Decision Tree

2105160

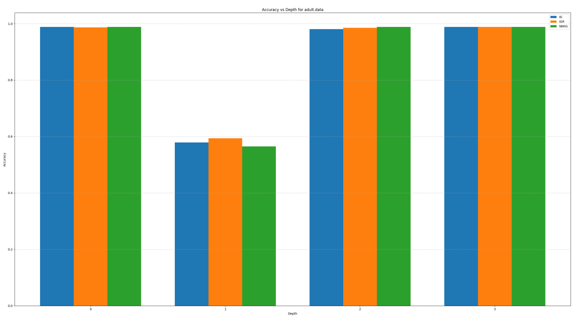
Introduction

A **Decision Tree** is a supervised learning algorithm used for classification and regression tasks. It recursively splits the dataset based on feature values to form a tree where each internal node represents a decision rule on a feature, and each leaf node represents a class label (or output value). The splits aim to maximize information gain or similar criteria, with common ones being:

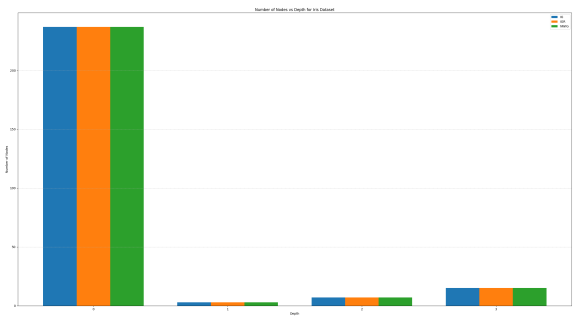
- **IG** – Information Gain
- **IGR** – Information Gain Ratio
- **NWIG** – Normalized Weighted Information Gain

Controlling the tree's depth or minimum samples per node helps avoid overfitting and improves generalization.

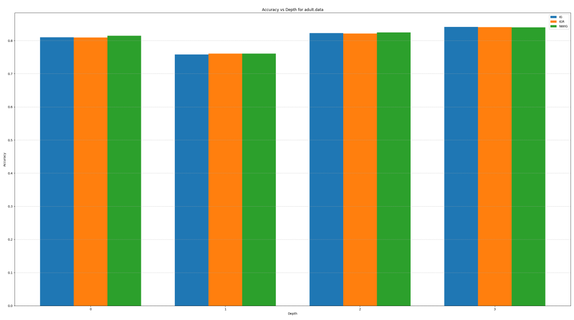
Performance Plots



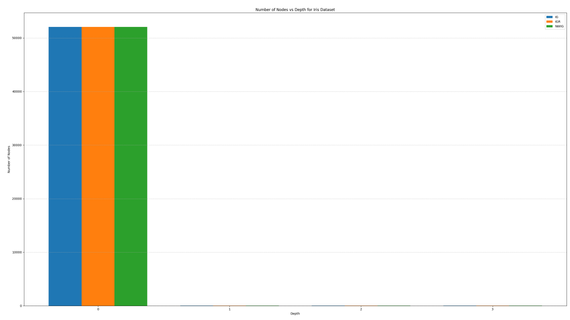
(a) Iris Dataset – Accuracy vs Depth



(b) Iris Dataset – Node Count vs Depth



(c) Adult Dataset – Accuracy vs Depth



(d) Adult Dataset – Node Count vs Depth

Figure 1: Decision Tree Performance on Iris and Adult Datasets for IG, IGR, NWIG

Observations and Analysis

- **Iris Dataset:**
  - All three criteria performed similarly, with small differences in accuracy.
  - IGR was slightly more stable across depths, and NWIG showed slightly better accuracy at depth 3.
  - Node count increased exponentially with depth, indicating greater complexity.
- **Adult Dataset:**
  - Depth 0 yielded high accuracy but also massive node counts ($\sim$52,000), likely due to fitting large leaf clusters.
  - NWIG slightly outperformed the others at deeper levels.
  - Accuracy plateaued after depth 2–3, suggesting early convergence.
- **General Insights:**
  - Tree depth directly impacts model complexity and accuracy.
  - Shallow trees generalize better but may underfit.
  - NWIG can balance gain and tree size better in noisy datasets.