Name: Amy Aumpansub   IS 467: Assignment 4
Due: June 1, 2018

Assignment 4

Problem 1

a)

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	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	V12, V8, V11, V7, V10, V14, V13, V9, V16, V15, V17, V4, V3, V5, V2, V6
	Number of Nodes	973
	Number of Terminal Nodes	487
	Depth	20

The holdout method is used to split data, in which the training set contains 70% of total observations or 14,011 tuples, whereas a test set contains 5,989 tuples or 30% of the total observations. The decision tree is built using CHART method with the GINI index of 0.000. The tree contains 973 nodes, which 487 nodes are terminal nodes. The error rate for a training set is 12.4% and for a test set is 19.4%.

b) The table shows different accuracies of different models. The best model is the first model, containing 10 tuples in parent and 5 tuples in child node with the level of depth of 20. Among 5 models, this first model is the best model because it has the highest accuracy rate (80.6% for a test set) and the lowest error rate (19.4% for a test set).

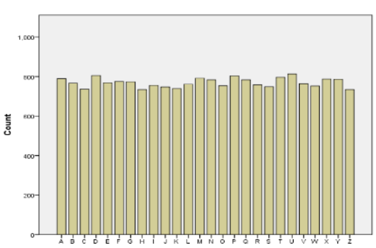
Model (Parent, Child)	Depth	Partition	Accuracy	Error Rate
(10, 5)	20	Train	87.6%	12.4%
		Test	80.6%	19.4%
(20, 10)	20	Train	82.4%	17.6%
		Test	78.0%	22.0%
(50, 25)	20	Train	73.3%	26.7%
		Test	71.8%	28.2%
(100, 50)	10	Train	62.9%	37.1%
		Test	62.9%	37.1%
(200, 100)	10	Train	58.0%	42.0%
		Test	58.1%	41.9%

c) The misclassification matrix is shown on the left. It comes for the best model selected in part b). The overall accuracy rate is high. The rate for a training set is 87.6% and for a test set is 80.6%. The accuracy is a good measure for this dataset because each class of target variable has approximately equal frequency, so there is no problem of imbalanced class.

Classification

Sample		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	nt Correc
Training	A	495	0	3	1	2	0	7	0	1	0	1	3	6	2	3	1	2	1	7	6	3	0	0	1	0	1	90.7%
	B	0	450	0	9	3	2	2	13	5	5	5	0	1	1	2	0	2	9	5	0	3	4	0	10	0	0	84.7%
	C	0	0	460	2	2	3	14	1	2	1	5	1	1	0	4	0	5	1	5	3	1	1	0	2	0	2	89.1%
	D	0	19	0	488	2	0	1	1	1	4	2	0	1	5	7	3	5	13	1	1	0	0	0	2	0	0	87.8%
	E	1	0	9	0	442	3	11	1	3	0	3	5	1	0	1	1	14	3	10	5	2	0	0	21	1	5	81.5%
	F	0	6	1	4	3	466	4	0	5	2	1	0	2	3	2	12	0	1	4	3	0	3	1	4	9	1	86.8%
	G	1	2	18	3	5	0	455	0	2	2	8	1	0	1	7	1	17	3	5	1	3	5	4	0	0	1	83.5%
	H	0	8	2	9	0	3	2	405	2	0	17	2	1	2	17	3	3	19	1	0	3	5	1	6	0	0	79.3%
	I	2	3	4	1	1	2	1	1	500	20	3	0	0	0	1	1	4	1	1	1	0	0	0	5	0	1	90.4%
	J	1	11	0	2	2	5	2	1	10	456	2	1	1	2	0	0	4	5	2	1	2	0	0	7	0	2	87.9%
	K	0	0	1	5	7	0	4	9	3	0	443	1	0	2	0	5	4	9	0	1	1	0	0	12	0	0	87.4%
	L	0	1	3	3	5	0	4	1	1	7	3	488	1	0	1	0	3	3	2	0	1	0	0	4	0	0	91.9%
	M	2	1	0	7	1	3	0	0	1	0	1	1	510	7	1	4	0	5	0	1	4	2	5	0	1	0	91.6%
	N	3	3	0	2	0	1	2	5	0	0	1	0	0	478	13	3	0	9	1	3	7	7	1	0	0	0	88.7%
	O	0	6	0	11	1	0	10	7	1	1	7	0	7	5	437	3	13	5	3	3	3	1	4	3	0	0	82.3%
	P	0	5	3	3	1	32	4	6	0	4	1	2	0	3	5	505	1	0	0	5	0	3	1	0	4	0	85.9%
	Q	2	4	7	1	4	1	6	3	5	0	2	0	1	1	6	1	480	6	7	0	0	0	0	0	3	3	88.4%
	R	0	11	2	3	1	0	5	10	1	3	8	3	1	2	5	2	1	470	3	2	1	4	0	2	0	0	87.0%
	S	0	9	0	0	12	6	3	5	1	1	5	2	1	1	3	0	5	3	471	0	3	2	0	10	3	7	85.2%
	T	0	0	2	1	2	9	4	0	2	2	6	1	0	0	1	3	0	0	1	471	1	6	0	6	7	5	88.9%
	U	1	3	4	5	4	0	1	1	0	1	0	1	7	8	2	0	10	2	1	0	506	10	0	1	0	0	89.1%
	V	0	5	0	0	0	12	1	0	1	0	1	0	1	3	1	2	3	0	0	5	4	476	3	0	2	0	91.5%
	W	2	0	0	2	0	2	3	1	0	0	0	0	5	9	1	4	0	1	0	1	6	9	489	0	3	0	90.9%
	X	0	2	1	2	13	2	1	4	3	0	6	3	0	5	2	0	0	6	2	2	2	2	0	502	1	0	89.5%
	Y	0	3	2	4	1	8	1	3	1	3	0	1	2	4	0	2	0	1	1	4	0	13	4	5	493	3	88.2%
	Z	0	3	0	2	8	3	1	2	4	3	1	0	0	1	3	0	7	5	7	2	0	1	0	3	0	434	88.6%
	Overall	3.6%	4.0%	3.7%	4.1%	3.7%	4.0%	3.9%	3.4%	4.0%	3.7%	3.8%	3.7%	3.9%	3.9%	3.7%	4.0%	4.2%	4.1%	3.9%	3.7%	4.0%	4.0%	3.7%	4.3%	3.8%	3.3%	87.6%
Test	A	210	0	4	0	0	0	2	0	0	1	4	10	4	0	0	0	1	0	1	1	2	0	0	3	0	0	86.4%
	B	1	178	0	5	1	0	1	7	3	3	1	0	0	0	4	0	1	5	2	0	1	8	0	14	0	0	75.7%
	C	0	0	185	0	6	0	9	1	0	5	2	3	0	0	3	0	2	3	1	0	0	0	0	0	0	0	84.1%
	D	0	16	0	194	0	0	1	3	0	0	0	0	2	0	11	2	5	9	1	1	2	0	0	1	0	1	77.9%
	E	2	0	7	0	162	2	4	3	2	0	4	7	0	0	1	1	5	1	9	0	2	1	0	8	0	5	71.7%
	F	1	4	1	3	1	171	4	1	2	3	0	2	1	2	0	15	0	0	5	3	0	3	1	4	9	2	71.8%
	G	0	4	15	2	3	3	163	0	1	0	6	2	0	0	4	0	11	1	1	0	0	5	3	3	0	1	71.5%
	H	0	5	0	9	3	1	2	155	1	0	13	0	1	1	5	3	1	17	0	0	2	1	0	1	1	1	69.5%
	I	0	1	1	0	0	3	1	1	182	6	0	0	0	0	0	2	0	0	1	0	1	0	0	2	1	0	90.1%
	J	0	1	0	3	0	1	2	0	10	192	2	0	1	0	1	0	7	2	0	1	1	0	0	2	2	0	84.2%
	K	0	0	5	4	2	0	4	7	0	1	186	0	0	1	0	1	2	9	1	1	2	0	0	6	0	0	80.2%
	L	0	1	1	1	4	0	5	0	0	1	1	206	0	1	1	0	5	1	1	0	0	0	0	1	0	0	89.6%
	M	3	1	1	5	0	3	1	0	0	0	1	0	200	6	0	0	0	1	1	1	3	2	6	0	0	0	85.1%
	N	1	0	0	2	0	0	1	3	0	0	1	0	4	200	4	7	0	9	0	2	3	4	2	0	1	0	82.0%
	O	0	2	0	3	0	0	3	1	0	0	1	0	3	0	176	2	13	3	6	1	5	0	2	1	0	0	79.3%
	P	2	0	0	4	2	8	1	5	1	0	0	0	0	3	2	171	1	3	0	5	1	0	2	0	4	0	79.5%
	Q	0	2	0	1	4	0	8	1	1	1	1	0	1	4	3	2	197	2	3	1	1	2	3	1	0	1	82.1%
	R	2	3	1	6	2	2	1	8	0	0	7	3	0	4	2	3	1	168	1	0	2	0	0	2	0	0	77.1%
	S	1	4	0	1	12	7	3	3	2	1	1	1	0	0	1	1	3	1	139	1	1	0	0	5	2	5	71.3%
	T	0	0	0	1	3	7	0	0	3	0	1	1	0	0	0	2	0	5	2	223	1	3	0	1	11	2	83.8%
	U	2	1	1	1	2	1	0	1	0	0	0	0	1	8	6	0	6	0	0	0	207	7	1	0	0	0	84.5%
	V	1	1	0	1	1	0	2	1	0	0	2	0	0	8	1	3	0	1	0	3	1	211	1	0	6	0	86.5%
	W	0	0	0	1	0	3	3	0	0	0	0	0	4	6	0	3	1	2	1	1	3	6	180	0	0	0	84.1%
	X	1	1	2	1	8	0	1	1	4	0	1	3	0	1	3	0	1	1	7	0	3	1	0	185	0	1	81.9%
	Y	0	1	1	1	0	2	1	2	0	2	0	0	0	4	0	0	2	0	2	6	1	6	2	4	188	2	82.8%
	Z	0	0	0	2	16	4	1	2	5	2	0	2	0	1	1	0	3	1	2	1	0	0	0	1	0	200	82.0%
	Overall	3.8%	3.8%	3.8%	4.2%	3.9%	3.6%	3.7%	3.4%	3.6%	3.6%	3.9%	4.0%	3.7%	4.2%	3.8%	3.6%	4.5%	4.1%	3.1%	4.2%	4.1%	4.3%	3.4%	4.1%	3.8%	3.7%	80.6%

Growing Method: CRT
Dependent Variable: V1



Independent Variable Importance

Independent Variable	Importance	Normalized Importance
V14	.344	100.0%
V11	.291	84.6%
V15	.275	80.0%
V12	.270	78.6%
V10	.264	77.0%
V9	.250	72.8%
V16	.244	71.2%
V13	.229	66.6%
V7	.223	64.9%
V8	.215	62.5%
V17	.205	59.8%
V2	.113	32.8%
V4	.111	32.2%
V6	.109	31.9%
V3	.104	30.4%
V5	.079	22.9%

Growing Method: CRT
Dependent Variable: V1

d) From the table above, the most important three attributes are ranked in order as following: V14 (x-ege mean edge count left to right), V11(xybar mean x y correlation), V15 (xegvy correlation of x-ege with y). The V14 attribute has the highest important rate and has the highest position in the tree, so the V14 is the most important attribute.

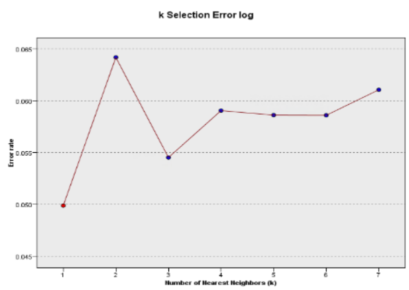
Name: Amy Aumpansub
Due: June 1, 2018

IS 467: Assignment 4

Problem 2

a) The KNN is performed using Euclidean metric. All 16 attributes (predictors) are standardized using z-scores to make sure that there is no problem regarding a difference in scale. This prevents distance measures from being dominated by attributes that have larger ranges. I performed PCA to reduce dimension and ran the KNN with new components, but the accuracy rate was significantly lower than original attributes. So, I decide to use 16 attributes (normalized) in part b.

b) The misclassification matrices for (K=1, 3, 5, and 7) are shown below. The best value of K is equal to 1 because it is the smallest K, which has the lowest error rate of 4.7%. The K=1 model has the highest accuracy rate of 95.3%.



K = 1

Error Summary

Partition	Percent of Records Incorrectly Classified
Training	4.7%
Holdout	4.7%

K=3

Error Summary

Partition	Percent of Records Incorrectly Classified
Training	5.0%
Holdout	5.0%

K= 5

Error Summary

Partition	Percent of Records Incorrectly Classified
Training	5.4%
Holdout	5.6%

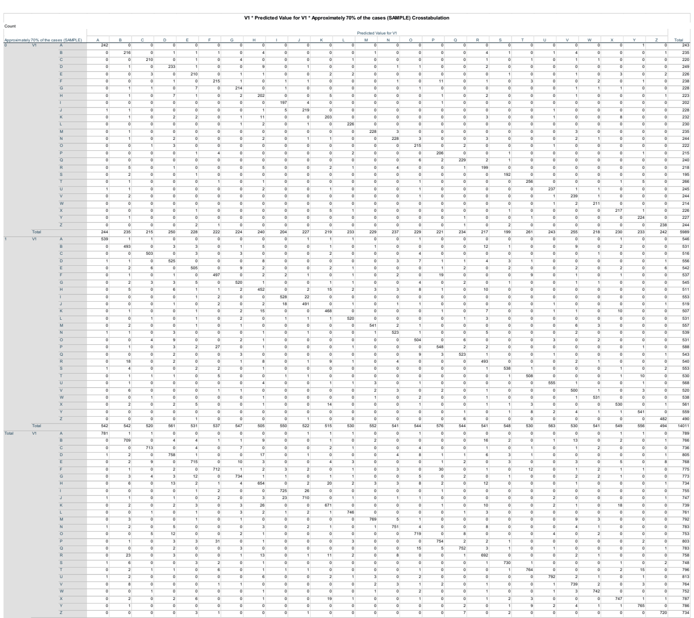
K= 7

Error Summary

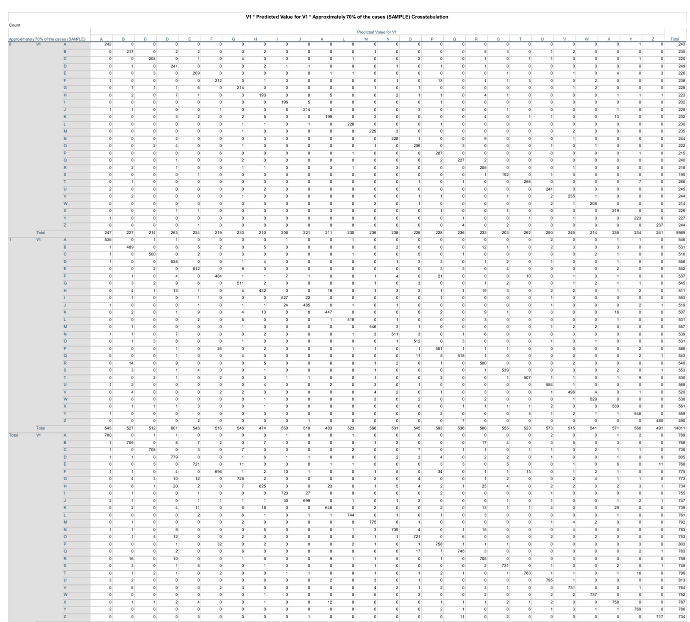
Partition	Percent of Records Incorrectly Classified
Training	5.7%
Holdout	5.6%

K	Partition	1	3	5	7
Accuracy	Train	95.3%	95.0%	94.6%	94.3%
	Test	95.3%	95.0%	94.4%	94.4%
Error Rate	Train	4.7%	5.0%	5.4%	5.7%
	Test	4.7%	5.0%	5.6%	5.6%

K = 1



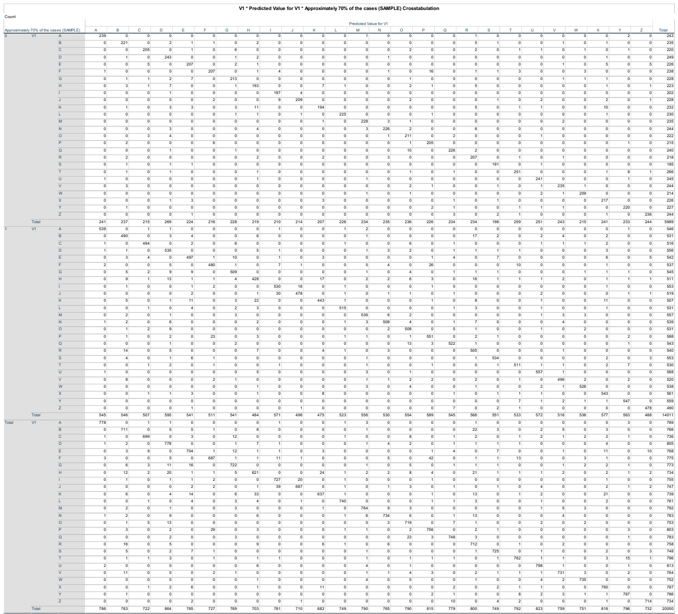
K=3



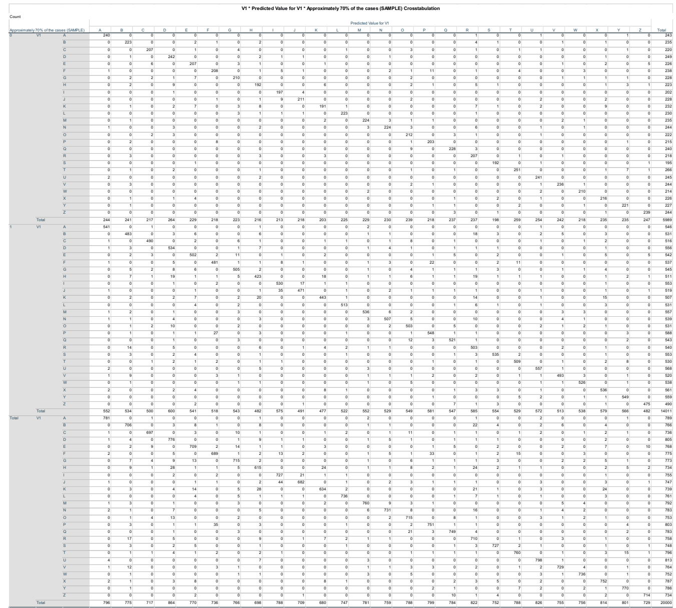
Name: Amy Aumpansub
Due: June 1, 2018

IS 467: Assignment 4

K = 5



K=7



c) The model from KNN (k=1) has a better performance than the model from decision tree because the KNN model has a relatively higher accuracy rate of 95.3% (test set), compared to an accuracy rate of 80.6% (test set) from decision tree. The KNN model also has a lower error rate of 4.7%, compared to the error rate of 19.4% from decision tree.

KNN Model

K	Partition	1	3	5	7
Accuracy	Train	95.3%	95.0%	94.6%	94.3%
	Test	95.3%	95.0%	94.4%	94.4%
Error Rate	Train	4.7%	5.0%	5.4%	5.7%
	Test	4.7%	5.0%	5.6%	5.6%

Decision Tree

Model (Parent, Child)	Depth	Partition	Accuracy	Error Rate
(10, 5)	20	Train	87.6%	12.4%
		Test	80.6%	19.4%

Problem 3

a)

1. For K-means, the cluster centers or the mean points are calculated based on average of each of the attributes, which does not need to be an object in the data set.

2. First measure is a Jaccard coefficient which is similarity measure for data that has asymmetric binary variables. Second measure is cosine similarity which is used for the data that has Term-frequency vectors (sparse numeric data).

3.

K= 3

i Final cluster centers

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

ii number of elements in each cluster

Number of Cases in each Cluster

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

iii class distribution

V8 * Cluster Number of Case Crosstabulation

Count

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

Actual Class

K= 4

i Final cluster centers

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

ii number of elements in each cluster

Number of Cases in each Cluster

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

iii class distribution

V8 * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case 1	2	3	4	Total
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

K= 5

i Final cluster centers

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

ii number of elements in each cluster

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

iii class distribution

V8 * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case 1	2	3	4	5	Total
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

K= 6

i Final cluster centers

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

ii number of elements in each cluster

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

iii class distribution

V8 * Cluster Number of Case Crosstabulation

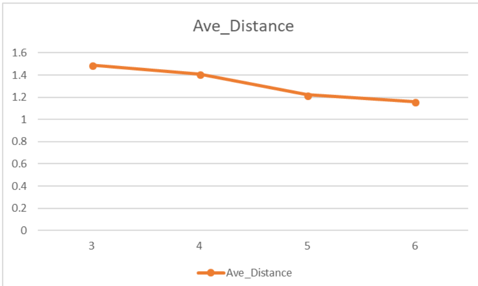
Count

		Cluster Number of Case 1	2	3	4	5	6	Total
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

Name: Amy Aumpansub
Due: June 1, 2018

IS 467: Assignment 4

4. The best value of K is 5 (elbow).



K	Ave_Distance
3	1.4915084
4	1.4089938
5	1.2235427
6	1.1620128

Best K = 5

5. From the class distribution in part iii, the best value of K would be 3 because the label is similar within the clusters, but it is different across clusters.

6. The normalization will influence the clustering results because the range of value changes. The attributes are normalized with z-scores before running K-mean (K=5). The number of elements in each cluster change after normalization. The cluster centers contain both positive and negative numbers. The distance between final clusters center changes. The frequency of data assigned to each cluster changes as well.

<u>After normalization:</u>

Final Cluster Centers

	Cluster 1	2	3	4	5
Zscore(V1)	1.45626	-1.08742	-.75669	.47893	-.20755
Zscore(V2)	1.44408	-1.04571	-.78371	.54612	-.26019
Zscore(V3)	.65695	-1.18706	-.18993	.14168	.55314
Zscore(V4)	1.43611	-.90606	-.77643	.54258	-.38641
Zscore(V5)	1.33994	-1.19871	-.63603	.44951	-.02707
Zscore(V6)	-.15042	.40134	1.61155	.16569	-.85882
Zscore(V7)	1.45420	-.63393	-.58677	.55343	-.72743

Number of Cases in each Cluster

Cluster	1	50.000
	2	55.000
	3	19.000
	4	28.000
	5	58.000
Valid		210.000
Missing		.000

Distances between Final Cluster Centers

Cluster	1	2	3	4	5
1		5.715	5.162	2.128	4.017
2	5.715		1.724	3.610	2.767
3	5.162	1.724		3.117	2.788
4	2.128	3.610	3.117		2.252
5	4.017	2.767	2.788	2.252	

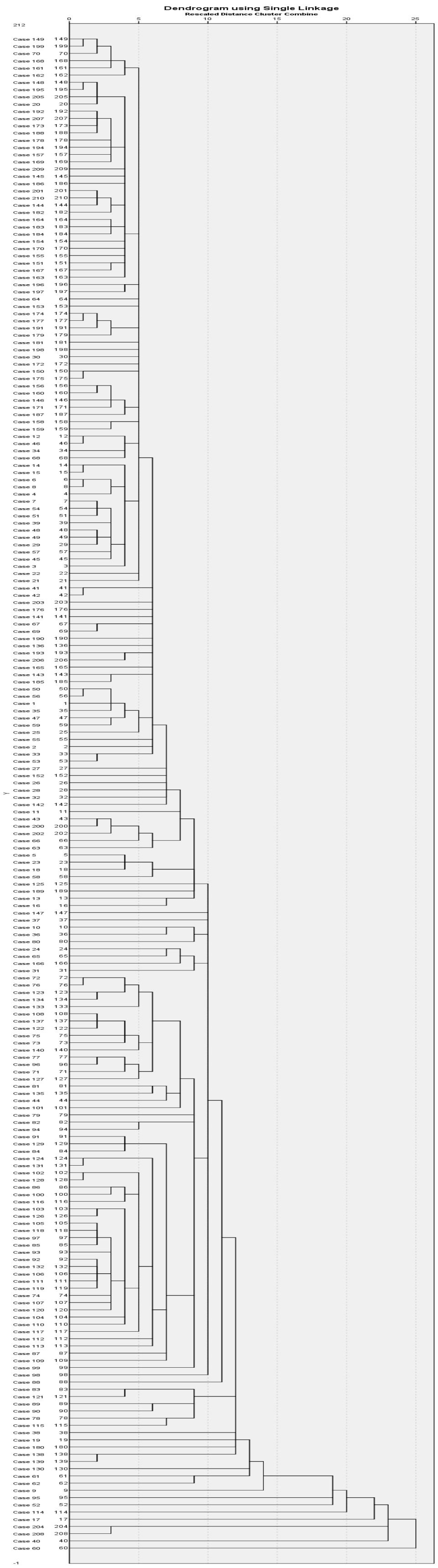
Name: Amy Aumpansub
Due: June 1, 2018

IS 467: Assignment 4

b)

1. Single linkage algorithm

i Dendogram



ii Class distribution

V8 * Single Linkage Crosstabulation

Count

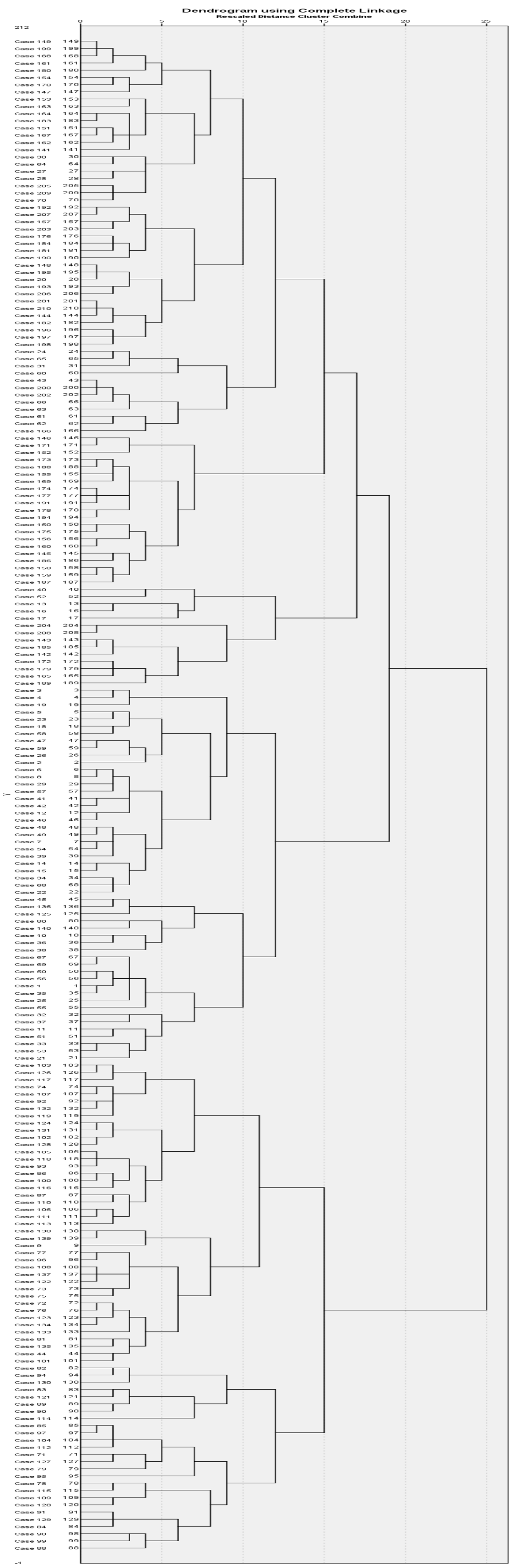
		Single Linkage			
		1	2	3	Total
V8	1.000	68	1	1	70
	2.000	70	0	0	70
	3.000	68	2	0	70
Total		206	3	1	210

Class

The dendogram is created from the nearest neighbors method. The class distribution is shown in the table above. The majority data falls in the first cluster.

2. Complete linkage and report

i Dendogram



ii Class distribution

V8 * Complete Linkage Crosstabulation

Count

		Complete Linkage			Total
		1	2	3	
V8	1.000	48	2	20	70
	2.000	4	66	0	70
	3.000	0	0	70	70
Total		52	68	90	210

Class

The dendogram is created from the furthest neighbors method. The class distribution from complete linkage method is clearer than the single linkage shown in the table above.

Name: Amy Aumpansub IS 467: Assignment 4
Due: June 1, 2018

C) **executive summary**

Overview

The seed dataset was retrieved from UCI Machine Learning Repository. It contains 210 random samples of kernels belonging to the three different varieties of wheat including Kama, Rosa and Canadian (class label). Each type contains 70 elements. The study combined a soft X-ray technique which used to detect the internal kernel structure with harvested wheat grain originating from experimental fields to measure kernels with seven geometric parameters (attributes) including area A, perimeter P, compactness C = 4*pi*A/P^2, length of kernel, width of kernel, asymmetry coefficient, and length of kernel groove. All seven attributes are numeric and continuously varied variables.

Problem

The seed dataset contains sample kernels that needs to be clustered into groups and labeled. Kernels will be clustered based on their seven geometric parameters.

Solutions

The raw dataset will be cleaned before doing the analysis in SPSS. We will implement a clustering analysis for this dataset by using two methods:

1. K-means clustering: This method divides the observations into k clusters, in which each observation belongs to the cluster with the nearest mean (center of cluster). All seven attributes will be used in this clustering, and we will run the multiple numbers of clusters (k =3, 4, 5, 6) to indicate the best number of clusters for our seed datasets.
2. Hierarchical clustering: This method clusters observations by building a hierarchy of clusters. We will perform the "bottom-up" approach which will start from single cluster to merged clusters. We will perform both single linkage approach in which distance between two clusters is distance between two closest records and complete linkages approach in which distance between two clusters is distance between two farthest records

Results and Recommendations:

The K-mean method suggests 5 clusters for 210 kernels in which each cluster contains kernels. Both single linkage and complete linkage approach suggest 3 clusters. However, the class distribution of complete linkage is clearer than single linkage. Among three methods, the class distribution is clearer and close to the actual class distribution when we use K-means with 3 clusters as it minimizes variability within the cluster, but maximizes variability between clusters. The graph shows the 7 attributes of each cluster. In recommendation, the 210 kernels should be clustered into 3 clusters using K-means method as shown in the left graph.