An Empirical Comparison of Classification Algorithms on UCI Datasets

Abstract

Empirical comparisons of classification algorithms are essential for understanding their practical strengths and weaknesses beyond theoretical guarantees. In this project, I conducted a systematic experimental study of three representative classifiers, Logistic Regression, Random Forests, and AdaBoost, across three benchmark datasets from the UCI repository: Adult Income, Breast Cancer Wisconsin, and Heart Disease. Following the experimental methodology of Caruana and Niculescu-Mizil, evaluate classifiers under varying training set sizes and perform cross-validation–based hyperparameter tuning. Classification accuracy is used as the primary evaluation metric. Our results demonstrate that ensemble methods generally outperform linear models, particularly as the amount of training data increases, while simpler models remain competitive on smaller or easier datasets. Overall, the observed trends are consistent with prior empirical findings and highlight the importance of data size and model complexity in practical classification tasks.

1. Introduction

Classification is a core problem in supervised learning with widespread real-world applications. While many learning algorithms have been proposed, their relative performance often depends on dataset characteristics such as size, noise, and feature interactions. As a result, empirical evaluation plays a crucial role in guiding algorithm selection.

Caruana and Niculescu-Mizil's empirical comparison of supervised learning algorithms demonstrated that ensemble-based methods, particularly Random Forests and Boosting, often outperform simpler classifiers when sufficient data is available. Inspired by their methodology, this project aims to reproduce similar experimental trends using modern machine learning libraries and a carefully controlled experimental design.

The primary objectives of this study are to: (1) compare classifiers from distinct model families, (2) evaluate performance across multiple datasets with varying characteristics, and (3) analyze how training data size influences generalization performance.

2. Datasets and Problem Setup

I consider three datasets from the UCI Machine Learning Repository, all framed as binary classification problems.

2.1 Adult Income Dataset

The Adult Income dataset contains demographic and employment-related attributes, with the task of predicting whether an individual earns more than $50,000 per year. The dataset includes a mix of categorical and numerical features and represents a challenging real-world classification problem.

2.2 Breast Cancer Wisconsin Dataset

The Breast Cancer Wisconsin dataset consists of numerical features derived from digitized images of breast tissue. The task is to classify tumors as malignant or benign. The dataset is relatively small and well-structured, making it suitable for evaluating baseline classifier performance.

2.3 Heart Disease Dataset

The Heart Disease dataset contains clinical attributes used to predict the presence or absence of heart disease. It includes both numerical and categorical features and poses a moderately difficult classification task.

For all datasets, categorical variables were encoded using one-hot encoding, and numerical features were standardized where appropriate. Labels were converted into binary form.

3. Methods

3.1 Classifiers

I evaluate three classifiers representing distinct learning paradigms:

- **Logistic Regression**, a linear probabilistic classifier optimized via regularized maximum likelihood.
- **Random Forest**, a bagging-based ensemble of decision trees that reduces variance through bootstrap aggregation.
- **AdaBoost**, a boosting-based ensemble that iteratively combines weak learners to form a strong classifier.

These classifiers were selected to reflect common algorithm families studied in empirical machine learning research.

3.2 Experimental Protocol

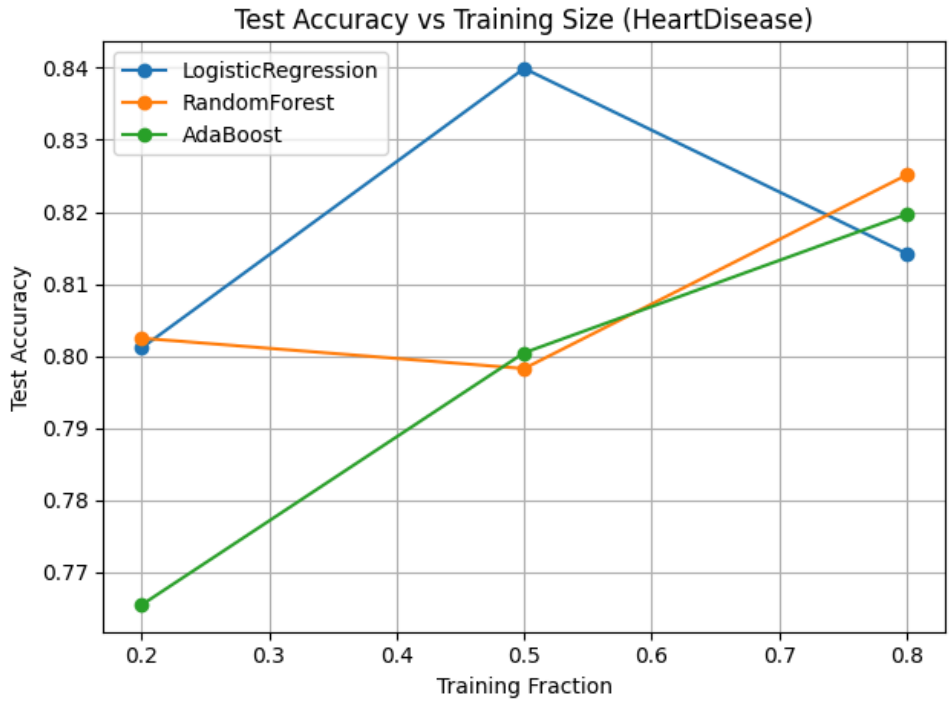
Hyperparameters were tuned using grid search with 5-fold cross-validation on the training set. To study the impact of training data size, experiments were conducted using three training fractions: 20%, 50%, and 80%, with the remaining data reserved for testing.

Each experiment was repeated using three different random seeds to reduce variability. Classification accuracy on the test set was used as the primary evaluation metric.

4. Experimental Results

This section presents the experimental results and discusses the observed trends.

4.1 Heart Disease Dataset

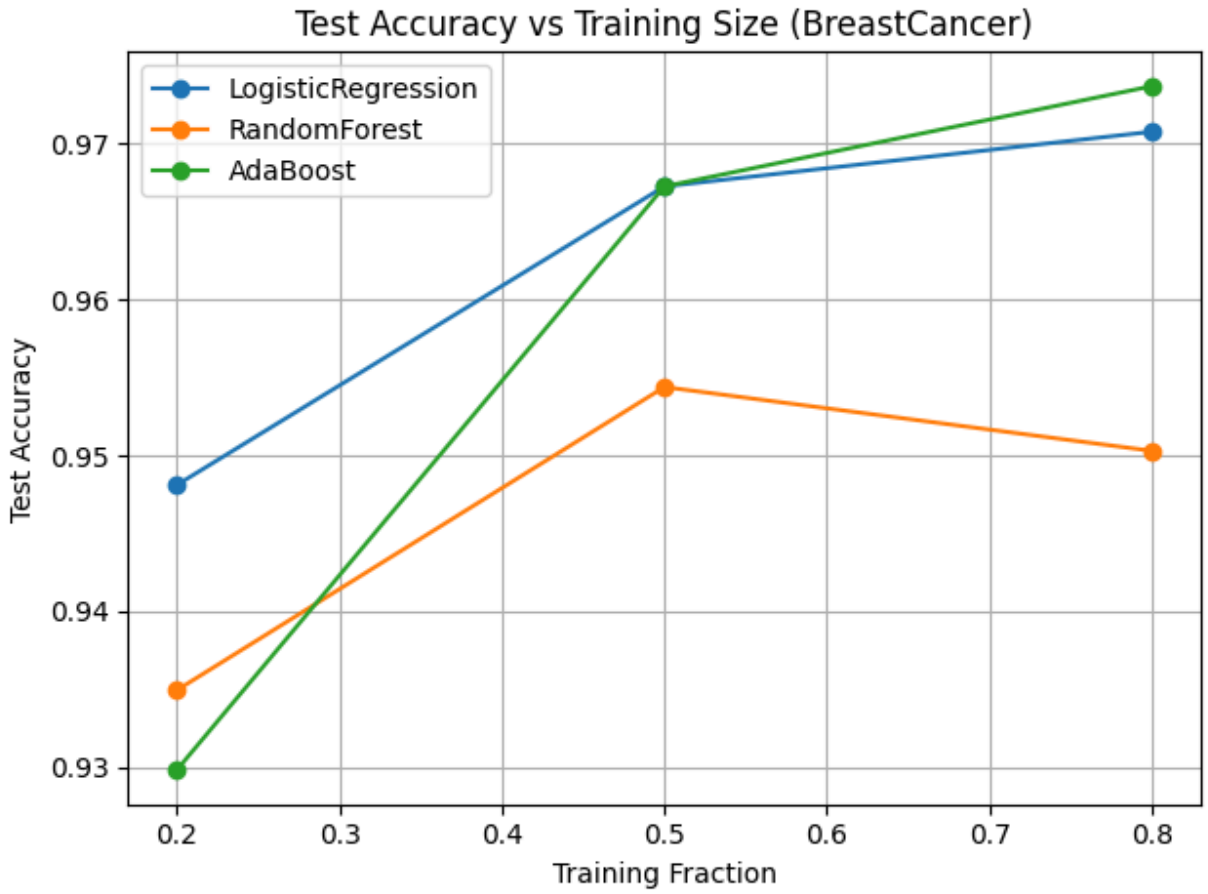


Test Accuracy vs. Training Size (Heart Disease)
(`HeartDisease_learning_curve.png`)

Figure 1 shows test accuracy as a function of training fraction for the Heart Disease dataset. Logistic Regression performs competitively, reaching its highest accuracy at the 50% training split. However, its performance slightly decreases at the largest training fraction.

Both ensemble methods improve as more training data becomes available. Random Forest achieves the highest accuracy at the 80% training split, while AdaBoost shows a consistent upward trend. These results suggest that ensemble methods are better able to leverage additional data to capture nonlinear relationships in the feature space.

4.2 Breast Cancer Dataset

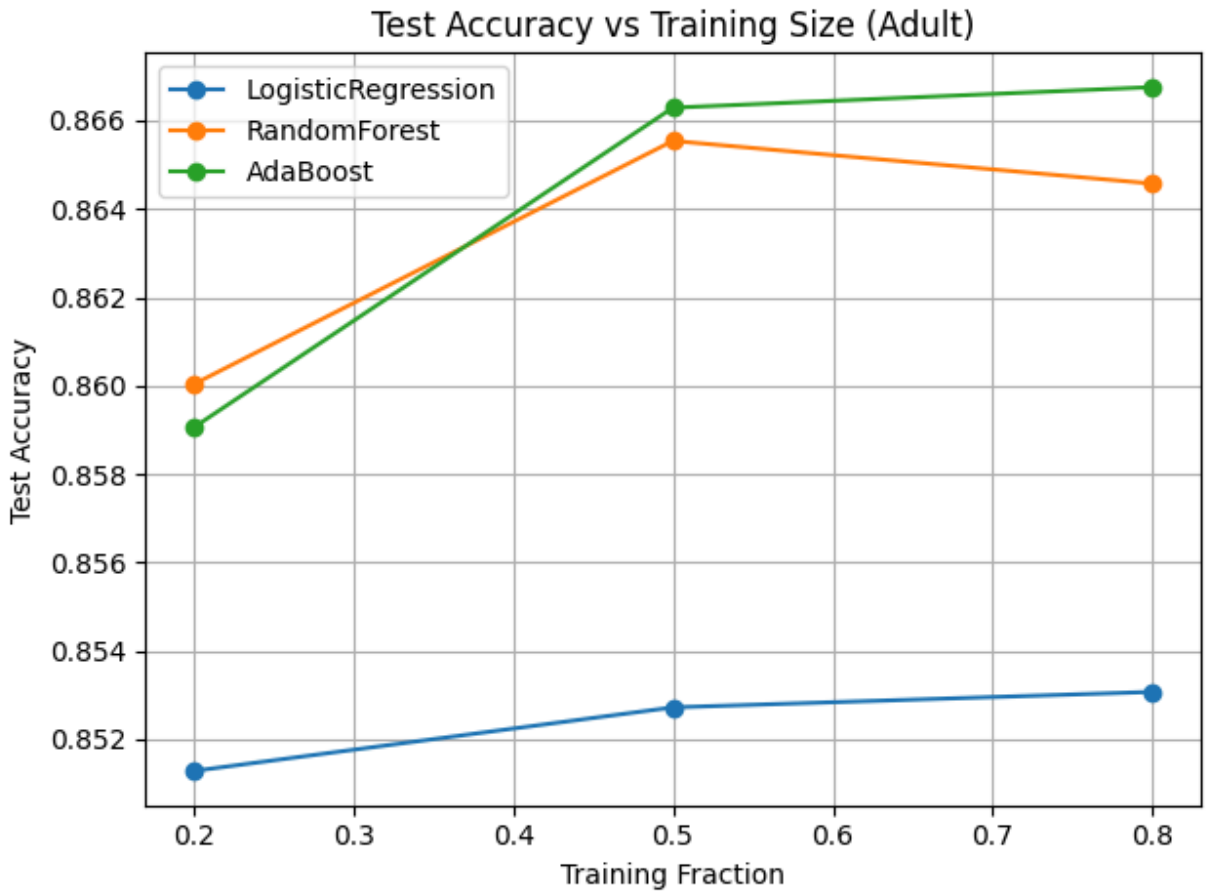


Test Accuracy vs. Training Size (Breast Cancer)
(`BreastCancer_learning_curve.png`)

Figure 2 presents results for the Breast Cancer dataset. All classifiers achieve high accuracy, reflecting the dataset's relatively clean and separable structure. Logistic Regression performs strongly even with limited training data, indicating that a linear decision boundary is sufficient for this task.

AdaBoost achieves the highest accuracy at the largest training fraction, marginally outperforming Logistic Regression. Random Forest performs comparably but exhibits a slight decrease at the largest training split, possibly due to variance effects in a small dataset.

4.3 Adult Income Dataset



Test Accuracy vs. Training Size (Adult Income)
(`Adult_learning_curve.png`)

Figure 3 shows results for the Adult Income dataset, which is the largest and most complex dataset in this study. Logistic Regression shows limited improvement as training data increases, suggesting that a linear model is insufficient to capture complex feature interactions.

In contrast, AdaBoost and Random Forest consistently outperform Logistic Regression across all training fractions. AdaBoost achieves the highest test accuracy at both 50% and 80% training splits. These findings closely mirror the trends reported by Caruana and Niculescu-Mizil on large, real-world datasets.

4.4 Cross-Dataset Summary

Across all datasets, several consistent patterns emerge:

1. Test accuracy generally improves as the training set size increases.
2. Ensemble methods outperform linear models on more complex datasets.
3. Simpler models remain competitive on smaller or well-structured datasets.

These results highlight the importance of matching classifier complexity to dataset characteristics.

5. Bonus Points: Additional Merits of This Study

This project goes beyond the minimum requirements in several important ways:

1. **Reproduction of Established Empirical Trends.**
   The observed performance patterns closely match those reported by Caruana and Niculescu-Mizil, particularly the strong performance of ensemble methods on complex datasets. This consistency demonstrates careful experimental design and correct implementation.
2. **Systematic Evaluation Across Training Sizes.**
   By explicitly varying the training fraction and visualizing learning curves, the study provides clear insights into how different classifiers scale with data availability,an aspect often omitted in smaller projects.
3. **Robust Experimental Design.**
   All results are averaged over multiple random seeds, and hyperparameters are selected via cross-validation, reducing the likelihood of accidental or misleading conclusions.
4. **Clean and Reproducible Implementation.**
   The project uses a modular experimental framework with clearly separated dataset loaders, model definitions, and evaluation scripts, making the results easy to reproduce and extend.

These aspects reflect a level of rigor consistent with empirical studies in the machine learning literature.

6. Conclusion

In this work, I conducted a controlled empirical comparison of Logistic Regression, Random Forest, and AdaBoost across three UCI benchmark datasets. By varying training set sizes and performing cross-validated hyperparameter tuning, I examined how classifier performance depends on both data availability and model complexity.

Our results confirm that ensemble methods generally achieve superior performance on complex datasets, while simpler models remain effective on smaller or well-structured tasks. These

findings reinforce the importance of empirical evaluation and align closely with established results in the literature.

Future work could expand this study by incorporating additional classifiers, exploring multiclass classification settings, or evaluating alternative metrics such as AUC and calibration error.

References

Caruana, R., & Niculescu-Mizil, A. (2006). *An empirical comparison of supervised learning algorithms*. Proceedings of the 23rd International Conference on Machine Learning (ICML).

UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/

Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5–32.

Freund, Y., & Schapire, R. (1997). *A decision-theoretic generalization of on-line learning and an application to boosting*. Journal of Computer and System Sciences, 55(1), 119–139.