IS 467 Assignment 5
Sarah Cummings

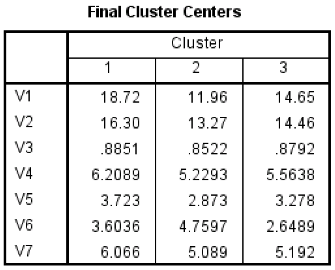
Problem 1— Clustering with seed data

i) Use k-means with all attributes using k=3-6

a) How the cluster centers were calculated with: 10 iterations.

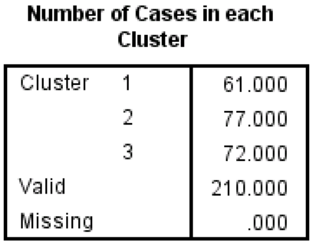
b) Similarity measure used: Euclidean distance/ default

c) For k=3, the final cluster centers were as given below, left. Their distributions can be seen at below, right. The number of cases in each in seen below, center.



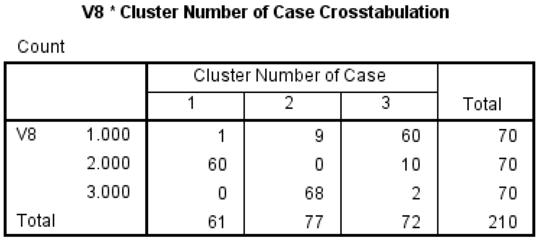
Final Cluster Centers

	Cluster		
	1	2	3
V1	18.72	11.96	14.65
V2	16.30	13.27	14.46
V3	.8851	.8522	.8792
V4	6.2089	5.2293	5.5638
V5	3.723	2.873	3.278
V6	3.6036	4.7597	2.6489
V7	6.066	5.089	5.192



Number of Cases in each Cluster

Cluster	1	61.000
	2	77.000
	3	72.000
Valid		210.000
Missing		.000



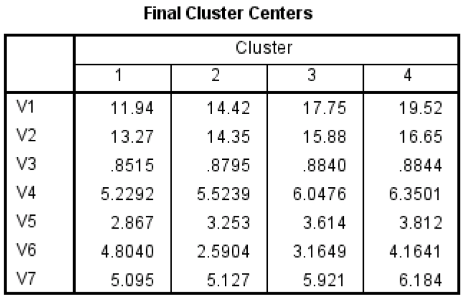
V8 * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case			Total
		1	2	3	
V8	1.000	1	9	60	70
	2.000	60	0	10	70
	3.000	0	68	2	70
Total		61	77	72	210

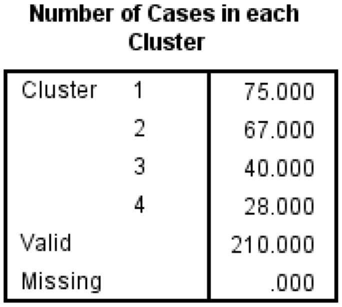
As you can see, these are fairly pure clusters with cluster one representing class two, cluster two representing class three, and cluster three representing class one.

For k=4, the final cluster centers were as given below, left. Their distributions can be seen at below, right. The number of cases in each in seen below, center.



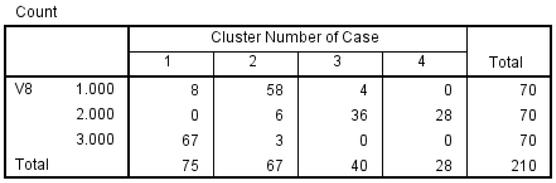
Final Cluster Centers

	Cluster			
	1	2	3	4
V1	11.94	14.42	17.75	19.52
V2	13.27	14.35	15.88	16.65
V3	.8515	.8795	.8840	.8844
V4	5.2292	5.5239	6.0476	6.3501
V5	2.867	3.253	3.614	3.812
V6	4.8040	2.5904	3.1649	4.1641
V7	5.095	5.127	5.921	6.184



Number of Cases in each Cluster

Cluster	1	75.000
	2	67.000
	3	40.000
	4	28.000
Valid		210.000
Missing		.000

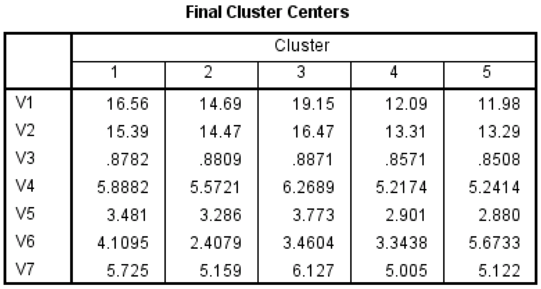


Count

		Cluster Number of Case				Total
		1	2	3	4	
V8	1.000	8	58	4	0	70
	2.000	0	6	36	28	70
	3.000	67	3	0	0	70
Total		75	67	40	28	210

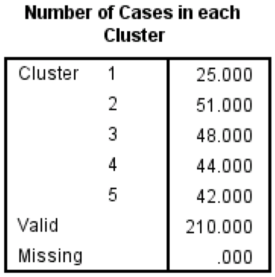
As you can see above, cluster one represents class three. Cluster two represents class one, and class two was split between cluster three and cluster four. These clusters are again pretty pure.

For k=5, the final cluster centers were as given below, left. Their distributions can be seen at below, right. The number of cases in each in seen below, center.



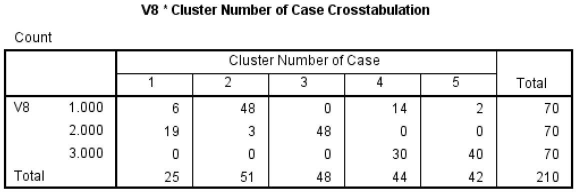
Final Cluster Centers

	Cluster				
	1	2	3	4	5
V1	16.56	14.69	19.15	12.09	11.98
V2	15.39	14.47	16.47	13.31	13.29
V3	.8782	.8809	.8871	.8571	.8508
V4	5.8882	5.5721	6.2689	5.2174	5.2414
V5	3.481	3.286	3.773	2.901	2.880
V6	4.1095	2.4079	3.4604	3.3438	5.6733
V7	5.725	5.159	6.127	5.005	5.122



Number of Cases in each Cluster

Cluster	1	25.000
	2	51.000
	3	48.000
	4	44.000
	5	42.000
Valid		210.000
Missing		.000

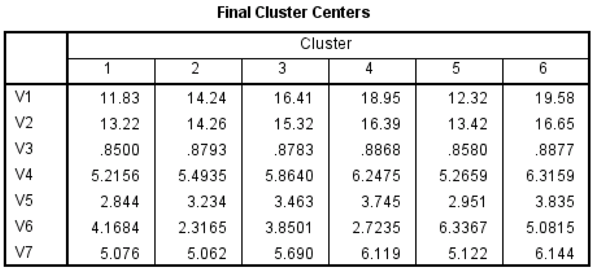


V8 * Cluster Number of Case Crosstabulation

Count

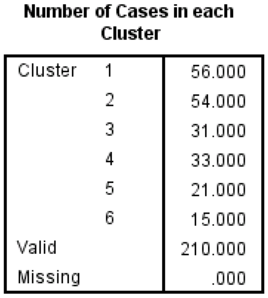
		Cluster Number of Case					Total
		1	2	3	4	5	
V8	1.000	6	48	0	14	2	70
	2.000	19	3	48	0	0	70
	3.000	0	0	0	30	40	70
Total		25	51	48	44	42	210

For k=6, the final cluster centers were as given below, left. Their distributions can be seen at below, right. The number of cases in each in seen below, center.



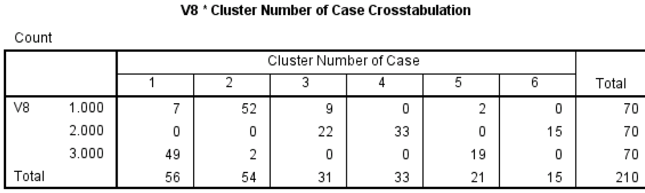
Final Cluster Centers

	Cluster					
	1	2	3	4	5	6
V1	11.83	14.24	16.41	18.95	12.32	19.58
V2	13.22	14.26	15.32	16.39	13.42	16.65
V3	.8500	.8793	.8783	.8868	.8580	.8877
V4	5.2156	5.4935	5.8640	6.2475	5.2659	6.3159
V5	2.844	3.234	3.463	3.745	2.951	3.835
V6	4.1684	2.3165	3.8501	2.7235	6.3367	5.0815
V7	5.076	5.062	5.690	6.119	5.122	6.144



Number of Cases in each Cluster

Cluster	1	56.000
	2	54.000
	3	31.000
	4	33.000
	5	21.000
	6	15.000
Valid		210.000
Missing		.000



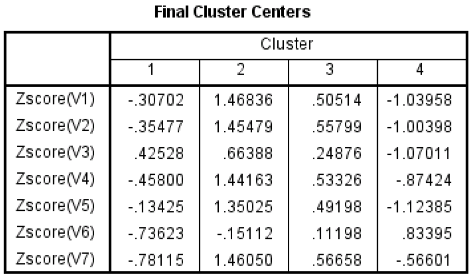
V8 * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case						Total
		1	2	3	4	5	6	
V8	1.000	7	52	9	0	2	0	70
	2.000	0	0	22	33	0	15	70
	3.000	49	2	0	0	19	0	70
Total		56	54	31	33	21	15	210

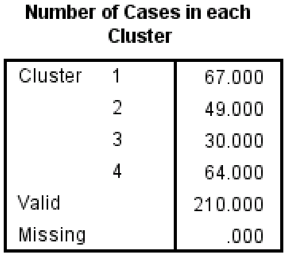
In my opinion, k=4 was best. Although there are technically only three types of seeds, having four clusters actually worked pretty well. Perhaps there is some subcategorization that can be done for class two, because the cases from class two were distinctly split between cluster three and cluster four.

After normalizing the attributes, we got the following results for k=4:



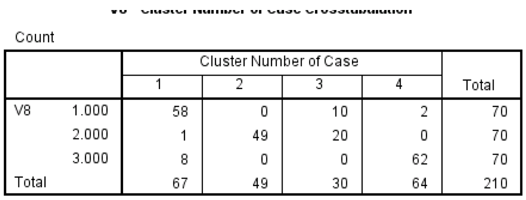
Final Cluster Centers

	Cluster			
	1	2	3	4
Zscore(V1)	-.30702	1.46836	.50514	-1.03958
Zscore(V2)	-.35477	1.45479	.55799	-1.00398
Zscore(V3)	.42528	.66388	.24876	-1.07011
Zscore(V4)	-.45800	1.44163	.53326	-.87424
Zscore(V5)	-.13425	1.35025	.49198	-1.12385
Zscore(V6)	-.73623	-.15112	.11198	.83395
Zscore(V7)	-.78115	1.46050	.56658	-.56601



Number of Cases in each Cluster

Cluster	1	67.000
	2	49.000
	3	30.000
	4	64.000
Valid		210.000
Missing		.000



V8 * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case				Total
		1	2	3	4	
V8	1.000	58	0	10	2	70
	2.000	1	49	20	0	70
	3.000	8	0	0	62	70
Total		67	49	30	64	210

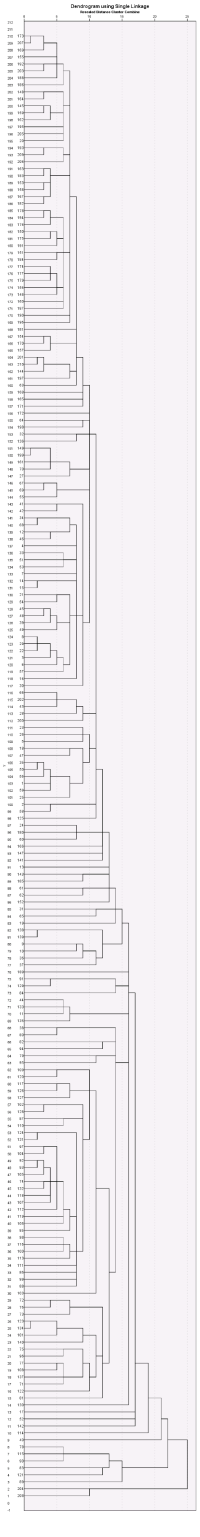
The normalization had very little affect on our clustering results. Again, class two was split into two different clusters, with 49 cases from class two going to cluster two, and 20 cases from class two going to cluster three.

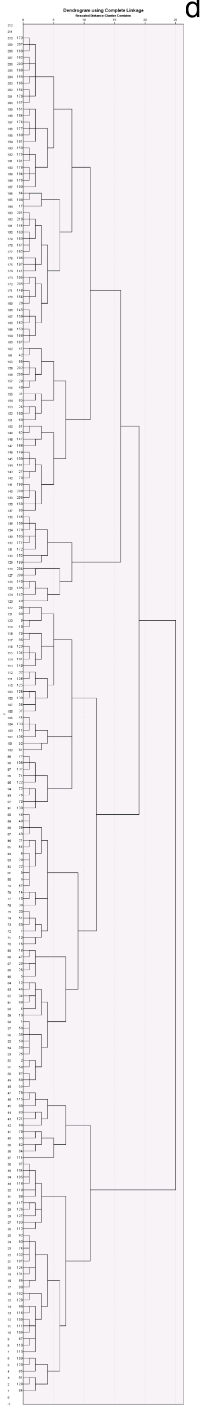
ii) Hierarchical clustering with seed data

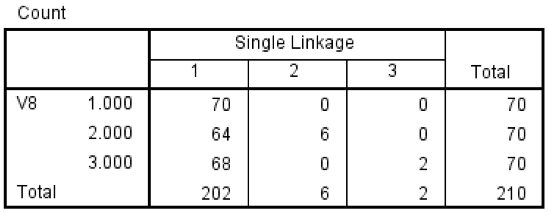
a) Single linkage dendogram and class distribution at the level of the dendogram when there are only three clusters.

b) Complete linkage dendogram and class distribution

The single linkage dendogram is at far left— the complete linkage dendrogram is to the right of it. have also included the distributions below. This method did not create very pure clusters for our data with the single linkage, but it improved with the complete linkage. Also the dendrograms are useless and too large to understand.

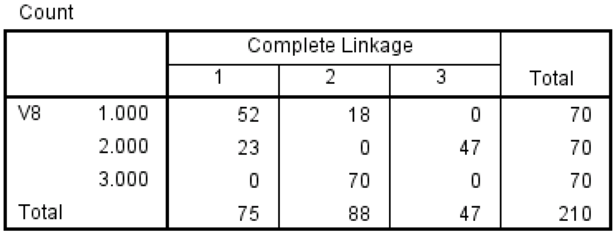






Count

		Single Linkage			Total
		1	2	3	
V8	1.000	70	0	0	70
	2.000	64	6	0	70
	3.000	68	0	2	70
Total		202	6	2	210



Count

		Complete Linkage			Total
		1	2	3	
V8	1.000	52	18	0	70
	2.000	23	0	47	70
	3.000	0	70	0	70
Total		75	88	47	210

c) Comparing hierarchical and k-means.

With the single linkage method, almost all of our cases were put into one cluster. This made the analysis almost useless. With complete linkage, however, our clusters were much more pure. Cluster one had 52/72 of its cases coming from class one. Cluster two had 70/88 cases from class three, and cluster three was all from class two. There were 41 misclassified cases with the complete linkage hierarchical clustering analysis overall. Comparing to our 3 cluster k-means analysis, with 22 misclassified cases, we an conclude that k-means analysis worked better for this data set than hierarchical.

d) Executive Summary: Using data from the Institute of Agrophysics of the Polish Academy of Sciences, our goal was to examine the difference between three types of wheat using cluster analysis. Our data contained 210 cases and the following attributes of the wheat kernels: area, perimeter, compactness length of kernel, width of kernel, asymmetry coefficient, and the length of the kernel groove. Our data also was labeled, meaning we had the actual class of wheat provided for each case.

In conducting cluster analysis, we attempted to learn intrinsic differences between the three types of wheat. The clustering will provide groupings of the cases. By leaving out the class label variable when running the analysis, we could see the true differences between the classes as the groupings arose.

Our first step was to clean the data, as a formatting issue in the text file caused several rows to be misaligned. Next, we ran k-means clustering analysis for the data, using k=3 as our parameter. This k value, which makes the most sense for our data, forced the algorithm to place the cases in one of three natural groups based on our seven variables. The process finds three “centers” or means for these groups and places each case in a group based on how close it is to each of the three centers. This method worked fairly well with only 22 misclassified cases, and our seven variables thus were proved useful to classify the type of wheat.

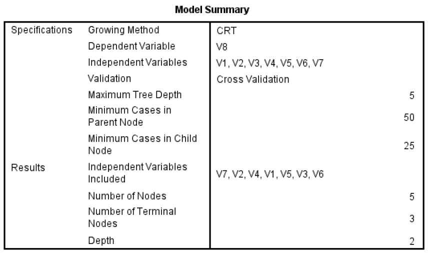
Next, we set k=4, 5, and 6 and repeated the process. Though we only have three types of wheat and there should only be 3 natural clusters, the k=4 results were interesting. With four clusters, our algorithm found a natural division in the class two wheat. The class two wheat was divided into two clusters. Perhaps this means there is strong variation in the class two wheat, or perhaps there is a way we could sub classify this type of wheat.

Finally, we used hierarchical clustering to analyze the data. The single linkage clustering was very unsuccessful, but the complete linkage clustering was much better. There were 41 misclassified cases with the complete linkage clustering analysis, again confirming three natural groupings of our data that correspond to the three classes of wheat. We could use this information to classify unlabeled data, or further understand the differences between our three classes of wheat.

Problem 2- Decision trees with seed data.

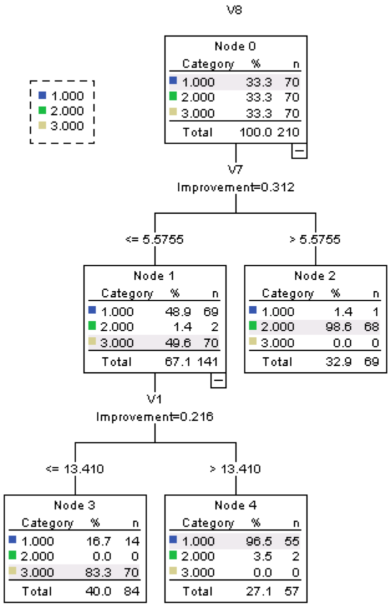
a) 5 trees with five configurations with results, and 10 fold validation:

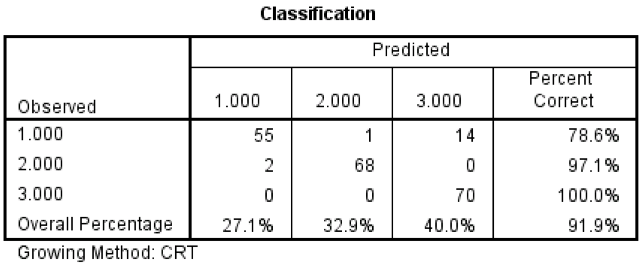
tree 1:



Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	V8
	Independent Variables	V1, V2, V3, V4, V5, V6, V7
	Validation	Cross Validation
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	50
	Minimum Cases in Child Node	25
Results	Independent Variables Included	V7, V2, V4, V1, V5, V3, V6
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2



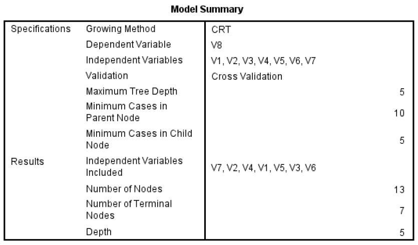


Classification

Observed	Predicted			
	1.000	2.000	3.000	Percent Correct
1.000	55	1	14	78.6%
2.000	2	68	0	97.1%
3.000	0	0	70	100.0%
Overall Percentage	27.1%	32.9%	40.0%	91.9%

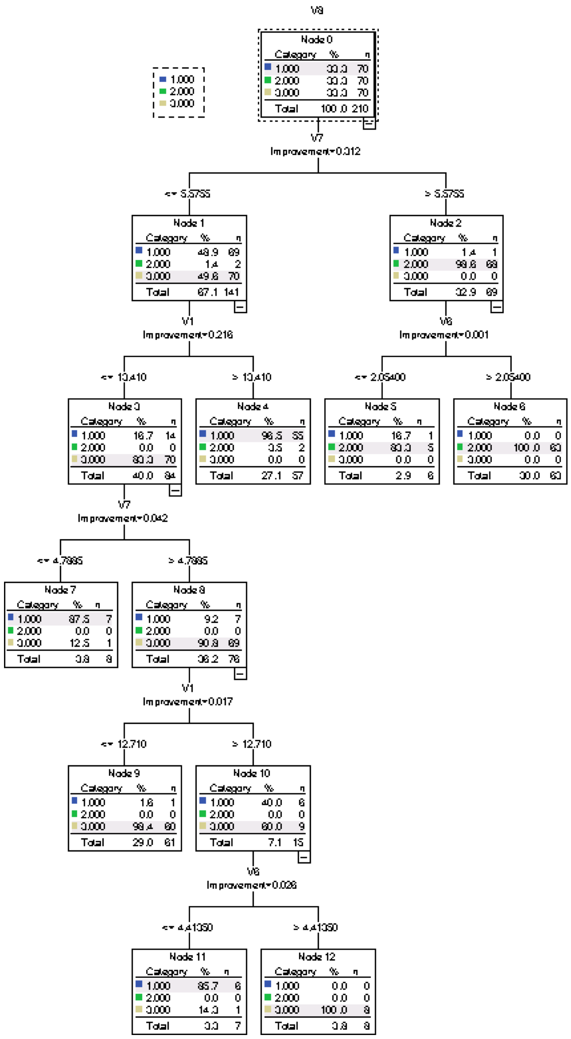
Growing Method: CRT

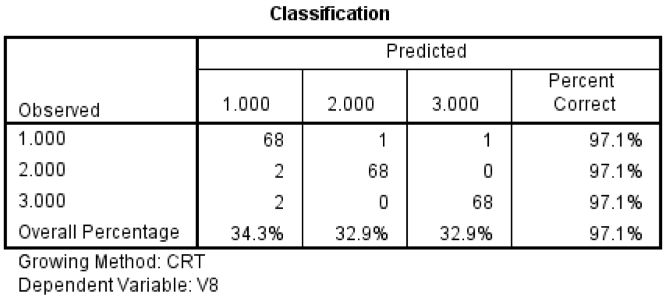
tree 2:



Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	V8
	Independent Variables	V1, V2, V3, V4, V5, V6, V7
	Validation	Cross Validation
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	V7, V2, V4, V1, V5, V3, V6
	Number of Nodes	13
	Number of Terminal Nodes	7
	Depth	5





Classification

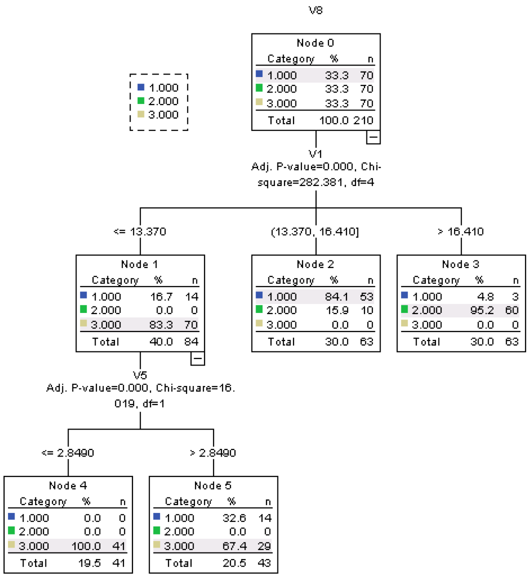
Observed	Predicted			
	1.000	2.000	3.000	Percent Correct
1.000	68	1	1	97.1%
2.000	2	68	0	97.1%
3.000	2	0	68	97.1%
Overall Percentage	34.3%	32.9%	32.9%	97.1%

Growing Method: CRT
Dependent Variable: V8

tree 3:

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	V8
	Independent Variables	V1, V2, V3, V4, V5, V6, V7
	Validation	Cross Validation
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	50
	Minimum Cases in Child Node	25
Results	Independent Variables Included	V1, V5
	Number of Nodes	6
	Number of Terminal Nodes	4
	Depth	2



Classification

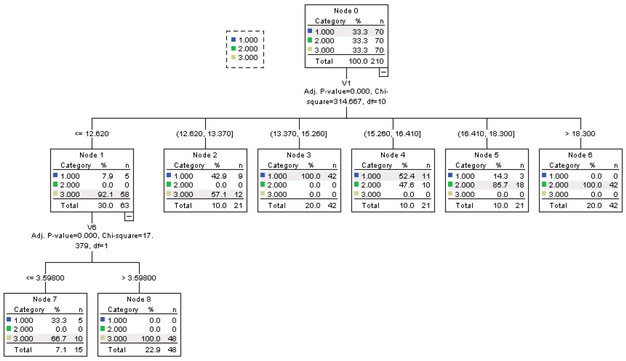
Observed	Predicted			
	1.000	2.000	3.000	Percent Correct
1.000	53	3	14	75.7%
2.000	10	60	0	85.7%
3.000	0	0	70	100.0%
Overall Percentage	30.0%	30.0%	40.0%	87.1%

Growing Method: CHAID
Dependent Variable: V8

tree 4:

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	V8
	Independent Variables	V1, V2, V3, V4, V5, V6, V7
	Validation	Cross Validation
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	20
	Minimum Cases in Child Node	10
Results	Independent Variables Included	V1, V6
	Number of Nodes	9
	Number of Terminal Nodes	7
	Depth	2



Classification

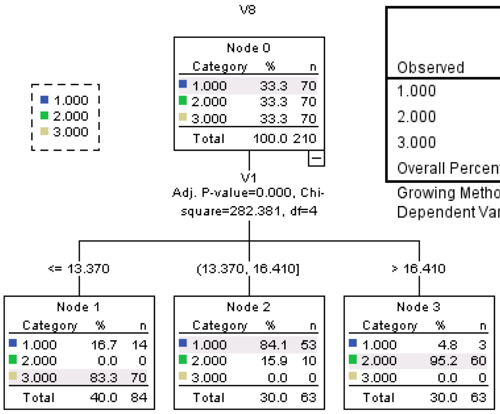
Observed	Predicted			
	1.000	2.000	3.000	Percent Correct
1.000	53	3	14	75.7%
2.000	10	60	0	85.7%
3.000	0	0	70	100.0%
Overall Percentage	30.0%	30.0%	40.0%	87.1%

Growing Method: CHAID
Dependent Variable: V8

tree 5:

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	V8
	Independent Variables	V1, V2, V3, V4, V5, V6, V7
	Validation	Cross Validation
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	V1
	Number of Nodes	4
	Number of Terminal Nodes	3
	Depth	1



Classification

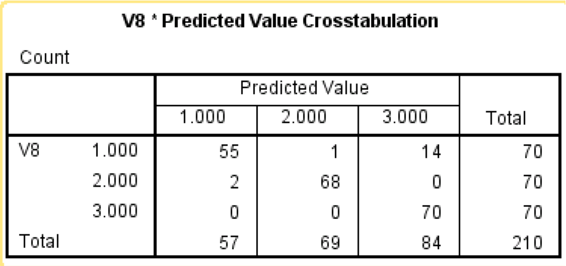
Observed	Predicted			
	1.000	2.000	3.000	Percent Correct
1.000	53	3	14	75.7%
2.000	10	60	0	85.7%
3.000	0	0	70	100.0%
Overall Percentage	30.0%	30.0%	40.0%	87.1%

Growing Method: CHAID
Dependent Variable: V8

Based on the results above, I actually like my first tree best out of the five. With 91.9% accuracy and only 5 total nodes, it is both very accurate and very easy to understand. While tree number 2 has 97.1% accuracy, there are thirteen nodes. The size makes the tree feel much more complicated and harder to explain. As per Occam's razor, tree number one is the most simple while still having an impressive accuracy greater the ninety percent.

b) Misclassification matrix for tree one:

As seen in the misclassification matrix, the tree did a particularly great job predicting class 2, and had a few issues predicting classes 1 and 3. It misclassified 14 class one cases as being in class three. Regardless, this tree is very impressive for its size with 91.9% accuracy.



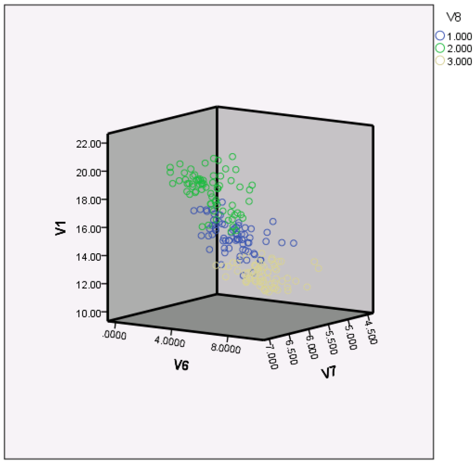
V8 * Predicted Value Crosstabulation

Count

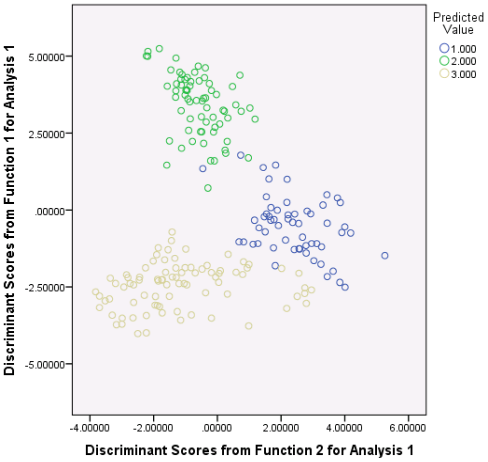
		Predicted Value			Total
		1.000	2.000	3.000	
V8	1.000	55	1	14	70
	2.000	2	68	0	70
	3.000	0	0	70	70
Total		57	69	84	210

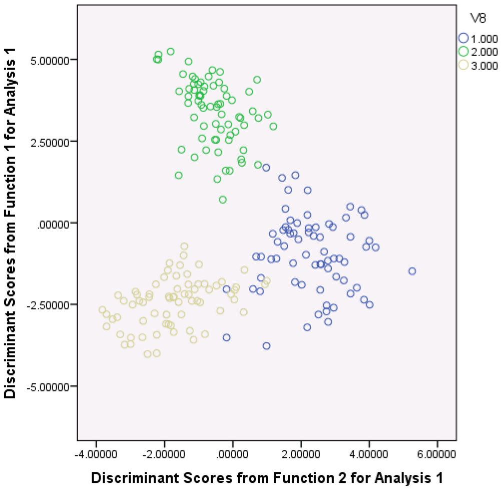
c) The three most important variables for this data set are v7, v1, and v6

d) At right is a 3D scatterplot of these three variables:



e) Other techniques we could use to identify variables for data visualization are feature extraction methods such as LDA. By reducing dimensionality with LDA first, we can visualize the clusters on a 2D scatterplot. See my visualizations below, which show our predicted classification clusters versus the actual class values.


Predicted Value
Discriminant Scores from Function 1 for Analysis 1
Discriminant Scores from Function 2 for Analysis 1


V8
Discriminant Scores from Function 1 for Analysis 1
Discriminant Scores from Function 2 for Analysis 1