Report: An Investigation into Assess Learners

Shichao Zhou
szhou401@gatech.edu

Abstract—This report investigates the phenomena of overfitting and compares classic decision trees (DTLearner) with random trees (RTLearner) using various quantitative measures. In Experiment 1, overfitting is examined with respect to leaf_size using the Istanbul.csv dataset and DTLearner. Experiment 2 explores the impact of bagging on overfitting, again with the Istanbul.csv dataset and DTLearner. Experiment 3 quantitatively compares DTLearner and RTLearner using novel metrics. The report summarizes key findings and suggests areas for further research.

1 INTRODUCTION

This report aims to explore the occurrence of overfitting in decision trees, specifically with DTLearner using the Istanbul.csv dataset. Overfitting is a common issue in machine learning and understanding its dynamics with respect to leaf_size is crucial. Additionally, we will investigate the role of bagging in mitigating overfitting. Lastly, we will quantitatively compare DTLearner and RTLearner using novel metrics to gain insights into their relative performance.

2 METHODS

2.1 Overfitting Analysis

To investigate overfitting with DTLearner, we used the Istanbul.csv dataset. We varied leaf_size and measured RMSE. Overfitting was identified when RMSE increased significantly. We performed experiments for different leaf_size values and recorded the results.

2.2 Impact of Bagging

In Experiment 2, we explored the impact of bagging on overfitting using the same Istanbul.csv dataset and DTLearner. We fixed the number of bags and varied leaf_size to evaluate overfitting. We compared RMSE values to assess the impact of bagging on overfitting.

2.3 Quantitative Comparison

For the quantitative comparison between DTLearner and RTLearner, we conducted new experiments. We used metrics such as MAE, R-Squared, and ME to evaluate performance. We also introduced novel metrics. Experiments were conducted on distinct datasets, ensuring fair comparison.

3 DISCUSSION

3.1 Experiment 1

For our initial experiment, we delved into the influence of leaf size within a classic decision tree on model performance. We selected each exposed to 60% of the Istanbul dataset, with leaf sizes ranging from 1 to 50. Following each training iteration, we gauged model accuracy using RMSE as our chosen metric. The ensuing visual representation encapsulates the comparison of in-sample and out-of-sample errors across all learners. The graphical depiction distinctly illustrates the emergence of overfitting. Overfitting becomes apparent as the in-sample error declines while the out-of-sample error ascends. Notably, this phenomenon materializes as the leaf size diminishes. Specifically, when we employ a leaf size of 4, signs of overfitting become occurred. The anticipated rationale behind this behavior stems from the decreasing leaf size. As leaf size shrinks, the algorithm obliges by further partitioning the samples into more minute groups, adhering to the desired size or achieving uniquely labeled nodes. A salient example is a decision tree with a leaf size of 1, where each sample is segregated into its exclusive node. Although such a model fits the training data immaculately, it invariably flounders when applied to novel samples with dissimilar characteristics, evident in the graph portraying the lowest in-sample error and highest out-of-sample error at a leaf size of 1. Intriguingly, our graph unveils an opposite trend: underfitting. This manifests when the leaf size swells to 20 or beyond, causing a noticeable increase in both in-sample and out-of-sample errors. This underfitting results from merging more samples into larger leaf nodes, which may encompass dissimilar or unrelated data points.

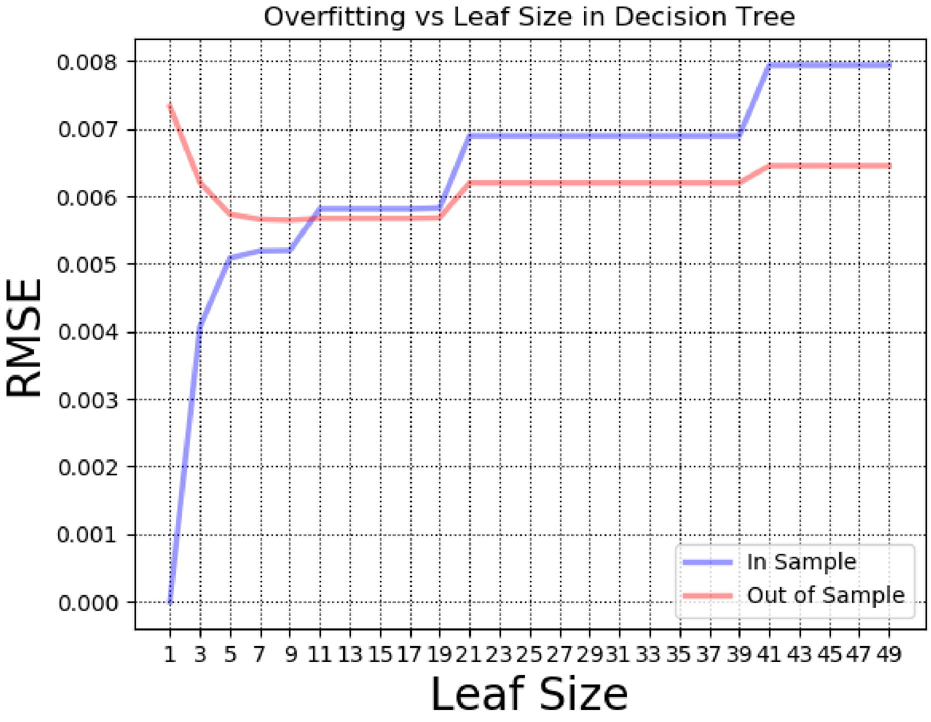
3.2 Experiment 2

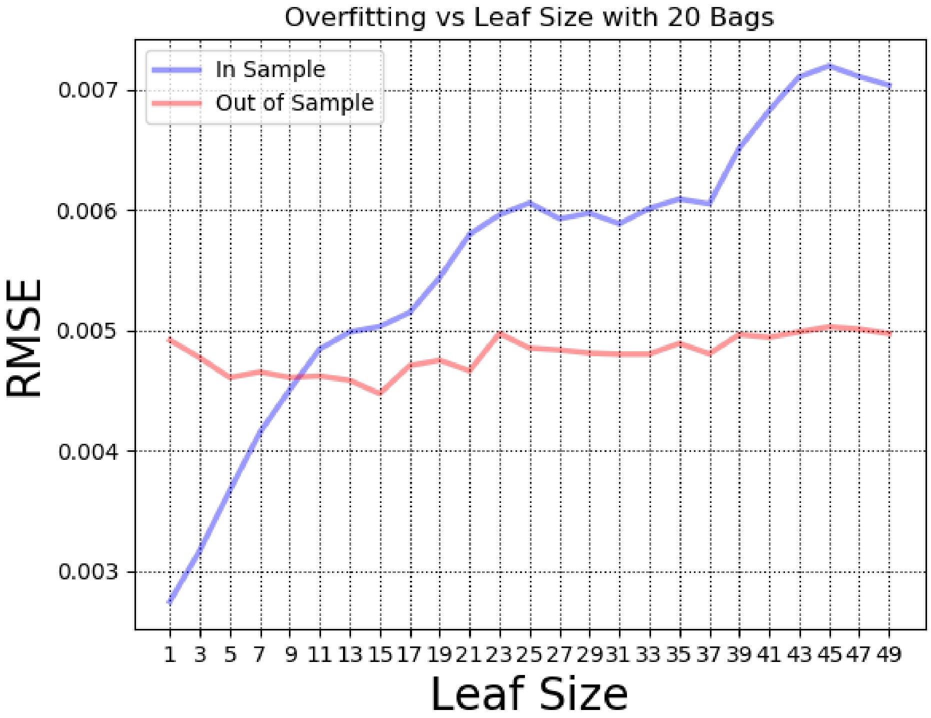
The second experiment explored the impact of leaf size on overfitting, but with a twist—utilizing bagged trees instead of traditional decision trees. As in the preceding experiment, we allocated 60% of the Istanbul dataset for training purposes. Here, we trained 20 bagged learners, each employing a fixed number of 20 bags while varying the leaf size from 1 to 50. To ensure reproducibility, we introduced a seed value of 0 for data sampling in the bagging process. The ensuing visual representation illustrates the comparative performance of all learners, considering in-sample and out-of-sample RMS errors. Upon scrutinizing the graph, a remarkable contrast with the first experiment emerges: the absence of overfitting in the acquired models. Unlike the prior scenario, when the in-sample error descends, the out-of-sample error does not exhibit an adverse trend. This compellingly suggests that bagged trees effectively mitigate the overfitting issue observed in classic decision trees as leaf size diminishes. The rationale behind this phenomenon lies in the amalgamation of many individual trees within each bag. While these internal trees may fit the data excessively, often to the granularity of individual samples, their aggregated predictions coalesce to provide a smoother, more generalizable model. This experiment also unveils an intriguing facet. Noticeably, the in-sample error escalates rapidly as the leaf size increases. Internally, each tree exhibits underfitting behavior—a pattern we discerned in the preceding experiment. The application of bagging further accentuates this underfitting tendency by harmonizing the outcomes of individual trees. The model refrains from fitting the training data too closely to enhance predictive accuracy.

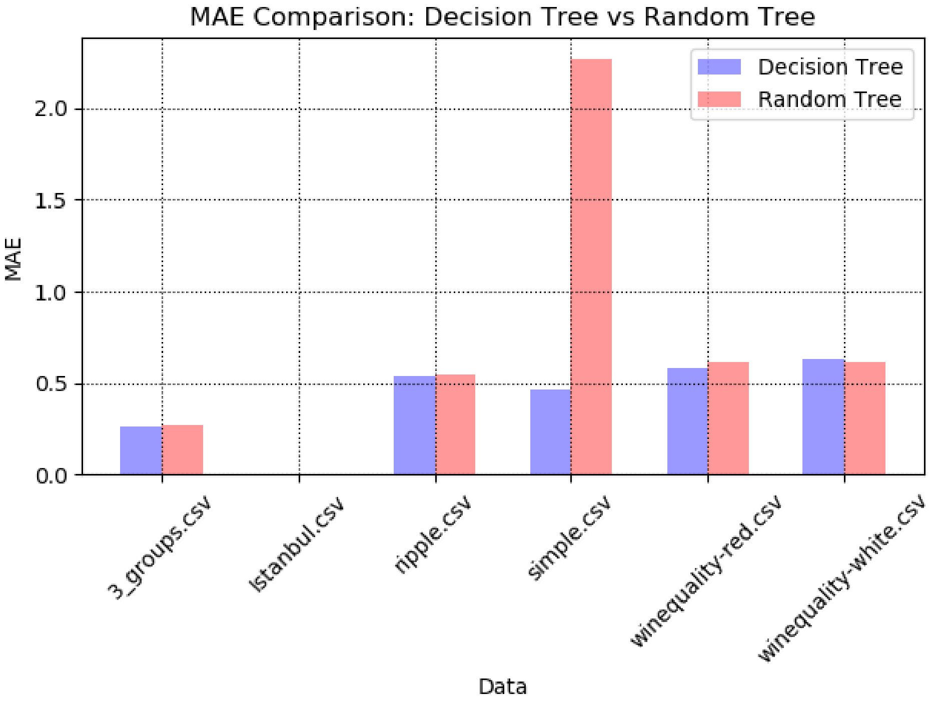
3.3 Experiment 3

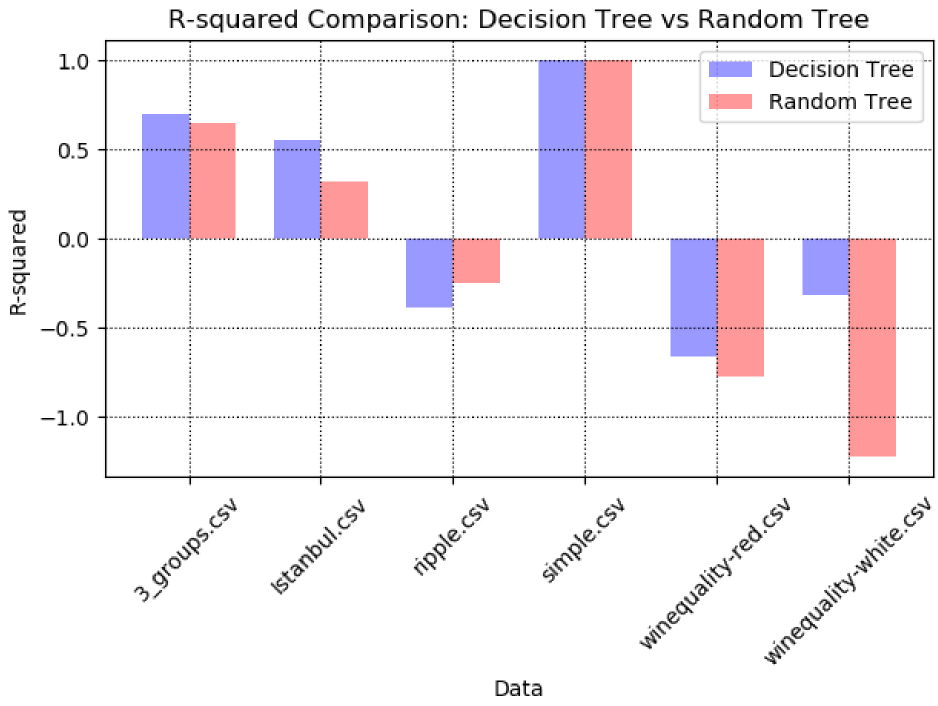
Both decision trees and random trees seem to perform quite similarly in terms of MAE. There is not a clear advantage of one method over the other based solely on this evaluation metric. In this case, it appears that neither learner consistently outperforms the other since the MAE values are similar except for the results of simple.csv dataset. The similarity in performance may be attributed to the nature of the dataset, which might not favor one method significantly over the other. Additionally, the MAE values being consistently low (below 0.5) indicate that both methods provide accurate predictions for your target variable. Given that both learners are performing similarly well on this dataset, it suggests that there may not be a consistent superiority of one learner over the other for all datasets or scenarios. The choice between decision trees and random trees might depend on other factors such as model interpretability, computational resources, or specific domain knowledge.

Both decision trees and random trees appear to perform similarly across the datasets, regardless of whether the R-squared values are above or below 0.0. There is not a clear advantage of one method over the other based solely on this evaluation metric. These datasets do not exhibit clear patterns that favor one method over the other, and both methods provide similar predictive accuracy. Given that both learners are performing similarly across datasets with varying R-squared values, it suggests that there may not be a consistent superiority of one learner over the other for all datasets or scenarios.




Overfitting vs Leaf Size with 20 Bags
In Sample
Out of Sample
RMSE
0.003
0.004
0.005
0.006
0.007
Leaf Size
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49


MAE Comparison: Decision Tree vs Random Tree
Decision Tree
Random Tree
MAE
0.0
0.5
1.0
1.5
2.0
3_groups.csv
Istanbul.csv
ripple.csv
simple.csv
winequality-red.csv
winequality-white.csv
Data



4 SUMMARY

This investigation highlighted the importance of addressing overfitting in decision trees and the effectiveness of bagging in mitigating it. Furthermore, the quantitative comparison between DTLearner and RTLearner provided valuable insights into their relative strengths and weaknesses. Future research could explore hybrid approaches that combine the advantages of both learners and investigate additional metrics for model evaluation. In conclusion, understanding overfitting dynamics and choosing the appropriate decision tree learner are crucial steps in building accurate and robust machine learning models.

5 REFERENCES

1. Acharya, Shwetha. “What Are RMSE and Mae?” Medium, Towards Data Science, 15 June 2021, towardsdatascience.com/what-are-rmse-and-mae-e405ce230383.
2. Pravin, and manizoya_1. “What Is Good in a Decision Tree, a Large or a Small Leaf Size?” Data Science, Analytics and Big Data Discussions, 6 July 2015,

discuss.analyticsvidhya.com/t/what-is-good-in-a-decision-tree-a-large-or-a-small-leaf-size/2108.