Report on
CSE 318 Assignment - 4: Decision Tree

Abdullah Al Mahmud
Student ID: 2105120

July 2025

1 Introduction

In this assignment, we implement a Decision Tree classifier that can handle both categorical and numerical attributes. The tree is constructed based on classical information-theoretic criteria such as Information Gain (IG), Information Gain Ratio (IGR), and Normalized Weighted Information Gain (NWIG). The goal is to recursively partition the dataset into subsets that are increasingly pure with respect to the class label, resulting in a tree structure that can be used for classification. For numerical attributes, the tree identifies optimal thresholds to split the data, integrating them seamlessly into the decision-making process. This implementation provides insight into how decision trees operate internally and how various attribute selection strategies influence the structure and performance of the model.

2 Average accuracy vs Max tree depth plots

2.1 Dataset: Iris.csv

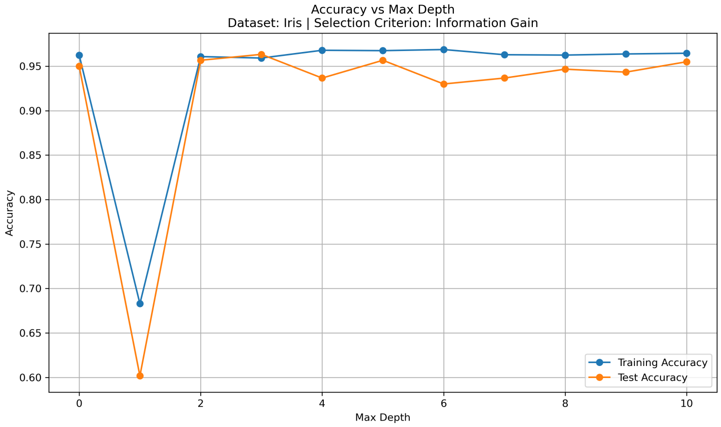

Figure 1: Accuracy vs Depth with Information gain criterion

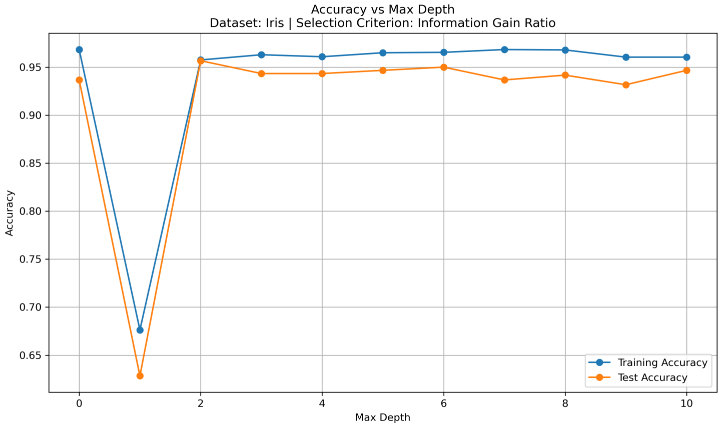

Figure 2: Accuracy vs Depth with Information gain ratio criterion

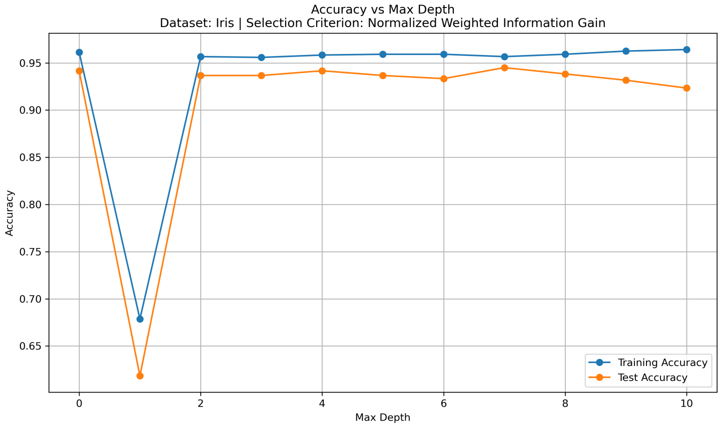

Figure 3: Accuracy vs Depth with Normalized weighted information gain criterion

2.2 Dataset: adult.data

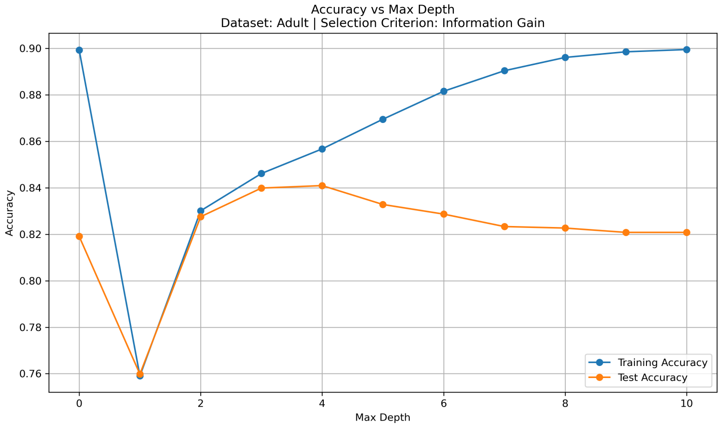

Figure 4: Accuracy vs Depth with Information gain criterion

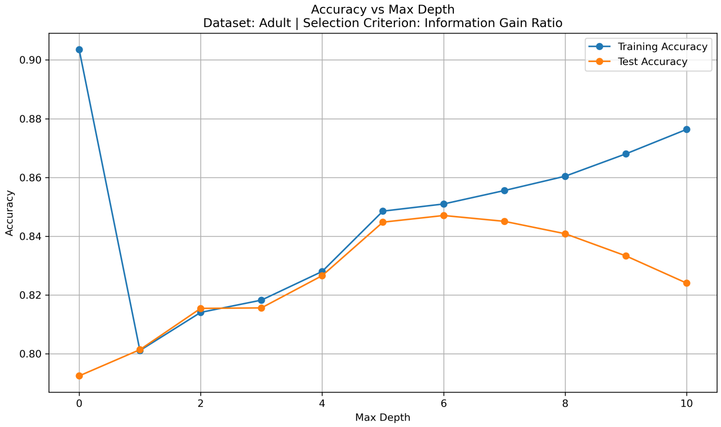

Figure 5: Accuracy vs Depth with Information gain ratio criterion

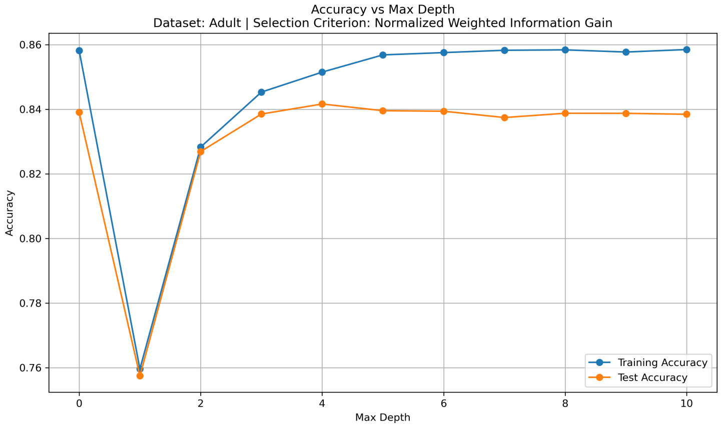

Figure 6: Accuracy vs Depth with Normalized weighted information gain criterion

3 Observations and Analysis

3.1 General Patterns Across All Criteria and Datasets

Training accuracy consistently increases as the tree depth increases or when no pruning is applied. However, test accuracy often improves initially but then slightly declines at deeper levels, indicating overfitting at high complexity. There is typically a sweet spot for maximum test accuracy, usually around depths 4 to 6, beyond which the model tends to overfit.

3.2 Results on `adult.data` Dataset

Information Gain (IG): When `max_depth` is 0 (no pruning), the tree grows fully, achieving the highest training accuracy of approximately 89.94%. However, test accuracy is lower (81.91%) compared to moderately pruned trees (e.g., depth 4), showing signs of overfitting. The best test accuracy observed is around 84.09% at depth 4. Beyond this depth, test accuracy slightly declines while training accuracy continues to increase, highlighting overfitting.

Information Gain Ratio (IGR): Similarly, with no pruning (`max_depth` = 0), training accuracy is highest (90.35%) but test accuracy is lower (79.25%), indicating overfitting. Moderate pruning at depth 6 yields the best test accuracy (84.71%). IGR tends to underfit at low depths but achieves good generalization at moderate depths, though it also starts to overfit at very high depths.

Normalized Weighted Information Gain (NWIG): With no pruning, NWIG achieves a high training accuracy (85.82%) and maintains relatively strong test accuracy (83.91%). This criterion shows less overfitting compared to IG and IGR even when the tree grows fully, suggesting its regularization properties help control complexity and stabilize performance across depths.

3.3 Results on `Iris.csv` Dataset

Information Gain (IG): The dataset is relatively simple, yielding high overall accuracy. At `max_depth` 0 (no pruning), the training accuracy is highest (96.25%) with test accuracy around 95%. Shallow trees (depth 1) perform poorly (60% test accuracy), but performance rapidly improves after depth 2, peaking around 96.33% test accuracy at depth 3. Some slight overfitting is observed beyond depth 4.

Information Gain Ratio (IGR): IGR achieves similarly high accuracy, with peak test accuracy (95–96%) occurring between depths 2 and 6. It is slightly more stable than IG, showing less fluctuation in test accuracy. The best generalization occurs near depth 6.

Normalized Weighted Information Gain (NWIG): NWIG performs almost as well as IG and IGR, maintaining consistent test accuracy between 93% and 94.5% from depth 2 onward. Like with `adult.data`, overfitting is minimal with NWIG, even without pruning.

3.4 Comparative Observations and Trade-offs

Criterion	Accuracy Peak Depth	Overfitting Risk
IG	4 (adult), 3 (Iris)	Medium
IGR	6 (adult), 6 (Iris)	High (if deep)
NWIG	4 (adult), 4–6 (Iris)	Low

Table 1: Comparison of splitting criteria based on accuracy peak depth and overfitting behavior.

Information Gain shows strong performance at low depths but overfits quickly. Information Gain Ratio helps handle multi-valued categorical attributes better but can underfit at low depths and overfit at high depths. NWIG offers better regularization, especially for real-world datasets like `adult.data`, by penalizing high branching factors and preventing overly complex splits.

3.5 Unexpected Observations

For the `adult.data` dataset, `max_depth` 0 (no pruning) already yields test accuracies between 79% and 83%, suggesting a dominant class structure in the data. In the `Iris.csv` dataset, test accuracy dramatically jumps from around 60% at depth 1 to over 95% at depth 2, indicating that the key attribute for class separation is selected early in the tree.

3.6 Tree Complexity Analysis (No Pruning)

The following table compares the size and depth of decision trees grown without pruning (`max_depth` = 0) for different splitting criteria on the `iris` and `adult` datasets:

Dataset	Selection Strategy	Node Count	Tree Depth
Iris	IG	12	3
Iris	IGR	12	3
Iris	NWIG	11	3
adult	IG	3391	12
adult	IGR	6679	14
adult	NWIG	906	9

Table 2: Comparison of decision tree size and depth for different splitting criteria without pruning (`max_depth` = 0).

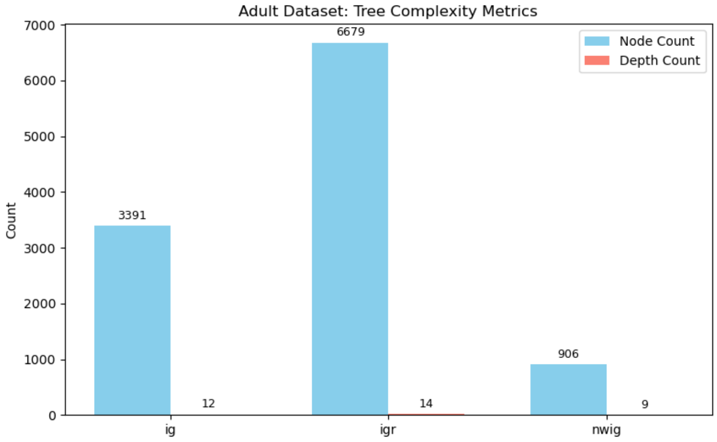

Figure 7: Node and depth count for adult.data

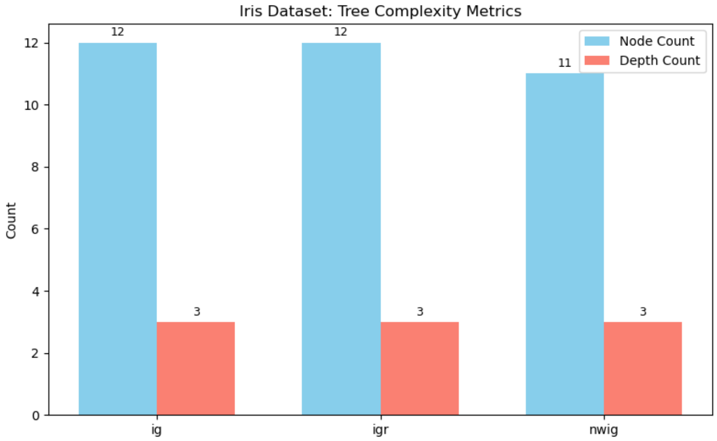

Figure 8: Node and depth count for Iris.csv

Analysis:

- The Iris dataset trees are relatively small and shallow across all criteria due to its simplicity and small size.
- On the adult dataset, trees grown using Information Gain Ratio (IGR) are the largest and deepest, reflecting more complex splitting possibly due to multi-valued categorical attributes.
- Trees built with Normalized Weighted Information Gain (NWIG) are substantially smaller and shallower, indicating that NWIG encourages simpler, more regularized trees.
- Information Gain (IG) produces moderate-sized trees, larger than NWIG but smaller than IGR.
- These differences in tree complexity affect both training time and risk of overfitting, with larger trees generally more prone to overfitting.

4 Discussion

The results show a clear trade-off between model complexity and generalization. When no pruning is applied (max_depth $= 0$), trees grow very deep and achieve high training accuracy but tend to overfit, causing test accuracy to plateau or decline.

Among the splitting criteria, Information Gain Ratio (IGR) produces the largest and deepest trees, which can lead to overfitting if uncontrolled. Normalized Weighted Information Gain (NWIG) yields simpler, smaller trees and generally maintains more stable test accuracy, indicating better regularization. Information Gain (IG) lies in between.

Limiting tree depth acts as an effective form of pre-pruning, improving test accuracy by reducing overfitting. The best generalization is usually achieved at moderate depths (around 4 to 6), balancing bias and variance well.

In summary, careful selection of splitting criteria and pruning depth is essential to build accurate yet robust decision trees.