IS 467 Assignment 4
Sarah Cummings

Problem 1: Decision trees for letter recognition data
Create a classification model for letter recognition using decision trees as a classification method with a holdout partitioning technique for splitting the data into training versus testing.

a) Five different configurations and the accuracy for training and testing for each

i. I first started with most of the automatic settings for the decision tree. I used CRT growing method with 5 levels as the maximum tree depth, GINI as the impurity measure, 100 minimum parent cases, 50 minimum child cases, with 50% in training and 50% in testing. I also had the automatic 0.0001 minimum change in improvement. As expected, this was not the most accurate tree. I ended with 35.7% accuracy in training and 36% accuracy in testing.

ii. Next, I decided to increase the maximum tree depth to 10 since we have so many variables. I kept all else the same. This tree had 63.2% accuracy in training and 61.4% accuracy in testing. This is a big improvement from our previous tree.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	V1
	Independent Variables	V2, V3, V4, V5, V6, V7, V8, V9, V10, V11, V12, V13, V14, V15, V16, V17
	Validation	Split Sample
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	V12, V8, V11, V7, V10, V14, V13, V9, V16, V15, V17, V3, V5, V2, V4, V6
	Number of Nodes	31
	Number of Terminal Nodes	16
	Depth	5

tree 1 summary

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	V1
	Independent Variables	V2, V3, V4, V5, V6, V7, V8, V9, V10, V11, V12, V13, V14, V15, V16, V17
	Validation	Split Sample
	Maximum Tree Depth	10
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	V12, V8, V11, V7, V10, V14, V13, V9, V15, V16, V17, V3, V5, V4, V2, V6
	Number of Nodes	127
	Number of Terminal Nodes	64
	Depth	10

tree 2 summary

iii. For my third tree, I decided to change the size of the training sample. With as many cases as we have for this data, I figured it would be fine to have a larger quantity of data in the training set. I put 66% in the training and 34% in the testing. I left all else the same as the second tree. To my surprise, this did not create much improvement in the classification. We had 64.0% accuracy in the trainmen and 63.0% accuracy in the testing.

iv. For the fourth tree, I went back to the idea of increasing the number of maximum levels of tree depth. I increased the maximum to 15, and I also put 80% in the training data and 20% in the testing data. This tree gave 66.0% accuracy in the training data and 63.6% accuracy in the testing data.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	V1
	Independent Variables	V2, V3, V4, V5, V6, V7, V8, V9, V10, V11, V12, V13, V14, V15, V16, V17
	Validation	Split Sample
	Maximum Tree Depth	10
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	V12, V8, V11, V7, V10, V14, V13, V9, V15, V16, V6, V17, V5, V3, V4, V2
	Number of Nodes	143
	Number of Terminal Nodes	72
	Depth	10

tree 3 summary

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	V1
	Independent Variables	V2, V3, V4, V5, V6, V7, V8, V9, V10, V11, V12, V13, V14, V15, V16, V17
	Validation	Split Sample
	Maximum Tree Depth	15
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	V12, V8, V11, V7, V4, V10, V14, V13, V9, V15, V16, V17, V3, V5, V6, V2
	Number of Nodes	165
	Number of Terminal Nodes	83
	Depth	15

tree 4 summary

v. Next, I decided to allow my tree to be huge. I decreased the parent and child node case minimums, and increased the maximum tree level depth to 20. I set the training data set back to 66% and testing at 34%. While this tree is most accurate so far with 72% accuracy in the training and 69% accuracy in testing, the tree being as huge as it is makes it difficult to interpret and understand. Perhaps it is not best in that it is so complicated and huge.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	V1
	Independent Variables	V2, V3, V4, V5, V6, V7, V8, V9, V10, V11, V12, V13, V14, V15, V16, V17
	Validation	Split Sample
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	50
	Minimum Cases in Child Node	25
Results	Independent Variables Included	V12, V8, V11, V7, V2, V4, V10, V14, V13, V9, V15, V16, V17, V5, V3, V6
	Number of Nodes	277
	Number of Terminal Nodes	139
	Depth	17

Risk

Sample	Estimate	Std. Error
Training	.290	.004
Test	.312	.006

Growing Method: CRT
Dependent Variable: V1

b) The misclassification matrix, interpretation, and discussion of what determines “best” model:

Test		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Percent Correct
	A	2...	0	5	0	1	0	1	2	0	0	7	2	3	0	0	0	6	5	4	3	0	0	0	1	4	0	83%
	B	0	171	0	5	2	0	18	21	3	0	1	0	0	0	0	2	13	12	2	5	0	0	0	10	1	2	64%
	C	0	0	173	0	3	4	6	1	0	0	16	1	0	0	16	2	2	0	4	4	0	0	0	4	9	0	71%
	D	0	29	0	192	0	0	0	14	1	0	4	0	0	2	6	4	1	29	2	0	0	0	0	4	1	0	66%
	E	0	0	3	0	171	0	12	3	0	0	14	7	0	0	0	2	10	0	20	4	0	0	3	6	6	4	65%
	F	0	13	0	3	0	170	11	3	4	0	2	0	0	2	0	31	0	1	4	8	0	2	0	3	11	0	63%
	G	0	3	13	2	1	0	179	2	8	0	4	0	0	0	15	1	18	4	7	4	0	0	6	5	5	2	64%
	H	0	7	0	14	0	0	6	116	0	0	15	1	0	0	20	12	0	11	4	1	2	0	1	8	3	0	52%
	I	0	6	3	1	0	3	0	1	208	2	1	0	0	1	1	5	3	3	7	6	0	0	0	5	2	2	80%
	J	0	4	2	5	2	4	1	1	13	199	1	0	0	0	1	0	4	6	6	4	0	0	0	3	1	0	77%
	K	2	0	1	0	5	0	9	11	0	0	180	3	0	11	0	3	0	9	4	0	10	0	0	11	0	0	69%
	L	0	0	18	1	4	0	3	5	4	3	19	176	0	0	2	0	1	7	3	1	0	0	0	5	0	0	70%
	M	10	2	0	7	0	0	8	10	2	0	12	11	165	15	5	0	3	8	1	1	4	0	6	2	6	0	59%
	N	2	4	1	4	5	1	0	11	1	0	5	3	0	205	2	14	1	12	1	0	2	6	1	1	0	0	73%
	O	0	1	1	15	0	0	9	1	2	0	8	0	0	0	178	6	13	11	0	7	0	0	3	2	0	0	69%
	P	0	4	2	3	0	7	5	1	1	1	0	0	0	1	1	203	0	1	0	7	0	2	0	1	4	0	83%
	Q	4	3	3	2	3	1	22	2	0	0	14	0	0	1	21	9	171	9	1	4	0	3	1	4	3	0	61%
	R	3	14	4	4	2	0	7	2	3	0	25	5	0	0	3	0	9	151	5	6	0	0	0	5	0	0	61%
	S	0	17	1	0	0	1	1	13	0	0	12	0	0	0	3	2	6	16	156	4	0	0	0	7	13	5	61%
	T	0	2	0	0	2	10	2	2	2	0	7	0	0	0	0	8	2	1	1	182	0	1	1	9	31	3	68%
	U	0	0	0	0	6	2	1	5	0	0	13	0	3	7	5	1	9	5	4	8	202	1	0	0	4	0	73%
	V	0	2	0	0	0	6	3	7	0	0	2	0	0	6	0	5	4	0	2	1	2	196	2	0	11	0	79%
	W	0	0	0	1	0	1	6	3	0	0	0	0	0	20	5	5	4	0	0	1	2	8	224	0	6	0	78%
	X	0	20	4	0	13	1	0	1	1	0	15	1	0	0	0	0	0	11	3	1	0	0	0	170	2	0	70%
	Y	0	4	0	0	0	6	3	1	1	0	4	0	0	0	1	20	11	1	1	11	0	9	2	0	193	0	72%
	Z	0	7	0	0	3	0	11	1	1	0	0	0	0	0	8	7	4	8	20	2	0	0	0	3	1	157	67%
	Overall Perc	3%	4.6%	3.4%	3.8%	3.3%	3.2%	4.8%	3.5%	3.8%	3.0%	5.6%	3.1%	2.5%	4.0%	4.3%	5.0%	4.3%	4.7%	3.9%	4.1%	3.3%	3.4%	3.7%	4.0%	4.7%	2.6%	69%

Sorry this is awful looking. I couldn’t screen shot an image this big and I had issues exporting it so I had to screen shot in pieces.

This table tells us that in testing, we are 69% accurate in predicting the letters. A and P were the best predicted letters at 83% accuracy. H had the worst accuracy at 52%. This tree had the best accuracy, but might not be “best” overall since it is kind of complicated with many branches and nodes. Simpler may be better, as per Occam’s razor theory.

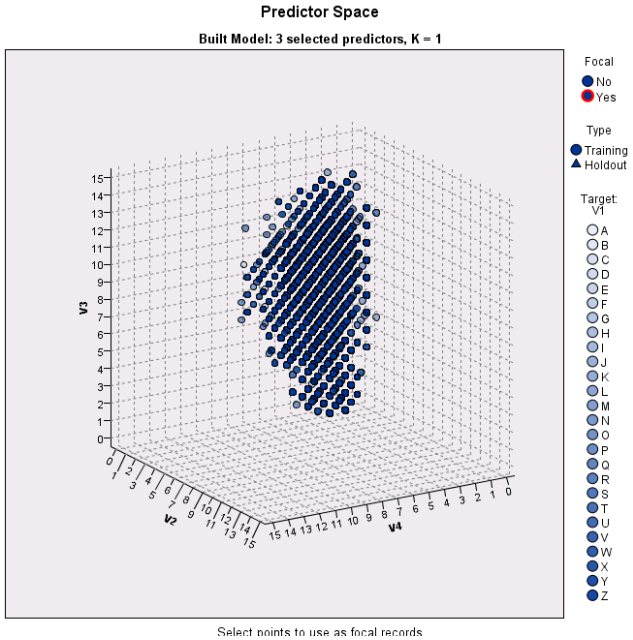
c) The most important attributes for determining the letters in my final tree (tree five) were variable 12 (0.022 improvement), variable 8 (0.019 improvement) and variable 10 (0.024 improvement). These also were to the most important variables in my fourth tree.

Problem 2: K-Nearest Neighbor classification for letter recognition data

a) I did not do any transformations. There are no missing values, no real outliers, and tall the attributes are of similar units.

b) Misclassification matrix and the appropriate performance metrics for k= 1,3, 7:

i. K=1



Left is the predictor space for k=1. Below we have a cross tabulation table formed with the actual letter variable and the predicted letter value for KNN n=1.

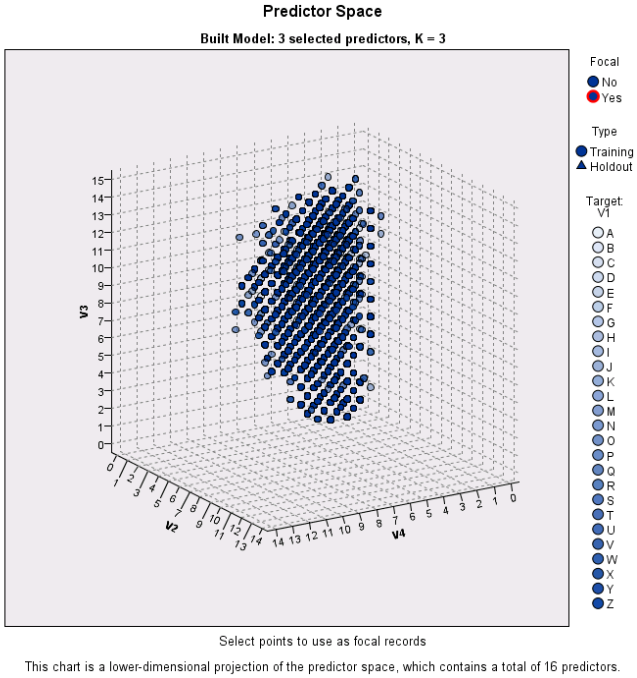
We have accuracies ranging from 93.48% for the letter A, to 76% for H. I cannot get the cross tabulation to compute overall total accuracy.

V1 * Predicted Value for V1 Crosstabulation

Statistics Count

		Predicted Value for V1																										
			A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Total
V1	A	0	732	3	1	6	2	2	2	2	0	0	2	5	5	2	4	2	0	2	5	0	4	2	2	2	0	2	789
	B	0	2	624	0	15	12	5	3	19	2	1	5	1	5	6	6	1	2	23	9	1	6	10	0	7	0	1	766
	C	0	0	2	649	2	12	2	22	1	2	1	4	2	1	0	3	2	4	3	3	3	9	0	2	4	3	0	736
	D	0	0	22	3	669	6	4	7	18	0	1	9	0	0	6	24	5	5	11	5	1	1	1	0	4	1	2	805
	E	0	0	8	9	3	654	4	11	12	2	0	6	8	3	1	3	0	0	7	12	1	2	1	0	15	1	5	768
	F	0	1	6	1	8	3	654	3	3	5	5	3	3	3	9	5	30	3	3	3	14	0	2	1	2	4	1	775
	G	0	0	6	12	6	18	2	665	9	3	3	5	5	6	3	7	1	2	1	3	3	4	0	7	1	1	0	773
	H	0	2	18	0	18	8	2	3	571	3	2	19	2	4	5	12	6	1	20	2	0	11	1	0	15	6	3	734
	I	0	0	2	1	2	2	11	0	1	692	21	1	7	0	0	3	3	1	2	2	1	1	0	0	2	0	0	755
	J	0	1	2	2	3	0	8	2	3	21	665	0	9	3	1	2	2	0	6	9	1	1	0	0	3	0	3	747
	K	0	3	9	2	11	12	6	3	30	2	0	589	1	4	4	4	1	4	19	4	1	1	2	2	22	1	2	739
	L	0	12	2	2	3	7	0	3	1	2	5	4	686	0	3	2	1	5	4	6	2	2	1	3	5	0	0	761
	M	2	4	4	1	0	2	2	2	4	0	2	1	0	716	12	5	0	0	10	2	1	5	11	3	2	1	0	792
	N	0	2	13	0	7	2	4	2	10	0	0	2	1	5	688	6	2	1	13	2	1	7	6	4	3	2	0	783
	O	0	3	5	8	16	4	3	10	9	0	1	2	1	1	3	636	3	18	1	8	0	5	4	3	6	1	2	753
	P	0	1	6	2	2	1	29	2	5	1	0	0	1	3	2	4	714	3	4	2	3	0	1	1	4	11	1	803
	Q	0	3	5	3	8	5	2	8	3	0	0	1	5	3	1	28	4	679	4	4	0	3	1	2	1	2	8	783
	R	0	3	28	0	9	2	7	5	12	0	4	14	4	3	13	1	0	3	635	5	1	3	2	0	3	0	1	758
	S	0	3	13	6	4	4	3	5	3	2	3	4	3	2	2	12	5	0	10	648	2	2	0	0	1	1	10	748
	T	0	0	3	4	1	3	8	5	3	1	2	6	2	2	4	2	3	0	2	5	711	1	6	0	1	20	1	796
	U	0	4	6	7	5	3	0	8	6	3	2	5	4	5	5	6	2	1	1	1	3	722	9	4	1	0	0	813
	V	0	0	8	1	1	2	3	2	1	0	0	2	0	4	4	1	3	0	0	2	9	5	702	9	0	5	0	764
	W	0	2	3	1	0	1	1	7	4	1	0	1	0	6	4	5	0	0	3	0	2	1	4	702	2	1	1	752
	X	0	1	6	5	11	11	0	4	15	2	0	14	3	3	0	11	2	3	5	2	6	4	2	0	668	1	8	787
	Y	0	2	2	4	0	2	2	1	2	1	2	0	1	3	0	1	6	2	2	1	26	7	12	2	4	698	3	786
	Z	0	2	3	1	4	13	0	2	2	2	6	3	2	1	1	2	1	6	2	7	1	0	0	0	4	3	666	734
Total		2	783	809	725	814	791	764	787	749	747	726	702	756	791	779	795	799	743	793	752	794	807	780	747	782	763	720	20000

ii. k=3

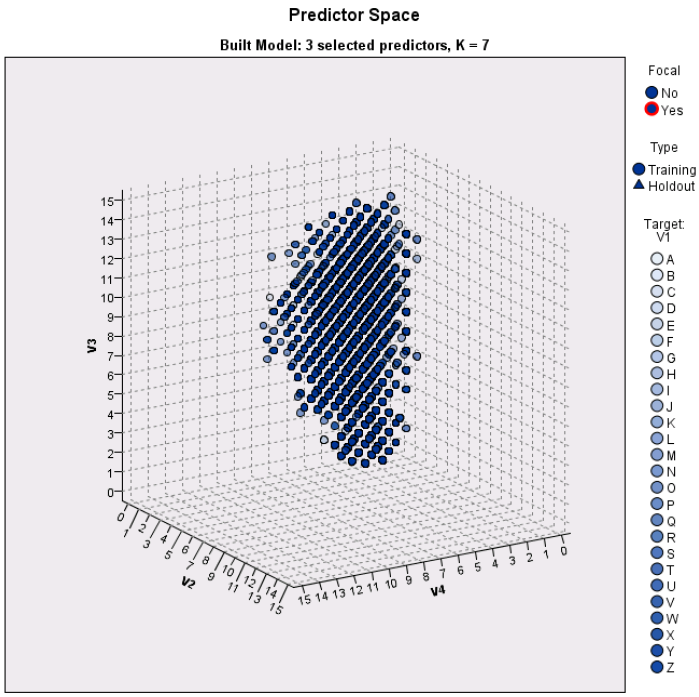


I have included the predictor space for k=3 at the left, and a cross tabulation table below. Again the predicted values are the columns and the actual values are the rows.

We had accuracies ranging in 97.7% for W to 66.3% for letter D. This percentages are significantly better than for k=1, with several letters having accuracies in the 90s.

	Predicted Value for V1																										
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Total
A	0	746	3	0	8	0	2	1	0	0	1	0	3	2	2	3	2	1	1	2	1	7	1	0	0	2	1	789
B	0	4	630	0	34	8	0	10	7	2	0	0	1	9	1	7	2	0	15	3	6	9	11	0	1	3	3	766
C	0	1	0	616	9	6	5	34	0	2	1	0	1	1	1	10	1	6	2	1	5	17	4	0	8	5	0	736
D	0	2	13	2	725	3	0	4	5	2	1	2	0	2	3	19	5	4	9	1	0	0	1	0	1	1	0	805
E	0	1	11	6	19	611	5	22	2	3	1	5	3	3	0	7	6	4	5	6	5	3	1	0	17	3	19	768
F	0	1	7	0	18	3	631	10	2	8	1	0	0	1	9	2	44	0	5	6	18	0	3	0	3	2	1	775
G	0	3	6	6	13	11	0	687	2	3	1	2	0	3	1	6	8	2	4	1	0	9	0	1	2	2	0	773
H	0	4	29	0	46	3	2	5	510	9	6	9	0	4	5	20	15	1	20	5	1	16	6	0	9	6	3	734
I	0	0	3	0	7	2	8	2	2	697	13	1	1	1	0	4	5	2	1	0	2	3	0	0	1	0	0	755
J	0	2	6	0	10	1	2	5	0	46	639	0	1	1	1	7	7	1	4	0	6	5	1	0	0	1	1	747
K	0	8	14	0	30	5	6	15	16	1	0	547	1	6	6	9	1	0	28	3	3	7	8	0	21	2	2	739
L	0	18	12	2	3	5	1	5	2	6	5	4	658	1	1	3	2	5	9	2	2	8	0	0	5	0	2	761
M	0	12	3	0	7	0	2	5	0	1	0	1	0	722	5	5	2	0	1	0	4	16	2	3	0	1	0	792
N	0	2	9	0	21	2	4	6	11	0	1	1	0	12	638	11	17	0	15	0	3	15	10	1	3	0	1	783
O	0	5	3	1	36	0	1	12	4	0	0	1	0	3	1	636	4	14	1	2	1	10	3	1	11	2	1	753
P	1	0	3	0	9	2	21	4	3	0	0	1	0	0	0	2	737	0	3	0	2	1	1	0	2	11	0	803
Q	0	10	3	1	11	3	3	23	0	2	1	1	0	3	0	31	10	654	6	3	1	5	1	1	1	6	3	783
R	0	4	37	0	21	4	2	6	3	0	1	3	1	9	5	4	2	1	636	3	5	2	6	0	1	2	0	758
S	0	9	19	1	10	5	6	13	5	3	1	1	2	5	3	14	7	5	10	599	2	6	6	0	3	4	9	748
T	0	3	5	0	11	1	8	9	1	4	0	1	0	1	1	2	8	0	3	1	701	8	3	0	2	22	1	796
U	0	3	2	1	10	1	1	6	2	2	1	2	1	6	1	7	3	1	1	0	0	754	3	2	1	2	0	813
V	0	5	6	1	3	1	3	5	2	0	0	1	0	5	1	7	14	0	2	0	6	9	682	7	0	4	0	764
W	0	3	5	0	2	0	2	13	5	1	0	0	2	8	2	6	2	0	4	0	1	7	6	677	1	5	0	752
X	0	13	8	0	19	9	4	10	2	6	4	8	2	2	0	11	8	2	14	0	5	7	1	0	643	3	6	787
Y	0	8	1	1	3	1	4	1	1	1	1	0	0	2	0	0	18	2	2	1	15	6	14	0	1	702	1	786
Z	0	3	7	0	9	5	1	5	0	12	4	0	0	0	0	6	3	11	2	3	3	0	0	0	4	5	651	734
	1	870	845	638	1094	692	724	918	587	811	683	591	677	812	687	839	933	716	803	642	798	930	774	693	741	796	705	20000

iii. k=7



The predictor space for k=7 is at left, dn the cross tabulation for actual versus predicted is below. The best accuracy was for L, with 95.8% and the worst was 73.5% for D.

It's important to note that the k=7 model was not seemingly better than the k=3 model. With our data set, using the 7 nearest neighbors in prediction lead to less accurate prediction than if we used the closed three points.

		Predicted Value for V1																										
			A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Total
V1	A	0	723	1	5	4	0	2	1	2	0	2	0	2	14	3	6	4	4	2	3	1	4	0	0	3	3	0	789
	B	0	1	610	0	24	10	5	6	15	0	2	3	1	7	2	2	0	1	32	10	1	11	7	2	8	1	5	766
	C	0	0	0	631	6	16	1	19	1	0	0	3	1	1	0	8	3	9	0	7	1	16	1	2	6	4	0	736
	D	0	0	18	2	696	1	1	2	12	1	0	3	0	1	7	18	4	6	13	5	0	7	0	0	8	0	0	805
	E	0	1	3	7	4	658	1	11	9	0	2	6	5	1	1	5	4	6	0	7	2	4	0	0	19	0	12	768
	F	0	0	4	0	17	3	648	1	0	2	4	1	0	2	11	1	38	1	2	5	23	1	0	0	5	4	2	775
	G	0	2	8	10	9	15	2	661	6	1	2	0	2	6	1	6	3	9	6	5	3	7	1	0	7	0	1	773
	H	0	2	16	2	21	2	2	1	571	0	3	14	0	6	6	8	10	2	27	5	1	13	1	4	12	4	1	734
	I	0	0	1	2	5	5	12	2	3	667	27	1	3	0	0	4	5	5	4	2	4	2	0	0	1	0	0	755
	J	0	1	7	0	4	3	1	2	1	17	676	0	3	5	0	1	2	2	4	2	5	2	0	0	5	0	4	747
	K	0	1	10	2	15	13	5	2	25	0	1	581	0	8	4	1	1	3	29	3	0	7	0	1	27	0	0	739
	L	0	14	5	3	2	10	1	6	4	3	4	6	661	1	1	3	5	8	6	5	0	3	0	0	8	1	1	761
	M	1	6	5	1	5	1	1	2	5	0	1	1	0	716	16	1	4	1	2	3	0	11	2	4	2	1	0	792
	N	0	0	4	0	20	3	3	0	12	0	4	1	0	9	657	11	6	0	26	1	1	10	9	2	2	1	1	783
	O	0	2	2	1	26	3	1	7	9	0	1	4	0	4	3	622	3	20	4	6	0	7	2	5	21	0	0	753
	P	0	0	4	2	5	4	33	4	0	1	0	0	1	0	2	3	715	2	3	3	3	1	0	1	8	8	0	803
	Q	0	3	3	3	13	3	1	9	0	0	0	2	1	1	1	26	3	691	4	2	1	5	2	2	4	3	0	783
	R	0	2	16	0	14	7	3	1	4	1	2	5	3	7	6	1	1	4	660	5	0	3	0	1	11	1	0	758
	S	0	7	13	5	8	10	4	3	5	2	4	1	0	2	0	7	2	5	7	644	2	3	3	0	2	1	8	748
	T	0	0	3	1	12	2	4	6	3	0	3	0	2	2	1	2	4	0	5	4	717	6	3	2	3	10	1	796
	U	0	0	1	4	3	0	1	2	4	0	2	2	1	12	5	9	3	2	4	0	0	750	2	1	4	1	0	813
	V	0	2	9	0	2	1	4	1	1	0	0	0	0	5	2	1	8	2	3	2	14	10	676	10	3	8	0	764
	W	0	1	4	1	4	3	2	4	11	0	0	0	2	10	5	7	1	2	4	0	0	4	9	673	2	2	1	752
	X	0	2	2	0	20	11	4	3	10	1	2	9	1	1	2	7	0	2	10	1	1	7	0	1	688	0	2	787
	Y	0	1	1	0	1	1	6	3	1	1	3	0	1	1	0	2	13	6	0	1	24	13	13	0	5	685	4	786
	Z	0	4	2	1	7	8	0	1	5	6	3	0	0	1	1	2	0	10	4	6	8	6	0	0	15	3	641	734
Total		1	775	752	683	947	793	748	760	719	703	748	643	690	823	737	764	842	803	861	737	812	913	731	711	879	741	684	2.E...

c) Interpretation of results and comparison to decision tree results:

The K-nearest neighbors method proved to be pretty successful in classifying the letters. With k=1, our accuracies for each letter ranged from 76-93%. With k=3, our accrues ranged from 66.3% to 97.7%. With k=7, our accuracies ranged from 73.5% to 95.8%. This method proved to be much more accurate than the decision trees, as our best decision tree had only 69% accuracy in testing. For this data set, using KNN was best.