CSE 318 Assignment-04: Decision Tree

MD. MAHMUD HASAN

July 2025

Contents

1 Overview

In this assignment, I have implemented a decision tree learning algorithm and applied it to the provided datasets. My algorithm supports the following three attribute selection criteria:

- Information Gain (IG)
- Information Gain Ratio (IGR)
- A custom variant of IG: Normalized Weighted Information Gain (NWIG)

2 Definitions of the Criteria

Let:

- S: the dataset
- A: an attribute in S
- S_v: the subset of S where attribute A has value v
- k: the number of unique values of attribute A
- $|S|$: the number of examples in dataset S

2.1 Information Gain (IG)

Information Gain measures the reduction in entropy after a dataset is split on attribute A. It is defined as:

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

Entropy is calculated as:

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

where $p(c)$ is the proportion of class c in dataset S.

2.2 Information Gain Ratio (IGR)

To correct IG's bias toward attributes with many values, the Information Gain Ratio normalizes IG by the Intrinsic Value (IV) of the attribute:

$$IGR(S, A) = \frac{IG(S, A)}{IV(A)}$$

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

This penalizes attributes that split the data into many small subsets, such as IDs or timestamps.

2.3 Normalized Weighted Information Gain (NWIG)

NWIG is a custom criterion that modifies IG to penalize attributes with high cardinality and adjusts for dataset size. It is defined as:

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

This reduces the score of attributes with many distinct values, especially in small datasets, helping to avoid overfitting due to over-splitting.

3 Tables

Criterion	**Max Depth**	**Avg. Accuracy (%)**	**Avg. Node Count**	**Avg. Depth**
Information Gain (IG)				
IG	0	93.33	11	5
IG	3	96.67	9	3
IG	5	96.67	13	5
IG	8	100.00	17	5
Gain Ratio (GR)				
GR	0	93.33	11	4
GR	3	96.67	9	3
GR	5	86.67	15	5
GR	8	93.33	9	4
Normalized Weighted IG (NWIG)				
NWIG	0	90.00	13	4
NWIG	3	96.67	9	3
NWIG	5	93.33	13	4
NWIG	8	96.67	19	6

Table 1: Iris Dataset

Criterion	Max Depth	Avg. Accuracy (%)	Avg. Node Count	Avg. Depth
Information Gain (IG)				
IG	0	76.65	10939	24
IG	3	84.03	657	3
IG	5	81.50	3617	5
IG	8	78.14	7619	8
Gain Ratio (GR)				
GR	0	78.00	10223	33
GR	3	82.97	234	3
GR	5	85.83	404	5
GR	8	83.89	1963	8
Normalized Weighted IG (NWIG)				
NWIG	0	75.74	11497	24
NWIG	3	84.02	706	3
NWIG	5	81.77	3289	5
NWIG	8	76.95	8061	8

Table 2: Adult Dataset

4 Graphical Representation

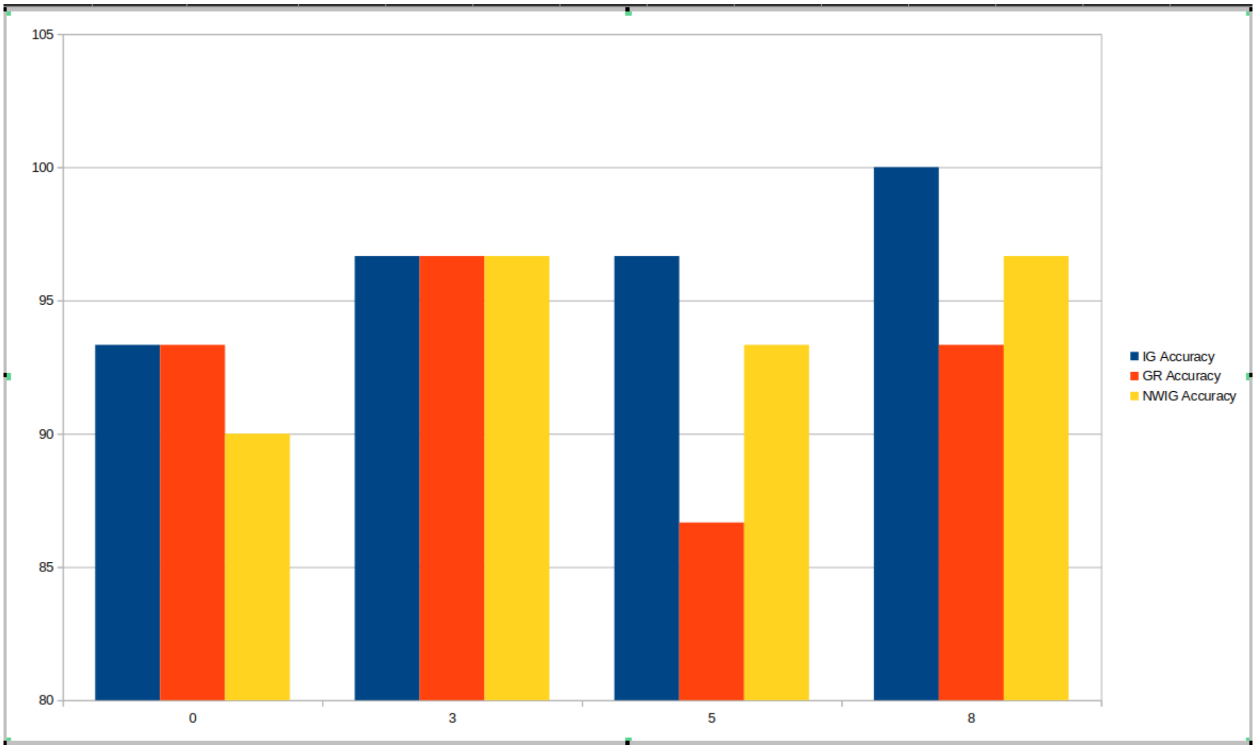

Figure 1: Average Accuracy vs. Tree Depth - Iris Dataset

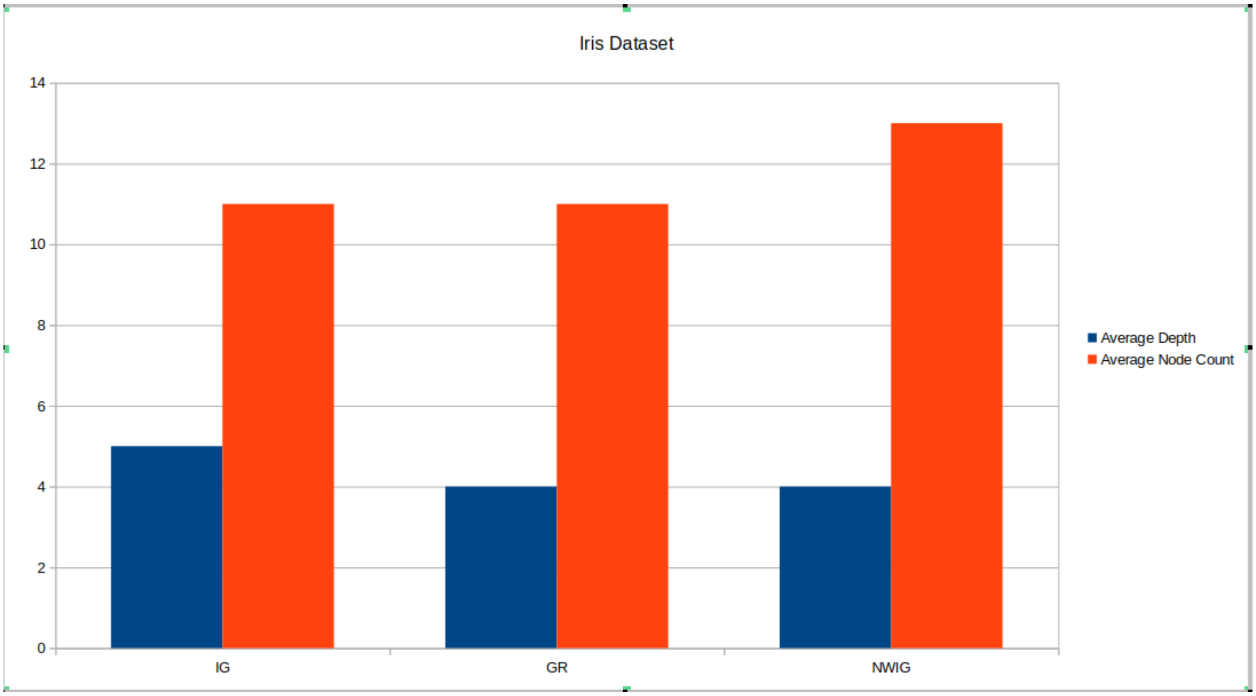

Figure 2: Number of Nodes vs. Tree Depth - Iris Dataset - without pruning

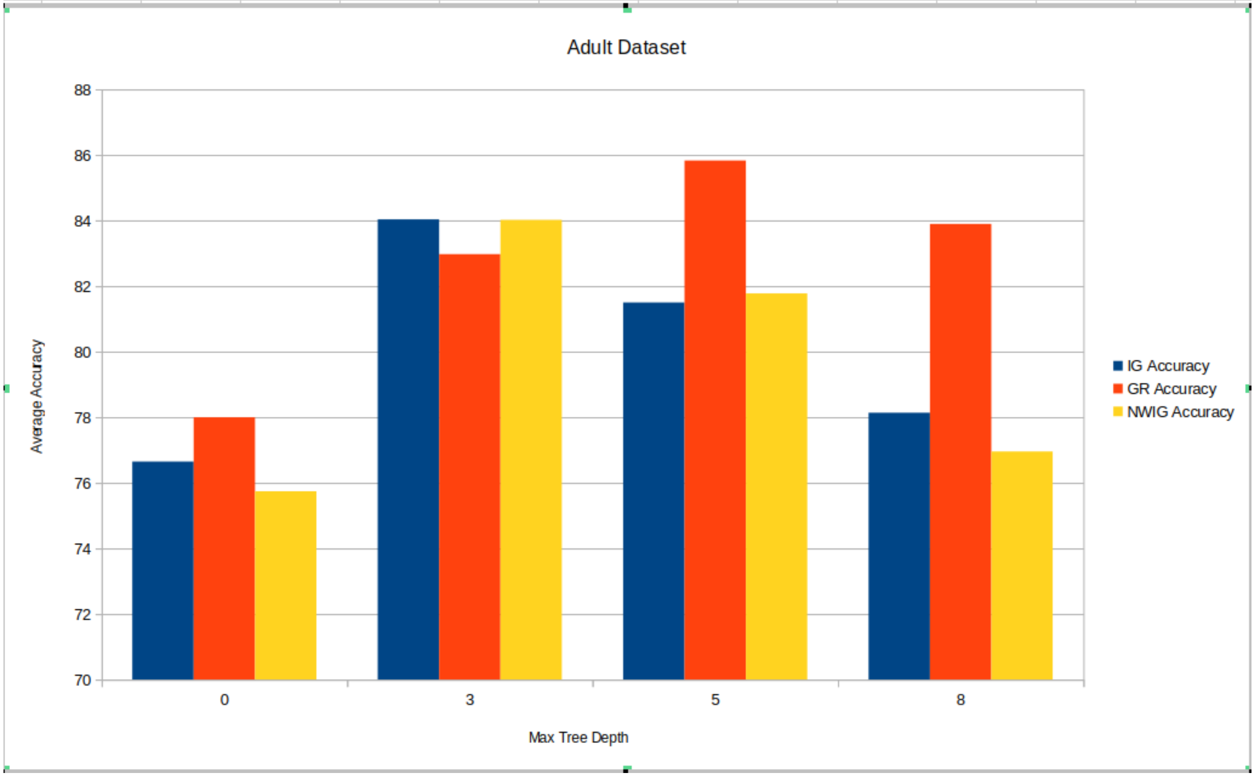

Figure 3: Average Accuracy vs. Tree Depth - Adult Dataset

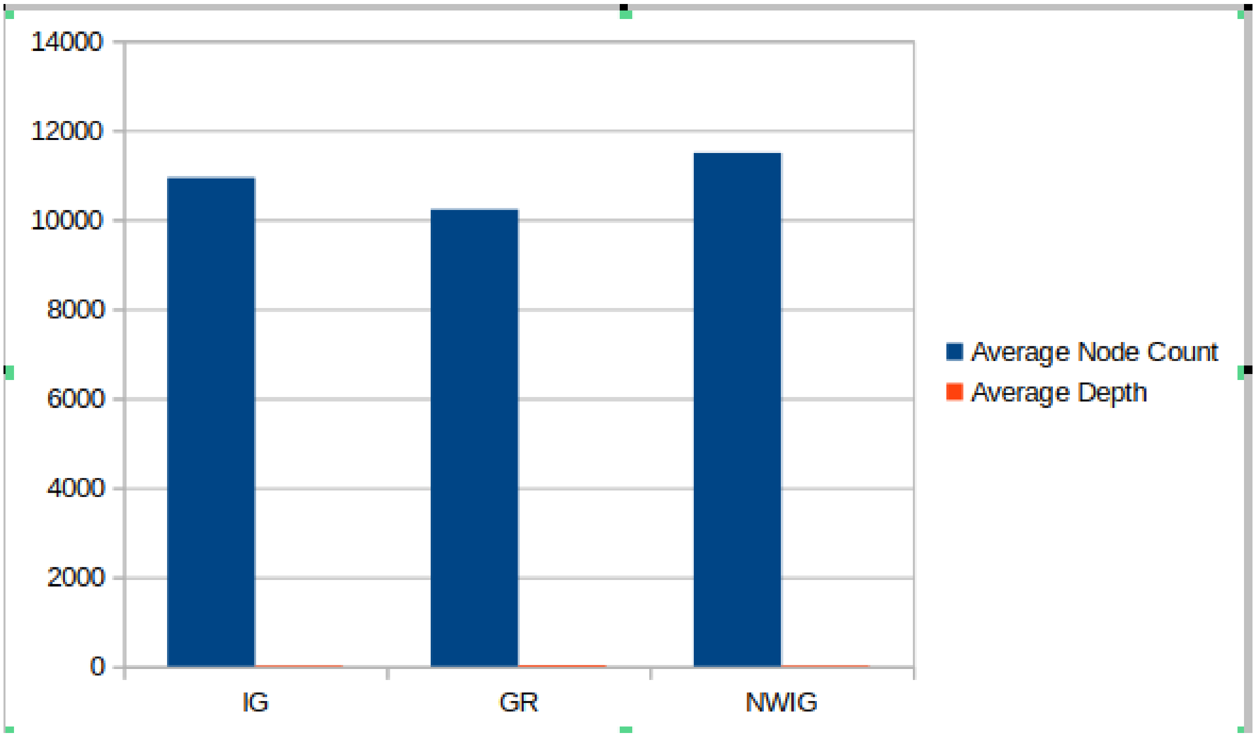

Figure 4: Number of Nodes vs. Tree Depth - Iris Dataset - without pruning

5 Performance Summary

Iris Dataset

- At **depth 0**, both **IG** and **GR** achieve the highest accuracy (93.33%).
- At **depth 3**, all three criteria (**IG**, **GR**, **NWIG**) perform equally well (96.67%).
- At **depth 5**, **IG** maintains the best performance (96.67%), outperforming GR and NWIG.
- At **depth 8**, **IG** achieves **100% accuracy**, the highest among all.
- **Observation**: **IG** consistently performs best as tree depth increases; pruning has less impact due to the simplicity of the dataset.

Adult Dataset

- At **depth 0**, **GR** leads with the highest accuracy (77.99%).
- At **depth 3**, **IG** achieves the highest accuracy (84.03%), followed closely by NWIG and GR.
- At **depth 5**, **GR** outperforms others (85.83%), indicating strong generalization with moderate depth.

- At **depth 8**, **GR** again gives the best accuracy (83.89%), while IG and NWIG drop slightly.
- **Observation**: **GR** handles the complexity of the adult dataset better and shows stronger resistance to overfitting at greater depths.

Overall Insights

- **IG** performs best for simpler datasets (Iris), especially at higher depths.
- **GR** is more effective for complex, high-cardinality datasets (Adult), where it avoids overfitting better.
- **NWIG** performs consistently well but does not lead at any specific depth; it offers a good balance with regularization.

Overfitting Reduction From the experimental results, the Gain Ratio (GR) criterion demonstrates the most consistent ability to reduce overfitting, particularly on the more complex Adult dataset. While Information Gain (IG) tends to overfit at greater depths, GR maintains higher accuracy with significantly fewer nodes and shallower trees, especially at depths 3 and 5. This suggests that GR's normalization effectively penalizes attributes with many distinct values, leading to more generalized trees. NWIG also helps reduce overfitting by penalizing high-cardinality splits, but its performance varies slightly more than GR. Overall, GR strikes the best balance between accuracy and model complexity across different pruning levels.

Effect of Pruning and Consistency of Performance Pruning generally improved the accuracy of the decision trees by preventing overfitting and simplifying the model. For both datasets, accuracy increased significantly when limiting the tree depth to moderate values (e.g., depth 3 or 5), compared to unpruned trees (depth 0). However, excessive pruning (e.g., depth 3 on complex datasets) sometimes led to underfitting and reduced performance.

Unexpected Patterns and Trade-offs During the analysis, some unexpected patterns emerged. For instance, in the Iris dataset, unpruned trees (depth 0) already performed quite well, indicating the dataset's simplicity. In contrast, the Adult dataset exhibited severe overfitting with unpruned trees, especially with IG and NWIG, as reflected by deep trees and high node counts. Interestingly, for some depths (e.g., depth 5), NWIG occasionally underperformed compared to GR, despite being designed to penalize high-cardinality attributes.