Analysis on Decision Tree Construction for Multi-class Classification of a Car Evaluation Dataset

Anik Saha, ID: 2005001

December 6, 2024

1 Introduction

Decision Tree is an tree-structure for conditional step-by-step decision making about particular data from given attributes. Decision trees are widely used for binary or multi-class classification problems. To select which attribute to choose while splitting a node in the tree, we consider using two different evaluation metrics, namely, Information Gain and Gini Impurity. In addition, we introduce the flexibility of choosing either the best performing attribute or one random from the top three performers at each level.

2 Problem Description

The problem is to classify cars as one of the 4 possible values "unacc", "acc", "good", "vgood".

The attributes provided are as follows:

- buying
- maint
- doors
- persons
- lugboot
- safety

3 Methodology

We first split the dataset into train and test set randomly as per the split ratio. Then we construct the decision tree from the train set and evaluate its performance on the test set.

The attribute selection strategies tested are as follows:

- Always taking the best attribute
- Selecting one randomly from the top three attributes

The evaluation metrics tested are as follows:

- Information Gain
- Gini Impurity

4 Results

4.1 Accuracy Comparison

	Average accuracy over 20 runs	
Attribute selection strategy	Information gain	Gini impurity
Always select the best attribute	93.28	92.54
Select one randomly from the top three attributes	81.49	81.42

4.2 Overall Accuracy Distribution

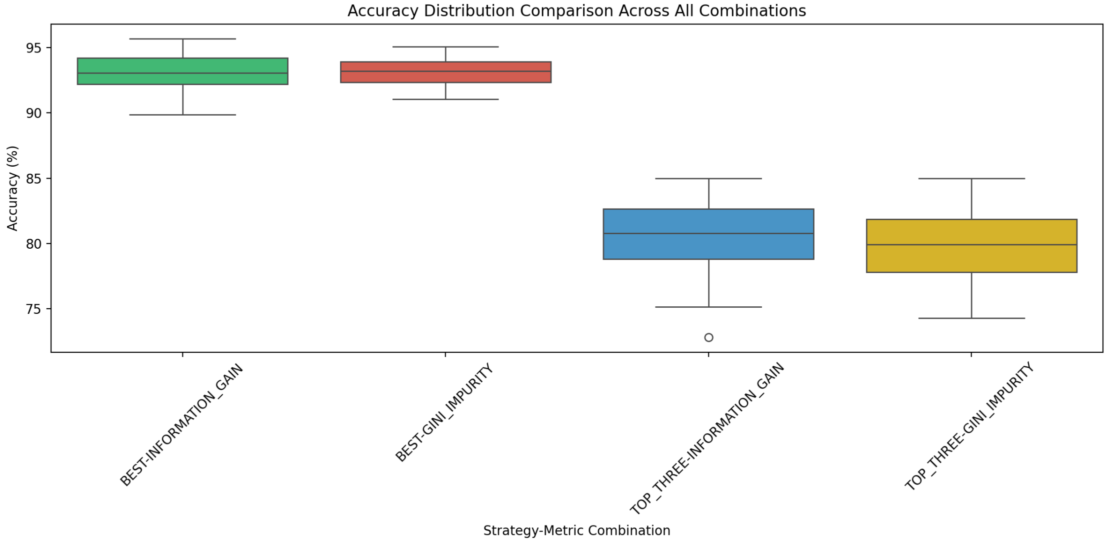

Figure 1: Comparison of accuracy distributions across different strategy and metric combinations.

The above box-whisker plot demonstrates the test-set accuracy of the decision tree built in case of 4 different combinations of selection strategy and evaluation metrics.

It is evident that the accuracy attains its maximum when we take the best attribute while testing at each node and the evaluation metric is set to be the Information Gain. The accuracy, in case of Gini Impurity is also very close to the previous case. However, when we pick one random attribute from the top three, not only does the mean accuracy drop, but also does the accuracy deviate more over several runs, thereby making it less consistent.

4.3 Confusion Matrices

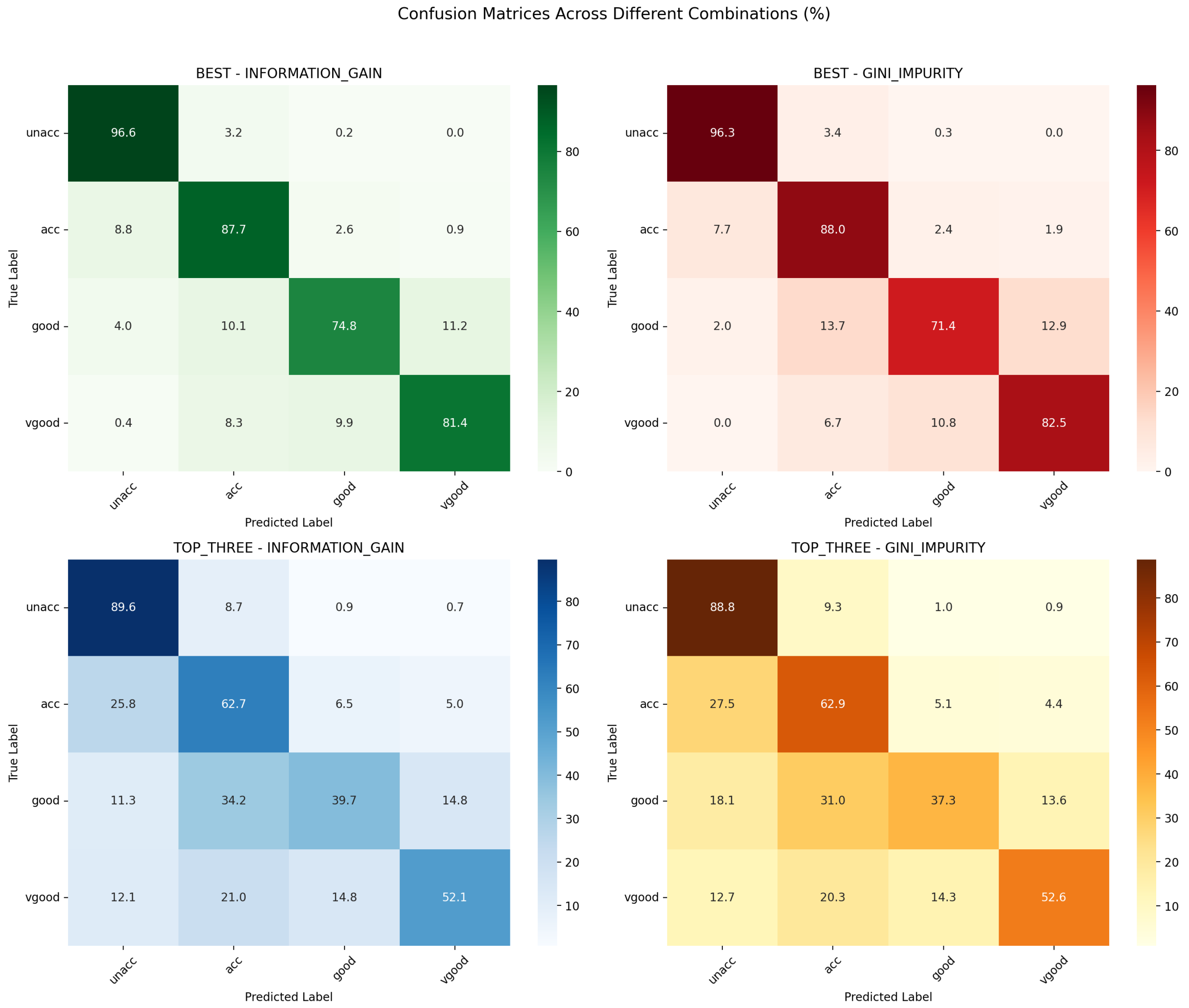

Figure 2: Confusion matrices showing classification performance across different strategy-metric combinations. Each subplot uses a different color gradient for better distinction.

The confusion matrices give insights about which class pairs the decision tree often confuse between.

For instance, it often classifies ”good” ones as ”acc”, as can be observed in all of the confusion matrices. And in case of random-choice from three attributes, it also often tends to classify ”vgood” classes as ”acc” or ”good”. Also, in case of random picking from top three, the confusion matrix appears more scattered and inconsistencies are higher.

4.4 Error Rate Analysis

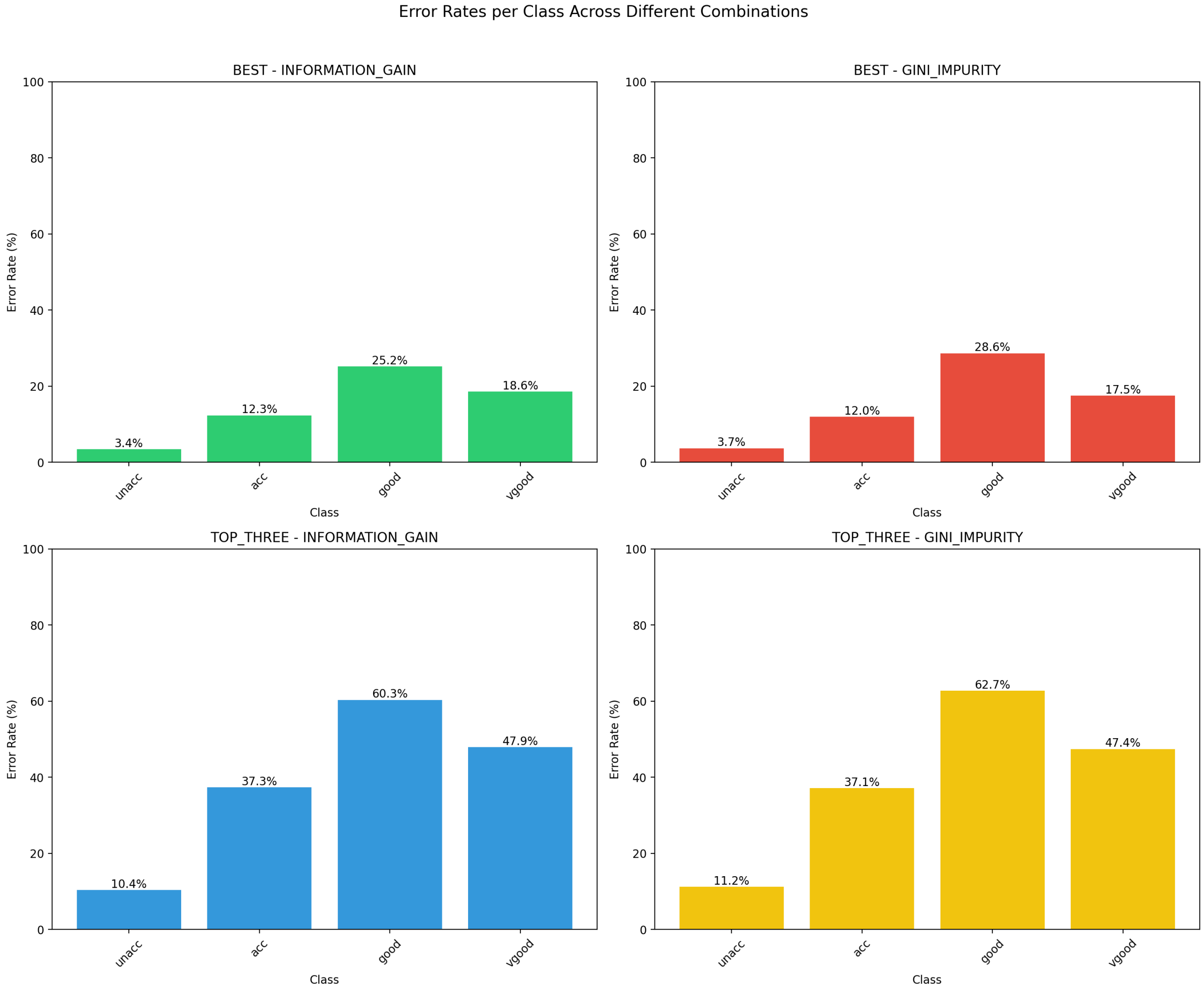

Figure 3: Error rates per class across different strategy-metric combinations, showing the misclassification percentages for each category.

These plots show that the decision tree, in almost all cases, classifies ”unacc” classes correctly. On the contrary, the maximum error rates are observed in case of ”good” class.

4.5 Attribute Depth Analysis

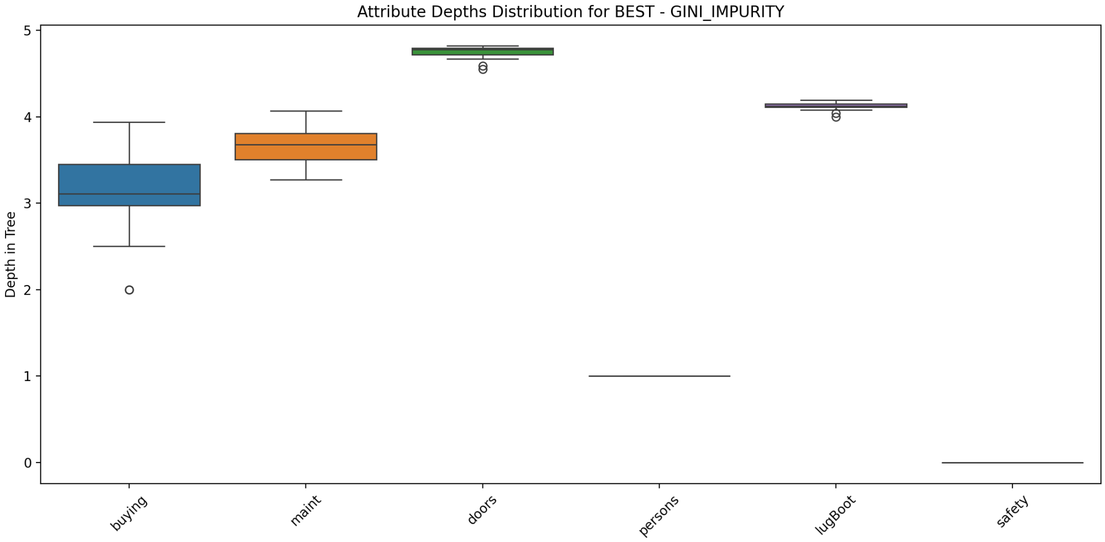

Figure 4: Distribution of attribute depths for BEST strategy with GINI IMPURITY metric.

This box whisker plot, shows an analysis of the attribute depths. An attribute with lower depth in the tree indicates that it is more crucial in the classification decision. Here, it can be seen that the attributes ”safety” and ”persons” are the most crucial having depths 0 and 1 consistently in the tree. This means they lie on the topmost decision nodes almost always. Others are placed at higher depths.

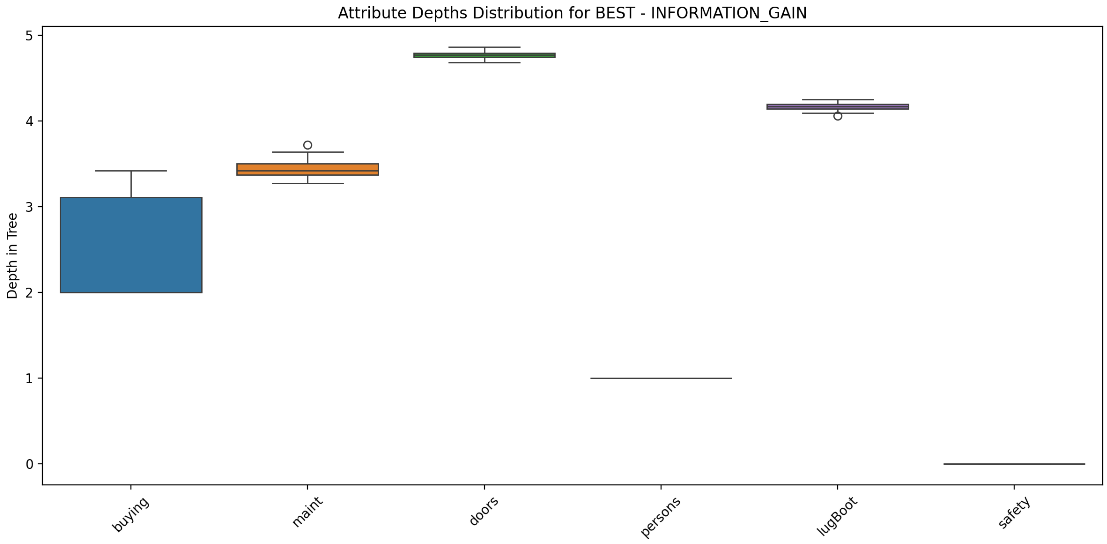

Figure 5: Distribution of attribute depths for BEST strategy with INFORMATION GAIN metric.

This plot for the case of Information Gain is very similar to the one above.

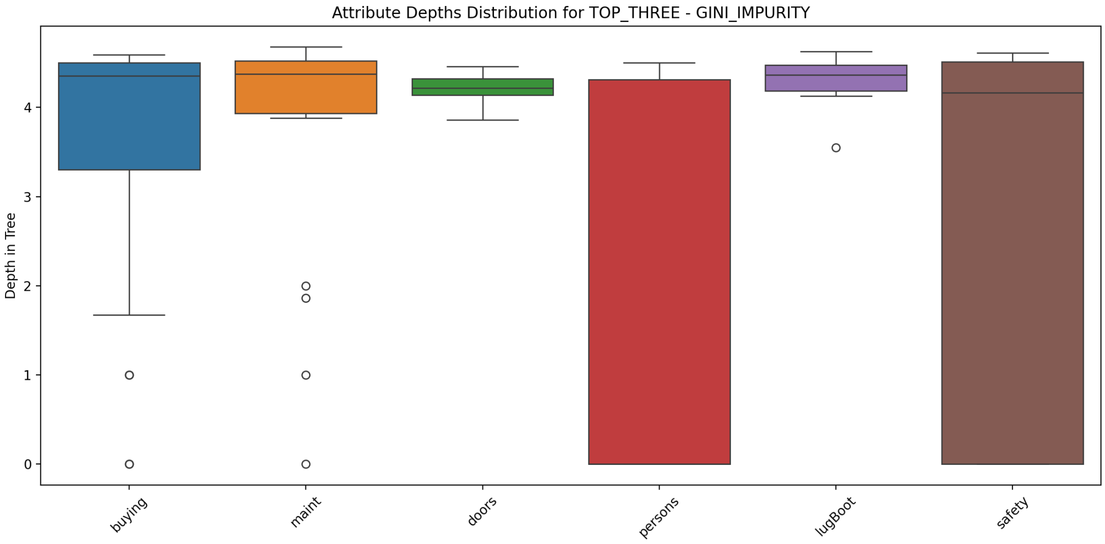

Figure 6: Distribution of attribute depths for TOP THREE strategy with GINI IMPURITY metric.

When we switch the selection strategy, these plots change drastically. The attributes ”persons” and ”safety” which were always placed closer to the root in the above cases, are now rather being placed at scattered depths ranging from 0 to 5. The reason is that, at the first nodes, due to random choices, often the optimal ones are rejected. And later, they are being placed at scattered positions in the tree. This also explains why the overall accuracy drops significantly in case of random choice.

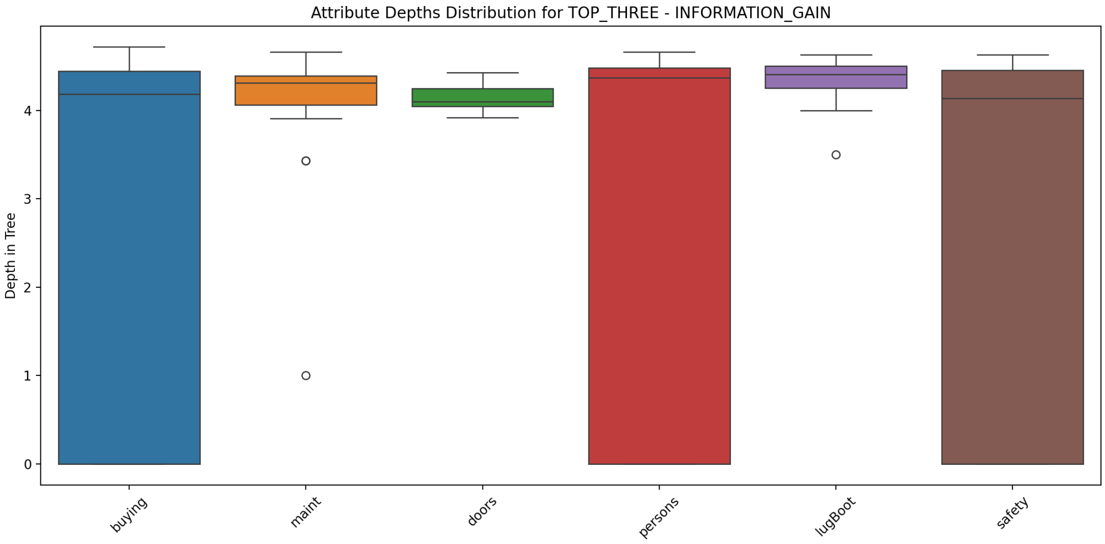

Figure 7: Distribution of attribute depths for TOP THREE strategy with INFORMATION GAIN metric.

This plot for the case of Information Gain is very similar to the one above as the depth distribution follows almost the same trend here.

4.6 Training Metrics Analysis

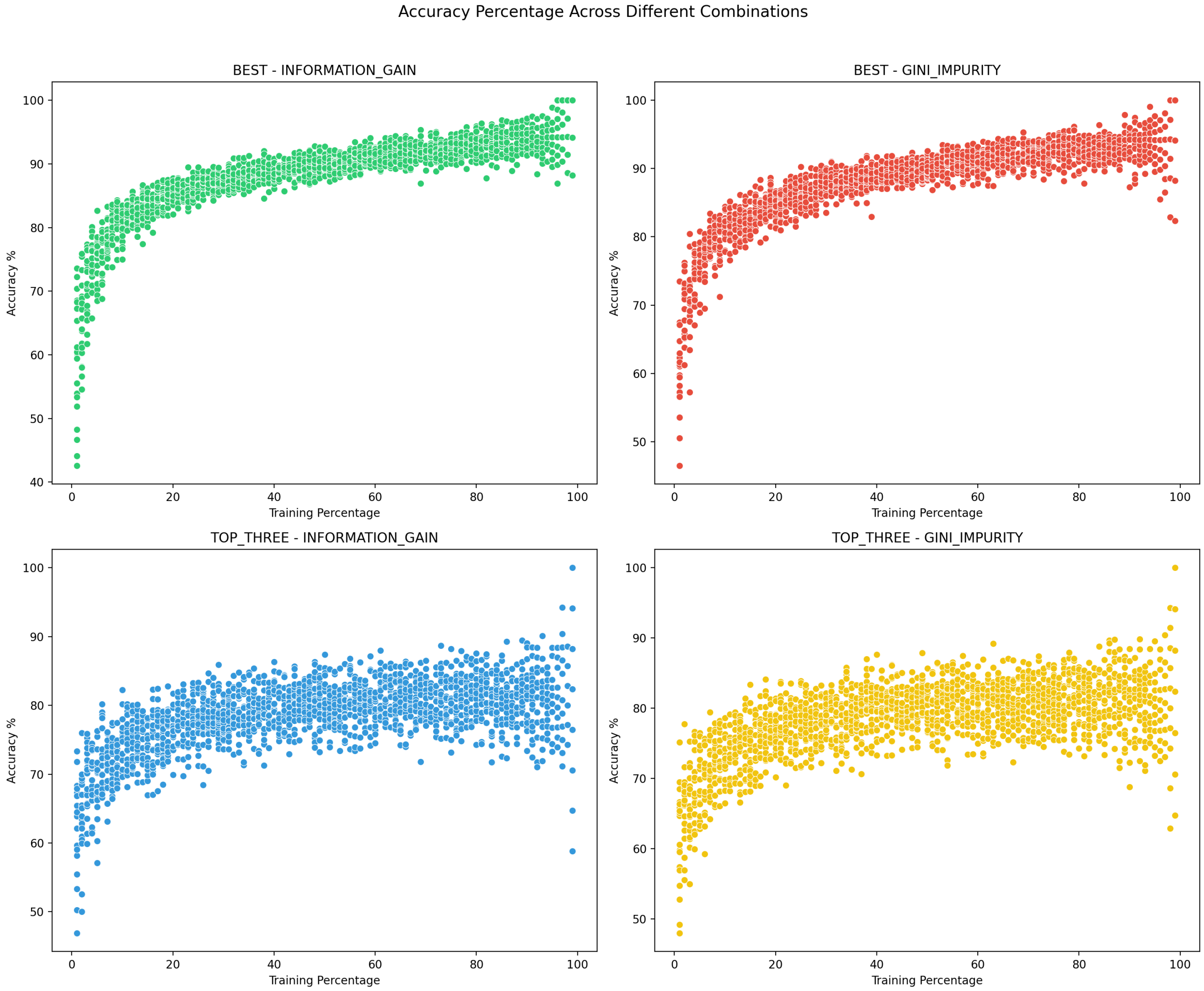

Figure 8: Accuracy percentage trends across different strategy-metric combinations showing how accuracy varies with training data percentage.

The above scatter plots demonstrate how the training accuracy varies over the training percentage. As we increase the percentage of train set in the split from 1% towards 99%, the accuracy first increase very fast and then grows slower. Moreover, the accuracy grows much faster and more consistently in the case of choosing the best attribute.

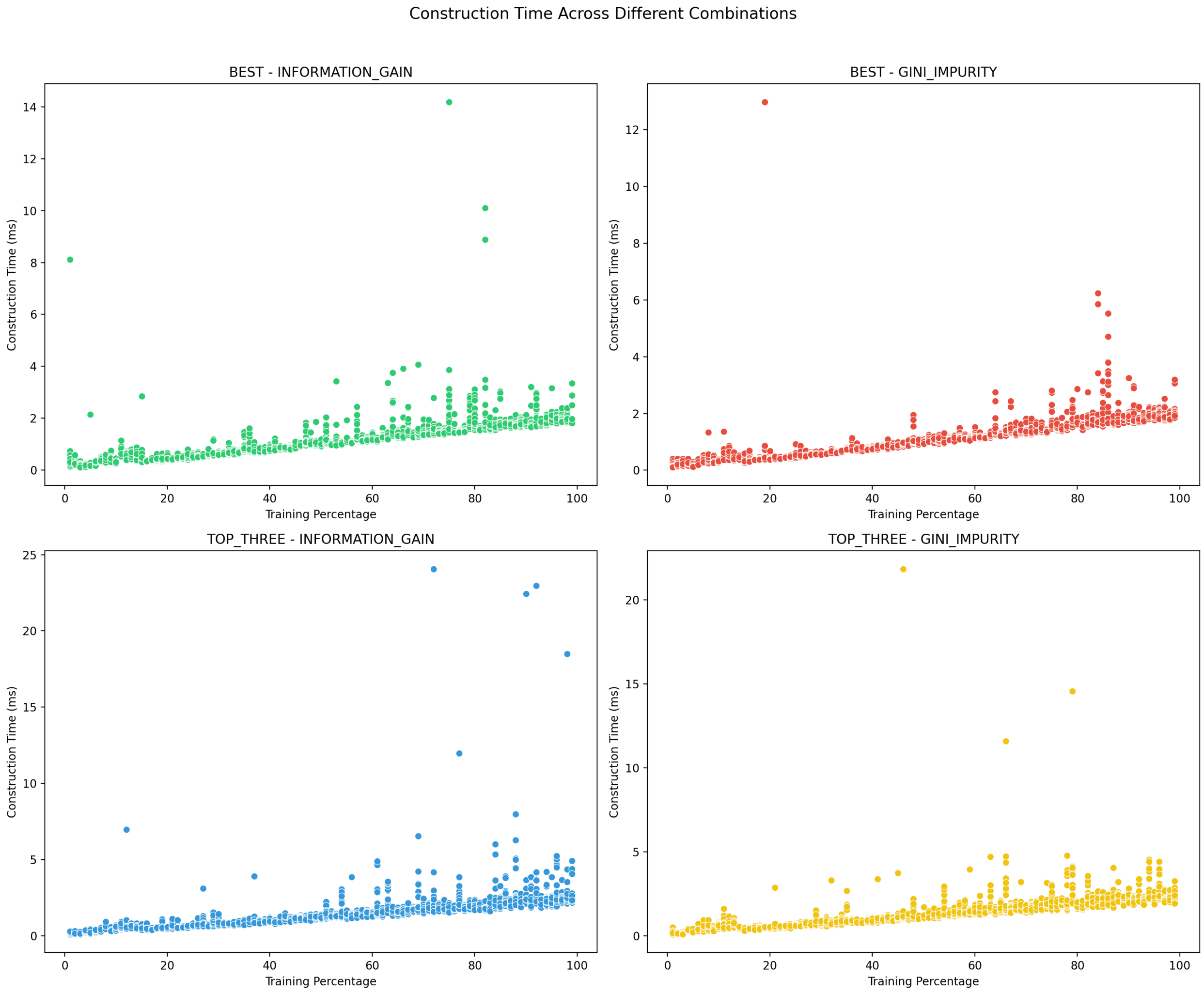

Figure 9: Construction time comparison across different strategy-metric combinations showing how build time varies with training data size.

The above scatter plots demonstrate how the construction time varies over the training percentage. As we increase the percentage of train set in the split from 1% towards 99%, the time almost always increases linearly, with some outliers. Naturally, the strategy and metrics do not much affect the construction time.

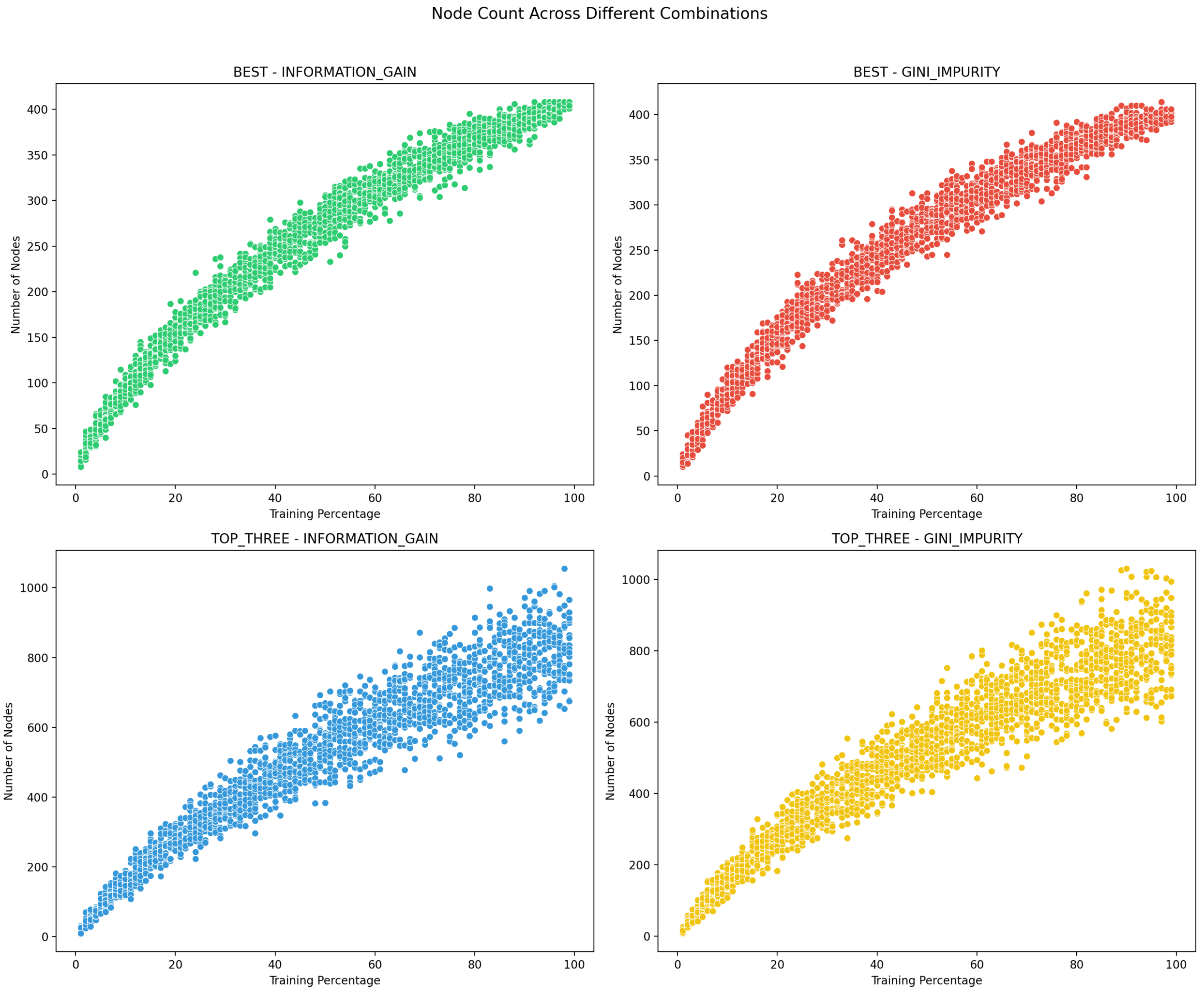

Figure 10: Node count comparison across different strategy-metric combinations showing how tree size varies with training data percentage.

The above scatter plots demonstrate how the tree node count varies over the training percentage. As we increase the percentage of train set in the split from 1% towards 99%, the node count consistently increases. In case of choosing top three and using Gini Impurity, the growth pattern appeared much more scattered.