Decision Tree

Aritra Debnath

1 Introduction

This report summarizes the experiments conducted with a decision tree learning algorithm on the **Adult (Sampled)**, **Adult (Full)**, and **Iris** datasets. The algorithm was implemented in C++ and tested using three attribute selection criteria: Information Gain (IG), Information Gain Ratio (IGR), and Normalized Weighted Information Gain (NWIG). Various maximum tree depths were explored to evaluate the impact of pruning on performance. The datasets were split into **80% training** and **20% testing**, with experiments repeated **20 times** for each configuration to compute average accuracy, node count, and tree depth.

2 Adult (Sampled) Dataset

2.1 Results

2.1.1 Information Gain (IG)

Max Depth	Avg Accuracy	Avg Nodes	Avg Depth
1	0.756875	3	1
2	0.8385	7	2
3	0.84975	13	3
5	0.84525	38.6	5
10	0.8255	171.7	10
0	0.798125	453.5	29

Table 1: Results for Adult (Sampled) with IG

2.1.2 Information Gain Ratio (IGR)

Max Depth	Avg Accuracy	Avg Nodes	Avg Depth
1	0.79925	3	1
2	0.807125	6.9	2
3	0.808375	12.2	3
5	0.841375	25.8	5
10	0.846375	88.5	10
0	0.8055	535.2	65.3

Table 2: Results for Adult (Sampled) with IGR

2.1.3 Normalized Weighted Information Gain (NWIG)

2.2 Graph

Max Depth	Avg Accuracy	Avg Nodes	Avg Depth
1	0.756875	3	1
2	0.83775	7	2
3	0.849125	13.1	3
5	0.84875	38.1	5
10	0.83975	170	10
0	0.810625	632.8	40.1

Table 3: Results for Adult (Sampled) with NWIG

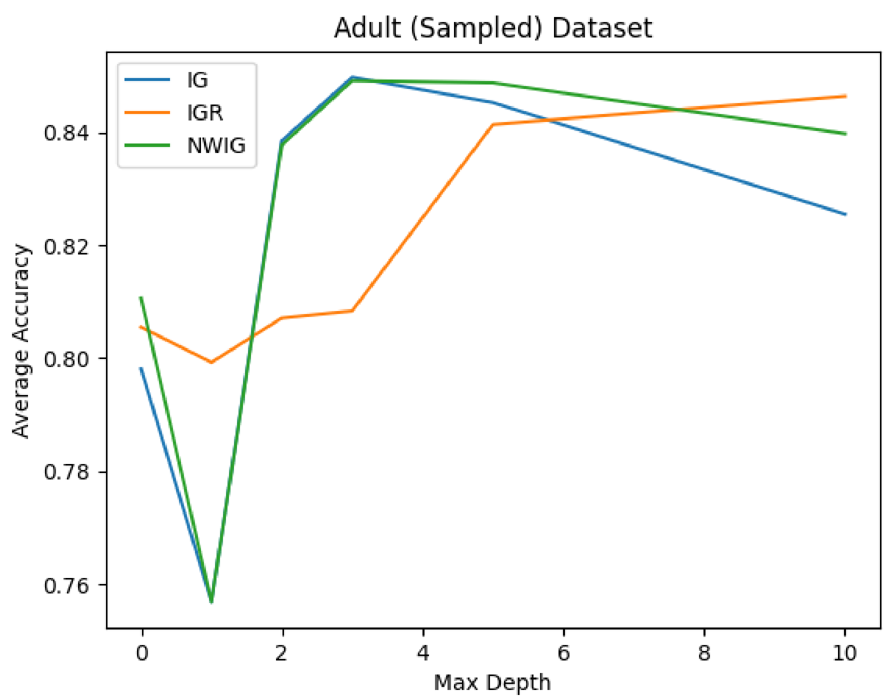

Figure 1: Average Accuracy vs Max Depth for Adult (Sampled)

3 Adult (Full) Dataset

3.1 Results

3.1.1 Information Gain (IG)

Max Depth	Avg Accuracy	Avg Nodes	Avg Depth
2	0.823338	7	2
3	0.84164	15	3
5	0.844142	53.4	5

Table 4: Results for Adult (Full) with IG

3.1.2 Information Gain Ratio (IGR)

Max Depth	Avg Accuracy	Avg Nodes	Avg Depth
2	0.800184	7	2
3	0.805681	13.8	3
5	0.84924	40.1	5

Table 5: Results for Adult (Full) with IGR

3.1.3 Normalized Weighted Information Gain (NWIG)

Max Depth	Avg Accuracy	Avg Nodes	Avg Depth
2	0.825979	7	2
3	0.841655	15	3
5	0.841962	54.5	5

Table 6: Results for Adult (Full) with NWIG

3.2 Graph

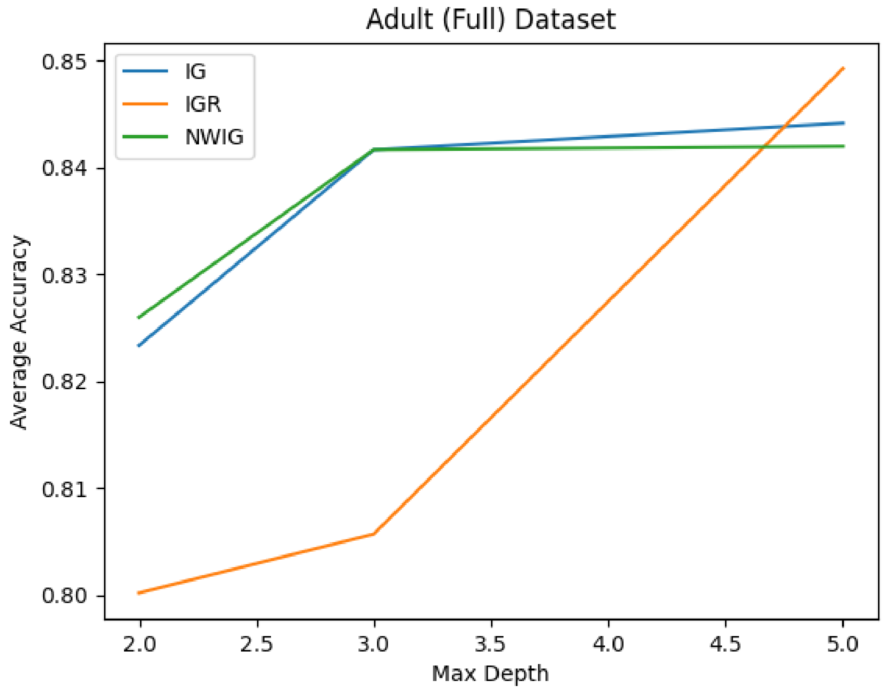

Figure 2: Average Accuracy vs Max Depth for Adult (Full)

4 Iris Dataset

4.1 Results

4.1.1 Information Gain (IG)

Max Depth	Avg Accuracy	Avg Nodes	Avg Depth
1	0.633333	3	1
2	0.948333	5	2
3	0.953333	8.5	3
5	0.946667	14.5	4.65
10	0.946667	15.2	5

Table 7: Results for Iris with IG

4.1.2 Information Gain Ratio (IGR)

Max Depth	Avg Accuracy	Avg Nodes	Avg Depth
1	0.633333	3	1
2	0.946667	5	2
3	0.946667	8.4	3
5	0.943333	14.4	4.65
10	0.943333	15.4	5.15

Table 8: Results for Iris with IGR

4.1.3 Normalized Weighted Information Gain (NWIG)

Max Depth	Avg Accuracy	Avg Nodes	Avg Depth
1	0.633333	3	1
2	0.97	5	2
3	0.946667	8.8	3
5	0.943333	14.7	4.7
10	0.946667	16.3	5.5

Table 9: Results for Iris with NWIG

4.2 Graph

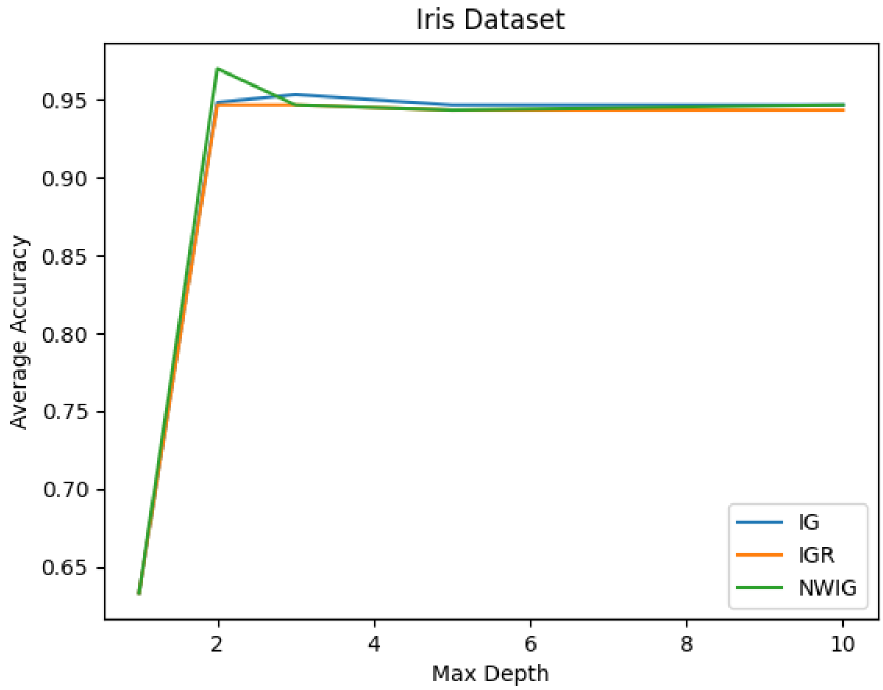

Figure 3: Average Accuracy vs Max Depth for Iris

5 Analysis and Insights

The experiments provide valuable insights into the performance of the IG, IGR, and NWIG criteria across the **Adult (Sampled)**, **Adult (Full)**, and **Iris** datasets, with **pruning** playing a significant role in balancing accuracy and tree complexity. Below are detailed findings:

- **Performance on Adult (Sampled)**: IGR consistently outperforms IG and NWIG at higher depths, achieving a peak accuracy of **0.846375** at depth **10**, compared to **0.8255** (IG) and **0.83975** (NWIG). This suggests IGR's normalization by intrinsic value effectively mitigates bias toward high-cardinality attributes, leading to more robust splits in this complex dataset. At unlimited depth (**0**), all criteria suffer from **overfitting**, with accuracy dropping (e.g., **0.798125** for IG) due to excessive tree growth (**453.5** nodes for IG). **Pruning** at depth **3** yields optimal accuracy for IG (**0.84975**) and NWIG (**0.849125**), demonstrating the importance of limiting tree depth.

- **Performance on Adult (Full)**: IGR achieves the highest accuracy at depth **5** (**0.84924**), slightly outperforming IG (**0.844142**) and NWIG (**0.841962**). At lower depths (**2** and **3**), IG and NWIG perform better, indicating they may select more discriminative features early in this larger dataset (**33k instances**). IGR produces smaller trees (**40.1** nodes at depth **5** vs. **53.4** for IG), reflecting its efficiency in avoiding over-splitting. The larger dataset size reduces **overfitting** compared to **Adult (Sampled)**, as accuracy improves steadily with depth.

- **Performance on Iris**: All criteria achieve high accuracy (**0.94–0.97**) at depths **2** and above, due to the **Iris** dataset's simplicity (**150 instances**, continuous features). NWIG excels at depth **2** (**0.97**), likely because its cardinality penalty favors the dataset's structure. Trees remain compact (**5–16.3** nodes), and **pruning** beyond depth **3** has minimal impact, as the dataset is easily separable. The small differences across criteria suggest that **Iris** is less sensitive to attribute selection strategies.

- **Impact of Pruning**: **Pruning** is critical for the **Adult** datasets, where unlimited depth leads to large trees (e.g., **535.2** nodes for IGR in **Adult (Sampled)**) and reduced accuracy due to **overfitting**. Optimal accuracy is often achieved at depths **3–5**, balancing complexity and generalization. In contrast, **pruning** has little effect on **Iris**, as its simplicity limits **overfitting** risks, with stable accuracy across depths.

- **Trade-offs and Notable Patterns**: IGR consistently produces smaller trees than IG and NWIG, enhancing computational efficiency, especially at higher depths. NWIG's conservative splits, driven by its cardinality penalty, can limit accuracy in complex datasets like **Adult (Sampled)** at higher depths. An unexpected pattern is IGR's excessive tree depth (**65.3** in **Adult (Sampled)** at unlimited depth), indicating overly granular splits when unconstrained. Overall, IGR offers a robust balance of accuracy and tree size for complex datasets, while NWIG is more effective for simpler datasets like **Iris**.