# Assessing Learners: CART-based Algorithms and Experimentation

Mohammed Fahad
mfahad7@gatech.edu

***Abstract*—This report explores the performance and behavior of several CART-based supervised learning algorithms: DTLearner, RTLearner, and BagLearner. Through a series of experiments, we investigate overfitting with DTLearner, the effect of bagging in BagLearner, and a comparison between DTLearner and RTLearner. The results are analyzed with respect to leaf size, training time, and prediction accuracy.**

## 1 INTRODUCTION

This report presents the assessment of three decision-tree-based regression learners: DTLearner (Decision Tree), RTLearner (Random Tree), and BagLearner. The goal is to analyze the impact of leaf size on overfitting, explore the effect of bagging in reducing overfitting, and quantitatively compare the performance of DTLearner and RTLearner using various metrics. Each learner has been evaluated using the Istanbul.csv dataset, which tracks returns from multiple indexes.

## 2 METHODS

The experimental setup involved three primary experiments:

- **Experiment 1:** We evaluated overfitting in DTLearner by varying the leaf size and analyzing both in-sample and out-of-sample performance.
- **Experiment 2:** We investigated how bagging, implemented in BagLearner, affects overfitting by running the learner with 10 bags, using DTLearner as the base learner.
- **Experiment 3:** We compared DTLearner and RTLearner based on training time and Mean Absolute Error (MAE) over a range of leaf sizes.

The experiments were conducted by training the learners on 60% of the dataset and testing on the remaining 40%. RMSE was used as the primary performance metric, with training time and MAE used for the DTLearner and RTLearner comparison.

## 3 DISCUSSION

### 3.1 Experiment 1: Overfitting in DTLearner

In this experiment, we measured the in-sample and out-of-sample RMSE for DTLearner as the leaf size increased from 1 to 20. As shown in Figure 1, overfitting is evident at smaller leaf sizes, where the in-sample RMSE is significantly lower than the out-of-sample RMSE. Overfitting reduces as the leaf size increases, and both RMSE values stabilize around a leaf size of 12. This indicates that larger leaf sizes help mitigate overfitting, at the cost of slightly reduced accuracy.

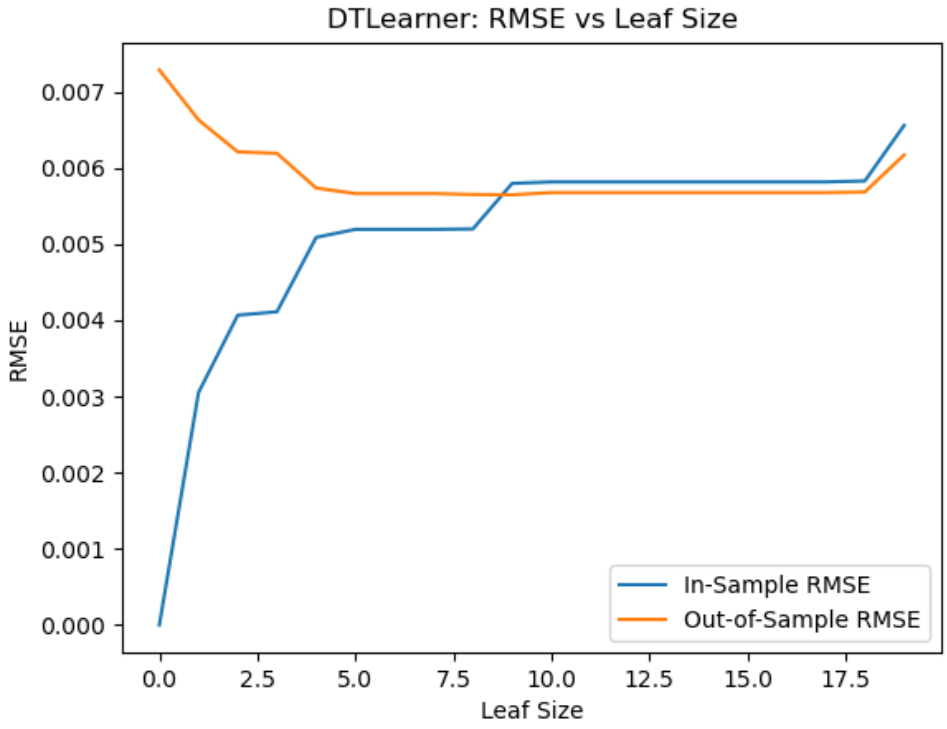


*Figure 1*—**DTLearner: RMSE vs Leaf Size** — RMSE decreases significantly in-sample with smaller leaf sizes, suggesting overfitting. As the leaf size grows, overfitting reduces and RMSE stabilizes.

### 3.2 Experiment 2: Bagging and Overfitting in BagLearner

The second experiment examines the effectiveness of bagging in reducing overfitting. Figure 2 shows that bagging improves performance, particularly at smaller leaf sizes, where the out-of-sample RMSE is reduced. This suggests that bagging

helps to stabilize the model and reduce overfitting, though it does not eliminate it entirely. Out-of-sample performance improves with bagging, making it an effective strategy in mitigating the negative effects of small leaf sizes.

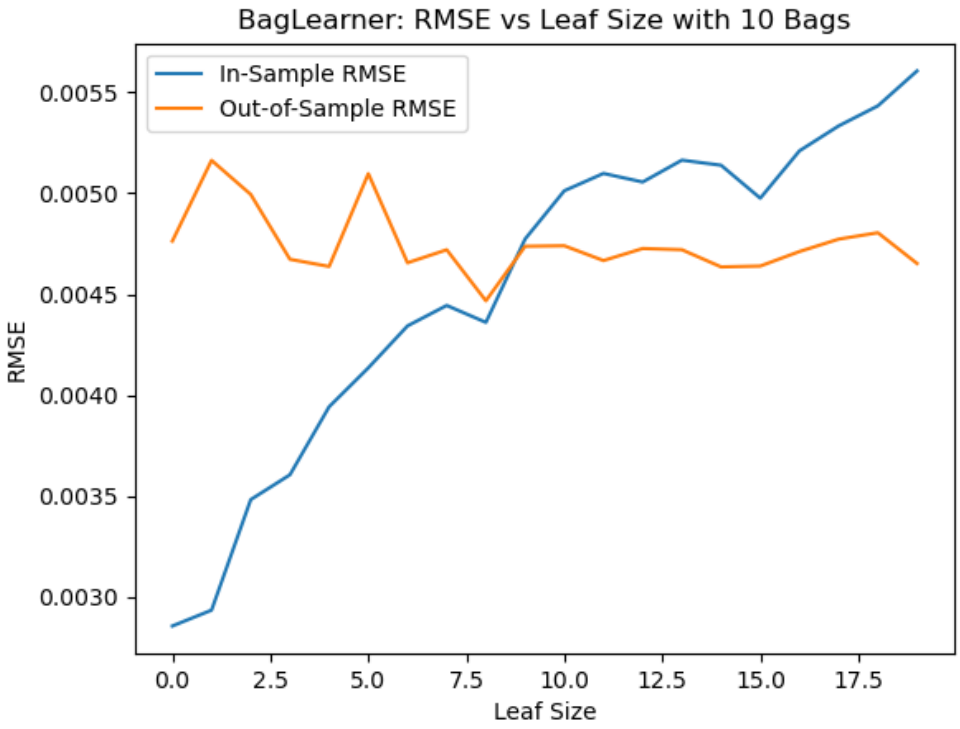


*Figure* 2—**BagLearner: RMSE vs Leaf Size with 10 Bags** — Bagging reduces overfitting, particularly for smaller leaf sizes, where the out-of-sample RMSE is significantly lower compared to DTLearner alone.

### 3.3 Experiment 3: DTLearner vs RTLearner

The final experiment compared DTLearner and RTLearner using two metrics: Mean Absolute Error (MAE) and training time. As shown in Figure 3, RTLearner consistently has a higher MAE than DTLearner across all leaf sizes, suggesting that DTLearner performs better in terms of accuracy. However, RTLearner outperforms DTLearner significantly in terms of training time, as shown in Figure 4. This makes RTLearner a faster option, particularly for larger datasets, though at the cost of accuracy.

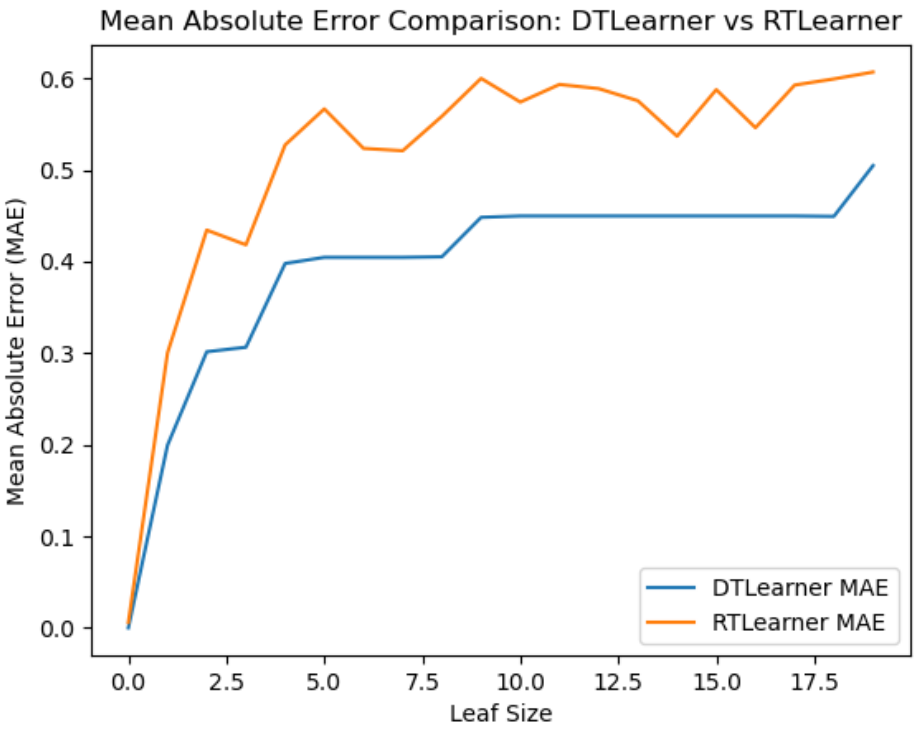


*Figure 3*—**Mean Absolute Error Comparison: DTLearner vs RTLearner** — DTLearner consistently has a lower MAE than RTLearner, indicating better prediction accuracy.

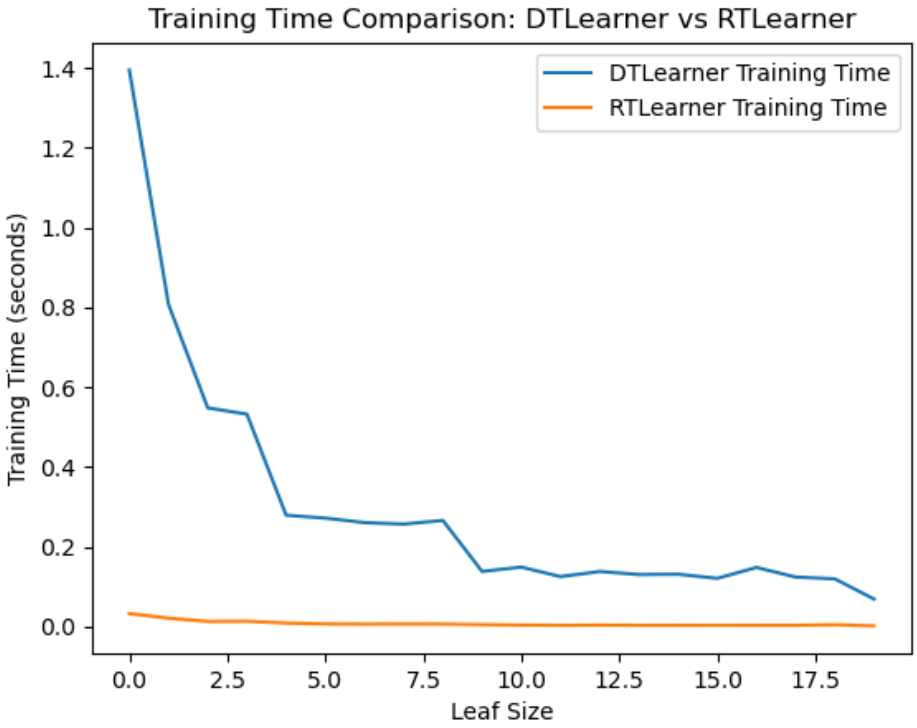


***Figure 4*—Training Time Comparison: DTLearner vs RTLearner** — RTLearner trains significantly faster than DTLearner, particularly at smaller leaf sizes, where DTLearner's training time spikes.

## 4 SUMMARY

The experiments demonstrated key differences between DTLearner, RTLearner, and BagLearner. DTLearner is more accurate but suffers from overfitting at small leaf sizes. BagLearner mitigates this by averaging the results of multiple learners, reducing overfitting and improving out-of-sample performance. RTLearner, while less accurate, trains significantly faster, making it suitable for applications where time is a critical factor. These findings suggest that DTLearner is preferable for accuracy, whereas RTLearner is more efficient for larger datasets or time-sensitive tasks.