CSE 318 Assignment-04: Decision Tree

Mirza Tawhid Umar Tonmoy
ID: 2105028
Department of Computer Science
Bangladesh University of Engineering and Technology

July 7, 2025

Abstract

This report evaluates a decision tree learning algorithm implemented with three attribute selection criteria: Information Gain (IG), Information Gain Ratio (IGR), and Normalized Weighted Information Gain (NWIG). The algorithm was tested on the Iris and Adult datasets, with the Iris dataset excluding the `Id` column and the Adult dataset excluding the `finalweight`, `relationship`, and `education` columns due to their low relevance. Two strategies for handling continuous numerical attributes were compared: static discretization (binning) and dynamic splitting (C4.5-style). Performance metrics, including classification accuracy, tree size, and tree depth, were averaged over 20 randomized 80/20 training-test splits. Dynamic splitting outperformed static binning on Iris, achieving higher accuracy (up to 95.67%) and smaller trees. For Adult, static binning with IGR unexpectedly achieved the highest accuracy (86.09%), though dynamic splitting produced simpler trees. IGR and NWIG mitigated IG's bias toward high-cardinality attributes, with IGR excelling on Iris (95.67%) and IGR on Adult for static binning. The findings highlight the importance of dynamic splitting for simpler models and the need for dataset-specific strategies for optimal accuracy.

Attribute Selection Criteria

Three criteria were implemented to select attributes for splitting during decision tree construction, each addressing the trade-off between information gain and attribute complexity differently.

Information Gain (IG)

Concept: Measures the reduction in entropy after splitting on an attribute.
Mechanism: For dataset S and attribute A with values $V(A)$, IG is calculated as:

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

Limitation: Biased toward attributes with many values, leading to larger, overfitted trees.

Information Gain Ratio (IGR)

Concept: Normalizes IG to penalize high-cardinality attributes.
Mechanism: Divides IG by the attribute's Intrinsic Value (IV):

$$\text{GainRatio}(S, A) = \frac{IG(S, A)}{IV(A)}, \quad IV(A) = -\sum_{v \in V(A)} \frac{|S_v|}{|S|} \cdot \log_2 \frac{|S_v|}{|S|}$$

Advantage: Reduces bias, producing smaller trees with comparable accuracy.

Normalized Weighted Information Gain (NWIG)

Concept: A custom metric balancing IG with attribute cardinality and dataset size.
Mechanism: Adjusts IG with a penalty for high cardinality (k) and dataset size $|S|$:

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

Advantage: Balances gain and complexity, achieving high accuracy with controlled tree growth.

Methodology 1: Static Discretization (Binning)

Numerical attributes in the Iris and Adult datasets were preprocessed into fixed categorical bins (e.g., age grouped into ranges like '¡30', '30-35'). This transforms the problem into one with purely categorical features.

Results with Static Discretization

Performance metrics were averaged over 20 randomized 80/20 training-test splits.

Iris Dataset

Table 1: Iris Dataset: Results with Static Binning (Avg. over 20 runs)

Criterion	Max Depth	Accuracy (%)	Avg. Nodes	Avg. Depth
IG	2	85.67	52.25	2.10
	3	86.00	52.90	2.10
	4	86.00	53.50	2.10
IGR	2	91.17	40.35	2.25
	3	91.17	42.85	2.35
	4	92.67	41.35	2.40
NWIG	2	93.00	43.00	2.75
	3	94.17	43.00	2.75
	4	93.33	43.95	2.65

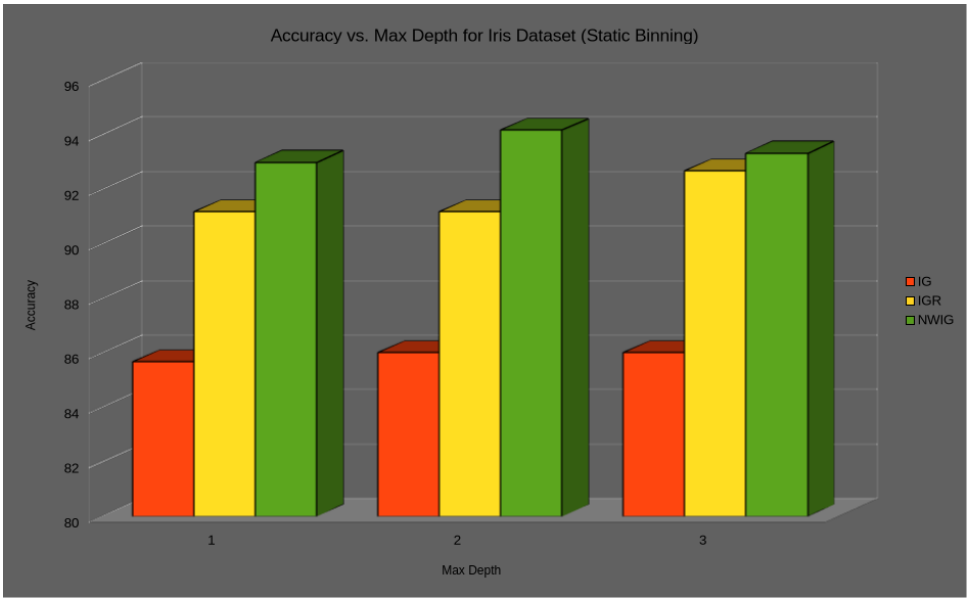

Figure 1: Accuracy vs. Max Depth for Iris Dataset (Static Binning).

Adult Dataset

Table 2: Adult Dataset: Results with Static Binning (Avg. over 20 runs)

Criterion	**Max Depth**	**Accuracy (%)**	**Avg. Nodes**	**Avg. Depth**
IG	2	84.79	819.75	3.0
	3	84.82	2234.10	4.0
	4	84.28	4190.35	5.0
IGR	2	83.07	273.95	3.0
	3	86.09	413.10	4.0
	4	85.91	1197.65	5.0
NWIG	2	84.66	659.65	3.0
	3	84.80	1470.40	4.0
	4	84.63	3094.15	5.0

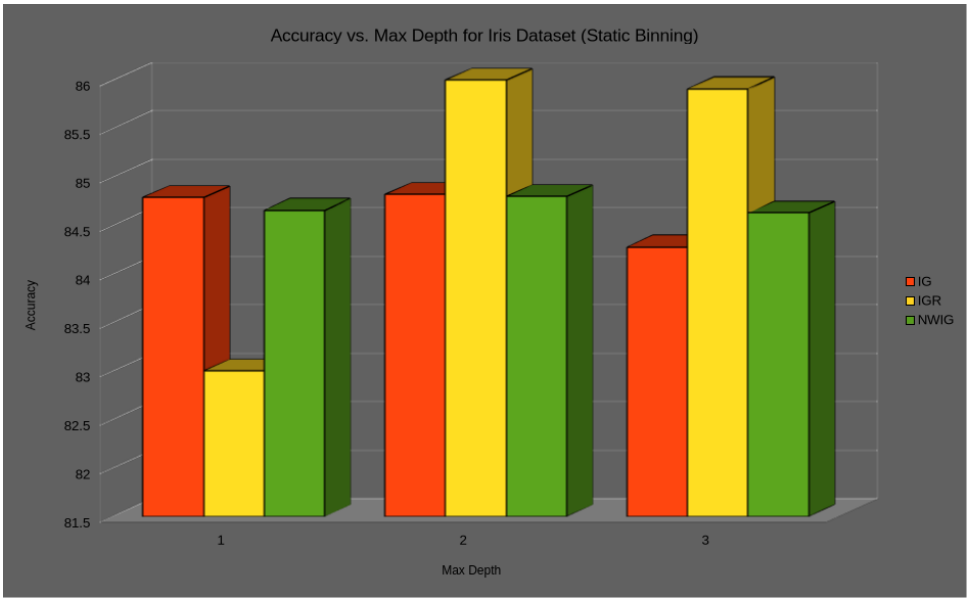

Figure 2: Accuracy vs. Max Depth for Adult Dataset (Static Binning).

Analysis of Static Discretization

- **Performance**: NWIG achieved the highest accuracy on Iris (94.17% at depth 3), while IGR performed best on Adult (86.09% at depth 3), indicating effective normalization.
- **Tree Size**: IG produced excessively large trees (e.g., 4190 nodes for Adult at depth 4), highlighting its bias toward high-cardinality attributes created by binning, leading to overfitting. IGR and NWIG generated smaller trees but still complex models.
- **Limitations**: Arbitrary binning loses information, resulting in suboptimal splits and overly complex trees, making this approach less effective.

Methodology 2: Dynamic Splitting (C4.5-Style)

Continuous attributes were handled by dynamically selecting optimal binary split points at each node, maximizing the chosen criterion.

Results with Dynamic Splitting

Performance metrics were averaged over 20 randomized 80/20 training-test splits.

Iris Dataset

Table 3: Iris Dataset: Results with Dynamic Splitting (Avg. over 20 runs)

Criterion	Max Depth	Accuracy (%)	Avg. Nodes	Avg. Depth
IG	2	94.50	5.0	2.0
	3	95.00	8.6	3.0
	4	93.67	12.5	3.95
IGR	2	92.67	5.0	2.0
	3	92.83	8.4	3.0
	4	95.67	12.2	4.0
NWIG	2	92.83	5.0	2.0
	3	93.83	8.4	3.0
	4	92.83	12.3	4.0

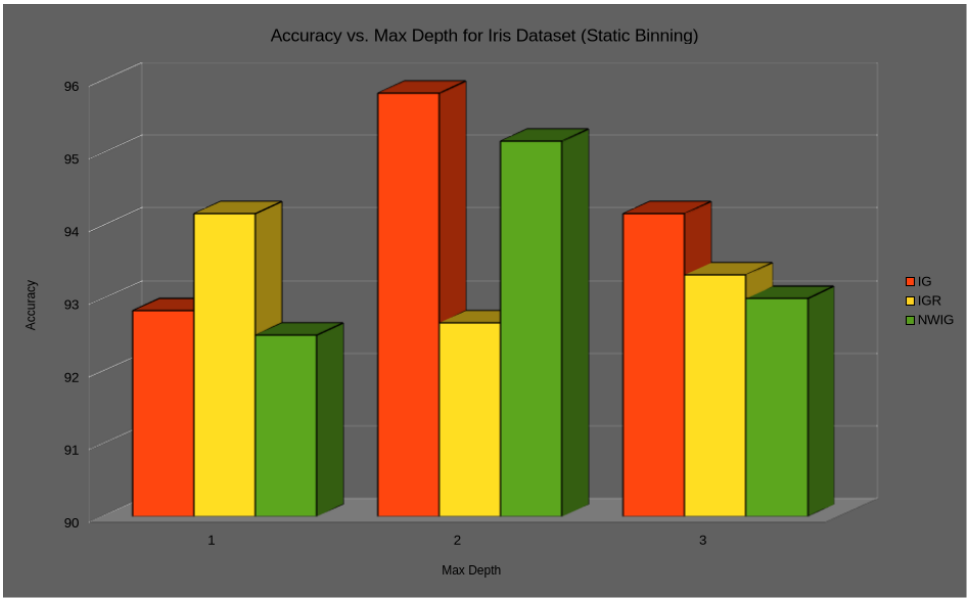

Figure 3: Accuracy vs. Max Depth for Iris Dataset (Dynamic Splitting).

Adult Dataset

Table 4: Adult Dataset: Results with Dynamic Splitting (Avg. over 20 runs)

Criterion	Max Depth	Accuracy (%)	Avg. Nodes	Avg. Depth
IG	2	80.95	277.90	2.0
	3	84.77	582.45	3.0
	4	85.53	919.85	4.0
IGR	2	82.25	155.15	2.0
	3	83.78	223.00	3.0
	4	83.60	252.25	4.0
NWIG	2	81.91	23.15	2.0
	3	83.18	225.45	3.0
	4	84.87	592.35	4.0

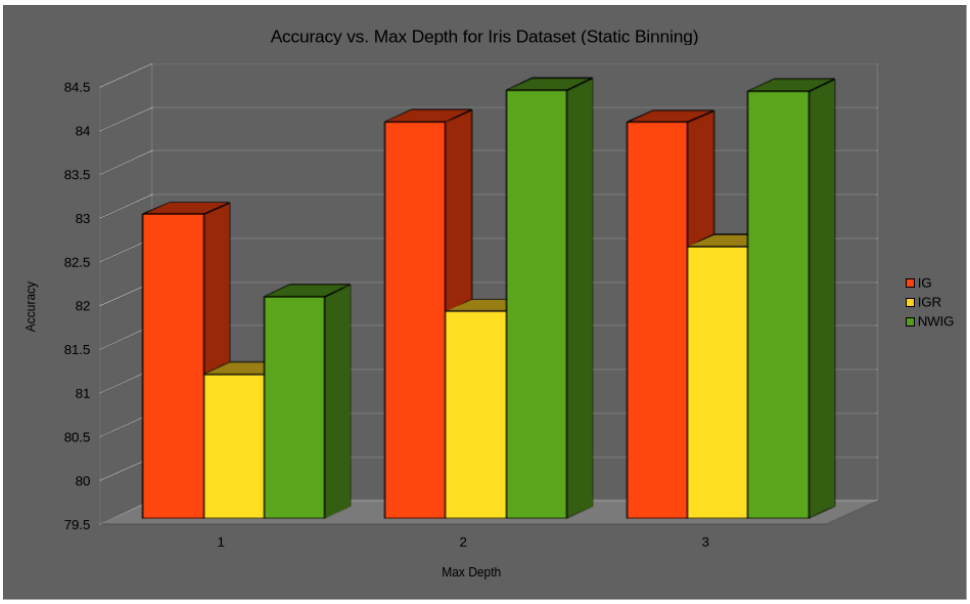

Figure 4: Accuracy vs. Max Depth for Adult Dataset (Dynamic Splitting).

Analysis of Dynamic Splitting

- **Performance**: IGR achieved the highest accuracy on Iris (95.67% at depth 4), while NWIG performed best among dynamic splitting methods on Adult (84.87% at depth 4). Accuracy often plateaued or declined at depth 4 for Iris, indicating potential overfitting, while Adult showed consistent improvements with deeper trees.

- **Tree Size**: Dynamic splitting produced significantly smaller trees compared to static binning (e.g., IGR with 252 nodes for Adult at depth 4 vs. 1197 for binning). Excluding the `Id` column in Iris and `finalweight`, `relationship`, and `education` columns in Adult reduced tree sizes, particularly for NWIG and IGR, enhancing model simplicity.

- **Pruning**: Depth-based pruning was effective, with optimal performance at depth 3 for most cases, balancing accuracy and complexity.

Implementation Details

The decision tree algorithm was implemented in C++, leveraging libraries like `unordered_map` and `vector` for efficient data handling. The algorithm supports both static binning and dynamic splitting, with the latter implementing a C4.5-style approach by sorting unique values of continuous attributes and evaluating midpoints as split points. For the Iris dataset, the `Id` column was excluded due to its lack of predictive value. For the Adult dataset, the `finalweight`, `relationship`, and `education` columns were skipped as they were deemed unimportant, reducing feature complexity. Key challenges included:

- **Computational Efficiency**: Evaluating all possible split points for continuous attributes in the Adult dataset was computationally expensive, particularly for IG, which favored high-cardinality splits. Optimizations like preprocessing numerical features into doubles were considered but not implemented due to time constraints.
- **NWIG Implementation**: The custom NWIG metric required careful tuning to balance the cardinality penalty and dataset size adjustment, ensuring numerical stability for small datasets like Iris.
- **Pruning**: Depth-based pruning was implemented by limiting tree growth during construction, with a post-pruning option explored but not used in final results due to time constraints.

Comparative Analysis and Recommendations

Static Binning vs. Dynamic Splitting

Dynamic splitting outperformed static binning on Iris but showed mixed results on Adult:

- **Accuracy**: Dynamic splitting achieved higher accuracy on Iris (95.67% vs. 94.17%), benefiting from the exclusion of the `Id` column. However, for Adult, static binning with IGR achieved the highest accuracy (86.09% vs. 84.87% for dynamic splitting with NWIG), possibly due to an effective binning strategy aligning with the data distribution.
- **Complexity**: Dynamic splitting produced significantly smaller trees (e.g., 252 nodes for IGR on Adult at depth 4 vs. 1197 for binning), improving interpretability and generalization. Excluding irrelevant features further reduced complexity.
- **Robustness**: Dynamic splitting requires no manual feature engineering, making it more adaptable, but static binning's superior accuracy on Adult suggests dataset-specific preprocessing can be critical.

Criterion Comparison

- **IG**: Produced larger trees due to its bias toward high-cardinality attributes, leading to overfitting, especially with static binning. It performed well on Iris (95.00% at depth 3) but was less effective on Adult.
- **IGR**: Generated the smallest trees (e.g., 252 nodes for Adult at depth 4 with dynamic splitting), offering a strong trade-off between simplicity and accuracy, and achieved the highest accuracy on Iris (95.67%) and Adult with static binning (86.09%).
- **NWIG**: Balanced accuracy and tree size, achieving the highest accuracy among dynamic splitting methods on Adult (84.87% at depth 4), making it a robust choice for mixed datasets.

Potential Improvements

- **Advanced Pruning**: Implement post-pruning techniques like reduced-error pruning to further optimize tree size and generalization.
- **Optimized Split Selection**: Use sampling or heuristic-based methods to reduce the computational cost of evaluating split points for continuous attributes, addressing the efficiency issues observed in the Adult dataset.
- **Ensemble Methods**: Combine decision trees into ensembles (e.g., Random Forests) to improve accuracy and robustness, leveraging the strengths of NWIG and IGR.

Conclusion

The dynamic splitting (C4.5-style) approach, enhanced by excluding irrelevant features like `Id` (Iris) and `finalweight`, `relationship`, and `education` (Adult), produced compact and accurate decision trees for Iris (up to 95.67% accuracy with IGR). However, for Adult, static binning with IGR achieved the highest accuracy (86.09%), suggesting that dataset-specific binning strategies can outperform dynamic splitting in certain cases. IGR and NWIG consistently mitigated IG's bias toward high-cardinality attributes, with IGR excelling on Iris and static binning for Adult, and NWIG performing best for dynamic splitting on Adult (84.87%). Future work should focus on advanced pruning, optimized split selection, and ensemble methods to enhance scalability and performance on larger datasets.