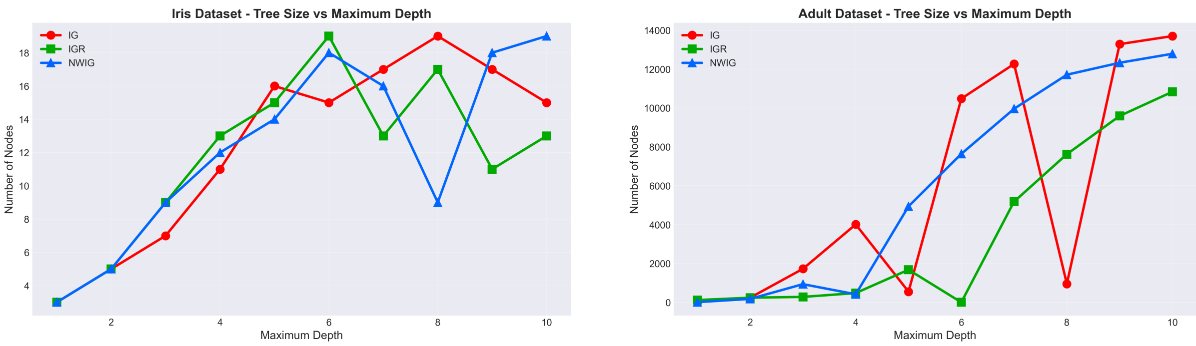


(a) Iris Dataset (b) Adult Dataset

Figure 3: Accuracy Distribution Across 20 Runs

NWIG demonstrated the most consistent performance across both datasets, with the lowest standard deviations (Iris: 0.0335, Adult: 0.0232). IGR showed concerning variability on the Adult dataset despite its high peak performance.

3.3 Tree Complexity



(a) Iris Dataset (b) Adult Dataset

Figure 4: Tree Complexity vs Maximum Depth

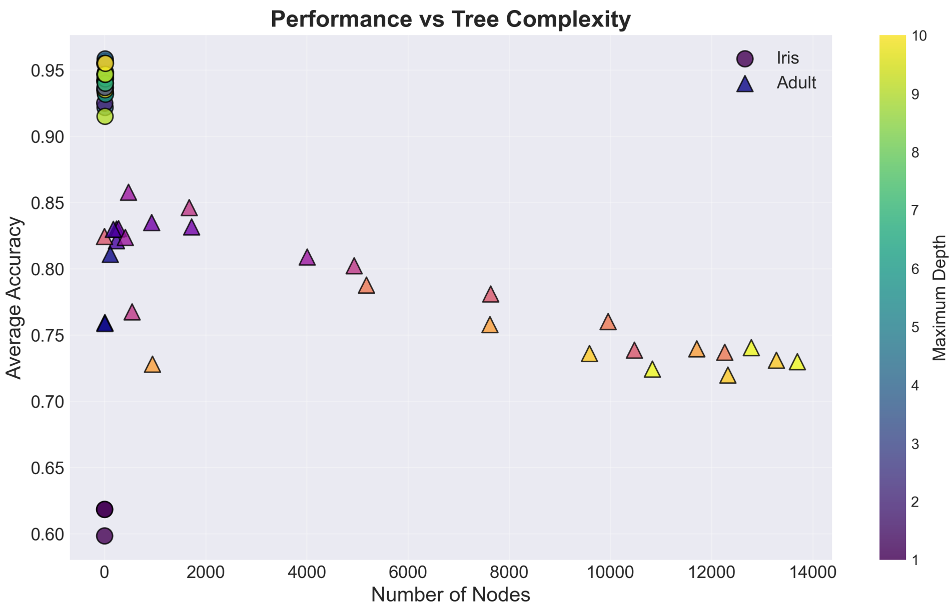
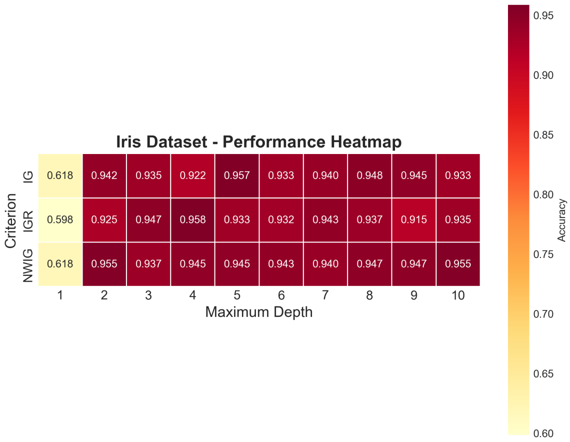


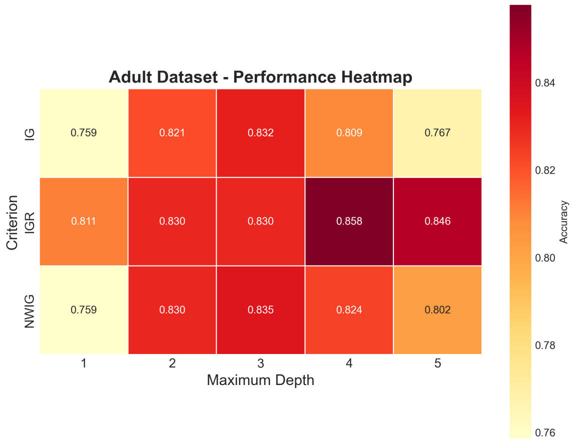
Figure 5: Performance vs Tree Complexity

NWIG consistently produced the most compact trees while maintaining competitive accuracy. The complexity-performance analysis revealed different patterns for each dataset: positive correlation for Iris (0.6038) and negative correlation for Adult (-0.7646), indicating overfitting on the Adult dataset.

3.4 Overfitting Analysis



(a) Iris Dataset



(b) Adult Dataset

Figure 6: Performance Heatmaps Across All Configurations

All criteria showed overfitting at maximum depth, but NWIG exhibited the most graceful degradation. On the Adult dataset, IG and IGR accuracy dropped significantly at depth 5, while NWIG maintained relatively stable performance.

4 Analysis and Discussion

4.1 Key Observations

Accuracy: IGR achieved the highest peak accuracy but with high variance, especially on complex datasets. NWIG provided competitive accuracy with superior consistency.

Consistency: NWIG demonstrated exceptional stability across all experimental runs, making it more reliable for practical applications. IGR's high variance on the Adult dataset raises concerns about its reliability.

Complexity: NWIG produced the most interpretable trees with fewer nodes, while IG created larger, potentially overfitted trees. The optimal depth range was 3-4 for most configurations.

Overfitting Resistance: NWIG showed the best resistance to overfitting, maintaining stable performance even at higher depths where other criteria degraded.

4.2 Trade-offs

The results reveal clear trade-offs between different aspects:

- IGR: Maximum accuracy vs. high variance
- NWIG: Balanced performance vs. slightly lower peak accuracy
- IG: Simplicity vs. tendency to overfit

4.3 Dataset Impact

The clean Iris dataset allowed all criteria to perform well with minimal differences. The complex Adult dataset amplified differences between criteria, highlighting the importance of careful selection for real-world applications.

5 Conclusions

Based on the experimental results:

For maximum accuracy: IGR is the best choice when peak performance is the primary concern, but requires careful validation due to high variance.

For reliability and consistency: NWIG offers the best balance of accuracy, stability, and interpretability, making it ideal for practical applications.

For simplicity: IG provides reasonable baseline performance with predictable behavior, suitable for simple scenarios or rapid prototyping.

Optimal depth: Depth constraints in the range of 3-4 effectively prevent overfitting while maintaining good performance across all criteria.

The custom NWIG criterion successfully addresses the limitations of traditional criteria by providing consistent performance with compact, interpretable trees. This makes it particularly valuable for real-world applications where model reliability and explainability are important.