Decision Trees and Random Forests

Hunter Garrison

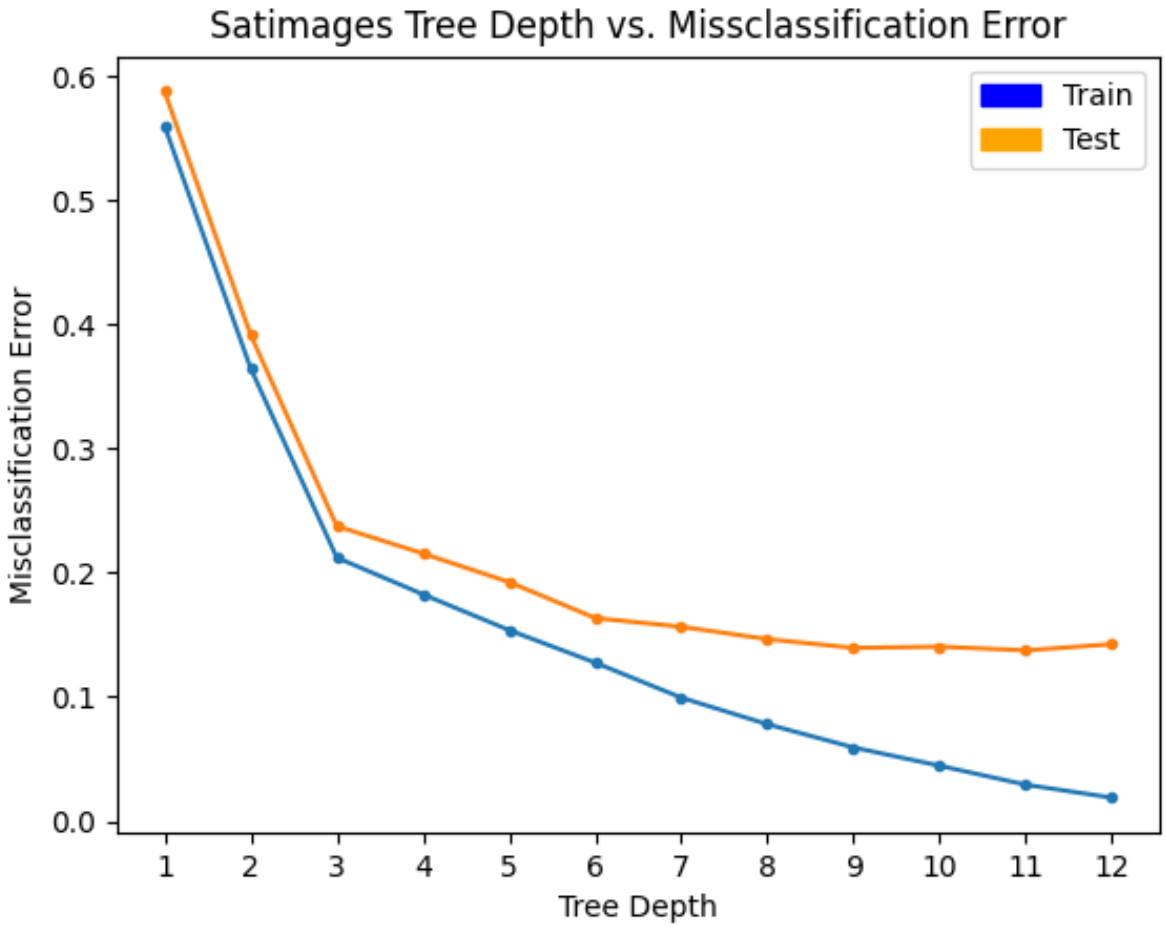
November, 2023

Introduction

This report will show the results of training and testing Decision Tree classifiers and Random Forest classifiers on both the satimage dataset and the madelon dataset. All sections will show error reports in the form of graphs.

Decision Trees with the satimage Dataset

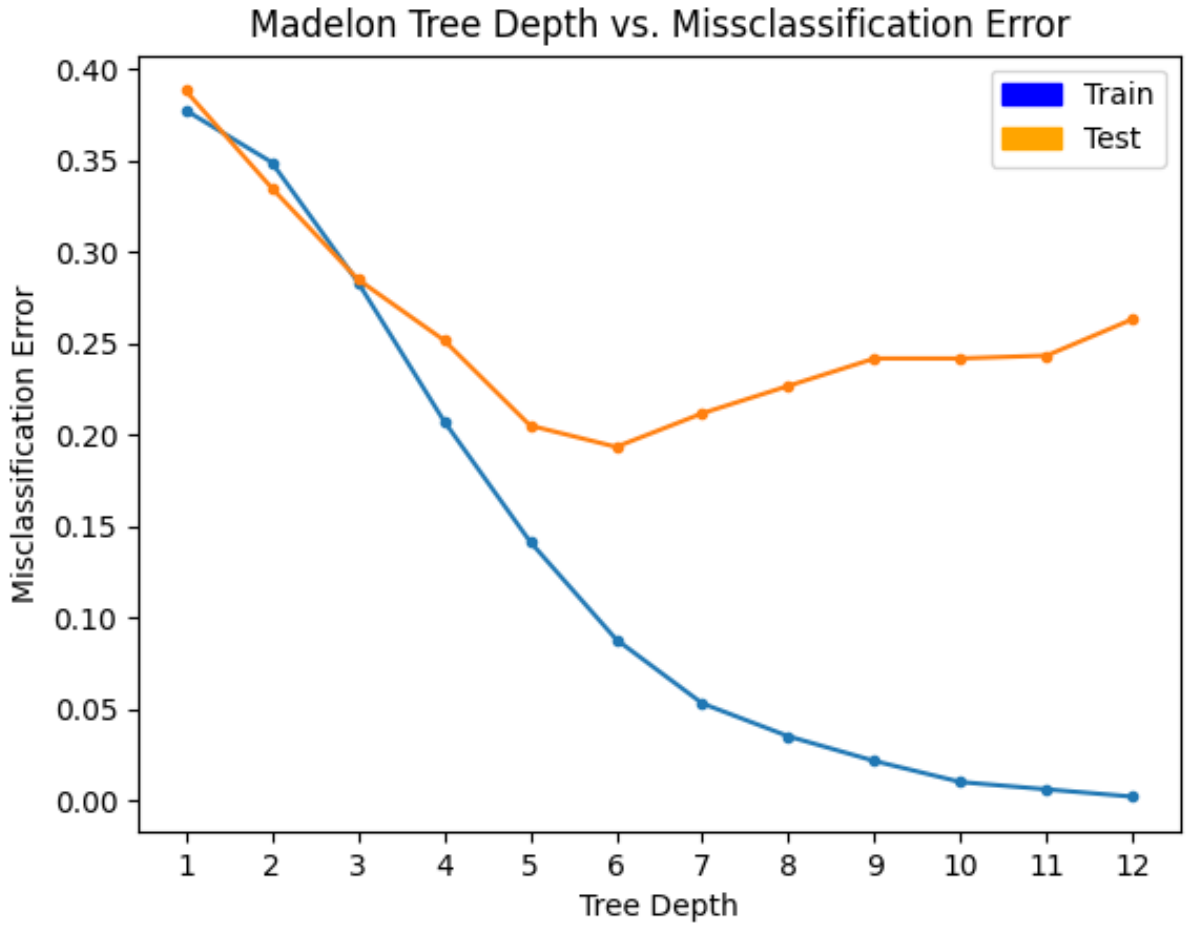
Using the satimage dataset (which has already been split into training and testing sets), we train default decision tree classifiers with max depths ranging from 1 to 12. We test our trees on both the training and testing sets and provide a graph showing their corresponding missclassification errors listed below.



We see that at after a maximum depth of 11 our testing error has hit a minimum and our training error continues to drop. This is a clear indication that after a maximum depth of 11, our decision tree starts to overfit the training data and provide worse estimates for our testing data.

Decision Trees with the madelon Dataset

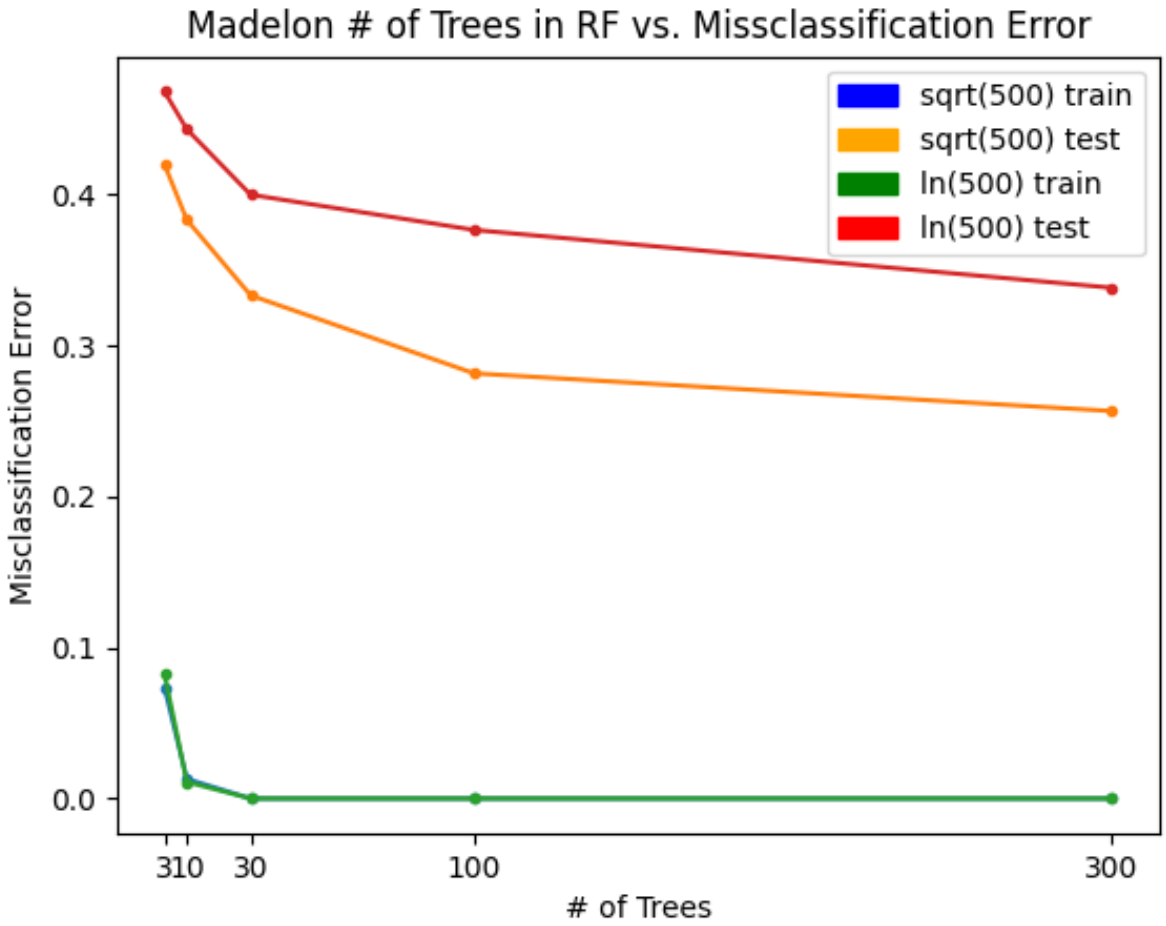
Using the madelon dataset (which has also already been split into training and testing sets), we again train default decision tree classifiers with max depths rangin from 1 to 12. We test our trees on both the training and testing sets and provide a graph showing their corresponding missclassification errors listed below.



Here, we see a more drastic example of overfitting our training data. After a maximum depth of 7, our testing error starts to rapidly increase as our training errors drops towards zero. We conclude that we should avoid creating trees that are too complex and that can learn our training data too welll in order to avoid overfitting and reducing our testing error.

Random Forests with the madelon Dataset

We again use the madelon dataset for classification but instead train random forest classifiers. Here, we want to find which amount of maximum features to use, $\sqrt{p}$ or $\ln p$ with p being the number of attributes of a data set. Since the madelon dataset has 500 attributes, we will test both values of max features against different values of total number of trees per forest. Below is a plot showing missclassification errors of all scenarios.



We see that both of our testing sets are continuously decreasing as the number of trees in the random forest increases, and that a maximum feature per tree value of $\sqrt{500}$ results in the best testing accuracy.

Conclusion

To conclude, we have shown the importance of properly choosing values for the hyperparameters of decision trees and random forests. The proper choosing of these parameters can help to reduce the likelihood of overfitting training data and making sure each model chooses the correct number of features for a given dataset.