EVALUATING ENSEMBLE METHODS ON THE IRIS DATASET

ZIKUN ZHENG

University of Washington, Seattle, WA
zzheng9@uw.edu

Abstract. This report presents a comprehensive analysis of various ensemble machine learning models, including CART, Bagging, Random Forest (RF), and Deep Forest (DF), applied to the classic Iris dataset. Conducted as part of the SPA DRP program, this study aims to elucidate the impact of ensemble methods on model accuracy and to explore the nuances of model performance through cross-validation and parameter variation.

1. Introduction

1.1. **Background.** The Iris dataset, a fundamental dataset in the field of machine learning and statistics, comprises 150 samples from three species of Iris flowers. Each sample is described by four features: sepal length, sepal width, petal length, and petal width. This study embarks on an exploratory journey through the application of machine learning models, starting with the Classification and Regression Tree (CART) model and extending to more sophisticated ensemble methods such as Bagging, Random Forest (RF), and Deep Forest (DF). These models represent a progression in the complexity and capability of algorithms to handle data variability and improve prediction accuracy.

1.2. **Motivation.** The pursuit of this investigation is driven by the hypothesis that ensemble methods can significantly enhance model performance over individual classifiers like CART by reducing variance, bias, or both. Furthermore, the study seeks to understand the dynamics of Bagging and Boosting within RF and DF models, respectively, and how manipulating parameters such as the number of trees and tree depth affects accuracy. These insights are crucial for both theoretical understanding and practical applications in data science.

1.3. **Understanding the Models.**
CART (Classification and Regression Trees): The CART model is a type of decision tree that can be used for both classification and regression tasks. It operates by recursively partitioning the data into subsets, which maximizes the homogeneity of the resultant subsets. CART models are simple to understand and interpret but can be prone to overfitting, especially in cases where the tree is allowed to grow complex without constraints.
Bagging: Bagging is an ensemble technique designed to improve the stability and accuracy of machine learning algorithms. It reduces variance and helps to avoid overfitting. Bagging involves creating multiple copies of the original training dataset using bootstrap sampling, training a model (often decision trees) on each copy, and then combining the models' outputs (e.g., by averaging) to make predictions.

Date: March 14, 2024.

Random Forest (RF):. RF is an extension of Bagging that incorporates decision trees as base learners. It introduces further randomness when growing trees; instead of searching for the most significant feature when splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity among the trees in the forest, which generally leads to a more robust overall model.

Deep Forest (DF):. The Deep Forest model is an innovative approach that extends the concept of ensemble learning. The Deep Forest model, with its intricate architecture of layered ensembles, mirrors the depth and complexity found in neural networks (NNs). It can capture complex data patterns without relying on the traditional neural network framework. It builds on the foundation of Random Forests and Bagging by layering these models to create a 'forest of forests'. The depth of the DF model is controlled by the number of layers of RF and Bagging models, while the width is determined by the number of trees within each Random Forest or Bagging model. The unique structure of the DF model allows it to effectively learn complex patterns and interactions in the data.

Difference and Intuition inside: The progression from CART to Deep Forest models encapsulates a journey from simplicity to complexity, aimed at addressing the challenges of overfitting and improving prediction accuracy across diverse datasets. CART's straightforward approach offers ease of interpretation but at the risk of overfitting, especially with complex data. To mitigate this, Bagging and Random Forest introduce techniques like bootstrap sampling and random feature selection, respectively, enhancing model stability and performance by aggregating multiple decision trees. Deep Forest extends this concept further by layering ensembles of trees, creating a 'forest of forests' that can capture intricate data patterns more effectively. This evolution reflects a trade-off between interpretability and accuracy, with simpler models like CART being more interpretable but potentially less accurate on complex tasks, while more complex models like Deep Forest offer higher accuracy at the cost of interpretability. The choice among these models typically hinges on the specific requirements of the task, balancing the need for accuracy, interpretability, and the ability to handle data complexity.

2. Methodology

2.1. **Data Preparation and Model Evaluation.** The Iris dataset was partitioned into training and testing sets, with 80% of the data allocated for training and the remaining 20% for testing. This study utilized R and several comprehensive packages for machine learning to prepare the data and implement the models.

2.2. **Model Implementation and Parameter Tuning.** Each model was applied to the dataset with a focus on understanding how different configurations impact performance. The CART model served as the baseline, with ensemble models like Bagging, RF, and DF introducing additional layers of complexity and prediction refinement.

Cross-Validation Technique: To ensure a robust evaluation, a 10-fold cross-validation technique was employed, offering a balanced view of model performance across different partitions of the data.

3. Results and Discussion

3.1. **Model Performance Comparison.** The comparison of model accuracies revealed insightful trends:

3.2. **Effect of Tree Depth and Number of Trees.** An in-depth analysis was conducted to understand the impact of varying the number of trees and their depth within the ensemble models, especially focusing on the RF and DF models. This analysis, illustrated in Figures 1 and 2, highlights the nuanced behavior of these models in response to changes in their configurations.

Model	Cross-Validation Accuracy
DeepForest	0.9333
RandomForest	0.9437
Bagging	0.9474
CART	0.3333

TABLE 1. Model performance comparison on the Iris dataset.

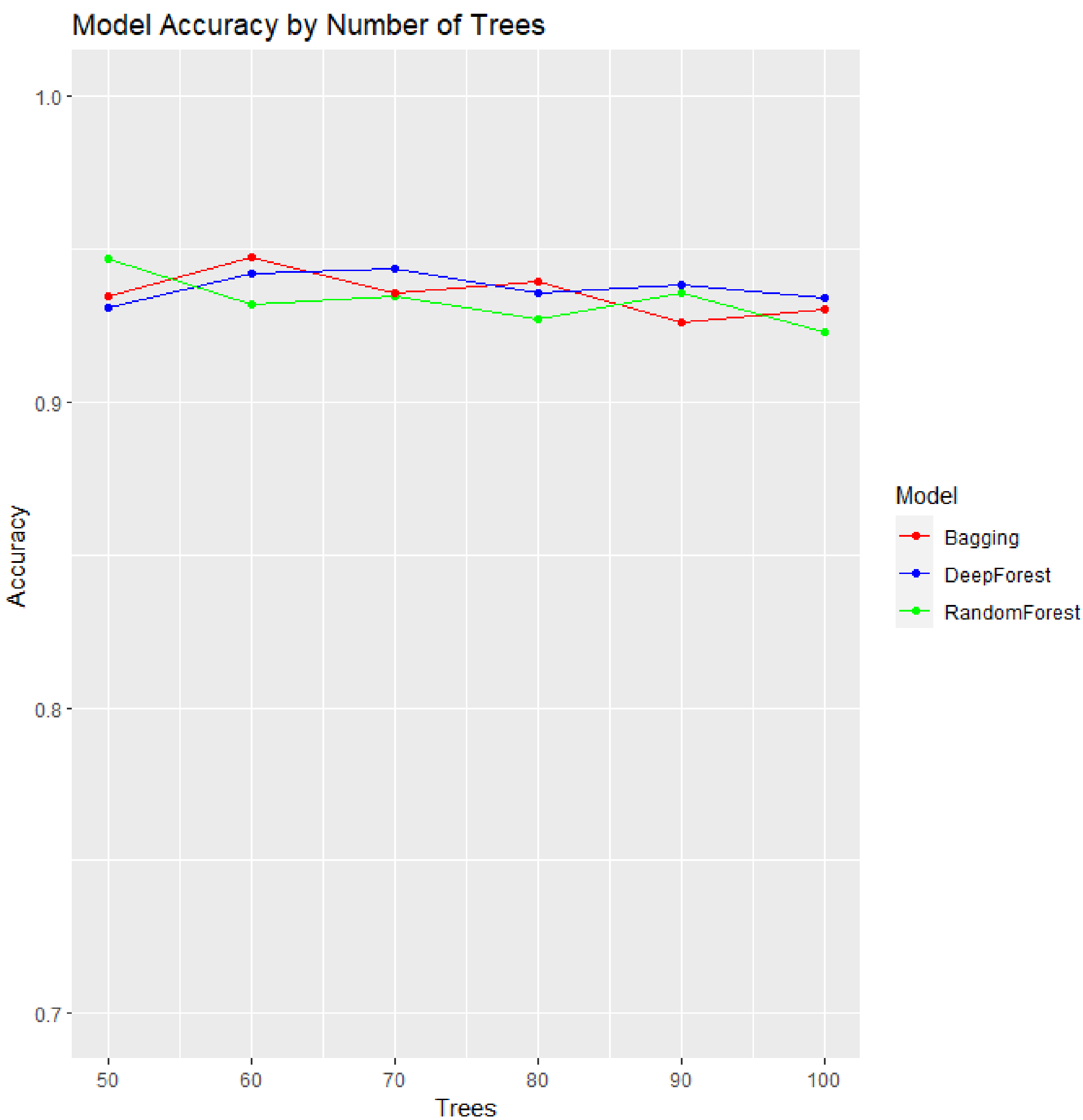

FIGURE 1. Impact of the number of trees on model accuracy.

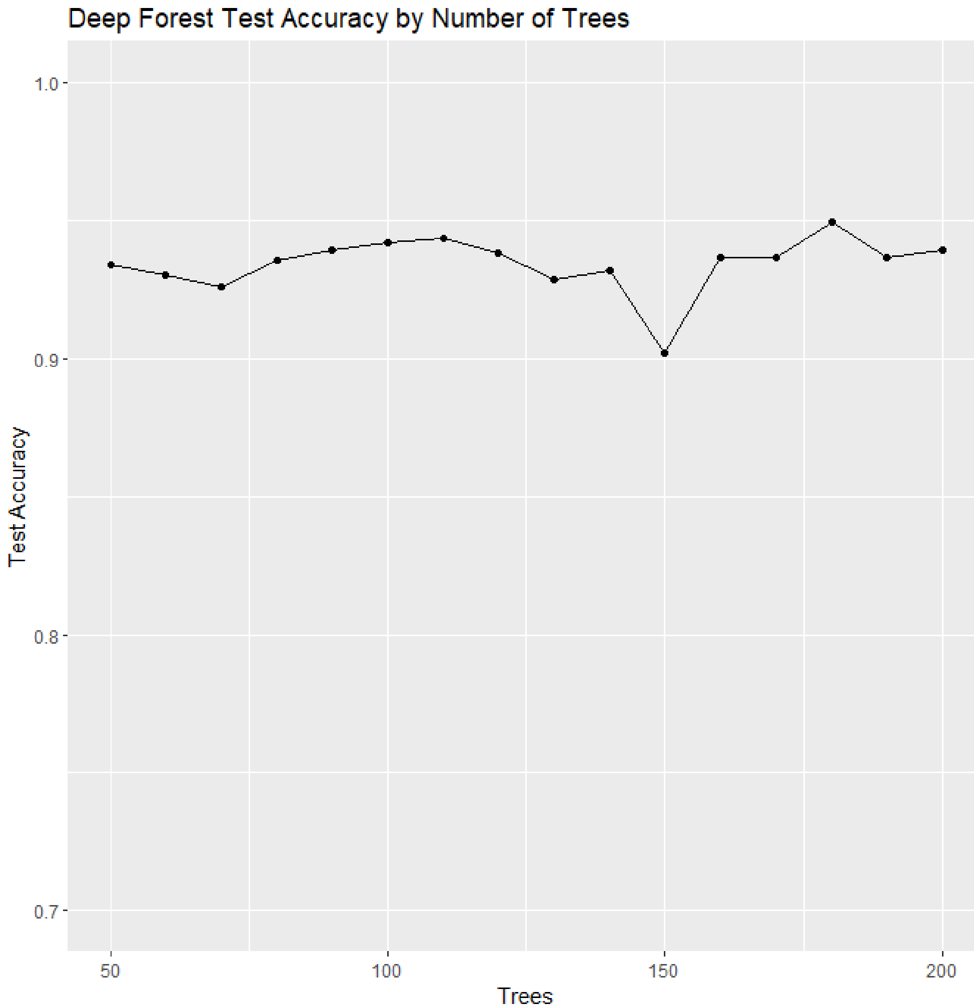

FIGURE 2. Detailed performance analysis of the Deep Forest model with varying tree counts.

4. Conclusion

This study provides a comprehensive analysis of ensemble methods applied to the Iris dataset, revealing the superior performance of ensemble models over the singular CART model. The findings underscore the effectiveness of Bagging and Random Forests in reducing prediction error and enhancing model robustness. The detailed exploration into the number of trees and their depth within the RF and DF models presents a critical insight: there exists an optimal configuration that balances model complexity with prediction accuracy, which is paramount for practical applications.

Future directions for this research include applying these insights to more complex datasets and exploring additional ensemble methods. The SPA DRP program has provided a fertile ground for such exploratory studies, offering valuable lessons on the intricacies of machine learning models and their practical implications.