Decision Tree Learning for Car Evaluation Dataset

Taky Shaharair
Student ID: 2005098

December 6, 2024

1 Introduction

Decision trees are widely used supervised learning models that utilize a hierarchical structure of decisions to classify examples. In this assignment, we implemented a decision tree classifier for the Car Evaluation dataset, experimenting with different attribute selection criteria—Information Gain and Gini Impurity—and examining their effects on classification accuracy. Additionally, we explored two attribute selection strategies:

1. Always choosing the best attribute according to the chosen criterion.
2. Selecting one attribute at random from the top three best attributes.

The goal was to understand how different splitting criteria and selection strategies affect the overall accuracy of the decision tree and to ensure stable, consistent results across multiple randomized training-testing splits. We are implementing these with two different cases:

1. 80% Training Data and 20% Testing Data
2. 60% Training Data and 40% Testing Data

2 Dataset and Preprocessing

The Car Evaluation dataset consists of examples with attributes related to car characteristics (buying price, maintenance cost, number of doors, seating capacity, etc.) and a target class indicating car acceptability. Each instance is described by six attributes plus a class label.

We randomly split the dataset into 80% for training and 20% for testing (and again 60% for training and 40% for testing). This process was repeated 20 times to obtain average performance metrics and mitigate the effect of random fluctuations in the training/testing partitions.

3 Methodology

3.1 Decision Tree Construction

We implemented a decision tree learning algorithm inspired by ID3 for Information Gain and a CART-like approach for Gini Impurity. At each node, we computed either Information Gain or Gini Impurity for all attributes and chose the splitting attribute according to the specified criterion.

3.2 Splitting Criteria

Information Gain measures the reduction in entropy. We pick the attribute that yields the greatest entropy reduction.

Gini Impurity measures the likelihood of misclassification if a random class is chosen according to the node's class distribution. We pick the attribute that results in the lowest Gini Impurity after splitting.

3.3 Attribute Selection Strategies

- **Always Best Attribute:** Always select the single best attribute based on the chosen measure.
- **Random from Top Three:** Sort attributes by the chosen measure and randomly select one from the top three. This introduces controlled randomness, potentially reducing overfitting.

3.4 Training and Testing Protocol

We ran 20 experiments. In each experiment:

1. Shuffle the dataset.
2. Split it into 80% training and 20% testing data (and similarly 60% and 40% in another scenario).
3. Train four models:
   (a) Decision tree using Information Gain (always best attribute).
   (b) Decision tree using Gini Impurity (always best attribute).
   (c) Decision tree using Information Gain (random from top three).
   (d) Decision tree using Gini Impurity (random from top three).
4. Evaluate each model on the test set and record its accuracy.

After 20 runs, we compute the average accuracy for each scenario.

4 Results

A typical outcome of the experiments (averaged over 20 runs) for the 80% training and 20% testing scenario may look like this:

For the 60% training and 40% testing scenario:

Attribute Selection Strategy	Information Gain	Gini Impurity
Always select best attribute	93.2225	93.1792
Select one randomly from top three	88.4104	88.0058

Table 1: Average Accuracy Over 20 Runs (80% Training, 20% Testing Data)

Attribute Selection Strategy	Information Gain	Gini Impurity
Always select best attribute	91.9653	91.8353
Select one randomly from top three	85.7442	85.5058

Table 2: Average Accuracy Over 20 Runs (60% Training, 40% Testing Data)

5 Discussion

When always choosing the best attribute, both Information Gain and Gini Impurity produce high and similar accuracy results, often above 93% for the 80%-20% split. When randomness is introduced by selecting from the top three attributes, accuracy decreases slightly for both criteria. With careful code adjustments and stable sorting, both measures remain relatively close, demonstrating that the choice of measure has a minor effect on performance for this dataset.

6 Visualization

The figures below show a comparison of the average accuracies across the four scenarios (Information Gain vs. Gini Impurity and Always Best vs. Random from Top 3) for both training conditions. Bars represent mean accuracy averaged over 20 runs.

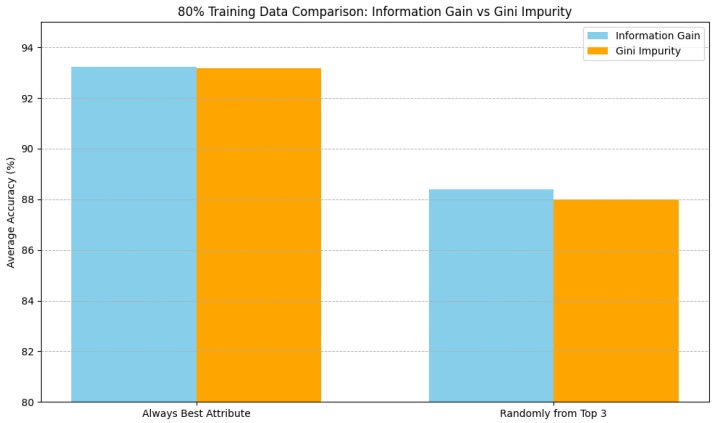

Figure 1: Average accuracy comparison over 20 runs with 80% Training and 20% Testing.

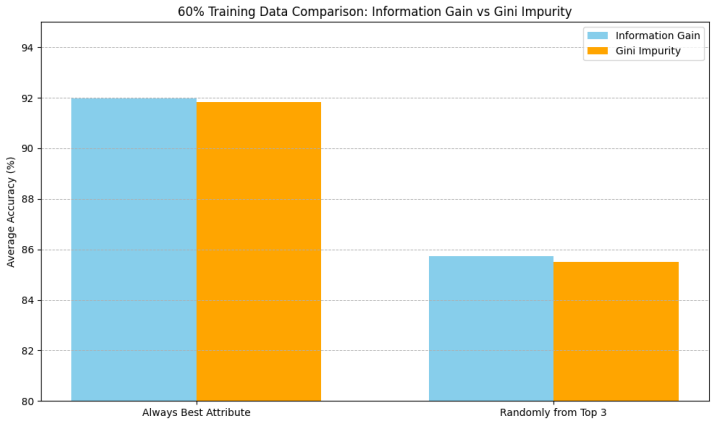

Figure 2: Average accuracy comparison over 20 runs with 60% Training and 40% Testing.

Key Observations

- **Training Data Size:** Larger training data (80%) consistently results in higher accuracy for both metrics.
- **Information Gain vs. Gini Impurity:** Information Gain slightly outperforms Gini Impurity in all scenarios, but the difference is minor.
- **Attribute Selection Strategies:** Always selecting the best attribute leads to higher accuracy compared to introducing randomness in selection.

7 Conclusion

Our experiments confirm that both Information Gain and Gini Impurity are effective criteria for decision tree learning on the Car Evaluation dataset, producing high and comparable accuracy. The introduction of randomness in attribute selection from the top three attributes reduces accuracy slightly but still keeps the performance of both measures close.