Keiland Pullen Assignment 4
DSC 441 – Fall 2021

Problem 1.A.)

1. Training a KNN classifier does take less computation time than testing. This is because KNN algorithms do more computation on test time instead of training time. The concept is to find a list of samples that are close to what needs to be classified.
2. The more training examples used in KNN, the higher the possibility of overfitting the data. In theory, each of these training examples are likely to produce the same output. Which would not be ideal for the test or validation data.
3. KNN can be used for both classification and regression. To predict the values of new data, KNN uses "feature similarity".
4. KNN performs better with a smaller number of features. Too many features may lead to a case of the model overfitting the data.

Problem 1.B.)

In this case, k=3 might prove to the best accuracy. However, this depends on the size of the data set. If the data is a binary set, the k=3 can be used to avoid a "tie". Ideally, we'd like to keep k as a small value, but in some cases the optimal value may be for k to equal the square root of n.

Keiland Pullen
DSC 441 – Fall 2021

Problem 2.a.)

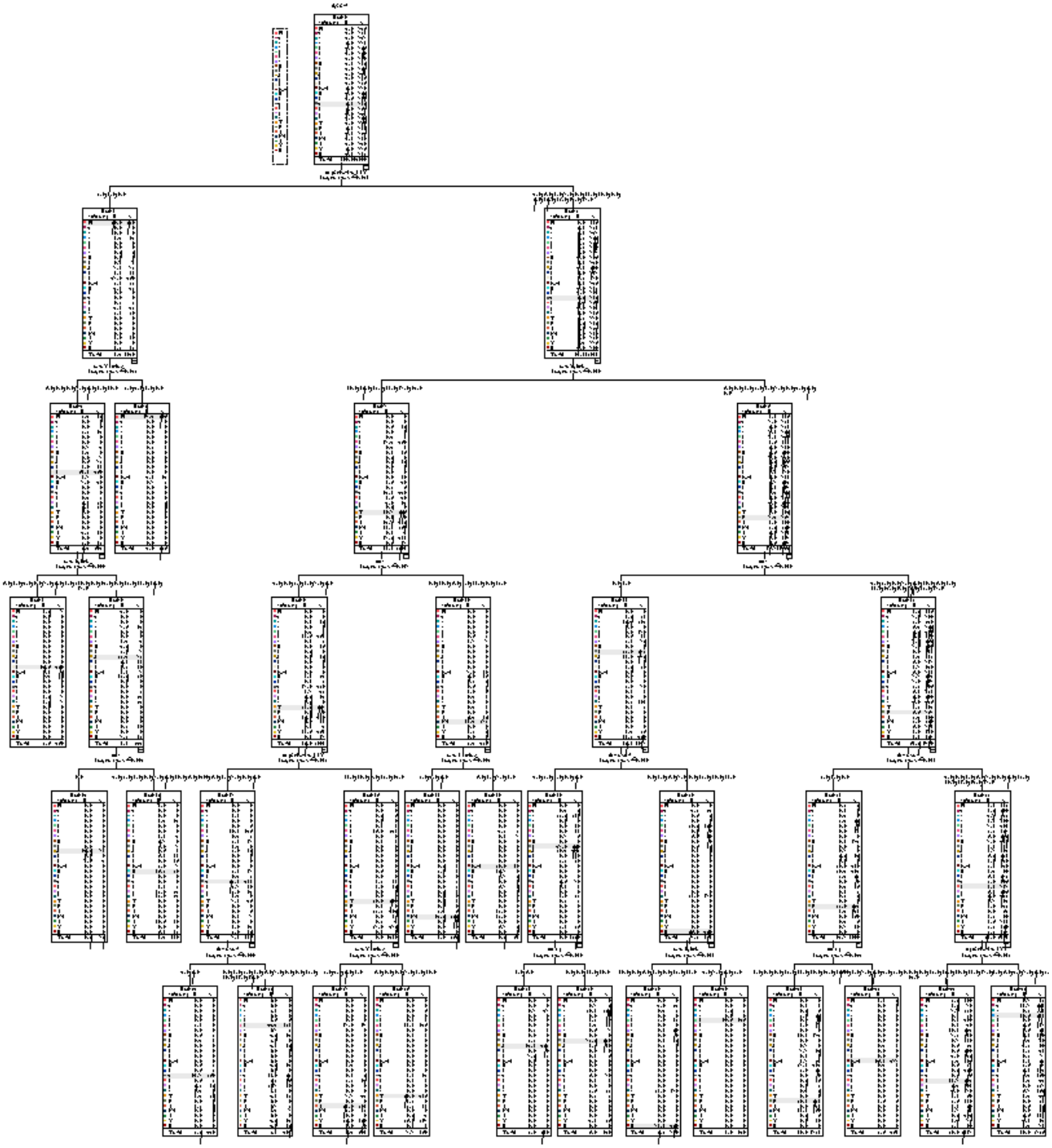
Model Summary #1

Specifications	Growing Method	CRT
	Dependent Variable	alphabet
	Independent Variables	horizonal, vertical, width, height, onpix, meanXpixels, meanYpixels, meanXvariance, meanYvariance, xybarmean, x2ybrMeanXXY, xy2brMeanXYY, xege, xegvy, edgeCount, yegvx
	Validation	Split Sample
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	x2ybrMeanXXY, meanYpixels, xybarmean, meanXpixels, meanYvariance, xege, xy2brMeanXYY, meanXvariance, xegvy, edgeCount, yegvx, width, vertical, height, onpix, horizonal
	Number of Nodes	35
	Number of Terminal Nodes	18
	Depth	5

Keiland Pullen
DSC 441 – Fall 2021

Assignment 4

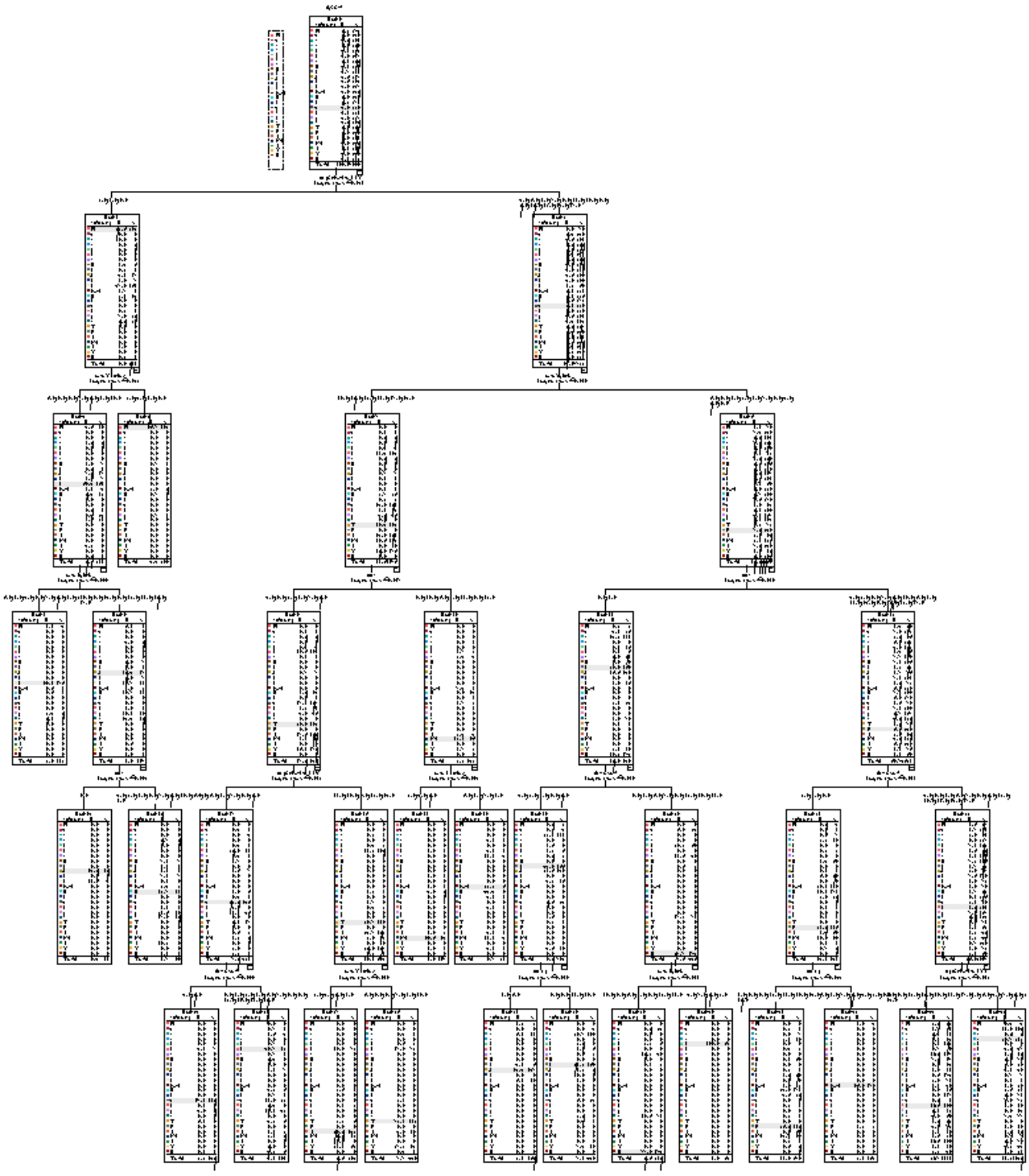
Training Sample #1



Keiland Pullen
DSC 441 – Fall 2021

Assignment 4

Testing Sample #1



Risk

Sample	Estimate	Std. Error
Training	.632	.004
Test	.632	.006

Keiland Pullen Assignment 4
DSC 441 – Fall 2021

Growing Method: CRT

Dependent Variable: alphabet

Classification #1

Classification

Sample	Observed	Predicted 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
Training	A	12	0	0	0	0	0	0	0	5	4	9	0	0	100	0	0	0	0	0	0	0	0	0	0	0	75.7%
	B	441	0	0	0	0	4	0	0	0	0	0	0	0	58	0	0	0	0	1	0	0	0	0	0	23	0.0%
	C	4	2	8	0	0	244	35	0	0	0	0	0	0	228	0	0	6	6	0	0	0	0	0	0	2	0.0%
	D	471	0	0	0	0	0	0	0	0	0	13	0	0	29	0	0	0	44	0	0	0	0	0	0	0	84.6%
	E	1	136	6	0	0	10	0	0	0	0	0	0	0	380	0	0	1	0	0	0	0	0	0	0	12	24.9%
	F	115	0	127	0	0	11	0	0	0	2	0	0	9	30	0	0	136	15	95	0	0	0	0	0	0	23.5%
	G	41	0	2	0	0	5	1	0	0	5	3	0	0	423	0	0	0	0	0	0	0	0	0	0	48	0.0%
	H	119	0	13	0	0	0	0	0	0	0	8	0	2	218	0	0	1	153	0	0	0	0	0	0	0	0.0%
	I	58	0	8	0	0	439	9	0	0	0	5	0	1	19	0	0	0	0	0	0	0	0	0	0	4	80.8%
	J	80	0	14	0	0	10	394	0	0	1	1	0	0	23	0	0	0	8	0	0	0	0	0	0	1	74.1%
	K	31	0	0	0	0	0	0	0	0	0	11	0	0	400	0	0	0	84	0	0	0	0	0	0	0	0.0%
	L	5	0	0	0	0	6	25	0	349	0	16	0	0	124	0	0	0	1	0	0	0	0	0	0	0	66.3%
	M	43	0	0	0	0	0	0	0	4	373	5	0	0	96	0	0	0	23	0	1	0	0	0	0	0	66.8%
	N	65	0	13	0	0	0	0	0	1	10	29	0	2	25	0	0	0	382	8	7	0	0	0	0	0	5.3%
	O	103	0	0	0	0	0	0	0	0	4	0	0	2	349	0	0	0	43	0	0	0	0	0	0	0	0.0%
	P	160	0	42	0	0	58	0	0	0	0	0	0	269	26	0	0	13	0	4	2	0	0	0	0	0	46.9%
	Q	33	0	16	0	0	1	0	0	0	4	2	0	2	482	0	0	1	6	13	2	0	0	0	0	0	85.8%
	R	225	0	78	0	0	0	0	0	5	0	23	0	14	181	0	0	0	2	0	1	0	0	0	0	0	0.0%
	S	164	0	8	0	0	22	0	0	5	0	23	0	1	158	0	0	0	4	0	0	0	0	0	0	159	0.0%
	T	33	0	12	0	0	2	0	0	0	0	0	0	24	69	0	0	376	0	33	1	0	0	0	0	0	68.4%
	U	22	0	0	0	0	0	0	0	0	6	0	0	0	110	0	0	9	427	0	0	0	0	0	0	0	74.4%
	V	14	0	2	0	0	0	0	0	0	8	0	0	4	48	0	0	17	137	287	6	0	0	0	0	0	54.9%
	W	19	0	5	0	0	0	0	0	0	8	0	0	0	61	0	0	0	169	12	244	0	0	0	0	0	47.1%
	X	117	0	0	0	0	0	0	0	0	0	19	0	0	394	0	0	3	6	0	0	0	0	0	0	0	0.0%
	Y	93	0	13	0	0	26	0	0	0	4	0	0	0	56	0	0	205	10	149	0	0	0	0	0	0	0.0%
	Z	57	0	14	0	0	0	0	0	0	0	12	0	0	88	0	0	1	0	0	0	0	0	0	0	340	66.4%
	Overall Percentage	18.0%	1.0%	2.7%	0.0%	0.0%	6.0%	3.3%	0.0%	2.6%	3.1%	1.3%	0.0%	2.4%	29.8%	0.0%	0.0%	5.5%	10.9%	4.3%	1.9%	0.0%	0.0%	4.2%	36.8%		

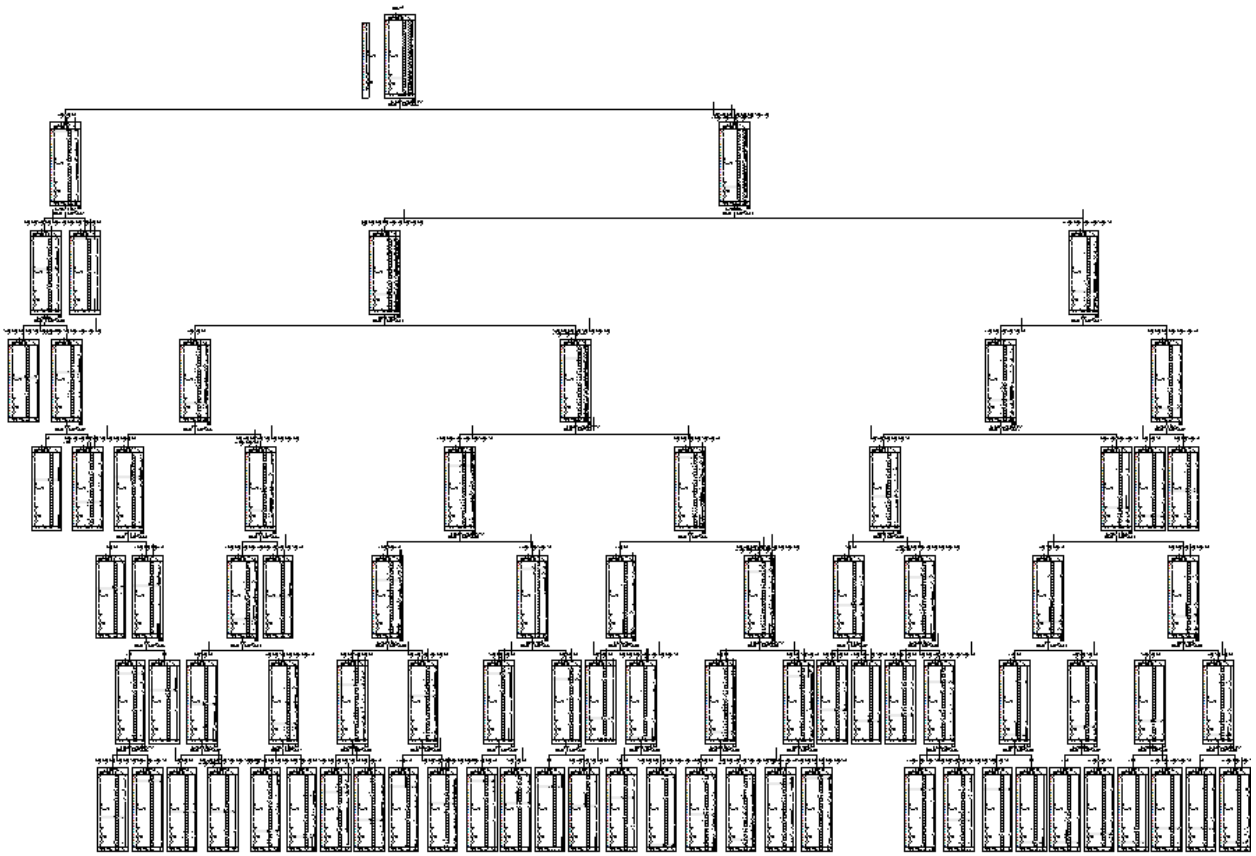
Sample	Observed	Predicted A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Percent Correct
Test	A	3	0	0	0	0	0	2	0	3	3	7	0	0	42	0	0	0	0	0	0	0	0	0	0	0	76.3%
	B	184	0	1	0	0	3	0	0	0	0	0	0	0	27	0	0	0	1	0	0	0	0	0	0	23	0.0%
	C	4	0	3	0	0	100	11	0	0	0	0	0	0	77	0	0	1	5	0	0	0	0	0	0	0	0.0%
	D	212	0	0	0	0	0	0	0	0	1	4	0	0	13	0	0	0	18	0	0	0	0	0	0	0	85.5%
	E	0	62	1	0	0	9	0	0	0	0	0	0	0	143	0	0	2	0	0	0	0	0	0	0	6	27.9%
	F	35	0	72	0	0	6	0	0	0	1	0	0	5	9	0	0	52	2	52	1	0	0	0	0	0	30.6%
	G	14	0	1	0	0	2	0	0	0	3	0	0	0	185	0	0	0	0	0	0	0	0	0	0	39	0.0%
	H	51	0	8	0	0	0	0	0	0	0	5	0	0	95	0	0	0	60	0	0	0	0	0	0	0	0.0%
	I	29	0	0	0	0	160	6	0	0	0	1	0	0	13	0	0	0	0	0	0	0	0	0	0	2	75.8%
	J	38	0	5	0	0	4	149	0	0	0	1	0	0	17	0	0	0	1	0	0	0	0	0	0	0	69.3%
	K	5	0	0	0	0	0	0	0	0	0	5	0	0	153	0	0	0	49	0	1	0	0	0	0	0	0.0%
	L	1	0	0	0	0	6	13	0	156	0	8	0	0	51	0	0	0	0	0	0	0	0	0	0	0	66.4%
	M	20	0	0	0	0	0	0	0	1	164	2	0	0	30	0	0	0	12	0	1	0	0	0	0	0	70.1%
	N	27	0	6	0	0	0	0	0	1	9	11	0	0	16	0	0	0	157	7	4	0	0	0	0	0	4.6%
	O	51	0	1	0	0	0	0	0	0	1	0	0	0	175	0	0	0	24	0	0	0	0	0	0	0	0.0%
	P	52	0	22	0	0	23	0	0	0	0	0	0	114	7	0	0	9	0	0	2	0	0	0	0	0	49.8%
	Q	19	0	8	0	0	0	0	0	0	1	1	0	1	188	0	0	0	1	2	0	0	0	0	0	0	85.1%
	R	93	0	32	0	0	0	0	0	4	0	19	0	4	72	0	0	0	0	0	0	0	0	0	0	0	0.0%
	S	55	0	0	0	0	8	0	0	7	0	14	0	0	68	0	0	0	1	0	0	0	0	0	0	51	0.0%
	T	8	0	6	0	0	1	0	0	0	0	0	0	9	43	0	0	172	0	6	0	0	0	0	0	0	70.2%
	U	10	0	0	0	0	0	0	0	0	2	0	0	0	53	0	0	0	174	0	0	0	0	0	0	0	72.8%
	V	14	0	2	0	0	0	0	0	0	1	0	0	0	20	0	0	10	60	132	2	0	0	0	0	0	54.8%
	W	5	0	4	0	0	0	0	0	0	3	0	0	0	28	0	0	0	89	12	93	0	0	0	0	0	39.7%
	X	50	0	0	0	0	0	0	0	0	0	9	0	0	188	0	0	0	1	0	0	0	0	0	0	0	0.0%
	Y	33	0	9	0	0	18	0	0	0	1	0	0	0	19	0	0	83	4	62	1	0	0	0	0	0	0.0%
	Z	21	0	6	0	0	0	0	0	0	0	1	0	1	39	0	0	1	0	0	0	0	0	0	0	153	68.9%
	Overall Percentage	17.2%	1.0%	3.1%	0.0%	0.0%	5.7%	3.0%	0.0%	2.9%	3.2%	1.5%	0.0%	2.2%	29.5%	0.0%	0.0%	5.5%	11.0%	4.6%	1.0%	0.0%	0.0%	4.6%	36.8%		

Model Summary #2

Specifications	Growing Method	CRT
	Dependent Variable	alphabet
	Independent Variables	horizonal, vertical, width, height, onpix, meanXpixels, meanYpixels, meanXvariance, meanYvariance, xybarmean, x2ybrMeanXXY, xy2brMeanXYY, xege, xegvy, edgeCount, yegvx
	Validation	Split Sample
	Maximum Tree Depth	7
	Minimum Cases in Parent Node	75
	Minimum Cases in Child Node	25
Results	Independent Variables Included	x2ybrMeanXXY, meanYpixels, xybarmean, meanXpixels, xegvy, meanYvariance, xege, xy2brMeanXYY, meanXvariance, edgeCount, yegvx, height, vertical, width, horizonal, onpix
	Number of Nodes	85
	Number of Terminal Nodes	43
	Depth	7

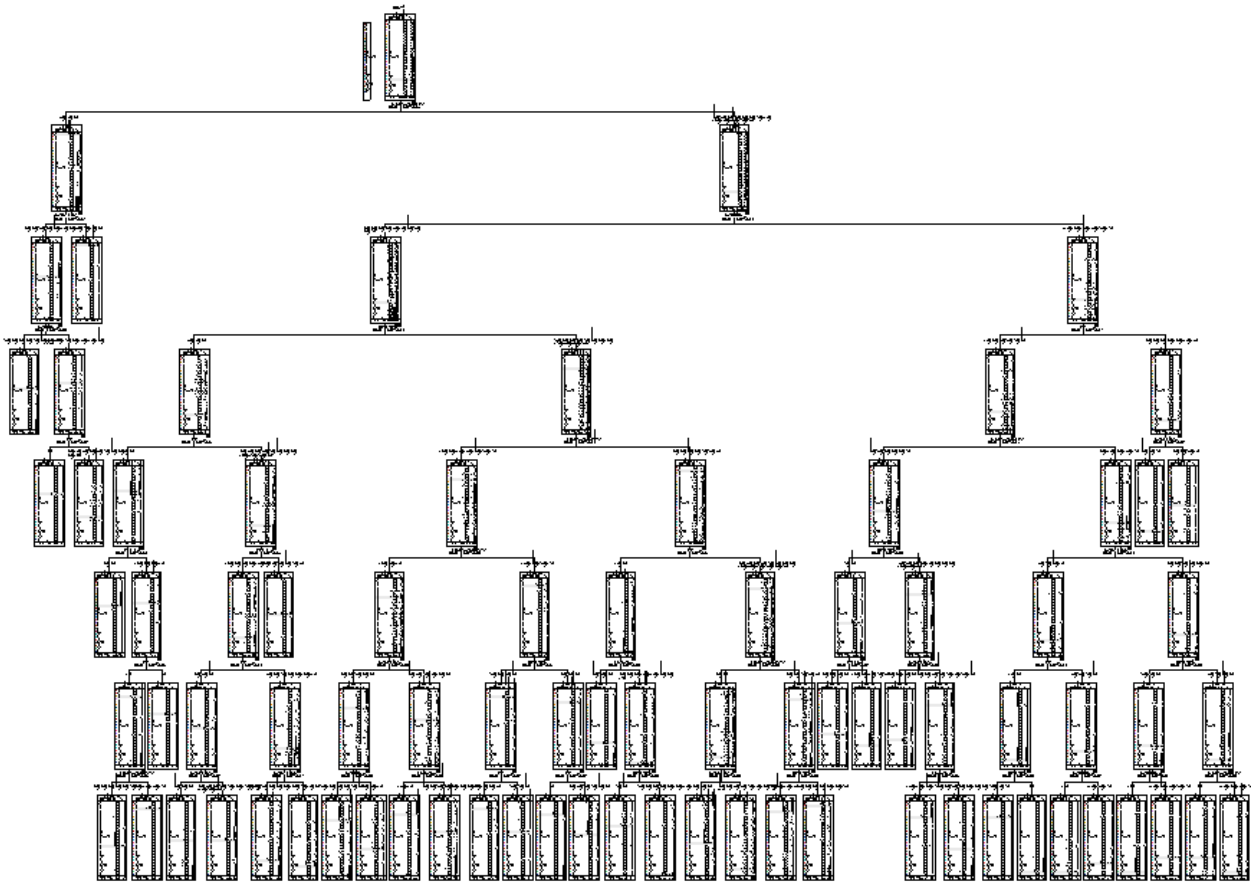
Keiland Pullen
Assignment 4
DSC 441 – Fall 2021

Training Sample #2



Keiland Pullen
DSC 441 – Fall 2021

Testing Sample #2



Risk

Sample	Estimate	Std. Error
Training	.450	.004
Test	.451	.006

Growing Method: CRT

Dependent Variable: alphabet

Keiland Pullen
DSC 441 – Fall 2021

Assignment 4

Classification #2

		Classification																							
		Predicted																							
Sample	Observed	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
Training	A	3	5	0	4	0	0	1	51	9	5	0	17	0	2	9	0	0	0	0	0	20	0	0	76.6%
	B	19	0	0	7	0	1	0	2	0	0	0	83	1	8	0	4	0	1	0	0	162	0	36	38.9%
	C	0	1	6	58	0	0	0	89	1	0	0	7	0	8	0	0	0	8	0	0	5	6	0	62.5%
	D	357	0	0	11	0	0	0	0	0	0	1	3	0	0	11	0	0	42	0	0	92	0	0	62.7%
	E	0	275	0	109	0	3	0	53	4	0	0	0	0	29	0	2	0	0	0	0	88	0	0	48.8%
	F	41	0	230	5	0	1	0	0	0	0	2	0	23	14	1	0	65	9	1	1	47	67	0	42.0%
	G	2	0	0	287	0	0	0	31	6	0	7	24	0	81	2	0	0	0	0	0	47	0	2	53.0%
	H	18	1	0	23	0	0	0	43	0	0	0	17	1	18	11	0	0	17	0	135	228	1	0	0.0%
	I	7	0	4	2	0	418	7	16	0	0	0	0	2	3	5	0	0	0	0	0	34	0	4	77.1%
	J	7	1	13	0	0	6	378	2	0	1	0	5	3	6	0	0	0	6	0	0	44	0	0	72.3%
	K	2	9	0	35	0	0	0	264	6	0	22	0	0	0	10	0	0	21	0	46	90	0	0	52.1%
	L	0	23	0	9	0	9	20	46	392	0	0	5	0	1	14	0	0	1	0	0	28	0	0	71.3%
	M	10	4	0	1	0	0	0	20	2	343	25	10	0	0	6	0	0	14	0	2	87	0	0	63.5%
	N	20	0	0	4	0	0	0	9	1	0	395	8	1	14	30	0	0	4	2	9	39	7	0	71.0%
	O	32	0	0	64	0	0	0	5	0	0	3	230	0	46	1	0	0	35	0	11	79	0	0	44.2%
	P	75	0	27	9	0	30	0	0	0	0	0	5	283	9	5	0	14	0	2	4	31	2	0	51.1%
	Q	1	14	0	43	0	0	0	30	3	0	3	83	1	311	4	0	2	0	1	7	53	8	0	54.8%
	R	40	0	0	20	0	0	0	32	4	0	0	62	2	6	102	0	7	0	0	1	213	0	0	19.8%
	S	4	2	0	39	0	12	0	29	8	0	0	58	1	22	24	137	0	2	0	1	190	0	2	24.3%
	T	5	0	9	12	0	1	0	34	0	0	0	14	0	3	0	1	384	0	3	0	28	46	0	69.6%
	U	2	0	0	44	0	0	0	0	3	5	41	3	0	24	0	0	0	393	0	2	56	7	0	67.3%
	V	13	0	2	4	0	0	0	3	0	0	14	1	3	4	0	0	8	1	272	132	47	20	0	51.9%
	W	5	0	0	0	0	0	0	0	0	0	29	3	0	9	0	0	0	3	12	363	67	8	0	72.7%
	X	5	22	0	0	0	0	0	86	0	0	0	0	0	0	21	0	0	6	0	0	365	1	0	69.7%
	Y	36	0	0	0	0	9	0	23	0	0	5	10	16	15	2	6	21	6	5	5	43	334	0	58.7%
	Z	0	66	3	33	0	0	0	9	0	0	0	0	0	11	10	19	0	0	0	0	79	1	277	53.2%
	Overall Percentage	5.0%	3.0%	2.1%	5.9%	0.0%	3.5%	2.9%	6.2%	3.1%	2.5%	3.9%	4.6%	2.4%	4.6%	1.9%	1.2%	3.6%	4.1%	2.1%	5.1%	16.0%	3.6%	2.3%	55.0%

Test	A	3	4	0	4	0	0	1	16	5	2	0	4	0	8	7	0	0	0	0	0	9	0	0	74.5%
	B	5	0	0	1	0	0	0	1	0	0	0	47	0	6	0	1	0	1	0	0	81	0	10	35.2%
	C	0	0	2	23	0	0	0	40	1	0	0	4	0	8	0	0	1	3	0	0	2	0	1	63.0%
	D	144	0	0	2	0	0	0	1	0	0	1	2	0	0	6	0	0	19	0	0	40	0	0	61.0%
	E	0	91	0	31	0	1	0	29	2	0	0	0	0	15	1	0	0	0	0	0	31	3	0	44.6%
	F	21	0	98	1	0	1	0	0	0	0	0	0	6	2	0	0	29	8	2	1	21	27	0	42.1%
	G	1	1	0	143	0	0	0	15	1	0	1	7	0	27	1	0	0	0	0	0	19	0	1	61.9%
	H	11	0	0	13	0	0	0	14	0	0	0	3	1	11	3	0	0	11	0	50	89	0	0	0.0%
	I	3	0	2	0	0	162	1	8	0	0	0	0	0	3	1	0	0	0	0	0	13	0	2	76.4%
	J	1	0	3	0	0	4	165	1	0	0	1	2	1	1	2	0	0	2	0	0	25	0	1	73.7%
	K	5	2	0	18	0	0	0	111	4	0	13	0	0	0	6	0	0	12	0	20	39	0	0	47.8%
	L	0	11	0	4	0	3	11	17	143	0	0	4	0	1	10	0	0	0	0	0	5	0	0	67.8%
	M	5	3	0	2	0	0	0	12	3	170	7	1	0	0	1	0	0	10	0	3	32	0	0	67.5%
	N	13	0	0	0	0	0	0	2	1	1	151	5	1	5	10	0	0	5	0	4	19	6	0	66.5%
	O	9	0	0	27	0	0	0	2	0	0	2	114	2	23	0	0	0	14	0	7	28	0	0	48.9%
	P	26	0	17	9	0	21	0	0	0	0	0	6	130	4	2	0	6	0	0	0	10	2	0	52.2%
	Q	0	8	0	9	0	0	0	18	1	0	2	26	0	111	1	0	1	0	1	2	28	5	0	51.6%
	R	14	0	0	4	0	0	0	16	5	0	0	23	3	3	50	0	6	2	0	0	110	0	0	20.5%
	S	4	1	0	11	0	4	0	6	4	0	0	17	0	10	13	37	0	1	0	1	59	0	6	20.0%
	T	4	0	6	6	0	0	0	12	0	0	0	7	0	1	1	1	171	0	1	1	15	15	0	70.4%
	U	4	0	0	23	0	0	0	2	1	3	18	1	0	6	0	0	0	145	0	2	22	2	0	63.3%
	V	7	0	0	0	0	0	0	1	0	0	9	0	1	1	0	0	3	0	129	58	18	13	0	53.8%
	W	1	0	0	1	0	0	0	0	0	0	18	2	0	2	0	0	0	7	0	186	32	4	0	73.5%
	X	3	7	0	0	0	0	0	44	0	0	0	0	0	0	7	0	0	1	0	0	173	2	0	67.6%
	Y	14	0	0	0	0	2	0	7	0	0	0	7	7	6	0	2	7	4	3	0	14	129	0	59.4%
	Z	0	32	3	12	0	0	0	3	0	0	0	0	0	4	3	5	3	0	0	0	33	0	106	49.8%
	Overall Percentage	5.0%	2.7%	2.2%	5.8%	0.0%	3.3%	3.0%	6.3%	2.9%	3.0%	3.7%	4.7%	2.5%	4.3%	2.1%	0.8%	3.8%	4.1%	2.3%	5.6%	16.3%	3.5%	2.1%	54.9%

Growing Method: CRT
Dependent Variable: alphabet

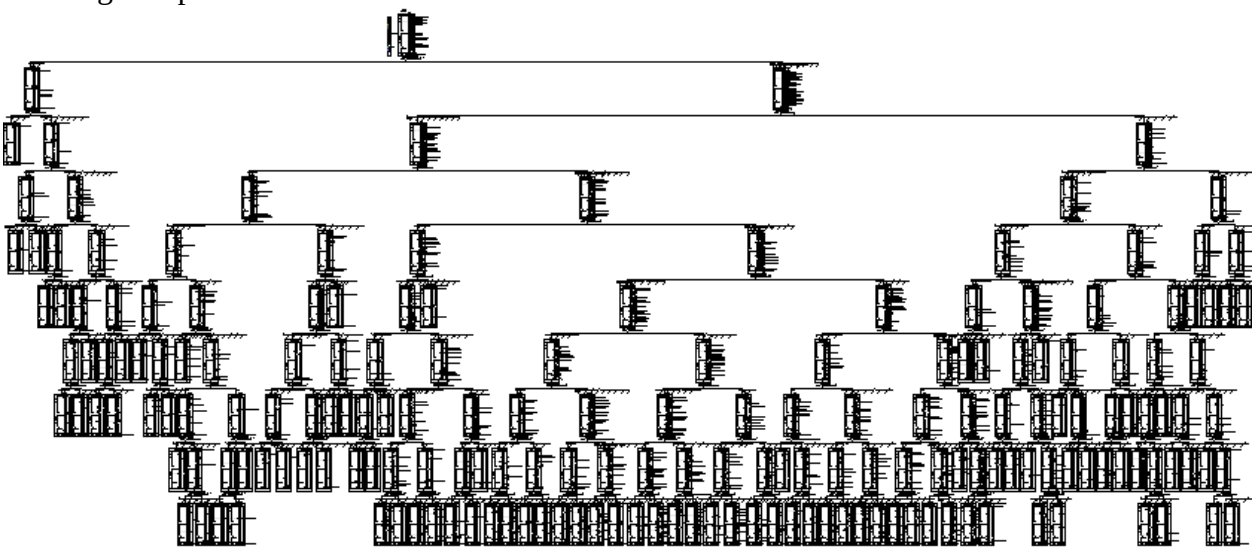
Model Summary #3

Specifications	Growing Method	CRT
	Dependent Variable	alphabet
	Independent Variables	horizonal, vertical, width, height, onpix, meanXpixels, meanYpixels, meanXvariance, meanYvariance, xybarmean, x2ybrMeanXXY, xy2brMeanXYY, xege, xegvy, edgeCount, yegvx
	Validation	Split Sample
	Maximum Tree Depth	9
	Minimum Cases in Parent Node	50
	Minimum Cases in Child Node	15
Results	Independent Variables Included	x2ybrMeanXXY, meanYpixels, xybarmean, meanXpixels, horizonal, meanYvariance, xege, xy2brMeanXYY, meanXvariance, xegvy, edgeCount, yegvx, width, height, vertical, onpix
	Number of Nodes	191
	Number of Terminal Nodes	96
	Depth	9

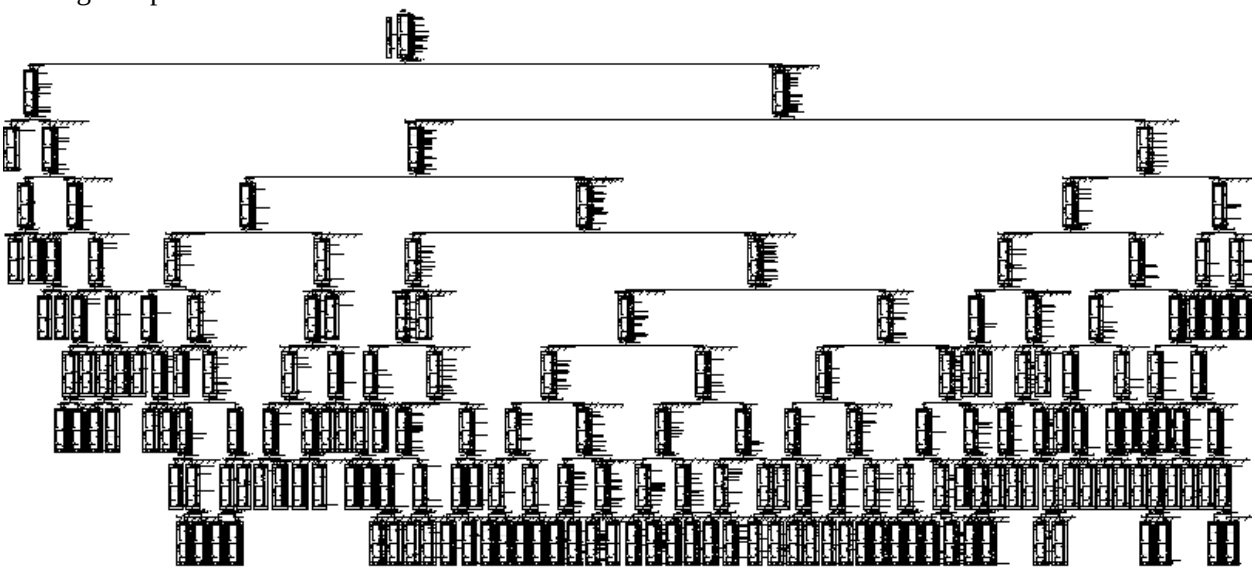
Keiland Pullen
DSC 441 – Fall 2021

Assignment 4

Training Sample #3



Testing Sample #3



Risk

Sample	Estimate	Std. Error
Training	.339	.004
Test	.356	.006

Growing Method: CRT

Dependent Variable: alphabet

Keiland Pullen
DSC 441 – Fall 2021

Assignment 4

Classification #3

		Classification																									
		Predicted																									
Sample	Observed	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
Training	A	0	7	6	0	10	0	0	0	0	0	5	0	2	0	13	64	12	0	0	0	0	3	2	0	77.6%
	B	0	67	3	3	11	0	0	0	0	0	0	0	2	6	1	35	33	0	0	0	0	0	0	0	71.5%
	C	278	4	8	7	139	1	0	5	0	3	0	0	5	0	4	69	5	0	7	0	0	4	0	0	51.4%
	D	0	477	6	2	8	2	0	0	1	0	3	1	8	2	0	18	4	0	0	0	0	0	0	6	82.7%
	E	1	0	296	3	82	3	0	0	0	14	0	0	0	0	11	55	45	0	2	0	0	25	0	4	54.6%
	F	0	73	0	401	6	0	0	2	0	0	0	2	0	17	0	23	19	11	1	1	0	0	13	0	69.7%
	G	6	22	5	2	393	0	0	0	0	0	0	0	2	2	39	49	19	0	0	0	0	0	0	1	71.5%
	H	0	60	3	24	21	233	0	0	8	0	0	0	88	1	1	52	0	0	2	0	0	11	0	2	44.6%
	I	13	23	1	16	2	1	414	10	0	0	1	0	0	6	2	4	11	0	0	0	0	5	0	3	77.7%
	J	2	52	3	25	1	0	3	388	1	5	0	1	7	0	5	5	8	0	0	0	0	6	0	0	73.9%
	K	0	17	2	4	64	84	0	0	183	4	2	1	27	0	0	76	0	0	0	47	0	16	0	1	34.2%
	L	1	1	42	1	10	5	0	0	2	406	5	0	0	0	4	31	4	0	0	0	0	12	0	1	76.6%
	M	0	49	0	1	4	0	0	0	0	0	419	9	19	1	5	43	7	0	1	0	4	0	0	0	73.1%
	N	0	53	0	15	3	0	0	0	0	0	1	449	11	6	0	20	1	0	2	1	4	0	0	0	78.9%
	O	0	46	7	0	65	0	0	0	0	0	0	0	295	1	16	40	14	0	0	12	0	0	0	0	57.5%
	P	0	60	0	47	14	0	1	4	0	0	0	3	0	391	0	16	6	2	2	0	1	0	3	0	69.3%
	Q	1	14	34	4	38	8	0	0	0	5	0	13	20	1	314	18	37	2	0	4	0	16	0	1	58.1%
	R	0	78	11	0	15	6	0	0	4	1	1	1	4	4	13	330	9	10	0	0	0	0	0	1	61.6%
	S	0	72	39	10	8	27	0	0	0	0	0	0	3	0	13	26	297	0	1	1	0	5	1	12	56.0%
	T	0	22	19	24	12	1	0	0	5	3	0	0	0	1	3	27	8	394	0	2	0	9	27	0	70.5%
	U	0	18	0	1	46	0	0	0	0	0	12	37	27	0	17	15	0	0	354	4	0	0	3	0	66.3%
	V	0	6	0	11	3	0	0	0	0	0	0	24	0	3	3	39	7	9	0	395	5	0	17	0	75.4%
	W	0	8	0	4	1	0	0	0	0	0	4	34	18	0	9	26	6	0	3	12	398	0	5	0	74.8%
	X	0	57	50	15	0	11	0	0	0	5	0	0	34	0	0	80	7	0	2	0	0	286	0	0	51.3%
	Y	1	29	0	24	3	0	0	0	0	1	0	3	7	2	14	22	16	22	0	6	0	2	383	0	71.6%
	Z	0	44	43	16	20	0	0	3	1	1	0	0	0	0	0	20	19	0	2	0	0	4	0	340	65.9%
	Overall Percentage	2.1%	9.6%	4.1%	4.7%	6.9%	2.7%	3.0%	2.9%	1.5%	3.2%	3.2%	4.1%	4.1%	3.1%	3.4%	8.5%	4.2%	3.2%	2.7%	3.4%	2.9%	2.9%	3.2%	2.6%	66.1%
Test	A	0	3	4	0	2	0	0	0	1	1	4	0	1	0	4	20	5	0	0	0	0	3	0	0	79.6%
	B	0	29	0	0	6	1	0	0	0	0	0	0	1	1	0	11	15	0	0	0	0	0	0	0	68.2%
	C	111	2	2	4	38	0	0	2	0	1	0	0	4	0	2	25	2	0	2	0	0	0	0	0	56.9%
	D	0	185	2	0	5	0	0	0	0	0	1	1	7	0	0	11	3	0	0	0	0	0	0	1	91.1%
	E	1	0	126	3	50	0	0	0	0	8	0	0	0	0	3	17	12	0	1	0	0	5	0	0	55.8%
	F	0	22	0	148	0	1	0	1	0	0	0	2	0	7	0	6	4	2	0	2	0	0	3	0	74.0%
	G	1	15	1	0	144	0	0	0	0	0	0	0	3	1	20	26	8	0	0	0	0	0	0	0	64.6%
	H	0	27	0	8	15	71	0	0	3	0	0	0	49	1	0	23	0	0	0	0	0	1	0	1	33.6%
	I	3	12	1	10	3	0	166	1	0	0	0	0	2	1	0	1	7	0	0	0	0	3	0	3	75.1%
	J	2	14	4	12	0	0	0	163	0	2	1	0	2	0	2	1	5	0	0	0	0	4	0	0	73.4%
	K	0	5	3	0	23	31	0	0	68	4	2	0	6	0	0	33	0	0	0	18	0	7	0	0	33.3%
	L	1	0	14	0	3	2	0	1	0	175	1	0	1	0	0	17	9	0	0	0	0	4	0	0	75.8%
	M	0	15	0	0	0	0	0	0	0	0	179	4	7	0	0	6	0	0	1	0	2	0	0	0	81.7%
	N	0	21	0	6	1	0	0	0	0	0	2	157	3	5	0	14	0	0	0	2	1	0	1	0	73.4%
	O	0	26	4	0	29	0	0	0	0	0	0	0	118	1	8	25	12	0	0	6	2	0	0	0	49.2%
	P	0	31	0	23	4	0	0	1	0	0	0	2	1	156	0	6	4	2	0	2	0	0	0	0	65.3%
	Q	0	9	12	5	17	5	0	0	0	2	0	4	7	0	134	12	17	0	0	5	0	6	0	0	55.1%
	R	0	34	5	0	9	7	0	0	1	0	1	0	0	1	3	119	7	3	0	0	0	0	0	0	53.6%
	S	0	28	20	4	4	10	0	0	0	0	0	0	3	1	6	11	107	0	1	1	0	1	1	5	49.1%
	T	0	4	18	9	7	0	0	1	1	1	0	0	0	0	0	7	6	162	0	2	0	0	17	0	69.6%
	U	0	8	0	1	26	0	0	0	0	0	11	22	12	0	10	7	1	0	180	0	0	0	0	0	64.5%
	V	0	3	0	3	6	0	0	0	0	0	0	11	1	1	3	11	0	0	0	188	1	0	7	0	78.3%
	W	0	3	0	6	0	0	0	0	0	0	1	16	9	0	4	11	5	0	3	2	156	0	1	0	70.9%
	X	0	28	19	7	0	1	0	0	0	1	0	0	9	0	0	32	5	0	2	0	0	120	0	1	52.4%
	Y	1	10	0	14	2	0	0	0	0	1	1	1	3	4	3	11	7	10	0	6	0	1	176	0	70.1%
	Z	0	14	22	4	10	0	0	1	0	3	0	0	0	0	0	11	7	0	0	0	0	1	1	142	65.1%
	Overall Percentage	2.0%	9.3%	4.4%	4.6%	6.9%	2.2%	2.8%	2.9%	1.3%	3.4%	3.5%	3.8%	4.2%	3.1%	3.4%	8.1%	4.2%	3.1%	3.2%	4.0%	2.8%	2.7%	3.5%	2.6%	64.4%

Growing Method: CRT

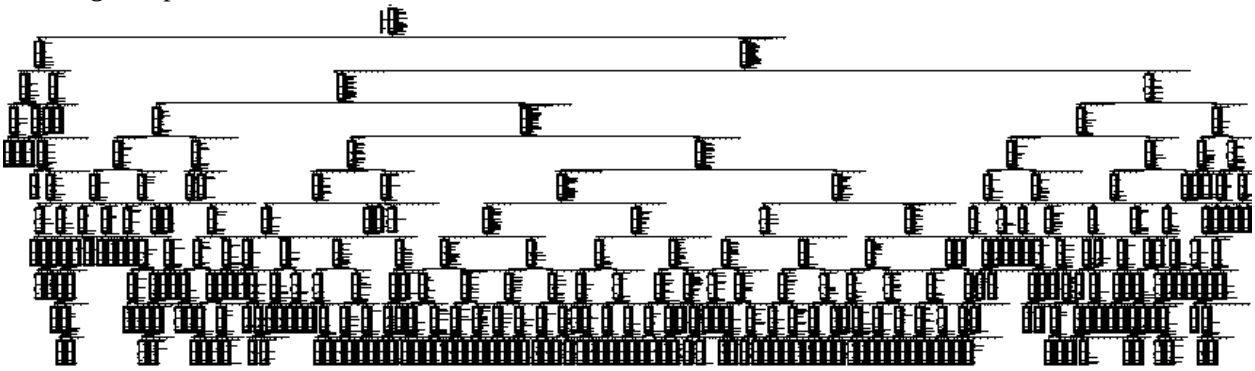
Model Summary #4

Specifications	Growing Method	CRT
	Dependent Variable	alphabet
	Independent Variables	horizonal, vertical, width, height, onpix, meanXpixels, meanYpixels, meanXvariance, meanYvariance, xybarmean, x2ybrMeanXXY, xy2brMeanXYY, xege, xegvy, edgeCount, yegvx
	Validation	Split Sample
	Maximum Tree Depth	10
	Minimum Cases in Parent Node	25
	Minimum Cases in Child Node	7
Results	Independent Variables Included	x2ybrMeanXXY, meanYpixels, xybarmean, meanXpixels, horizonal, width, meanYvariance, xege, xy2brMeanXYY, meanXvariance, edgeCount, xegvy, yegvx, onpix, vertical, height
	Number of Nodes	317
	Number of Terminal Nodes	159
	Depth	10

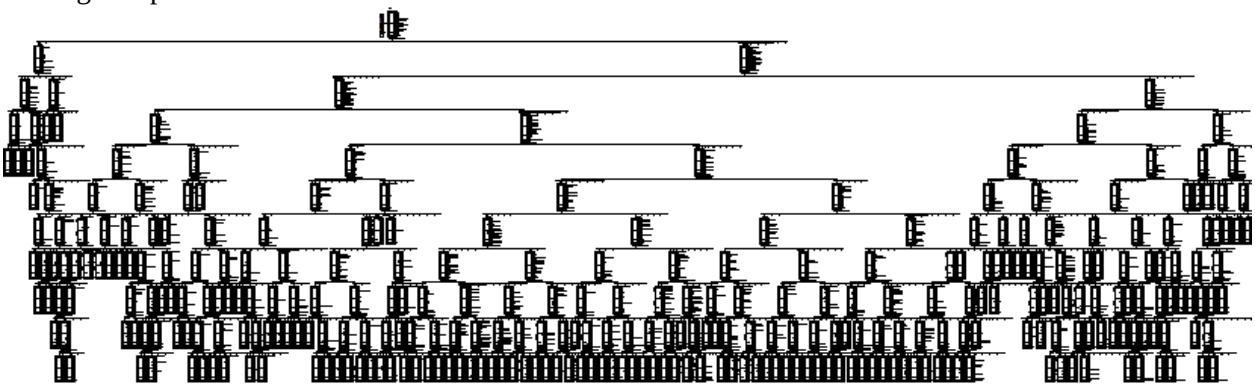
Keiland Pullen
DSC 441 – Fall 2021

Assignment 4

Training Sample #4



Testing Sample #4



Keiland Pullen
DSC 441 – Fall 2021

Classification #4

Classification

Sample	Observed	Predicted 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
Training	A	0	0	1	1	14	4	0	0	1	3	0	2	15	6	5	65	7	0	0	0	0	0	0	0	77.7%
	B	0	13	2	3	0	24	0	0	1	0	0	2	48	7	20	33	10	0	0	0	0	10	1	0	67.5%
	C	371	0	2	7	13	12	1	4	3	0	0	0	37	2	3	68	5	0	1	0	0	0	0	0	70.0%
	D	0	425	0	0	0	21	0	0	10	0	0	6	46	1	1	21	19	0	0	0	1	25	0	0	72.6%
	E	1	0	324	2	33	18	1	0	29	0	0	0	24	1	24	49	12	2	0	0	0	6	5	9	60.0%
	F	0	6	0	393	3	19	0	1	1	0	0	2	12	24	2	2	15	4	0	1	3	33	18	0	72.1%
	G	12	5	0	0	327	69	0	0	1	1	1	2	39	4	36	30	13	0	0	0	0	3	1	0	59.7%
	H	1	41	0	2	0	310	0	0	39	0	1	0	36	5	1	53	11	0	0	0	0	20	0	0	58.5%
	I	0	0	1	8	2	4	411	1	2	0	0	0	17	8	6	4	6	0	0	0	0	13	1	4	81.2%
	J	0	6	0	6	2	0	6	392	11	7	1	2	37	11	8	14	4	0	0	0	0	11	1	0	75.1%
	K	0	2	23	0	0	58	0	0	322	0	2	0	1	0	0	60	5	1	1	1	0	27	0	0	63.9%
	L	0	1	0	0	6	13	2	3	20	399	0	0	9	0	4	70	5	0	0	0	0	4	0	0	73.9%
	M	1	2	0	0	6	28	0	0	11	4	441	4	10	1	6	11	0	0	2	0	4	0	0	0	92.7%
	N	0	7	0	2	0	10	0	0	8	1	5	399	37	8	1	18	1	0	15	1	7	2	13	0	74.4%
	O	0	27	0	0	8	47	0	0	15	0	0	1	364	1	33	14	2	0	0	0	0	3	0	0	68.8%
	P	3	1	0	23	0	11	1	5	0	0	1	5	38	452	0	1	0	0	0	4	3	4	1	0	79.9%
	Q	2	5	1	0	8	37	0	1	35	0	0	9	27	3	370	40	6	0	0	2	1	11	0	0	65.7%
	R	0	26	1	0	8	43	0	0	13	0	1	6	59	1	13	344	6	0	0	1	0	0	0	0	63.4%
	S	0	4	20	0	7	19	0	0	10	0	0	6	10	3	4	48	327	0	1	0	0	30	4	7	63.9%
	T	0	1	4	4	0	6	0	1	22	1	0	0	22	7	3	9	2	433	0	2	1	24	10	0	78.2%
	U	0	4	24	1	10	29	0	0	8	0	22	7	35	0	3	8	0	0	400	1	0	0	6	0	71.4%
	V	1	1	0	6	1	21	0	0	0	0	0	1	7	5	11	12	0	2	1	438	18	0	7	0	82.0%
	W	0	2	0	6	5	42	0	0	4	0	5	5	4	6	9	1	0	0	0	6	427	0	0	0	81.2%
	X	0	7	34	3	0	38	0	0	14	0	0	1	16	3	0	64	7	0	0	0	0	373	0	0	66.4%
	Y	0	5	0	1	13	3	0	0	3	2	0	3	10	15	17	21	10	13	0	10	0	0	429	0	77.3%
	Z	0	1	7	1	2	2	0	1	5	1	0	3	23	7	6	13	44	1	0	0	0	8	6	388	74.3%
	Overall Percentage	2.8%	4.2%	3.2%	3.3%	3.3%	6.3%	3.0%	2.9%	4.2%	3.0%	3.4%	3.3%	7.0%	4.1%	4.2%	7.6%	3.7%	3.2%	3.0%	3.3%	3.3%	4.3%	3.6%	2.9%	71.6%
Test	A	0	0	0	0	6	1	0	0	1	0	1	0	2	0	7	27	3	0	0	0	0	0	0	0	79.4%
	B	0	9	0	2	1	9	0	0	2	0	0	2	19	4	7	7	6	0	0	0	0	5	0	0	68.4%
	C	125	1	3	1	8	9	0	3	1	0	0	0	9	1	3	34	4	0	2	0	0	0	1	0	60.7%
	D	0	152	3	0	0	4	0	0	3	0	0	8	23	0	1	7	6	0	0	0	0	8	0	0	69.1%
	E	1	0	125	1	10	13	1	0	12	0	0	0	10	0	11	25	11	0	0	0	0	4	1	2	54.8%
	F	1	3	0	156	0	13	0	2	0	0	1	0	6	15	2	0	5	4	0	1	0	15	4	0	67.8%
	G	5	1	1	0	131	33	0	0	0	0	1	0	17	3	9	14	6	0	0	0	0	0	1	0	58.2%
	H	0	13	1	1	0	121	0	0	17	0	1	0	13	3	0	13	7	0	0	0	0	12	0	0	59.3%
	I	4	1	0	3	0	2	191	0	0	0	0	0	10	6	4	8	2	0	0	0	0	7	1	2	77.0%
	J	0	2	0	0	3	0	3	170	0	5	0	3	13	5	1	6	1	0	1	0	0	8	1	1	75.6%
	K	0	0	12	0	0	30	0	0	158	0	1	0	0	0	0	24	3	0	0	0	0	7	0	0	67.2%
	L	0	1	0	0	2	5	0	3	11	165	0	0	5	0	2	24	2	0	0	0	0	1	0	0	74.7%
	M	0	1	0	0	1	9	0	0	7	1	221	1	4	1	1	1	0	0	3	0	3	2	0	0	85.3%
	N	0	6	0	1	0	4	0	0	5	1	1	178	14	8	0	7	0	0	6	1	8	0	4	0	72.1%
	O	0	10	1	0	2	21	0	0	4	0	0	0	153	0	13	10	1	0	0	0	0	0	0	0	68.3%
	P	0	4	0	12	0	5	0	0	0	0	0	3	15	184	0	0	0	0	0	2	1	1	3	0	77.6%
	Q	0	4	0	0	6	15	0	1	15	0	0	8	11	0	123	18	4	0	0	1	0	2	1	4	55.9%
	R	0	7	1	0	4	16	0	0	10	0	1	3	23	1	6	132	5	0	0	0	0	0	0	0	61.4%
	S	0	3	7	0	3	14	0	0	9	0	0	4	4	0	6	19	137	0	1	0	0	17	4	4	58.1%
	T	0	1	2	1	3	0	0	0	7	1	0	0	10	4	0	0	1	190	0	0	2	9	9	0	78.8%
	U	0	2	7	0	14	10	0	0	7	0	7	2	10	0	1	11	0	0	178	1	1	0	2	0	70.4%
	V	0	2	0	1	0	6	0	0	0	0	0	4	1	5	1	16	0	1	1	178	6	0	5	0	77.4%
	W	0	1	0	1	9	20	0	0	0	0	3	4	2	4	2	1	0	0	0	2	175	0	1	0	77.4%
	X	0	2	14	2	0	15	0	0	7	0	0	0	3	1	1	25	4	1	0	0	0	147	1	0	65.3%
	Y	3	1	0	0	9	2	0	0	5	0	1	2	4	6	4	11	4	10	0	6	0	0	163	0	70.6%
	Z	0	0	3	0	0	1	0	0	2	2	0	0	6	0	3	9	12	2	0	0	0	2	11	157	74.1%
	Overall Percentage	2.3%	3.8%	3.0%	3.1%	3.6%	6.3%	3.3%	3.0%	4.7%	2.9%	4.0%	3.7%	6.5%	4.2%	3.5%	7.5%	3.8%	3.5%	3.2%	3.2%	3.3%	4.1%	3.6%	2.9%	69.7%

Growing Method: CRT
Dependent Variable: alphabet

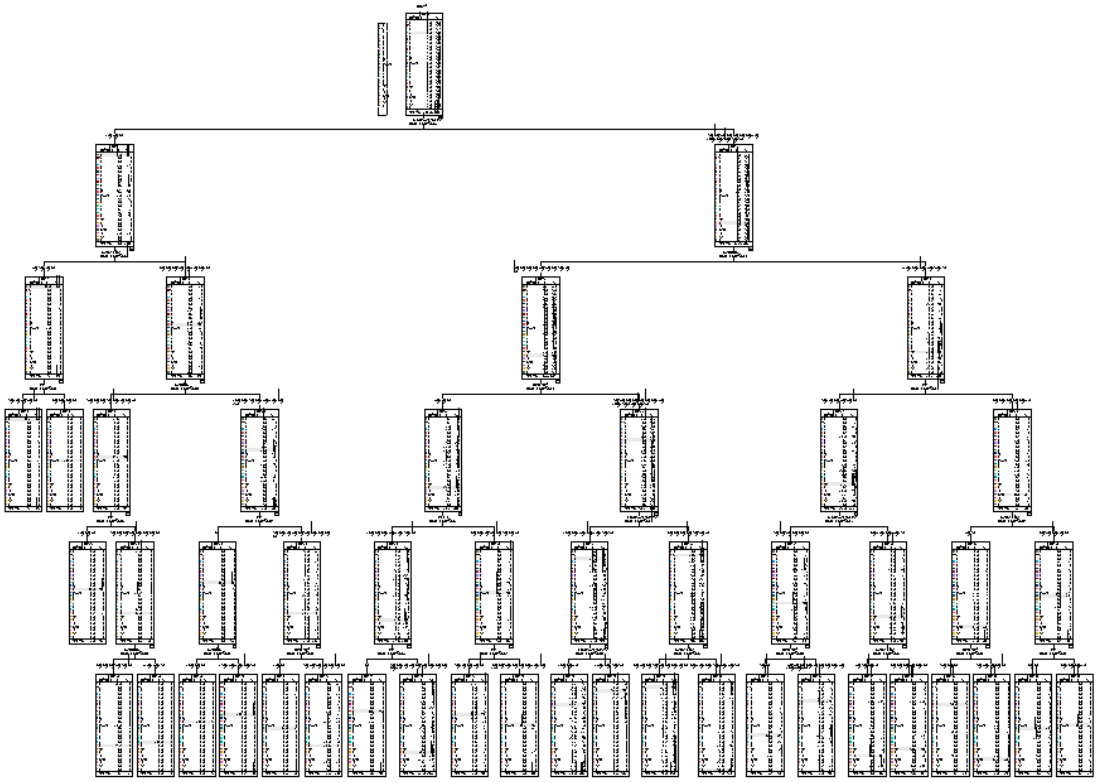
Model Summary #5

Specifications	Growing Method	CRT
	Dependent Variable	alphabet
	Independent Variables	horizonal, vertical, width, height, onpix, meanXpixels, meanYpixels, meanXvariance, meanYvariance, xybarmean, x2ybrMeanXXY, xy2brMeanXYY, xege, xegvy, edgeCount, yegvx
	Validation	Split Sample
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	15
	Minimum Cases in Child Node	4
Results	Independent Variables Included	x2ybrMeanXXY, meanYpixels, xybarmean, meanXpixels, horizonal, width, meanYvariance, xege, xy2brMeanXYY, meanXvariance, edgeCount, xegvy, yegvx, vertical, height, onpix
	Number of Nodes	49
	Number of Terminal Nodes	25
	Depth	5

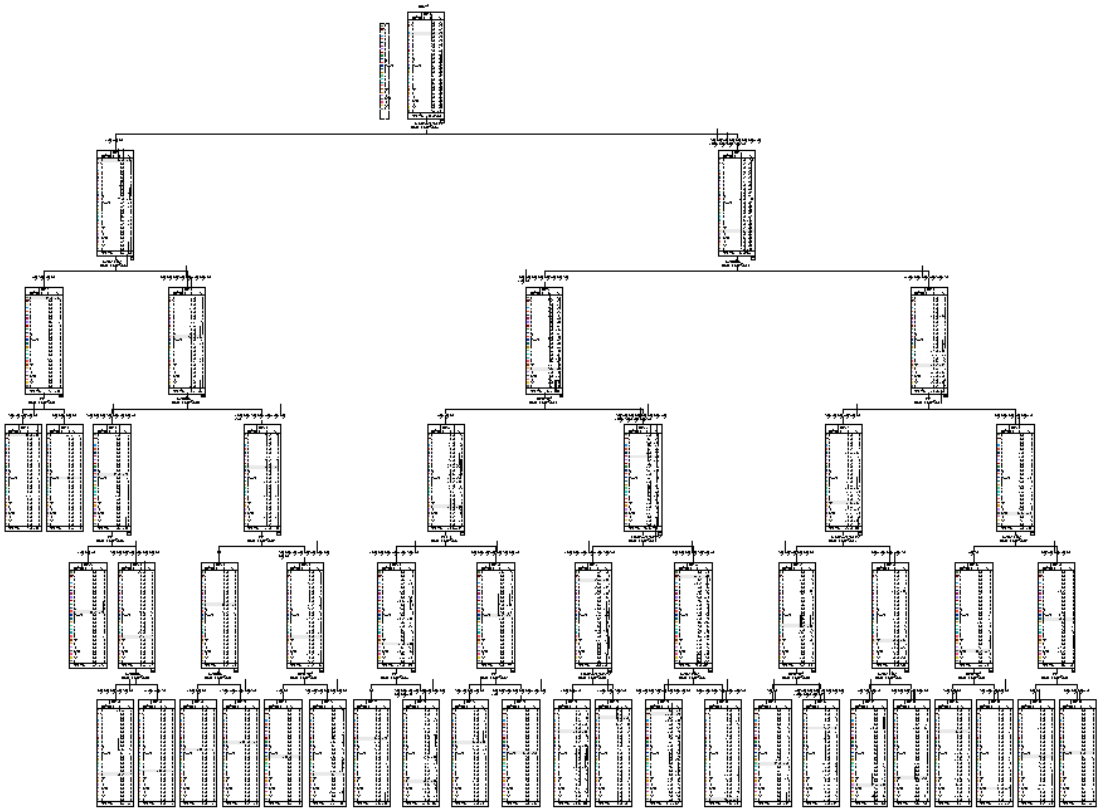
Keiland Pullen

DSC 441 – Fall 2021

Training Sample #5



Testing Sample #5



Classification #5

Classification

Sample	Observed	Predicted 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
Training	A	20	0	0	0	0	51	0	0	4	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	76.1%
	B	0	0	0	1	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	95.8%
	C	309	0	0	6	126	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	43	0	0	0	0	0	60.2%
	D	0	0	0	0	12	0	0	1	0	0	0	0	2	0	0	0	0	0	12	0	41	0	0	0	0	49	0.0%
	E	138	0	0	5	302	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	12	0.0%
	F	0	0	0	138	6	0	0	0	0	0	0	1	0	9	0	0	0	0	0	126	24	103	0	0	0	0	25.6%
	G	46	0	0	2	398	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	72.8%
	H	0	0	0	15	78	0	0	2	0	0	0	0	0	1	0	0	8	1	153	0	0	0	0	0	0	0	0.0%
	I	3	0	0	5	23	0	335	5	0	0	0	0	0	1	0	0	3	0	1	0	0	0	0	0	0	3	83.6%
	J	0	0	0	11	5	0	8	383	0	0	0	0	0	0	0	0	1	0	5	0	0	0	0	0	0	0	73.0%
	K	40	0	0	0	280	0	0	6	0	0	0	1	0	0	0	0	12	0	83	0	0	0	0	0	0	0	0.0%
	L	28	0	0	0	78	0	4	16	0	369	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	68.7%
	M	0	0	0	0	27	0	0	1	0	0	403	2	0	0	0	0	11	0	8	0	3	0	0	0	0	0	73.4%
	N	0	0	0	12	11	0	0	1	0	0	11	40	0	1	0	0	9	0	362	9	3	0	0	0	0	0	7.4%
	O	0	0	0	1	123	0	0	0	0	0	0	0	0	0	0	0	0	0	46	0	1	0	0	5	0.0%		
	P	0	0	0	44	11	0	0	0	0	0	0	3	0	263	0	0	0	0	18	16	3	0	0	0	0	0	47.0%
	Q	4	0	0	19	379	0	0	0	0	0	0	5	0	3	0	0	1	1	6	10	0	0	0	0	0	0	0.0%
	R	0	0	0	81	51	0	0	2	0	0	0	1	0	15	0	7	28	0	0	0	0	0	0	0	0	0	1.3%
	S	0	0	0	4	80	0	0	0	0	1	0	0	0	0	0	0	38	0	3	0	0	0	0	0	55	7.2%	
	T	0	0	0	8	52	0	0	0	0	0	0	0	0	23	0	0	0	389	0	26	0	0	0	8	70.1%		
	U	4	0	0	0	65	0	0	1	0	0	13	0	0	0	0	0	0	7	408	0	0	0	0	2	72.2%		
	V	0	0	0	2	9	0	0	0	0	0	0	2	0	4	0	0	0	20	146	311	4	0	0	0	56.3%		
	W	0	0	0	7	2	0	0	0	0	0	1	2	0	0	0	0	0	0	173	17	228	0	0	0	45.2%		
	X	0	0	0	0	105	0	0	1	0	0	0	0	0	0	0	0	0	18	3	2	0	0	0	0	82	0.0%	
	Y	0	0	0	17	24	0	0	0	0	0	0	1	0	0	0	0	0	0	197	29	146	0	0	0	8	0.0%	
	Z	0	0	0	19	126	0	0	0	0	0	0	0	0	0	0	0	0	9	2	0	0	0	0	0	294	54.1%	
	Overall Percentage	4.2%	0.0%	0.0%	2.8%	17.5%	0.0%	2.5%	3.0%	0.0%	2.6%	3.1%	0.4%	0.0%	2.3%	0.0%	0.0%	1.3%	5.5%	11.0%	4.4%	1.7%	0.0%	0.0%	3.7%	37.3%		
Test	A	8	0	0	0	21	0	0	5	0	0	1	0	0	0	0	0	7	0	0	0	0	0	0	0	75.0%		
	B	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97.5%		
	C	129	0	0	5	60	0	0	0	0	0	0	0	0	0	0	0	0	2	21	0	0	0	0	0	57.8%		
	D	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	4	0	19	0	0	0	0	16	0.0%		
	E	66	0	0	2	127	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	5	0.0%		
	F	0	0	0	61	0	0	0	0	0	0	0	3	0	5	0	0	0	62	8	44	0	0	0	0	25.8%		
	G	23	0	0	1	169	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	74.8%		
	H	0	0	0	6	25	0	0	0	0	0	0	0	0	1	0	0	5	0	58	0	0	0	0	0	0.0%		
	I	2	0	0	3	7	0	137	3	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	2	60.4%		
	J	0	0	0	8	3	0	2	164	0	0	0	0	0	0	0	0	1	0	5	0	0	0	0	0	73.9%		
	K	22	0	0	0	109	0	0	2	0	0	0	0	0	0	0	0	4	0	42	0	0	0	0	0	0.0%		
	L	8	0	0	0	32	0	8	10	0	135	0	0	0	0	0	0	10	0	0	0	0	0	0	0	60.3%		
	M	0	0	0	0	16	0	0	1	0	0	167	0	0	0	0	0	1	0	3	0	0	0	0	0	68.7%		
	N	0	0	0	7	4	0	0	0	0	0	7	16	0	1	0	0	1	0	160	6	1	0	0	0	6.7%		
	O	0	0	0	0	47	0	0	0	0	0	0	0	0	2	0	0	0	0	21	0	1	0	0	0	0.0%		
	P	0	0	0	20	7	0	1	0	0	0	0	1	0	120	0	0	0	4	14	1	0	0	0	0	49.4%		
	Q	2	0	0	5	143	0	0	0	0	0	0	1	0	0	0	0	2	0	1	5	1	0	0	0	0.0%		
	R	0	0	0	29	30	0	0	0	0	0	0	0	0	3	0	2	14	0	0	0	0	0	0	0	1.0%		
	S	0	0	0	4	32	0	0	0	0	1	0	0	0	1	0	0	9	0	2	0	0	0	0	27	4.1%		
	T	0	0	0	10	15	0	0	0	0	0	0	0	0	10	0	0	0	159	0	13	0	0	0	4	66.3%		
	U	1	0	0	0	33	0	0	3	0	0	6	0	0	0	0	0	0	2	178	0	0	0	0	0	71.8%		
	V	0	0	0	2	6	0	0	0	0	0	0	4	0	0	0	0	0	7	51	108	2	0	0	0	50.9%		
	W	0	0	0	2	1	0	0	0	0	0	1	2	0	0	0	0	0	0	85	8	113	0	0	0	45.6%		
	X	0	0	0	0	54	0	0	0	0	0	0	0	0	0	0	0	0	10	0	4	0	0	0	39	0.0%		
	Y	0	0	0	5	8	0	0	0	0	0	0	0	0	0	0	0	0	0	91	10	65	4	0	0	1	0.0%	
	Z	0	0	0	3	31	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	116	60.7%		
	Overall Percentage	4.4%	0.0%	0.0%	2.9%	16.6%	0.0%	2.5%	3.2%	0.0%	2.3%	3.1%	0.5%	0.0%	2.4%	0.0%	0.0%	1.3%	5.5%	11.5%	4.2%	2.1%	0.0%	0.0%	3.5%	36.8%		

For this problem, the best configuration was #4. The Classification matrix indicates that this configuration has the best overall percent of correct predictions for both Training and Test data.

Problem 2.b.)

For this problem, the best configuration was #4. The classification is found above. In the Test sample, the smallest number of correct predictions was 55% and the largest was 85%. In the Training sample, the smallest was 58.5% and the largest was 82%. Accuracy is indeed a good indicator of the performance of a model. However, it can be misleading. To correct this, it would ideal to adjust the algorithms and execute multiple tests until the ideal model is found.

Problem 2.c.)

The most important attributes were “x2ybr mean of x*x*x”, “y-bar mean y of pixels in box”, and “y2bar mean y variance. In my models, they were named “x2ybrMeanXXY”, “meanYvariance” and “MeanYpixels”.

Keiland Pullen
DSC 441 – Fall 2021

Assignment 4

Problem 3. 1.)

For this data set, the data was standardized because the original numeric values were of various units, from "x-box horizontal position of box" to means and variances. The z-scores of each transformed variable to be used.

Problem 3.2.)

K=1

Initial Cluster Centers

	Cluster	
	1	2
Zscore(horizonal)	-2.10303	3.64638
Zscore(vertical)	-2.12896	2.41027
Zscore(width)	-2.04613	3.91050
Zscore(height)	-2.37561	1.60409
Zscore(onpix)	-1.60059	1.59507
Zscore(meanXpixels)	.05065	3.01214
Zscore(meanYpixels)	1.07510	-2.36564
Zscore(meanXvariance)	-1.34405	.50787
Zscore(meanYvariance)	-.91505	-1.33506
Zscore(xybarmean)	-.51518	1.49421
Zscore(x2ybrMeanXXY)	1.72788	-2.07300
Zscore(xy2brMeanXYY)	.03415	.51476
Zscore(xege)	-.87728	2.12373
Zscore(xegvy)	1.72041	-4.74457
Zscore(edgeCount)	-1.43819	-1.04864
Zscore(yegvx)	.12298	.12298

Iteration History[a]

	Change in Cluster Centers	
Iteration	1	2
1	5.422	7.028
2	.426	.778
3	.288	.408
4	.194	.242
5	.130	.149
6	.089	.096

7	.059	.061
8	.040	.041
9	.029	.028
10	.018	.018
11	.014	.013
12	.010	.010
13	.005	.005
14	.002	.002
15	.001	.001
16	.002	.002
17	.001	.001
18	.001	.001
19	.001	.001
20	.001	.001
21	.001	.001
22	.001	.001
23	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 23. The minimum distance between initial centers is 14.459.

Final Cluster Centers

	Cluster	
	1	2
Zscore(horizonal)	-.73007	.69584
Zscore(vertical)	-.70460	.67157
Zscore(width)	-.73195	.69764
Zscore(height)	-.70029	.66746
Zscore(onpix)	-.72202	.68818
Zscore(meanXpixels)	-.11580	.11037
Zscore(meanYpixels)	.03295	-.03140
Zscore(meanXvariance)	-.00245	.00234
Zscore(meanYvariance)	.00815	-.00777
Zscore(xybarmean)	-.04329	.04126
Zscore(x2ybrMeanXXY)	.06167	-.05878
Zscore(xy2brMeanXYY)	.03420	-.03260
Zscore(xege)	-.47848	.45605

Zscore(xegvy)	-.01876	.01788
Zscore(edgeCount)	-.38039	.36255
Zscore(yegvx)	.04924	-.04693

Number of Cases in each Cluster

Cluster	1	9759.000
	2	10239.000
Valid		19998.000
Missing		.000

K=3

Initial Cluster Centers

	Cluster		
	1	2	3
Zscore(horizonal)	1.03301	3.64638	-1.58036
Zscore(vertical)	-.01065	2.41027	-2.12896
Zscore(width)	1.42857	3.91050	-2.04613
Zscore(height)	2.48847	1.60409	-2.37561
Zscore(onpix)	2.05159	1.59507	-1.60059
Zscore(meanXpixels)	1.53139	3.01214	-2.41727
Zscore(meanYpixels)	2.79548	-2.36564	-2.79573
Zscore(meanXvariance)	.13748	.50787	.50787
Zscore(meanYvariance)	-1.75506	-1.33506	-.49505
Zscore(xybarmean)	-2.12269	1.49421	-3.32832
Zscore(x2ybrMeanXXY)	.58762	-2.07300	-1.31282
Zscore(xy2brMeanXYY)	1.95657	.51476	-1.88827
Zscore(xege)	.83758	2.12373	-1.30600
Zscore(xegvy)	4.30640	-4.74457	-.21909
Zscore(edgeCount)	.50956	-1.04864	-1.43819
Zscore(yegvx)	1.35948	.12298	.12298

Iteration History[a]

	Change in Cluster Centers		
Iteration	1	2	3

1	6.546	6.714	6.254
2	.289	.550	.307
3	.141	.296	.101
4	.117	.208	.049
5	.157	.215	.026
6	.216	.242	.032
7	.263	.216	.043
8	.276	.176	.053
9	.349	.170	.062
10	.303	.121	.077
11	.239	.090	.079
12	.204	.076	.075
13	.153	.063	.073
14	.125	.051	.056
15	.083	.039	.043
16	.079	.034	.043
17	.062	.027	.034
18	.049	.026	.031
19	.040	.024	.026
20	.029	.018	.020
21	.023	.013	.017
22	.025	.015	.017
23	.016	.014	.016
24	.018	.009	.010
25	.013	.005	.008
26	.019	.007	.010
27	.014	.005	.006
28	.012	.006	.006
29	.016	.007	.008
30	.018	.009	.011
31	.018	.007	.012
32	.016	.009	.009
33	.016	.008	.009
34	.018	.010	.008
35	.023	.009	.011
36	.018	.006	.009
37	.019	.009	.010
38	.011	.008	.010
39	.009	.007	.006

40	.009	.004	.004
41	.007	.005	.003
42	.013	.003	.006
43	.014	.006	.008
44	.009	.005	.005
45	.008	.004	.004
46	.004	.002	.002
47	.005	.001	.002
48	.003	.002	.002
49	.003	.002	.001
50	.001	.001	.001
51	.000	.001	.001
52	.002	.000	.001
53	.002	.001	.000
54	.001	.000	.000
55	.002	.001	.000
56	.000	.001	.001
57	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 57. The minimum distance between initial centers is 12.600.

Final Cluster Centers

	Cluster		
	1	2	3
Zscore(horizonal)	-.05201	.74306	-.75618
Zscore(vertical)	-.09098	.70945	-.69947
Zscore(width)	-.06983	.77114	-.77612
Zscore(height)	-.16773	.71928	-.66804
Zscore(onpix)	-.40354	.84829	-.67579
Zscore(meanXpixels)	-.72195	.29674	.07998
Zscore(meanYpixels)	1.23086	-.20882	-.45002
Zscore(meanXvariance)	-.49727	.06490	.20236
Zscore(meanYvariance)	-.16107	-.01643	.10509
Zscore(xybarmean)	.49954	-.05242	-.21677
Zscore(x2ybrMeanXXY)	1.12767	-.23239	-.36893
Zscore(xy2brMeanXYY)	-.49636	.02805	.24077

Zscore(xege)	-.31921	.58534	-.44411
Zscore(xegvy)	.97529	-.15156	-.37125
Zscore(edgeCount)	-.69294	.53787	-.19041
Zscore(yegvx)	-.61629	.07927	.25202

Number of Cases in each Cluster

Cluster	1	4189.000
	2	8119.000
	3	7690.000
Valid		19998.000
Missing		.000

K = 5

Initial Cluster Centers

	Cluster				
	1	2	3	4	5
Zscore(horizonal)	.51034	-1.58036	-.01234	2.07836	3.64638
Zscore(vertical)	-.01065	-2.12896	.59458	1.19981	2.41027
Zscore(width)	.93218	-2.04613	-.55698	1.42857	3.91050
Zscore(height)	2.04628	-2.37561	.71971	1.16190	1.60409
Zscore(onpix)	3.42116	-1.60059	-1.14407	.68202	1.59507
Zscore(meanXpixels)	.54423	2.51856	-3.40443	.05065	3.01214
Zscore(meanYpixels)	-.64527	-1.50546	-3.22583	1.50520	-2.36564
Zscore(meanXvariance)	.13748	.50787	.50787	-.23290	.50787
Zscore(meanYvariance)	-1.33506	-.91505	.34496	1.18498	-1.33506
Zscore(xybarmean)	-.91705	1.89608	-3.32832	1.49421	1.49421
Zscore(x2ybrMeanXXY)	-.17256	-.55265	-2.45309	.96771	-2.07300
Zscore(xy2brMeanXYY)	.03415	1.47597	-1.40767	-1.88827	.51476
Zscore(xege)	3.40988	-1.30600	-1.30600	-.44857	2.12373
Zscore(xegvy)	2.36690	-.86559	-.21909	2.36690	-4.74457
Zscore(edgeCount)	2.45731	-1.43819	-1.43819	.12001	-1.04864
Zscore(yegvx)	4.45071	.12298	.12298	-2.35001	.12298

Iteration History[a]

Keiland Pullen
DSC 441 – Fall 2021

Change in Cluster Centers

Iteration	1	2	3	4	5
1	5.180	4.643	4.830	4.458	5.478
2	.761	.457	.643	.385	1.108
3	.479	.257	.550	.513	.765
4	.295	.198	.315	.494	.475
5	.199	.123	.146	.303	.297
6	.157	.087	.080	.207	.187
7	.114	.073	.053	.154	.134
8	.100	.055	.036	.121	.086
9	.076	.044	.031	.092	.068
10	.060	.032	.027	.071	.054
11	.053	.021	.021	.048	.052
12	.041	.019	.018	.038	.040
13	.029	.018	.016	.029	.029
14	.020	.013	.013	.025	.018
15	.016	.009	.010	.015	.018
16	.013	.015	.015	.011	.017
17	.012	.019	.017	.016	.015
18	.009	.008	.007	.011	.010
19	.006	.007	.006	.011	.007
20	.005	.003	.002	.005	.006
21	.004	.003	.002	.004	.006
22	.002	.003	.002	.002	.005
23	.002	.006	.005	.002	.004
24	.001	.004	.003	.003	.002
25	.002	.002	.001	.002	.002
26	.002	.003	.002	.002	.003
27	.003	.003	.002	.003	.004
28	.004	.002	.002	.004	.004
29	.003	.001	.001	.002	.000
30	.001	.001	.000	.002	.000
31	.002	.002	.002	.002	.000
32	.001	.003	.001	.003	.000
33	.000	.002	.001	.002	.000
34	.000	.001	.001	.001	.000
35	.000	.001	.000	.001	.000
36	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 36. The minimum distance between initial centers is 9.947.

Final Cluster Centers

	Cluster				
	1	2	3	4	5
Zscore(horizonal)	1.05025	-1.14934	-.07985	.03358	.20958
Zscore(vertical)	.78617	-1.30366	.10069	.02075	.49377
Zscore(width)	1.14130	-1.18117	-.16600	.01830	.28719
Zscore(height)	.84468	-1.33603	.25816	-.05700	.29459
Zscore(onpix)	1.33549	-.99071	-.04453	-.35596	.01641
Zscore(meanXpixels)	.14154	.02415	-.18899	-.79046	1.01764
Zscore(meanYpixels)	.02731	-.23325	-.21320	1.27250	-.88312
Zscore(meanXvariance)	-.09069	-.14402	.93713	-.49643	-.61902
Zscore(meanYvariance)	-.22650	-.05974	.25107	-.11263	.11078
Zscore(xybarmean)	-.16728	-.16650	-.62037	.54274	.81272
Zscore(x2ybrMeanXXY)	.03807	-.19370	-.07756	1.15092	-1.02680
Zscore(xy2brMeanXYY)	-.05818	.03812	.36974	-.52508	.04818
Zscore(xege)	1.11087	-.54270	-.12118	-.33159	-.18366
Zscore(xegvy)	.18151	-.21485	-.19323	.98038	-.80268
Zscore(edgeCount)	.65182	-.41960	.30922	-.65887	-.04096
Zscore(yegvx)	-.35819	.12764	.40509	-.67482	.45246

Number of Cases in each Cluster

Cluster	1	4154.000
	2	4117.000
	3	5001.000
	4	3643.000
	5	3083.000
Valid		19998.000
Missing		.000

Keiland Pullen
DSC 441 – Fall 2021

K = 7

Initial Cluster Centers

	Cluster						
	1	2	3	4	5	6	7
Zscore(horizonal)	5.73707	.51034	-1.05768	.51034	3.64638	-1.58036	.51034
Zscore(vertical)	2.41027	-.01065	-1.22111	-.01065	2.41027	-2.12896	1.19981
Zscore(width)	4.90327	.93218	-1.05336	.93218	3.91050	-2.04613	-.06059
Zscore(height)	1.16190	2.04628	-.16467	-.16467	1.60409	-2.37561	1.16190
Zscore(onpix)	1.59507	3.42116	-1.14407	-.23102	1.59507	-1.60059	-.23102
Zscore(meanXpixels)	1.03781	.54423	3.50573	-.44294	3.01214	-2.41727	-1.92368
Zscore(meanYpixels)	1.50520	-.64527	-2.79573	2.79548	-2.36564	-2.79573	.21492
Zscore(meanXvariance)	.50787	.13748	.87825	.13748	.50787	.50787	.50787
Zscore(meanYvariance)	-.49505	-1.33506	-.07504	-1.33506	-1.33506	-.49505	2.86500
Zscore(xybarmean)	-1.72081	-.91705	2.29796	1.49421	1.49421	-3.32832	-.51518
Zscore(x2ybrMeanXXY)	-.17256	-.17256	-1.69291	-.93273	-2.07300	-1.31282	-.17256
Zscore(xy2brMeanXYY)	.99536	.03415	1.47597	-3.33009	.51476	-1.88827	3.39839
Zscore(xege)	3.40988	3.40988	-1.30600	-1.30600	2.12373	-1.30600	-1.30600
Zscore(xegvy)	3.01340	2.36690	-.86559	1.07391	-4.74457	-.21909	-.21909
Zscore(edgeCount)	-.65909	2.45731	-1.43819	-.26954	-1.04864	-1.43819	1.28866
Zscore(yegvx)	-1.11351	4.45071	.12298	.12298	.12298	.12298	-1.11351

Iteration History[a]

	Change in Cluster Centers						
Iteration	1	2	3	4	5	6	7
1	5.126	5.045	4.421	4.397	5.410	4.622	4.457
2	1.296	.792	.525	.734	1.085	.729	.338
3	.658	.392	.233	.620	.586	.248	.341
4	.470	.277	.138	.310	.429	.068	.252
5	.286	.184	.101	.182	.298	.057	.155
6	.192	.111	.070	.126	.154	.064	.102
7	.156	.072	.040	.096	.075	.076	.075
8	.123	.036	.021	.082	.051	.062	.058
9	.097	.029	.032	.065	.025	.065	.051
10	.085	.024	.011	.066	.019	.046	.033
11	.073	.019	.009	.055	.011	.034	.028
12	.071	.021	.008	.053	.009	.026	.017

13	.069	.024	.016	.046	.006	.017	.011
14	.069	.021	.017	.051	.009	.026	.009
15	.070	.026	.014	.048	.007	.026	.007
16	.081	.033	.012	.062	.014	.033	.005
17	.072	.026	.015	.055	.017	.027	.009
18	.067	.026	.008	.055	.015	.027	.005
19	.067	.025	.009	.057	.016	.026	.006
20	.059	.020	.004	.039	.012	.021	.007
21	.040	.017	.006	.033	.010	.021	.007
22	.035	.012	.006	.028	.012	.011	.004
23	.031	.007	.004	.023	.010	.004	.003
24	.017	.006	.003	.012	.009	.007	.004
25	.013	.005	.001	.009	.005	.004	.002
26	.008	.005	.002	.006	.004	.002	.003
27	.010	.003	.000	.009	.002	.003	.002
28	.008	.003	.001	.006	.003	.002	.000
29	.009	.002	.001	.009	.002	.002	.001
30	.006	.001	.002	.005	.003	.002	.001
31	.003	.002	.002	.004	.002	.002	.002
32	.003	.002	.001	.003	.002	.002	.002
33	.003	.001	.003	.004	.000	.003	.002
34	.004	.001	.002	.004	.000	.003	.001
35	.004	.001	.002	.003	.000	.002	.000
36	.001	.000	.002	.002	.001	.001	.000
37	.002	.000	.002	.000	.001	.001	.001
38	.000	.001	.000	.000	.000	.000	.001
39	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 39. The minimum distance between initial centers is 9.084.

Final Cluster Centers

	Cluster						
	1	2	3	4	5	6	7
Zscore(horizonal)	.98747	.67096	-.74320	-.61123	1.01679	-1.14133	-.07864
Zscore(vertical)	.75615	.51225	-.33156	-.63764	.98220	-1.42631	.11961
Zscore(width)	.76537	.92732	-.50173	-.50811	.90013	-1.30091	-.16932
Zscore(height)	.55100	.76795	-.44117	-.68164	.67040	-1.37439	.26798
Zscore(onpix)	.23059	1.50815	-.59562	-.77650	.53482	-1.00169	-.07639

Zscore(meanXpixels)	-.97613	.25988	.77027	-.50726	.90371	-.15876	-.20325
Zscore(meanYpixels)	.92023	-.03229	-.80195	1.35551	-.64292	-.33586	-.20149
Zscore(meanXvariance)	-.40236	.08486	-.77810	-.54549	-.47470	.28029	.94501
Zscore(meanYvariance)	.17913	-.48517	.15749	-.37375	.21394	-.14366	.31942
Zscore(xybarmean)	.85586	-.55744	.83830	.20387	.69163	-.65801	-.60198
Zscore(x2ybrMeanXXY)	.91257	.02965	-.84220	1.22855	-.87903	-.24739	-.07836
Zscore(xy2brMeanXYY)	-.41056	.04843	.18037	-.46491	-.17018	.05454	.38256
Zscore(xege)	.05485	1.30844	-.66361	-.45277	.26901	-.44713	-.17618
Zscore(xegvy)	.86165	.13963	-.72609	.97672	-.63464	-.25253	-.17879
Zscore(edgeCount)	-.28308	.76769	-.51297	-.89126	.44488	-.28635	.32268
Zscore(yegvx)	-.99491	-.36342	.02151	-.36672	.68135	.26103	.42469

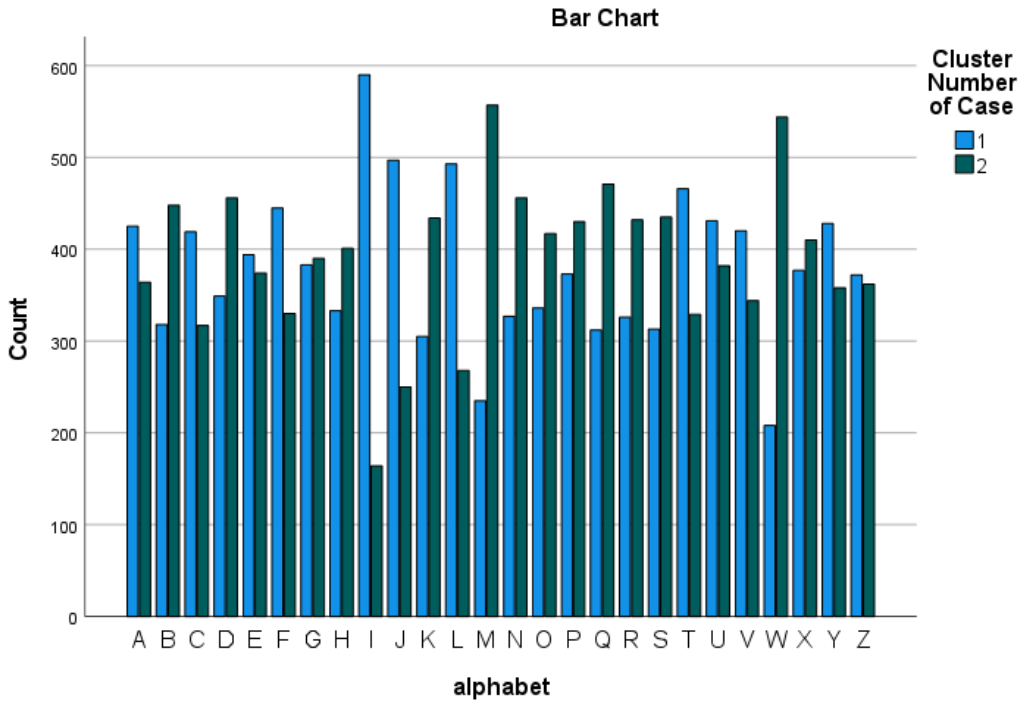
Number of Cases in each Cluster

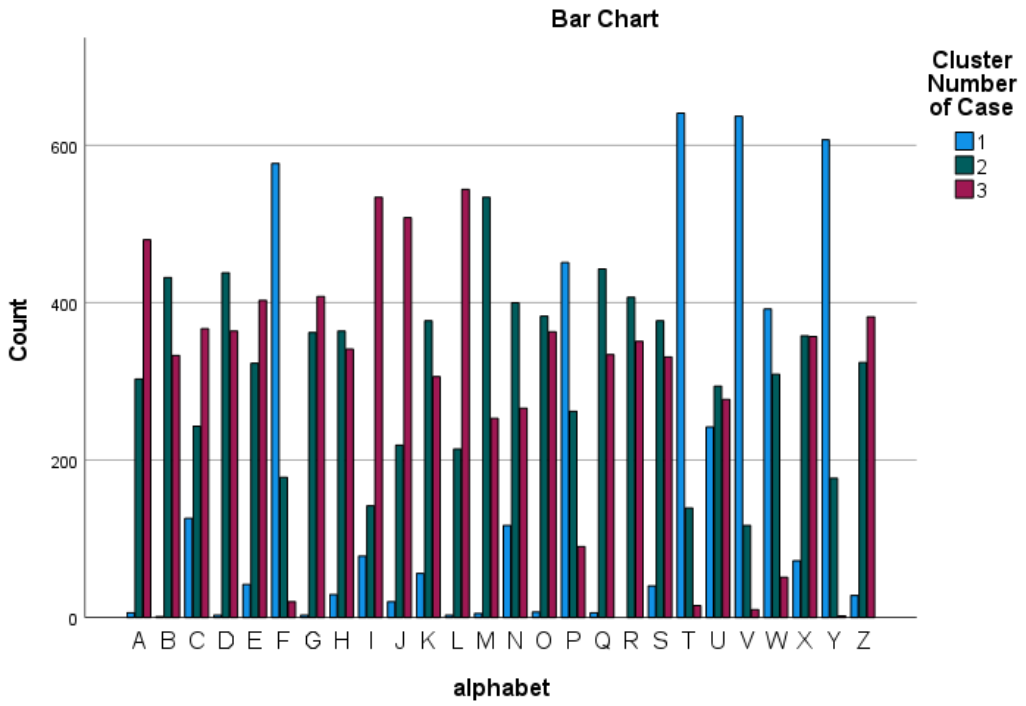
Cluster	1	2330.000
	2	3062.000
	3	2498.000
	4	2448.000
	5	2385.000
	6	2687.000
	7	4588.000
Valid		19998.000
Missing		.000

Keiland Pullen
DSC 441 – Fall 2021

Problem 3.3.)

To analyze the results, a “crosstabs” Bar Chart was used to display the results of the Clusters. A visual comparison between the crosstabs and decision trees proves a bit difficult to determine, however, it seems that the Cluster may be more accurate.





Keiland Pullen
DSC 441 – Fall 2021