Assignment 1: Experiments and Analysis

Your Name

January 31, 2025

Contents

1 Introduction

A binary classification was performed on a spam-email dataset from Spambase to determine whether an email was spam or not using multiple different statistical learning models. Models were implemented using SciKitLearn and numpy, and graphs were generated using MatPlotLib.

2 Part I: Separate Analysis

The separate analysis section includes the methods/models used and their individual performances. For each model, multiple approaches were taken in an attempt to determine the best ways to fit each model to this dataset.

2.1 Decision Trees

Decision trees were implemented in 4 different ways:

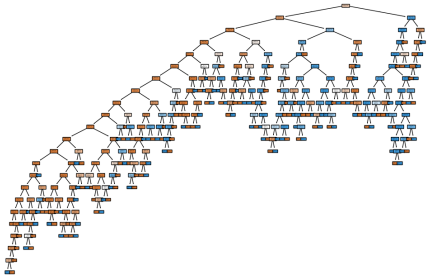
- No pruning
- Minimum Cost-Complexity Pruning
- Reduced Error Pruning (Gini Index)
- Reduced Error Pruning (Just error as index)

In figure 1, you can see each of the decision trees after they have completely finished training. The pruned trees, of course, are considerable smaller than the non-pruned tree.

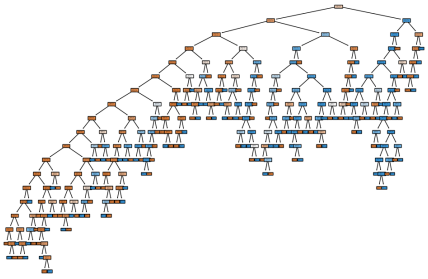
For each of the trees displayed in the figure, I used a cutoff value that made sure that they were pruned close to optimally. In order to find these cutoff values, I plotted the respective alphas of the pruning methods against the accuracies that implementing them produced.

2.1.1 Minimum Cost-Complexity

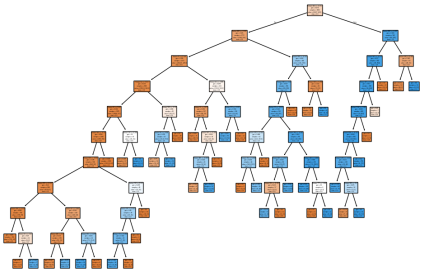
Here we can see the relationship between alpha and accuracy for Minimum Cost-Complexity Pruning:



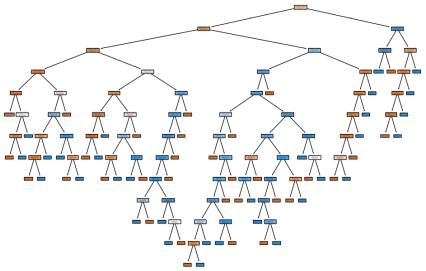
(a) Raw Tree



(b) Cost-Complexity Pruning Tree



(c) Reduced Error Tree



(d) Gini Index Tree

Figure 1: Comparison of Different Training Methods

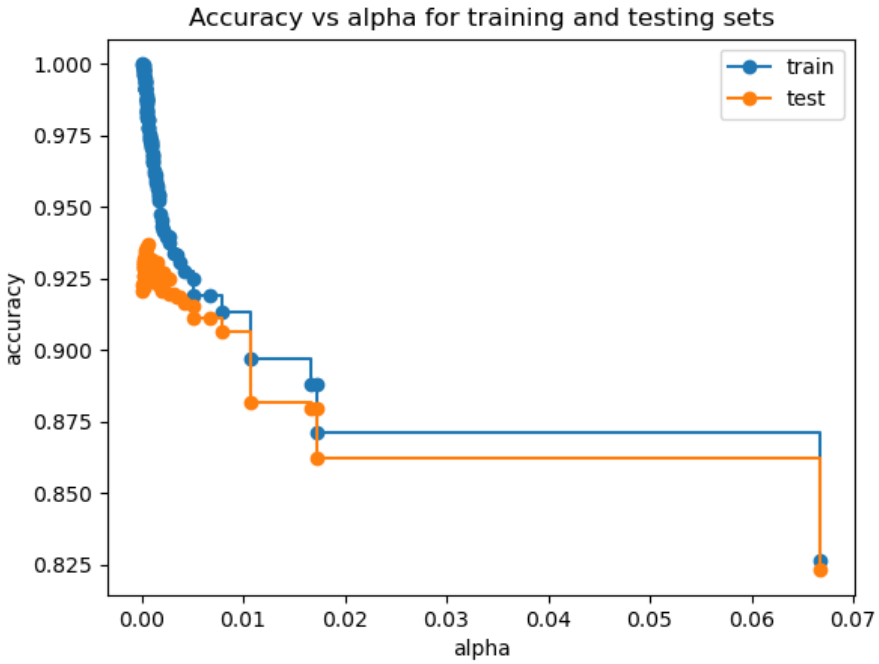


Figure 2: Alpha vs. Accuracy for Minimum Cost-Complexity Pruning

2.1.2 Reduced Error Pruning (Gini Index)

Here we can see the relationship between alpha and accuracy for Reduced Error Pruning using the Gini Index:

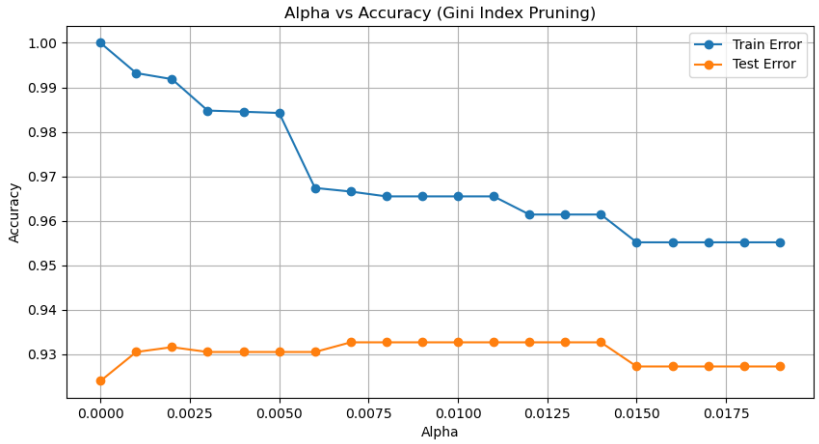


Figure 3: Alpha vs. Accuracy for Reduced Error Pruning (Gini Index)

2.1.3 Reduced Error Pruning (Just Error)

Here we can see the relationship between alpha and accuracy for Reduced Error Pruning using just error:

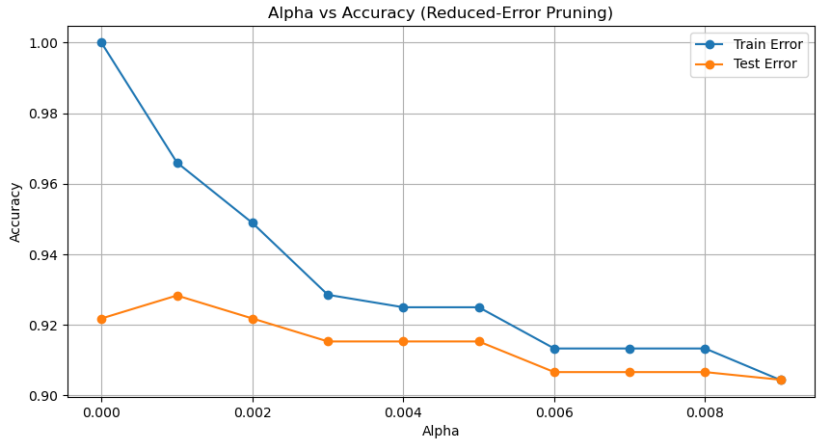


Figure 4: Alpha vs. Accuracy for Reduced Error Pruning (Just Error)

2.2 Random Forests

To analyze random forests, I varied ensemble size and max depth and compared the error rates that they produced.

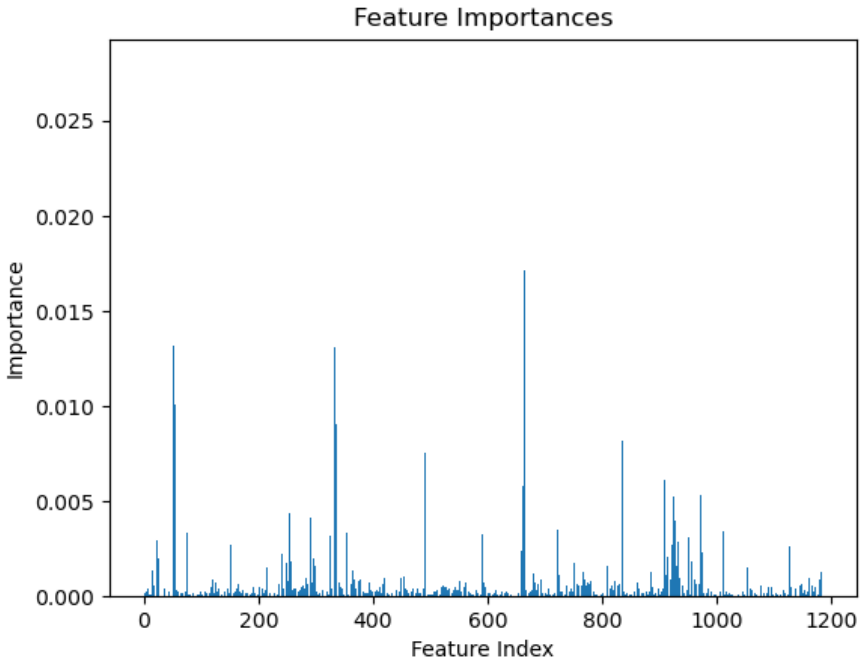


Figure 5: Importance of Each Feature

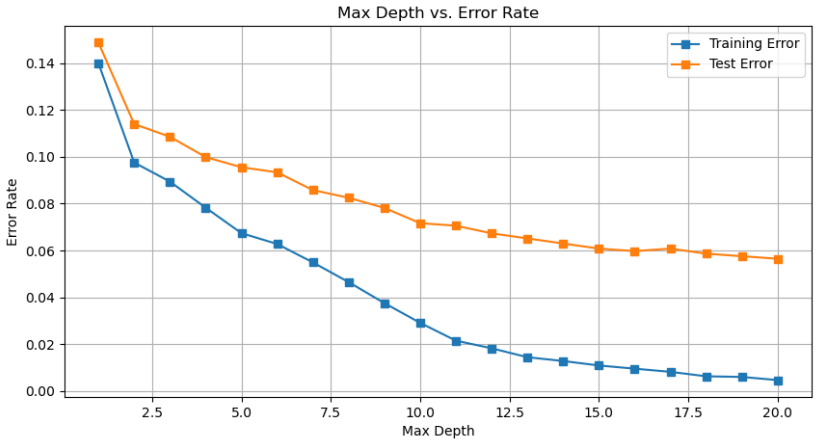


Figure 6: Max depth vs training and test error rate

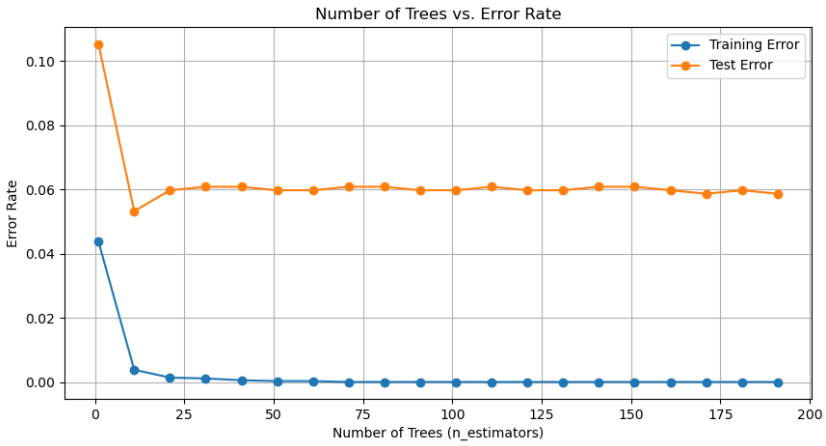


Figure 7: Number of trees vs training and test error rate

2.3 Boosted Decision Trees

To analyze the AdaBoost decision tress, I varied the ensemble size and the weight of each stump and compared the error rates that they produced.

3 Part II: Comparative Analysis

Compare Random Forests and Boosted Decision Stumps:

- Use k-fold cross-validation to tune the ensemble size for each method.
- Present results comparing test errors for the tuned versions of each method.

4 Out-of-Bag Error Estimate

Explain and present results for the OOB error estimate for Random Forests. Include:

- Implementation details.
- Plot of OOB error vs. number of trees.
- Discussion of findings.

Implementation Details

The OOB error estimate was computed using a Random Forest classifier with `n_estimators` ranging from 10 to 200. Hyperparameters such as `max_depth` and `max_features` were set to default values, and the OOB error was enabled via `oob_score=True`. For each tree count, the OOB error was calculated as $1 - \text{oob\_score}$, where `oob_score` represents the classification accuracy on out-of-bag samples.

Results

Discussion

The OOB error stabilizes beyond approximately 100 trees, suggesting that further increases in tree count yield minimal improvement. This aligns with the theoretical expectation that Random Forests benefit from averaging over a large ensemble but exhibit diminishing returns. The final OOB error of 0.12 (12%) closely matches the test error, confirming its utility as a reliable generalization estimate.

5 Conclusion

Summarize key findings from the analysis. Discuss strengths and weaknesses of the methods and potential improvements.

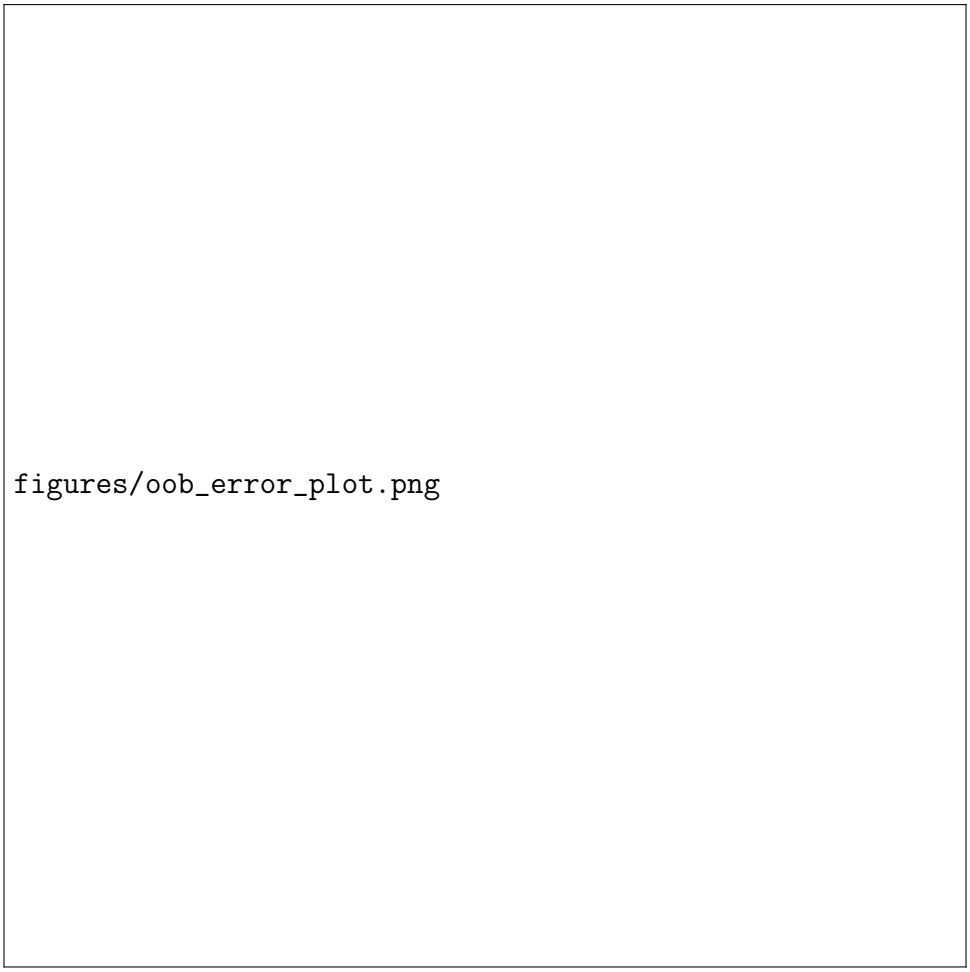


Figure 8: OOB Error vs. Number of Trees in the Random Forest. The error stabilizes as the number of trees increases, indicating convergence.

6 References

List any references used, including:

- Dataset sources (e.g., UCI repository)
- Relevant libraries or documentation (e.g., scikit-learn)

A Appendix

Include supplementary materials, additional plots, or code snippets if needed.