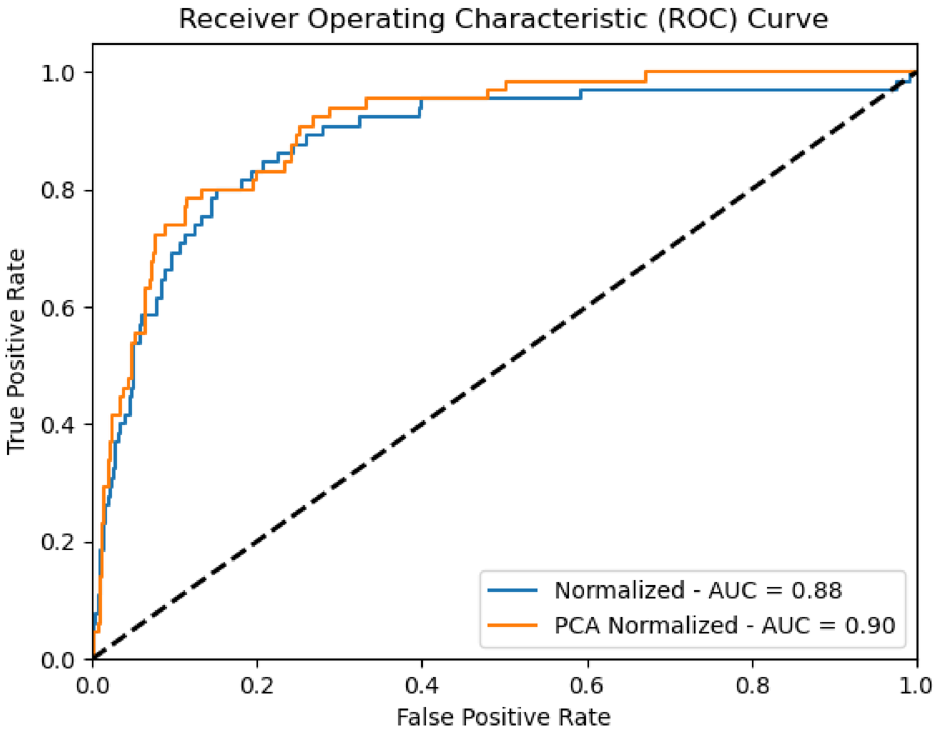
1. (d) (Written)

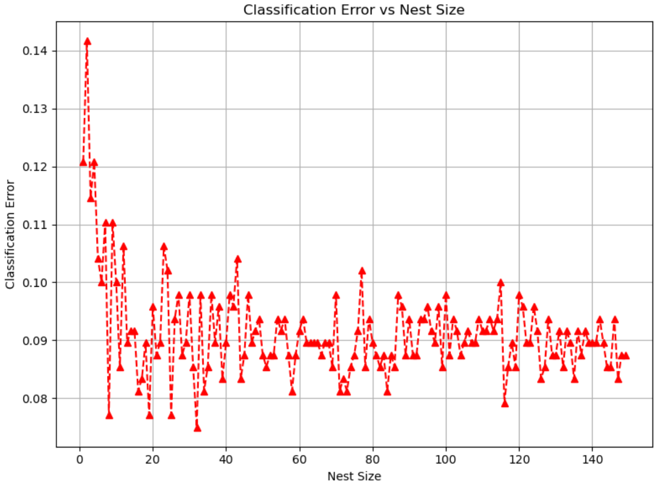


The ROC curves for both models are very similar, with the curves for the normalized and PCA datasets overlapping each other. This implies that neither of the two datasets is explicitly better than the other to use for logistic regression. The deciding factor could be the time complexity of the two models then.

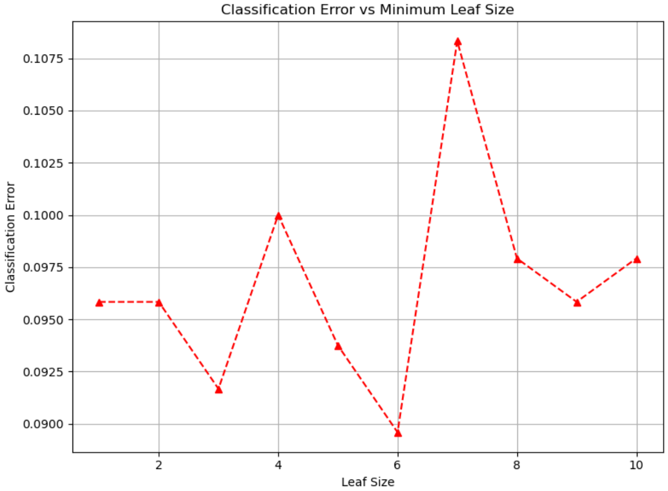
2. (b) (written)

Based on the four plots below, the optimal parameters are:

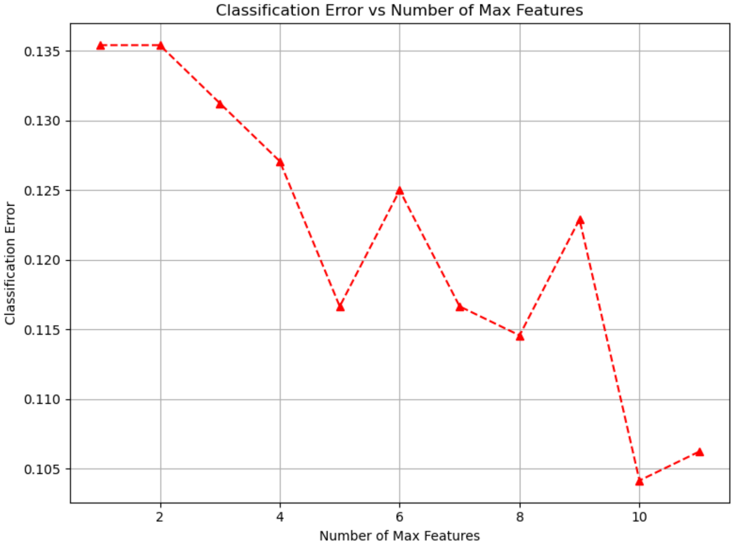
- Nest size = 9 (cap due to overfitting)
- Minimum leaf size = 6 (Not too little leaves, not too many)
- Max features = 10 (looks into all of the features)
- Tree depth = 12 (cap due to overfitting)



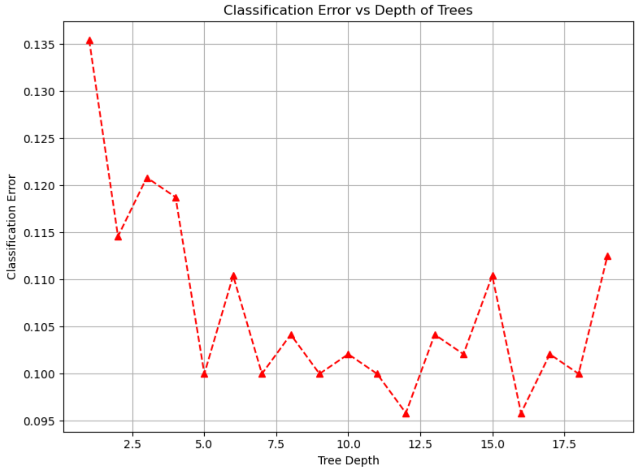
Based on the above plot, the optimal nest size would be 9.



Based on the above plot, the optimal minimum leaf size would be 6.



Based on the above plot, the optimal number of max features is 10.



Based on the above plot, the optimal tree depth is 12.

2. (c)  (Written)

OOB: 0.793

Test accuracy: 0.902

Using the optimal parameters, my version of the random forest was 90.2% accurate, with an error of 9.8%. This is better than the average OOB accuracy of 79.3%. This is because the test dataset is more similar to the training dataset than the out-of-bag samples, so the model performs better on the test set.