Classification of Leaf Species

The data set used for this project is a set of features of the leaves of 30 different plant species. The features were extracted from digital images of each leaf specimen. The original data set contains data for 40 species with both simple and complex leaves, but here we will only use those with simple leaves. I will attempt to find the best classifier for leaf species.

The first classification method I will use is discriminant analysis. In trying quadratic discriminant analysis, I found that the within class covariance matrices were singular, so I will just use linear discriminant analysis. Since there are 30 different species, I won't include confusion matrices as they each take up a few pages. Below are the resubstitution error counts for LDA followed by the cross-validated error rates.

Resubstitution Summary for Linear Discriminant Function

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.0833	0.1000	0.0000	0.1250	0.0000	0.0000	0.3000	0.0000	0.0000	0.0769	0.0000	0.1667	0.1538	0.3333	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0909	0.1538	0.2222	0.5000	0.1818	0.2500	0.0000	0.0833	0.1818	0.1818	0.0909	0.0000	0.3636	0.0000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1235
Priors	

Cross-validation Summary for Linear Discriminant Function

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.1000	0.1250	0.0833	0.2500	0.5000	0.0000	0.0714	0.3846	0.0000	0.1667	0.1538	0.3333	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.2727	0.3846	0.2222	0.6667	0.1818	0.2500	0.0000	0.0833	0.1818	0.1818	0.2727	0.0000	0.4545	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1912
Priors	

A total cross-validated error rate of .1912 is decent, but we can do better.

Next, I will use k-Nearest Neighbor methods for k values from 8 down to 2. The resubstitution and cross-validated error counts are presented below beginning with k=8.

Resubstitution Summary using 8 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.1000	0.3750	0.0000	0.2500	0.4000	0.0000	0.0714	0.2308	0.0000	0.1667	0.2308	0.4167	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.3077	0.2222	0.7500	0.1818	0.4167	0.0000	0.1667	0.2727	0.3636	0.1818	0.0000	0.4545	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1971
Priors	

Cross-validation Summary using 8 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.2500	0.1000	0.1000	0.3750	0.0000	0.2500	0.4000	0.0000	0.0714	0.3846	0.0000	0.1667	0.1538	0.5000	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.0769	0.2222	0.7500	0.2727	0.4167	0.0000	0.0833	0.1818	0.3636	0.1818	0.0000	0.4545	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1941
Priors	

Resubstitution Summary using 7 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.0000	0.3750	0.0000	0.1250	0.4000	0.0000	0.0714	0.2308	0.0000	0.0833	0.2308	0.3333	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.2308	0.1111	0.6667	0.2727	0.4167	0.0000	0.0833	0.1818	0.3636	0.1818	0.0000	0.4545	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1735
Priors	

Cross-validation Summary using 7 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.1000	0.3750	0.0000	0.2500	0.4000	0.0000	0.0714	0.2308	0.0000	0.1667	0.2308	0.4167	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.3077	0.2222	0.7500	0.1818	0.4167	0.0000	0.1667	0.2727	0.3636	0.1818	0.0000	0.4545	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1971
Priors	

Resubstitution Summary using 6 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.0000	0.3750	0.0000	0.1250	0.3000	0.0000	0.0714	0.2308	0.0000	0.1667	0.2308	0.2500	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.1538	0.2222	0.6667	0.1818	0.3333	0.0000	0.1667	0.1818	0.3636	0.1818	0.0000	0.4545	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1676
Priors	

Cross-validation Summary using 6 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.0000	0.3750	0.0000	0.1250	0.4000	0.0000	0.0714	0.2308	0.0000	0.0833	0.2308	0.3333	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.2308	0.1111	0.6667	0.2727	0.4167	0.0000	0.0833	0.1818	0.3636	0.1818	0.0000	0.4545	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1735
Priors	

Resubstitution Summary using 5 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.0000	0.3750	0.0000	0.1250	0.3000	0.0000	0.0714	0.2308	0.0000	0.1667	0.1538	0.4167	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.1538	0.1111	0.6667	0.2727	0.5000	0.0000	0.0000	0.1818	0.3636	0.1818	0.0000	0.4545	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1706
Priors	

Cross-validation Summary using 5 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.0000	0.3750	0.0000	0.1250	0.3000	0.0000	0.0714	0.2308	0.0000	0.1667	0.2308	0.2500	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.1538	0.2222	0.6667	0.1818	0.3333	0.0000	0.1667	0.1818	0.3636	0.1818	0.0000	0.4545	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1676
Priors	

Resubstitution Summary using 4 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.0000	0.5000	0.0000	0.2500	0.3000	0.0000	0.2143	0.2308	0.0000	0.0833	0.1538	0.3333	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.0769	0.1111	0.5000	0.1818	0.4167	0.0000	0.0000	0.1818	0.3636	0.1818	0.0000	0.5455	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1647
Priors	

Cross-validation Summary using 4 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.0000	0.3750	0.0000	0.1250	0.3000	0.0000	0.0714	0.2308	0.0000	0.1667	0.1538	0.4167	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.1538	0.1111	0.6667	0.2727	0.5000	0.0000	0.0000	0.1818	0.3636	0.1818	0.0000	0.4545	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1706
Priors	

Resubstitution Summary using 3 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.0833	0.1000	0.0000	0.2500	0.0000	0.1250	0.2000	0.0000	0.1429	0.0769	0.0000	0.0833	0.1538	0.2500	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.0769	0.1111	0.5833	0.1818	0.2500	0.0000	0.0833	0.1818	0.1818	0.1818	0.0000	0.2727	0.0000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1206
Priors	

Cross-validation Summary using 3 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.0000	0.5000	0.0000	0.2500	0.3000	0.0000	0.2143	0.2308	0.0000	0.0833	0.1538	0.3333	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.0769	0.1111	0.5000	0.1818	0.4167	0.0000	0.0000	0.1818	0.3636	0.1818	0.0000	0.5455	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1647
Priors	

Resubstitution Summary using 2 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.2500	0.3000	0.0000	0.5000	0.0833	0.1250	0.3000	0.0000	0.2143	0.2308	0.0000	0.1667	0.1538	0.2500	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.1667	0.0000	0.3077	0.1111	0.7500	0.1818	0.3333	0.0000	0.0833	0.1818	0.1818	0.1818	0.0000	0.2727	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1794
Priors	

Cross-validation Summary using 2 Nearest Neighbors

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.0833	0.1000	0.0000	0.2500	0.0000	0.1250	0.2000	0.0000	0.1429	0.0769	0.0000	0.0833	0.1538	0.2500	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0000	0.0769	0.1111	0.5833	0.1818	0.2500	0.0000	0.0833	0.1818	0.1818	0.1818	0.0000	0.2727	0.0000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1206
Priors	

The error rates for k-Nearest Neighbors for most of the different k values were about the same or slightly better than that of LDA. The cross-validated error rate of .1206 for k=2, however, was by far the best of all the cross-validated error rates so far.

At this point I realize I should do some variable selection process. I would guess that the error rates would improve slightly by doing so.

Backward Elimination: Step 1

Statistics for Removal, DF = 29, 296			
Variable	**Partial R-Square**	**F Value**	**Pr > F**
SpecimenNumber	0.0784	0.87	0.6651
Eccentricity	0.5346	11.73	<.0001
AspectRatio	0.7998	40.77	<.0001
Elongation	0.6295	17.34	<.0001
Solildity	0.7957	39.76	<.0001
StochasticConvexity	0.5378	11.88	<.0001
IsoperimetricFactor	0.6476	18.76	<.0001
MaximalIndentationDepth	0.5438	12.17	<.0001
Lobedness	0.4595	8.68	<.0001
AveIntensity	0.3246	4.91	<.0001
AveContrast	0.5677	13.41	<.0001
Smoothness	0.3025	4.43	<.0001
ThirdMoment	0.2381	3.19	<.0001
Uniformity	0.3769	6.17	<.0001
Entropy	0.4596	8.68	<.0001

Variable SpecimenNumber will be removed.

Variable(s) That Have Been Removed
SpecimenNumber

Backward Elimination: Step 2

Statistics for Removal, DF = 29, 297			
Variable	**Partial R-Square**	**F Value**	**Pr > F**
Eccentricity	0.5347	11.77	<.0001
AspectRatio	0.7998	40.91	<.0001
Elongation	0.6284	17.32	<.0001
Solildity	0.7979	40.43	<.0001
StochasticConvexity	0.5420	12.12	<.0001
IsoperimetricFactor	0.6450	18.61	<.0001
MaximalIndentationDepth	0.5359	11.83	<.0001
Lobedness	0.4521	8.45	<.0001
AveIntensity	0.3241	4.91	<.0001
AveContrast	0.5606	13.07	<.0001
Smoothness	0.3009	4.41	<.0001
ThirdMoment	0.2387	3.21	<.0001
Uniformity	0.3790	6.25	<.0001
Entropy	0.4599	8.72	<.0001

No variables can be removed.

No further steps are possible.

Using Backward Elimination the variable SpecimenNumber was removed in the first step and no variables could be removed after that. It makes sense that SpecimenNumber would be removed. It is simply a label for the specimens of each species and has no importance for the analysis. I probably should have realized that and not included it in the first place. Nonetheless, the resubstitution and cross-validated error rates for LDA and k-NN with k=2 without SpecimenNumber are below. Note that from here on out all the analyses will start without SpecimenNumber included.

Resubstitution Summary using Linear Discriminant Function without SpecimenNumber

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.0833	0.1000	0.0000	0.1250	0.0000	0.0000	0.3000	0.0000	0.0000	0.0769	0.0000	0.1667	0.1538	0.3333	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.0909	0.1538	0.2222	0.5000	0.1818	0.2500	0.0000	0.0833	0.1818	0.1818	0.0909	0.0000	0.3636	0.0000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1235
Priors	

Cross-validated Summary using Linear Discriminant Function without SpecimenNumber

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.1000	0.1000	0.1250	0.0833	0.2500	0.5000	0.0000	0.0714	0.3846	0.0000	0.1667	0.1538	0.3333	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.0833	0.2727	0.3846	0.2222	0.6667	0.1818	0.2500	0.0000	0.0833	0.1818	0.1818	0.2727	0.0000	0.4545	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1912
Priors	

Resubstitution Summary using 2 Nearest Neighbors without SpecimenNumber

Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.1667	0.3000	0.1000	0.5000	0.0000	0.1250	0.3000	0.0000	0.1429	0.3846	0.0000	0.1667	0.1538	0.2500	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.2500	0.0000	0.1538	0.1111	0.4167	0.2727	0.2500	0.0833	0.0833	0.1818	0.1818	0.1818	0.0000	0.3636	0.1000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.1706
Priors	

Cross-validation Summary using 2 Nearest Neighbors without SpecimenNumber

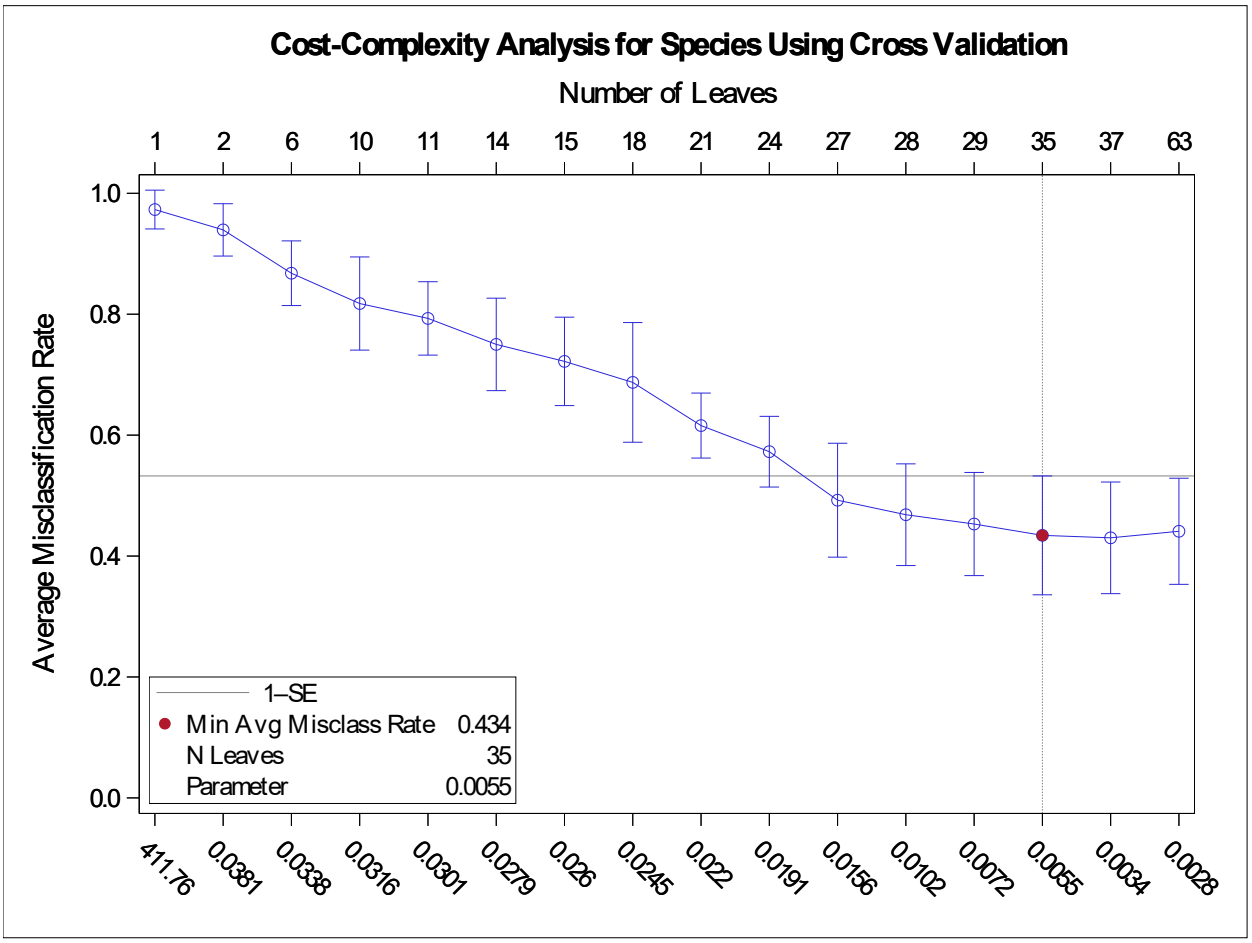
Error Count Estimates for Species															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Rate	0.0833	0.1000	0.0000	0.2500	0.0000	0.1250	0.2000	0.0000	0.0714	0.0769	0.0000	0.0833	0.1538	0.1667	0.0000
Priors	0.0353	0.0294	0.0294	0.0235	0.0353	0.0235	0.0294	0.0324	0.0412	0.0382	0.0471	0.0353	0.0382	0.0353	0.0294

	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Rate	0.1667	0.0000	0.0000	0.0000	0.3333	0.1818	0.1667	0.0000	0.0833	0.1818	0.0909	0.1818	0.0000	0.2727	0.0000
Priors	0.0353	0.0324	0.0382	0.0265	0.0353	0.0324	0.0353	0.0353	0.0353	0.0324	0.0324	0.0324	0.0324	0.0324	0.0294

	Total
Rate	0.0971
Priors	

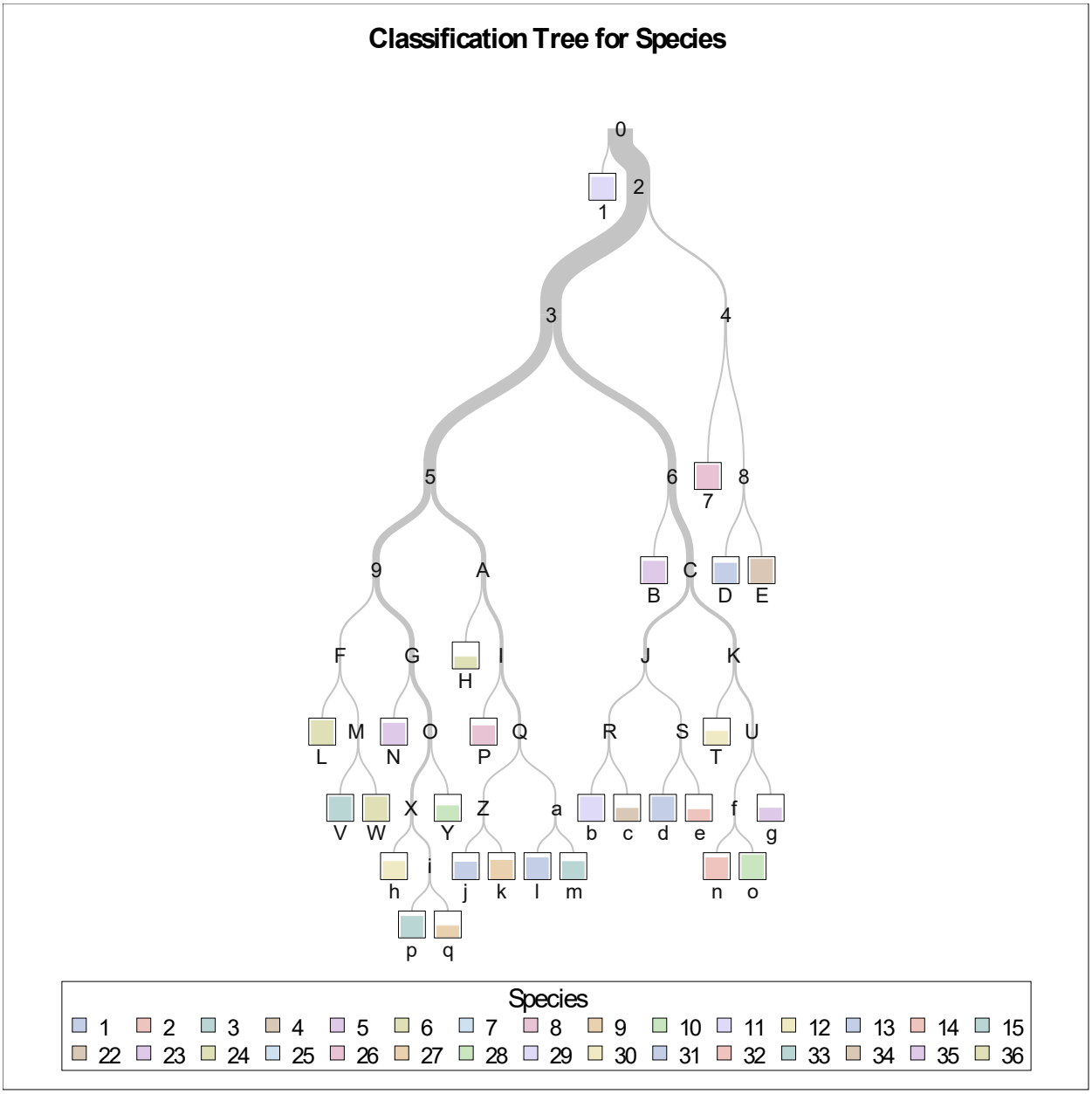
The error rates for LDA were exactly the same as before the variable selection. The resubstitution error rate for k-NN improved slightly, but the cross-validated error rate improved drastically to .0971. That's far and away the best error rate we've gotten so far.

Next, I will do a classification tree. Below is the Cost-Complexity plot from fitting a fully-grown tree.



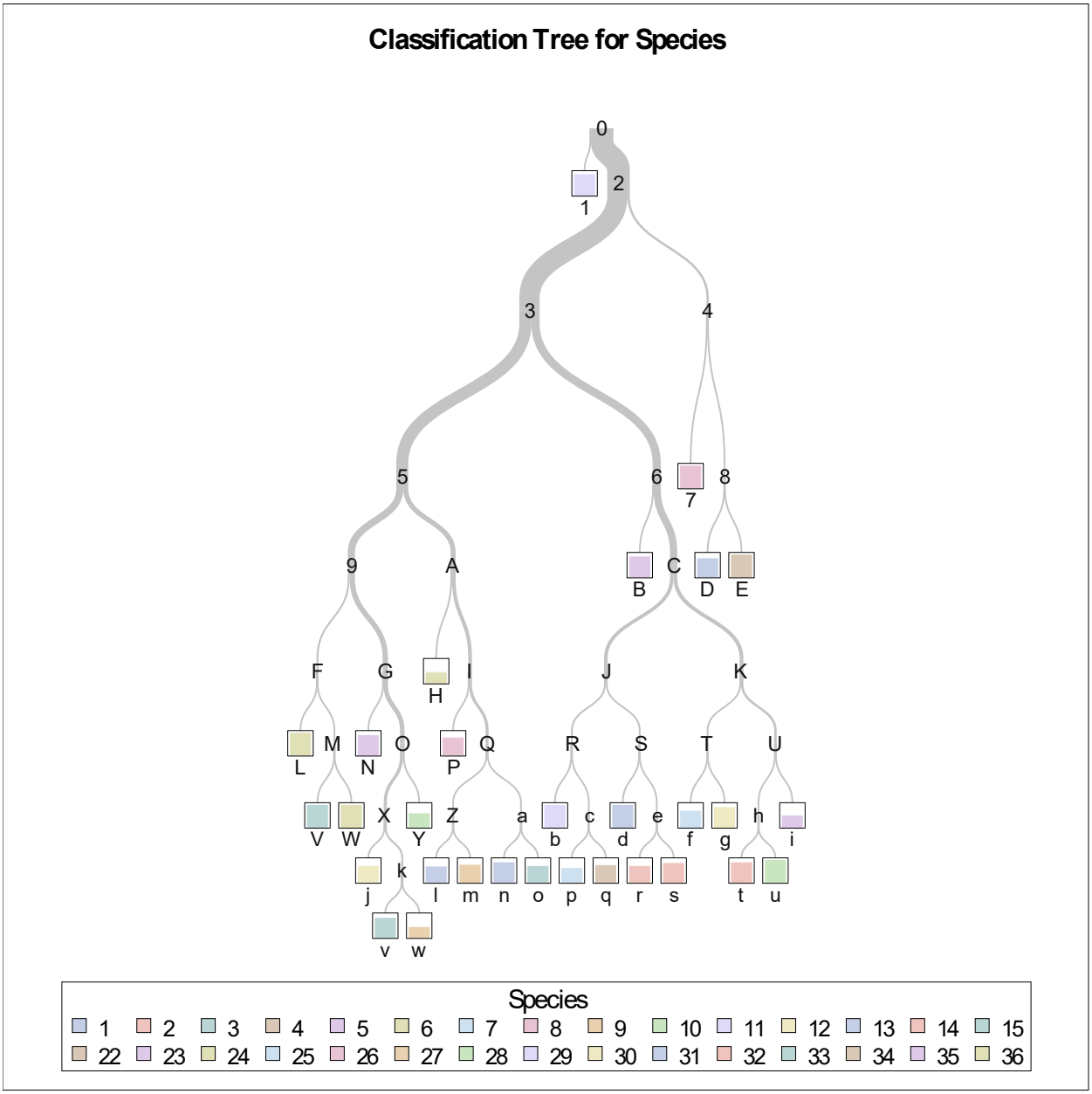
The minimum average misclassification rate is at 35 terminal nodes. Using the 1-SE rule, I will try a tree with 27 terminal nodes. I will also try a tree with 30 terminal nodes because that's how many species there are in the data.

Here is the tree with 27 leaves and the fit statistics.



Fit Statistics for Selected Tree						
	N Leaves	ASE	Mis-class	Entropy	Gini	RSS
Model Based	27	0.0116	0.2441	0.9516	0.3481	118.3
Cross Validation	27	0.0246	0.4870			

A cross-validated misclassification rate of .4870 is really bad. I'll try the tree with 30 leaves, but it probably won't be much better.



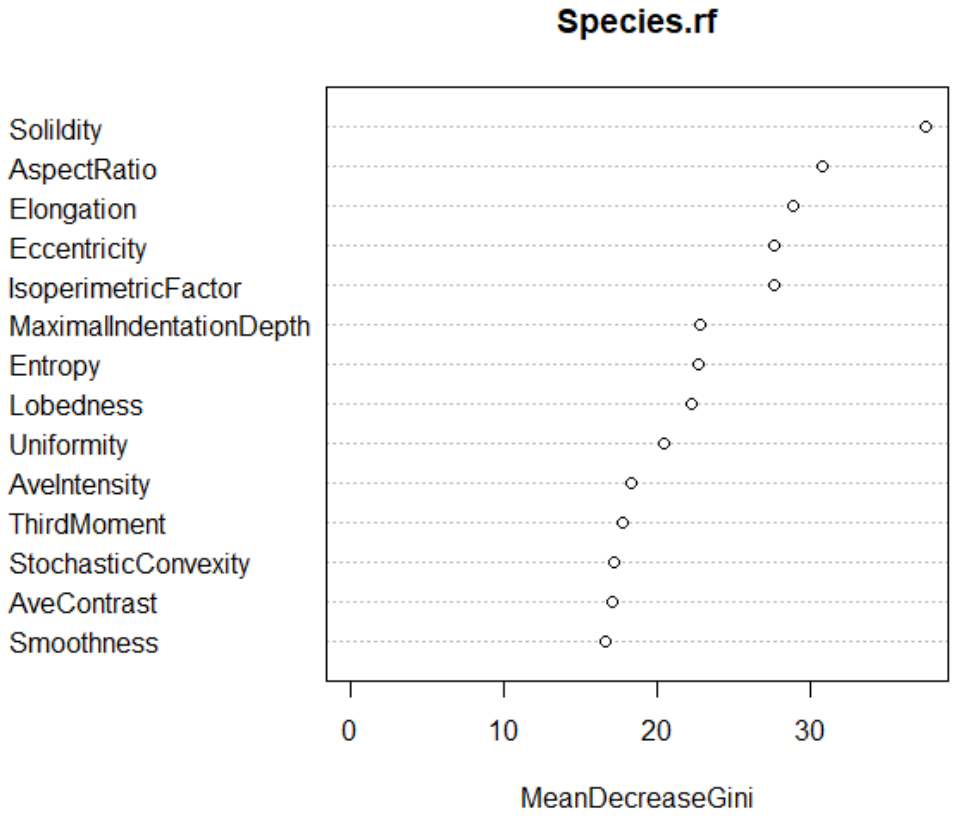
Fit Statistics for Selected Tree						
	N Leaves	ASE	Mis-class	Entropy	Gini	RSS
Model Based	30	0.0102	0.2000	0.8441	0.3049	103.7
Cross Validation	30	0.0246	0.4558			

Again, .4558 is a bad misclassification rate. I don't know why it's so bad. Maybe it's because of the number of species or maybe I'm doing something wrong. I won't bother with any sort of interpretation though because it's so bad.

Next, I will try random forests. Again, note that we will begin without including SpecimenNumber. Below is the out-of-bag estimate of the error rate followed by the variable importance plot.

```
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 22.94%
```



Based on this plot, none of the variables are extremely unimportant, but there is a clear break. I will try a random forest using only Solidity, AspectRatio, Elongation, Eccentricity, and IsoperimetricFactor. The out-of-box error rate is below.

```
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 41.76%
```

That is one of the worst error rates we've gotten so far and it was much better before removing the variables. I thought that might happen since the variables I removed weren't actually much lower on the variable importance plot. I just thought I'd try in case it was better. Even the better error rate of .2294 for the first random forest is worse than everything we got using LDA and k-Nearest Neighbors.

After running all of these classifiers, one clearly sticks out above the rest. K-Nearest Neighbors with 2 neighbors had a cross-validated error rate of .0971, which is quite remarkable. None of the other methods got any better than about .12. It's pretty clear that k-NN with k=2 gives the best classification of this data.

References

1. 'Evaluation of Features for Leaf Discrimination', Pedro F.B. Silva, Andre R.S. Marcal, Rubim M. Almeida da Silva (2013). Springer Lecture Notes in Computer Science, Vol. 7950, 197-204.