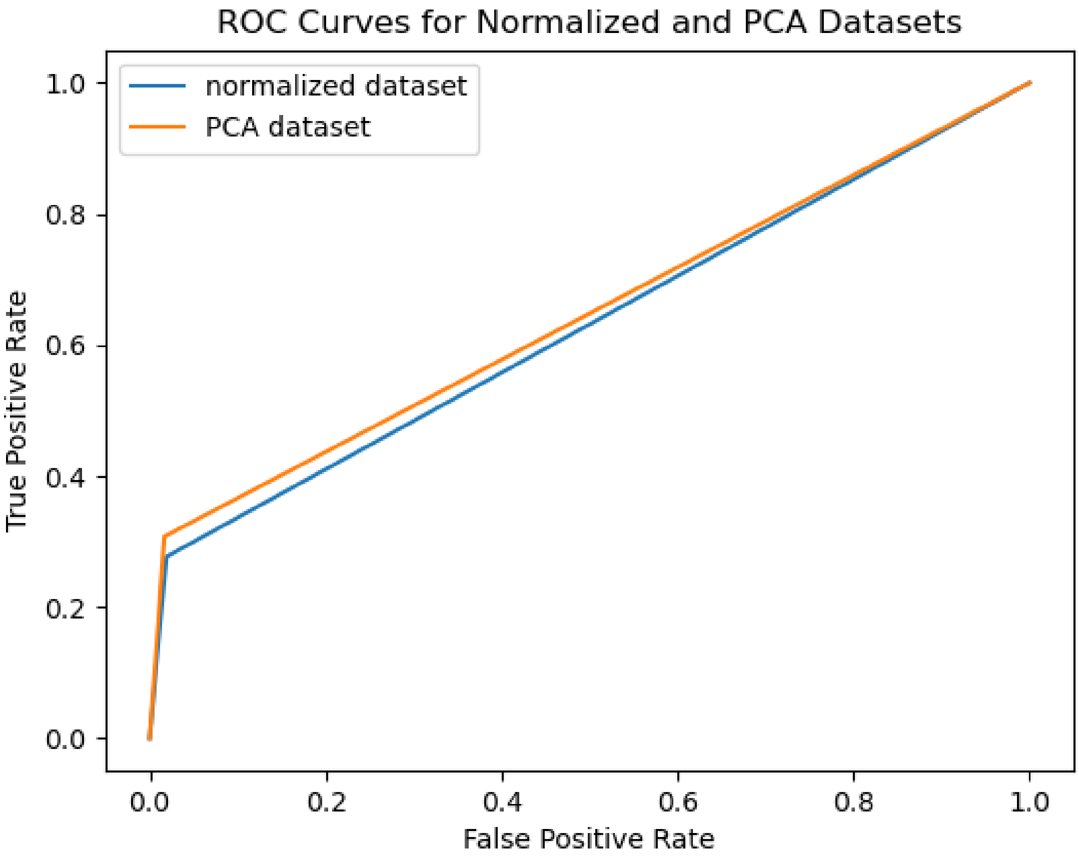
/* THIS CODE IS MY OWN WORK, IT WAS WRITTEN WITHOUT CONSULTING CODE
WRITTEN BY OTHER STUDENTS OR LARGE LANGUAGE MODELS SUCH AS CHATGPT.
Tommy Skodje */

I collaborated with the following classmates for this homework:
None

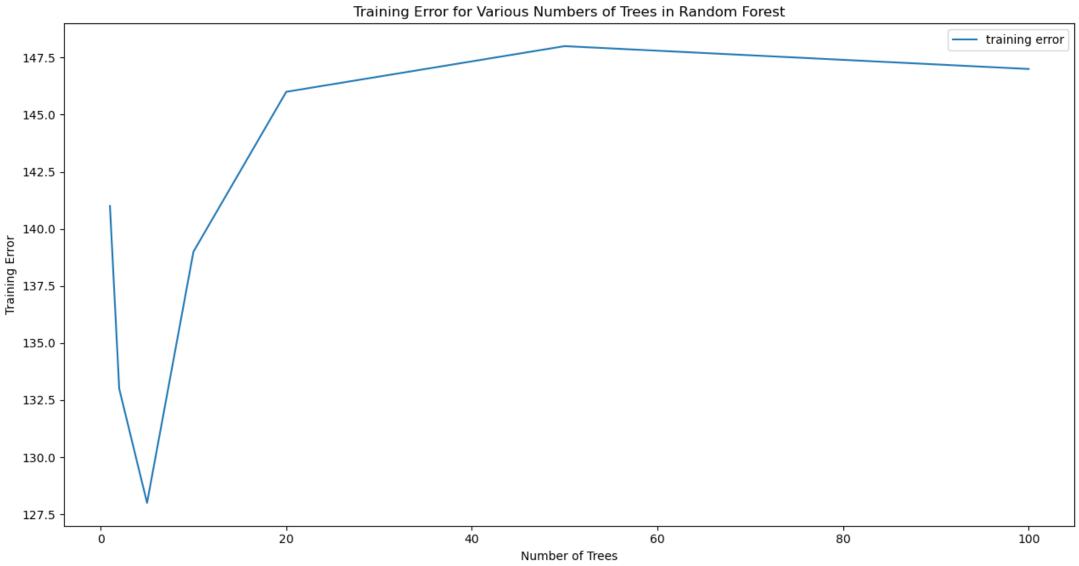
1d.

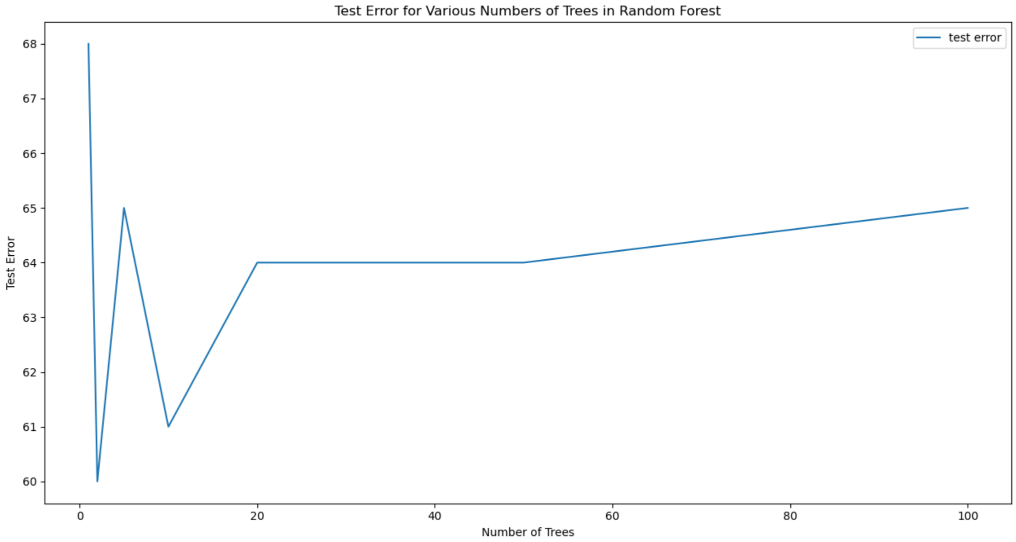


The PCA dataset's ROC curve seems to have a consistently larger AUC value throughout the entire graph. When printing the AUC values for both datasets in my code, the AUC for the normalized dataset is about 0.6288, while the AUC for the PCA dataset is about 0.6454. This means that the PCA dataset has more optimal true positive and false positive rates for all thresholds.

2b.

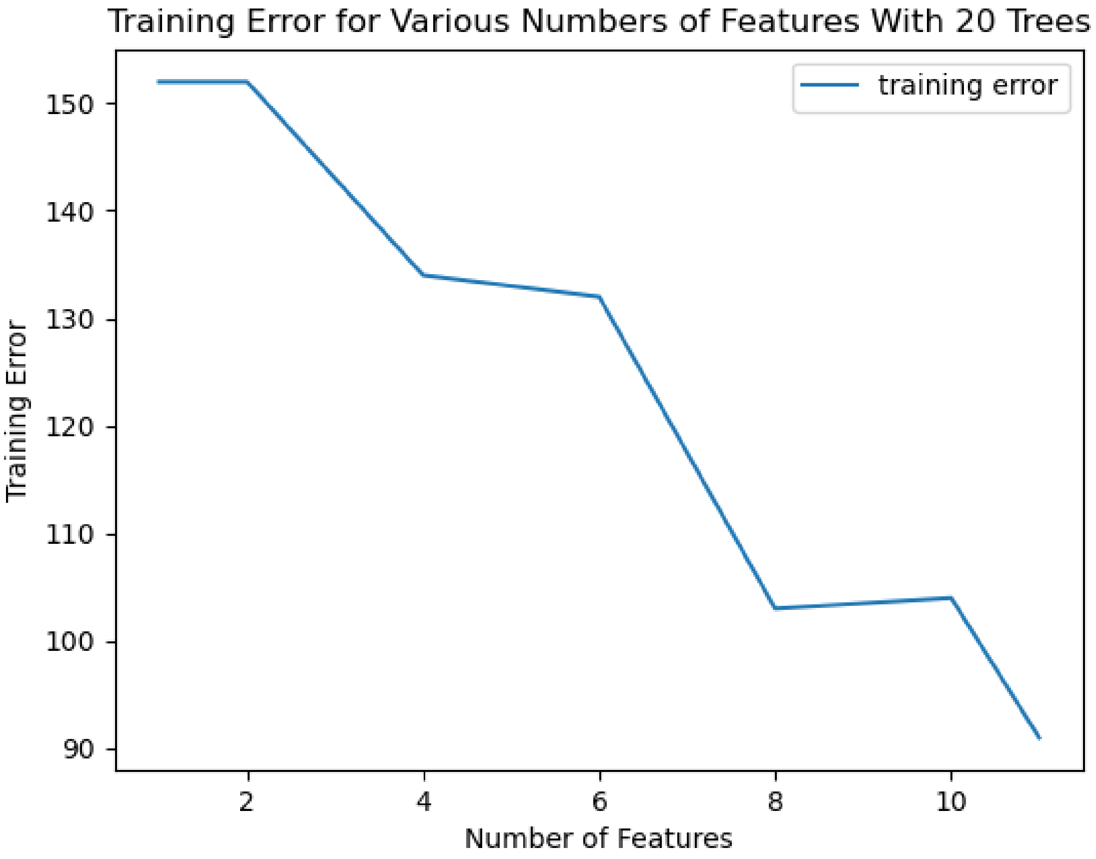
First, I determined the optimal number of trees by looking at the training and test error for various amounts of trees. I set features to 3, depth to 5, and min leaf samples to 10 as baselines. I also used entropy to split the tree nodes.

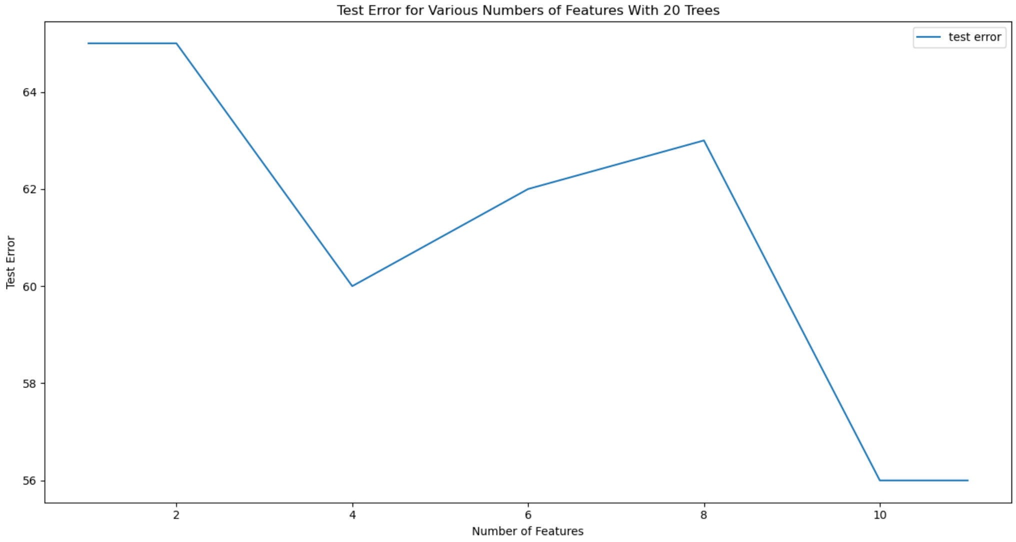




The test error stabilizes around 20 trees, so I decided that 20 trees was the optimal number of trees in the random forest.

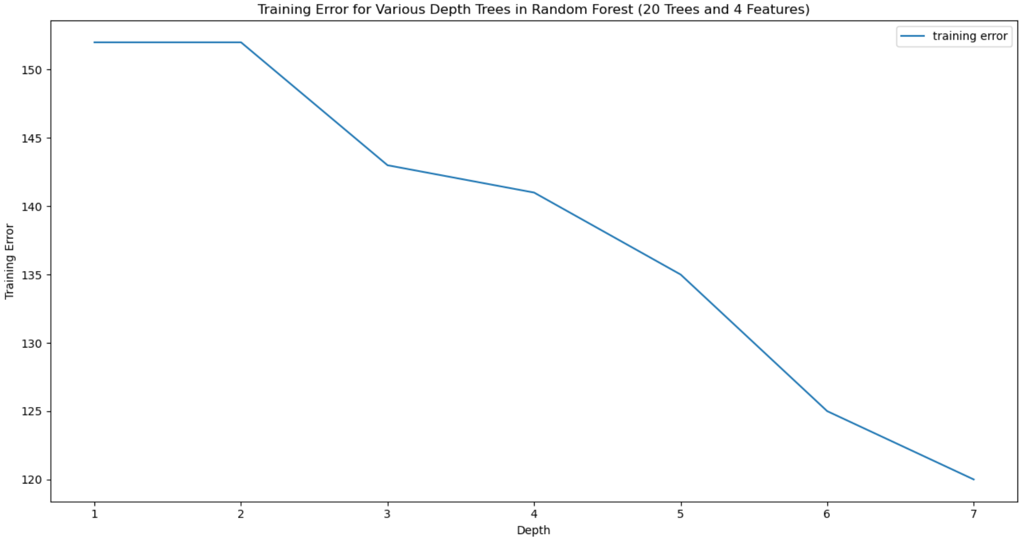
Next, I found the optimal number of features for a random forest of 20 trees.

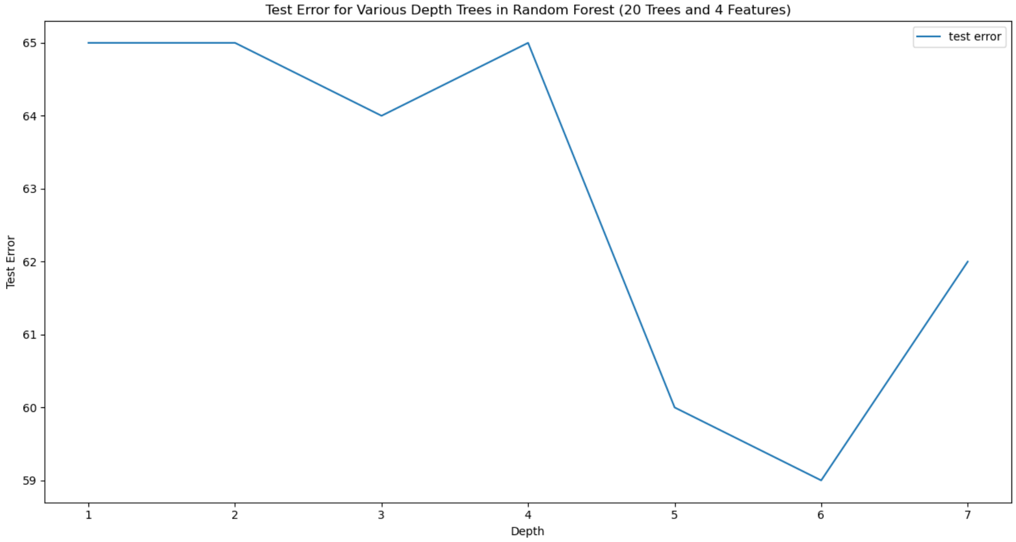




Technically, 10 and 11 features produced the lowest test error, but 4 features is “good enough,” and produces a much simpler and more interpretable model. Therefore, due to Occam’s Razor, I chose 4 features.

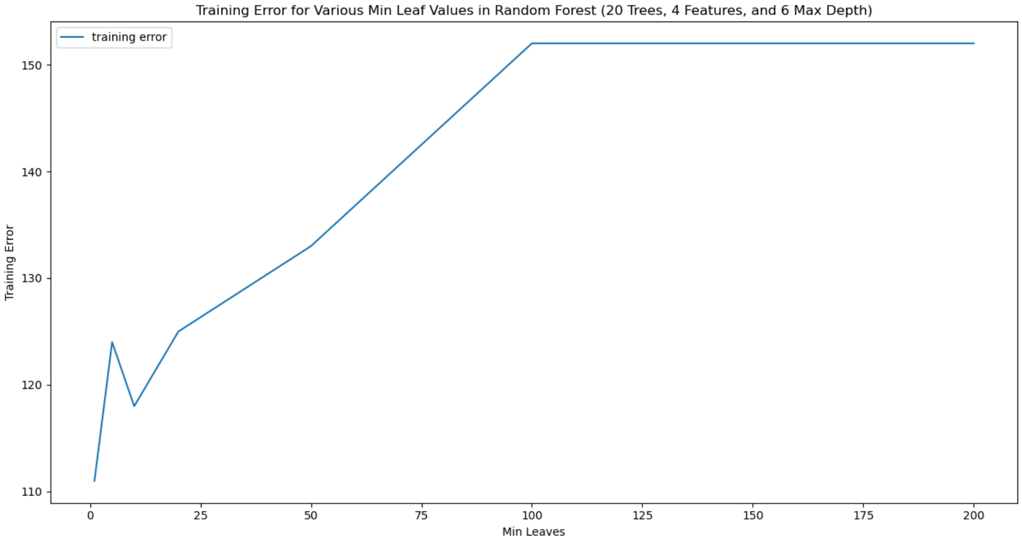
Next, I found the optimal depth. I kept the depths fairly shallow since random forest tends to work better with smaller trees.

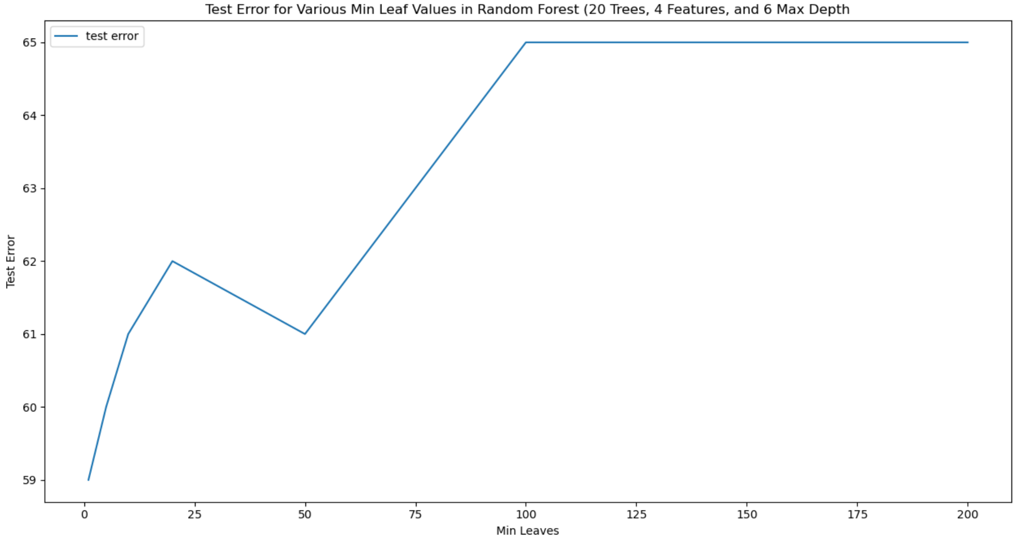




The test error is minimized at max depth = 6, so that is the optimal depth for the random forest.

Next, I found the optimal minimum leaf samples for 20 trees, 4 features, and a max depth of 6.





It seems that 1 min leaf sample creates the lowest test error. However, this leads to a very complex and slow model. Therefore, I picked 50 minimum leaf samples as the optimal number.

Therefore, the best parameters on the wine quality dataset are:
20 trees
4 features

6 maximum depth
50 minimum leaf samples

2c.
Using the optimal parameters of 20 trees, 4 features, 6 maximum depth, and 50 minimum leaf samples, the random forest produces 61 errors for the test data.

The OOB error for the training data is as follows:
{1: 127, 2: 127, 3: 133, 4: 132, 5: 132, 6: 123, 7: 131, 8: 127, 9: 128, 10: 129, 11: 127, 12: 125, 13: 127, 14: 128, 15: 128, 16: 128, 17: 128, 18: 123, 19: 126, 20: 126}

For 20 trees, the OOB error for the training data is 126.

The training dataset contains 1120 rows, while the test dataset contains 480 rows.

The OOB error rate was 126 / 1120 = 0.1125
The test data error rate was 61 / 480 = 0.1271

So the test error rate is slightly higher than the OOB error rate.