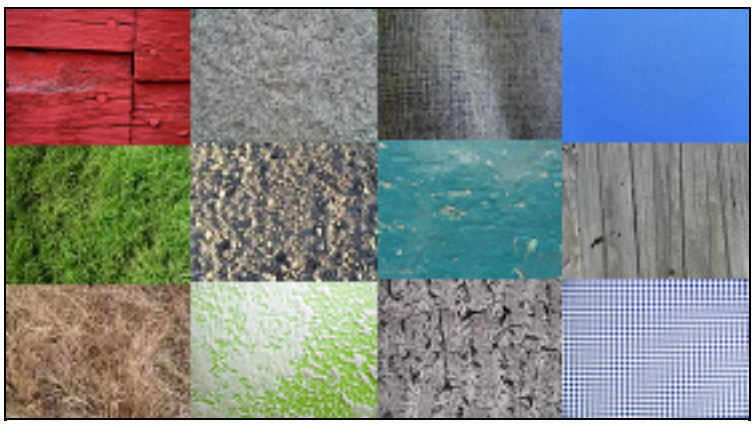
Analysis of Classification Data

In this example of data mining for knowledge discovery we consider a classification problem with a large number of objects to be classified based on many attributes. A set of 40 characters or attributes are measured on 5500 items which belong to 11 different categories of varied textures. Textures include a grass lawn, pressed calf leather, handmade paper, cotton canvas, etc. All of the attributes are measured on a continuous scale. Data are obtained from (http://sci2s.ugr.es/keel/dataset.php?cod=72#sub2 [1])



Objective of the Analysis

Pattern recognition and Classification of 5500 objects into 11 classes based on 40 attributes

Data Files for this case (*right-click and "save as"*) :

Texture.csv [2] - full dataset

TXTrain1.csv [3]	TXTest1.csv [13]
TXTrain2.csv [4]	TXTest2.csv [14]
TXTrain3.csv [5]	TXTest3.csv [15]
TXTrain4.csv [6]	TXTest4.csv [16]
TXTrain5.csv [7]	TXTest5.csv [17]
TXTrain6.csv [8]	TXTest6.csv [18]
TXTrain7.csv [9]	TXTest7.csv [19]
TXTrain8.csv [10]	TXTest8.csv [20]
TXTrain9.csv [11]	TXTest9.csv [21]
TXTrain10.csv [12]	TXTest10.csv [22]

Texture.zip [23] - all data files above together in a .zip file for convenience

Overview of Classification Problem and Cross-Validation

Classification problem may be treated as a special type of regression problem where, based on the values of the predictors, each observation is placed into one and only one of the categories. Probability that the i^{th} object will be placed into one of the j categories is 1, for all $i = 1, \ldots n$. Each object has a different probability to be placed into different classes and is put into the class which

maximizes this probability.

Performance of a classification rule is measured through the mis-classification probability. Following techniques of classification are applied here

- Linear Discriminant Analysis
- K Nearest Neighbour
- Classification Tree
- Random Forest

CD.1: Exploratory Data Analysis (EDA) and Data Pre-processing

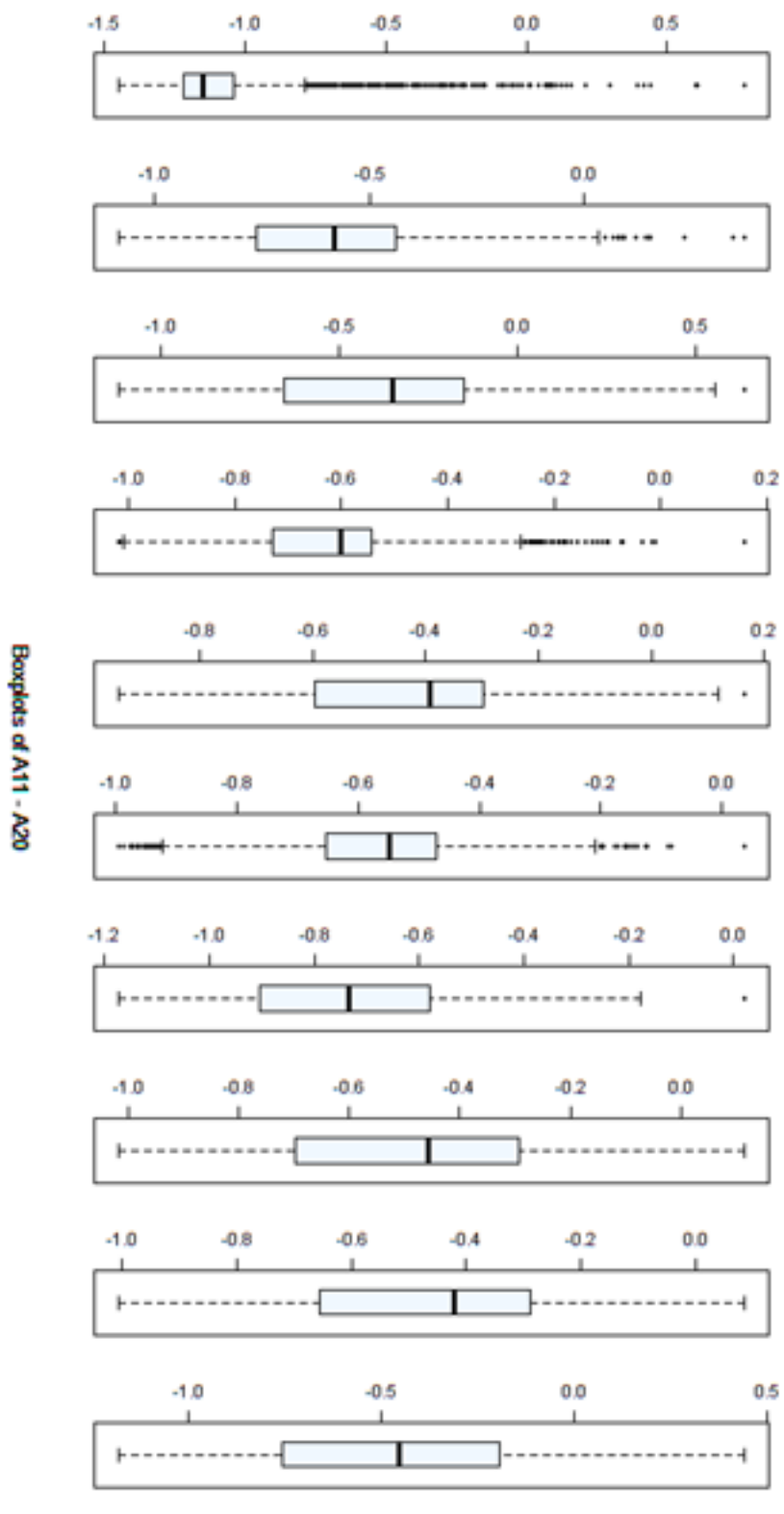
There are 40 predictors in the data. Univariate descriptive statistics and the box-plots are shown below.

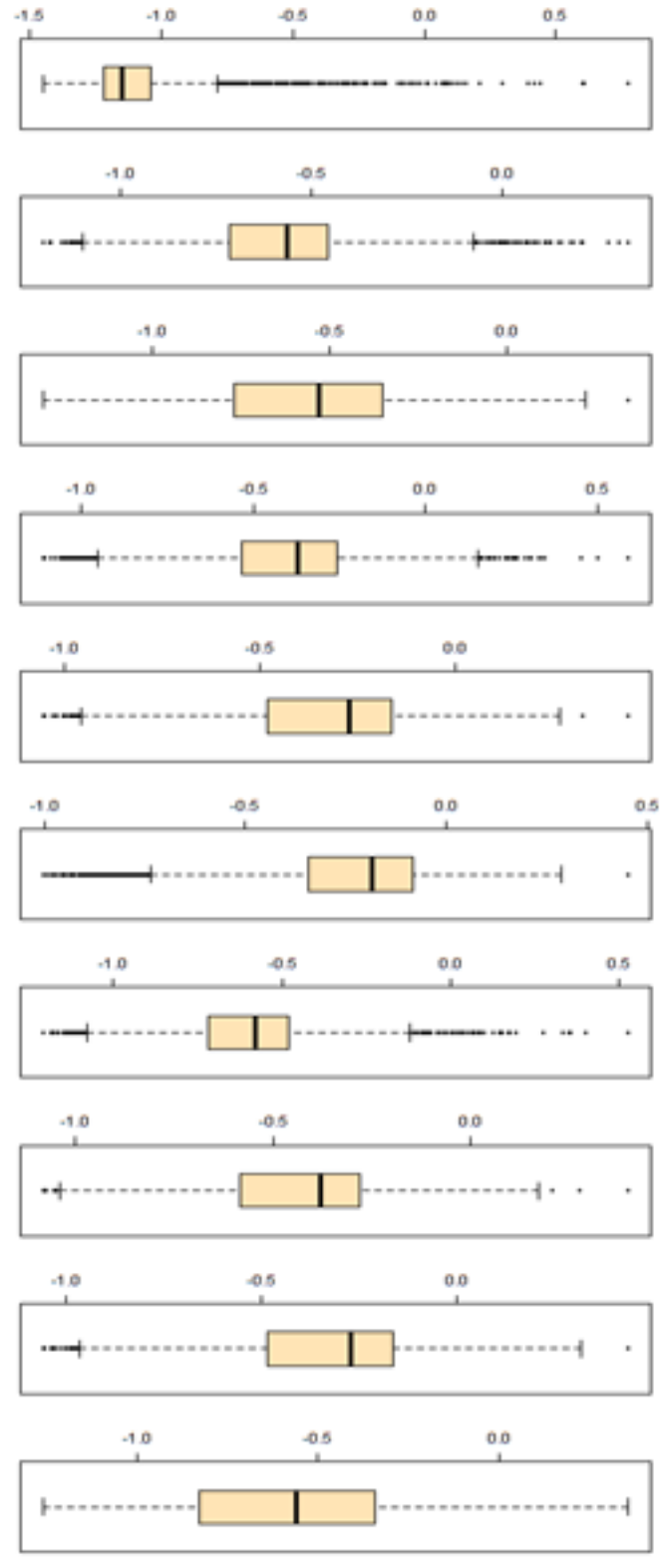
R codes for Data Preparation and Exploratory Data Analysis

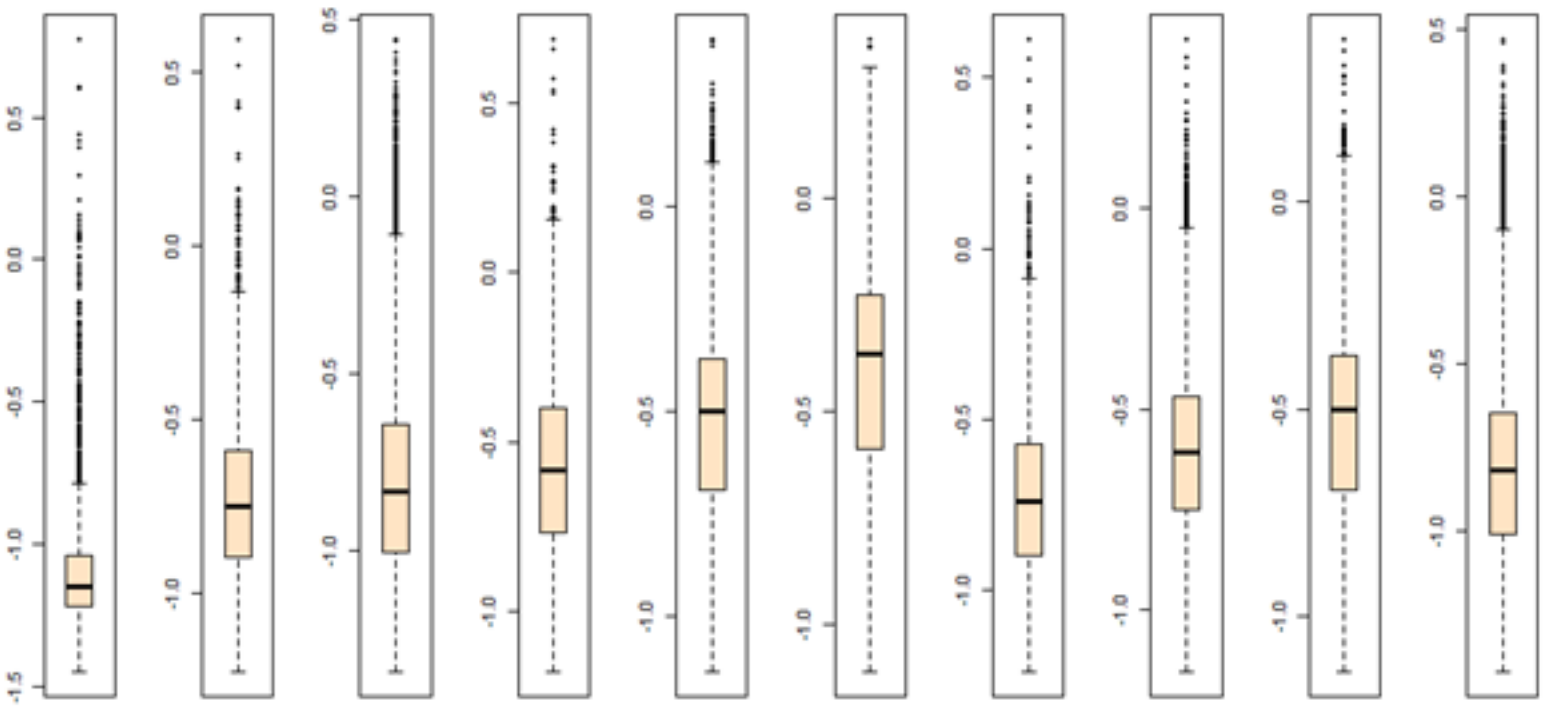
Descriptive statistics

Attribute Name	No of Obs	Min	Q1	Median	Q3	Max	Range	IQR	MAD	Mean	StDev	StErr
A1	5500	-1.45	-1.22	-1.15	-1.04	0.77	2.22	0.17	0.12	-1.10	0.20	0.003
A2	5500	-1.20	-0.71	-0.56	-0.46	0.33	1.53	0.26	0.18	-0.59	0.21	0.003
A3	5500	-1.31	-0.77	-0.53	-0.35	0.34	1.65	0.42	0.30	-0.58	0.31	0.004
A4	5500	-1.11	-0.54	-0.37	-0.26	0.59	1.70	0.28	0.20	-0.40	0.23	0.003
A5	5500	-1.05	-0.48	-0.27	-0.16	0.44	1.49	0.32	0.20	-0.33	0.24	0.003
A6	5500	-1.00	-0.35	-0.18	-0.08	0.45	1.46	0.26	0.18	-0.24	0.22	0.003
A7	5500	-1.21	-0.72	-0.58	-0.48	0.53	1.73	0.24	0.17	-0.60	0.20	0.003
A8	5500	-1.08	-0.58	-0.38	-0.28	0.40	1.48	0.30	0.19	-0.43	0.22	0.003
A9	5500	-1.06	-0.48	-0.27	-0.16	0.44	1.49	0.32	0.20	-0.33	0.24	0.003
A10	5500	-1.26	-0.83	-0.56	-0.34	0.36	1.61	0.49	0.36	-0.60	0.33	0.004
A11	5500	-1.45	-1.22	-1.15	-1.04	0.77	2.22	0.17	0.12	-1.10	0.20	0.003
A12	5500	-1.08	-0.76	-0.58	-0.44	0.37	1.46	0.32	0.24	-0.59	0.21	0.003
A13	5500	-1.12	-0.66	-0.35	-0.15	0.64	1.75	0.50	0.35	-0.40	0.34	0.005
A14	5500	-1.02	-0.73	-0.60	-0.54	0.16	1.18	0.19	0.12	-0.63	0.14	0.002
A15	5500	-0.94	-0.60	-0.39	-0.30	0.16	1.11	0.30	0.17	-0.45	0.20	0.003
A16	5500	-0.99	-0.65	-0.55	-0.47	0.04	1.03	0.18	0.13	-0.58	0.16	0.002
A17	5500	-1.17	-0.90	-0.73	-0.58	0.02	1.19	0.33	0.24	-0.73	0.20	0.003
A18	5500	-1.02	-0.70	-0.46	-0.29	0.12	1.13	0.40	0.28	-0.49	0.23	0.003
A19	5500	-1.00	-0.65	-0.42	-0.29	0.08	1.09	0.36	0.24	-0.47	0.23	0.003
A20	5500	-1.18	-0.76	-0.46	-0.19	0.44	1.62	0.57	0.41	-0.48	0.35	0.005
A21	5500	-1.45	-1.22	-1.15	-1.04	0.77	2.22	0.17	0.12	-1.10	0.20	0.003
A22	5500	-1.23	-0.90	-0.75	-0.59	0.60	1.82	0.31	0.23	-0.74	0.22	0.003
A23	5500	-1.34	-1.00	-0.83	-0.64	0.45	1.79	0.36	0.27	-0.78	0.33	0.004
A24	5500	-1.18	-0.77	-0.58	-0.40	0.69	1.87	0.37	0.28	-0.58	0.26	0.004
A25	5500	-1.14	-0.69	-0.50	-0.37	0.41	1.55	0.32	0.24	-0.51	0.25	0.003
A26	5500	-1.11	-0.59	-0.36	-0.23	0.37	1.48	0.36	0.25	-0.40	0.25	0.003
A27	5500	-1.24	-0.90	-0.74	-0.57	0.61	1.85	0.33	0.24	-0.73	0.23	0.003
A28	5500	-1.15	-0.75	-0.61	-0.47	0.42	1.58	0.28	0.21	-0.59	0.24	0.003
A29	5500	-1.13	-0.69	-0.50	-0.37	0.39	1.52	0.32	0.24	-0.51	0.25	0.003
A30	5500	-1.42	-1.01	-0.82	-0.65	0.47	1.89	0.37	0.27	-0.77	0.33	0.004
A31	5500	-1.45	-1.22	-1.15	-1.04	0.77	2.22	0.17	0.12	-1.10	0.20	0.003
A32	5500	-1.18	-0.78	-0.64	-0.54	0.57	1.74	0.24	0.18	-0.65	0.19	0.003
A33	5500	-1.15	-0.75	-0.53	-0.25	0.68	1.82	0.50	0.37	-0.49	0.33	0.004
A34	5500	-1.12	-0.74	-0.63	-0.56	0.31	1.44	0.18	0.13	-0.64	0.14	0.002
A35	5500	-1.02	-0.64	-0.49	-0.34	0.34	1.36	0.30	0.22	-0.49	0.19	0.003
A36	5500	-1.03	-0.72	-0.58	-0.46	0.16	1.19	0.26	0.18	-0.59	0.17	0.002
A37	5500	-1.25	-0.91	-0.77	-0.69	0.09	1.34	0.22	0.16	-0.78	0.16	0.002
A38	5500	-1.10	-0.72	-0.56	-0.40	0.19	1.29	0.32	0.24	-0.55	0.21	0.003
A39	5500	-1.08	-0.67	-0.51	-0.37	0.20	1.28	0.30	0.22	-0.52	0.20	0.003
A40	5500	-1.22	-0.79	-0.59	-0.35	0.47	1.68	0.44	0.33	-0.57	0.31	0.004

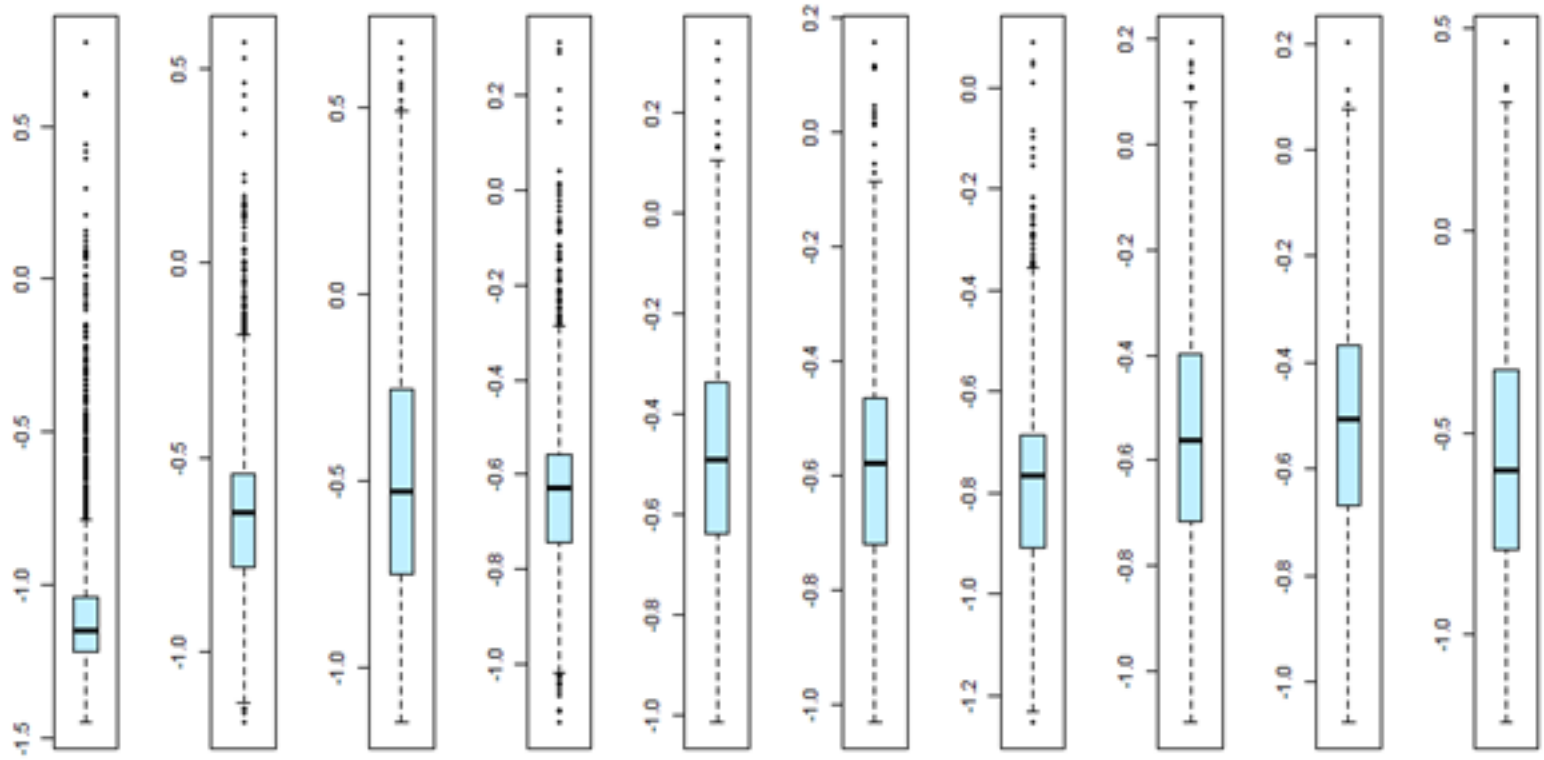
Boxplots


Boxplots of A11 - A20


Boxplots of A1 - A10



Boxplots of A21 - A30



Boxplots of A31 - A40

From the above univariate exercises, it is clear that several of the 40 attributes have outliers of varied proportion. In order to include as many rows as possible, but eliminating the extreme outliers, all data points (rows) were included, which do not contain any outlier value in any of the 40 predictors, outliers being defined as a value outside of [Q1-3IQR, Q3+3IQR] limits. This eliminates 128 rows. A more conservative approach is where the outlier is defined to be a point outside of the limit [Q1-1.5IQR, Q3+1.5IQR]. In this case, 1060 rows would have been removed.

While looking at the correlation matrix it was found that there is a very high degree of dependency among the predictor variables. The red highlighted cells all show very high degree of dependency and it is all in the positive direction.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20
A1	1.00																			
A2	0.56	1.00																		
A3	0.34	0.78	1.00																	
A4	0.54	0.96	0.77	1.00																
A5	0.40	0.91	0.89	0.95	1.00															
A6	0.38	0.85	0.76	0.93	0.95	1.00														
A7	0.60	0.96	0.78	0.96	0.89	0.84	1.00													
A8	0.44	0.91	0.93	0.92	0.97	0.90	0.92	1.00												
A9	0.42	0.89	0.89	0.94	0.97	0.95	0.91	0.97	1.00											
A10	0.33	0.72	0.96	0.74	0.83	0.73	0.77	0.90	0.88	1.00										
A11	1.00	0.56	0.34	0.54	0.40	0.38	0.60	0.44	0.42	0.33	1.00									
A12	0.50	0.87	0.83	0.81	0.79	0.66	0.86	0.85	0.79	0.80	0.50	1.00								
A13	0.41	0.81	0.94	0.81	0.87	0.75	0.82	0.91	0.88	0.91	0.41	0.90	1.00							
A14	0.62	0.89	0.81	0.88	0.86	0.78	0.87	0.87	0.86	0.77	0.62	0.91	0.86	1.00						
A15	0.42	0.84	0.93	0.84	0.90	0.79	0.84	0.93	0.90	0.89	0.42	0.91	0.98	0.91	1.00					
A16	0.49	0.72	0.85	0.77	0.85	0.80	0.74	0.85	0.85	0.80	0.49	0.70	0.89	0.83	0.91	1.00				
A17	0.39	0.81	0.86	0.77	0.78	0.65	0.84	0.85	0.80	0.86	0.39	0.97	0.91	0.87	0.91	0.71	1.00			
A18	0.34	0.81	0.93	0.80	0.86	0.74	0.82	0.91	0.87	0.91	0.34	0.93	0.98	0.87	0.98	0.83	0.96	1.00		
A19	0.32	0.81	0.93	0.80	0.86	0.75	0.81	0.91	0.88	0.91	0.32	0.91	0.97	0.86	0.98	0.85	0.94	0.99	1.00	
A20	0.29	0.77	0.93	0.77	0.84	0.73	0.79	0.89	0.86	0.93	0.29	0.89	0.98	0.82	0.96	0.83	0.93	0.99	0.99	1.00
A21	1.00	0.56	0.34	0.54	0.40	0.38	0.60	0.44	0.42	0.33	1.00	0.50	0.41	0.62	0.42	0.49	0.39	0.34	0.32	0.29
A22	0.68	0.65	0.72	0.62	0.59	0.46	0.70	0.69	0.62	0.74	0.68	0.84	0.81	0.80	0.78	0.70	0.82	0.78	0.76	0.76
A23	0.48	0.50	0.49	0.61	0.61	0.61	0.51	0.57	0.59	0.47	0.48	0.44	0.60	0.60	0.55	0.66	0.40	0.51	0.47	0.52
A24	0.60	0.64	0.77	0.63	0.63	0.50	0.68	0.72	0.66	0.78	0.60	0.83	0.86	0.80	0.82	0.75	0.84	0.83	0.80	0.82
A25	0.54	0.66	0.78	0.69	0.72	0.62	0.68	0.76	0.73	0.77	0.54	0.77	0.87	0.82	0.83	0.81	0.77	0.83	0.80	0.83
A26	0.49	0.67	0.82	0.68	0.71	0.58	0.70	0.77	0.74	0.82	0.49	0.83	0.90	0.82	0.88	0.81	0.84	0.88	0.87	0.88
A27	0.67	0.65	0.74	0.62	0.59	0.46	0.70	0.70	0.63	0.76	0.67	0.83	0.82	0.79	0.79	0.71	0.83	0.79	0.77	0.78
A28	0.58	0.62	0.74	0.66	0.68	0.59	0.65	0.72	0.69	0.73	0.58	0.73	0.83	0.79	0.78	0.79	0.72	0.77	0.73	0.77
A29	0.55	0.65	0.78	0.70	0.72	0.63	0.69	0.76	0.74	0.78	0.55	0.76	0.87	0.81	0.83	0.82	0.76	0.82	0.79	0.82
A30	0.49	0.50	0.52	0.62	0.62	0.62	0.52	0.59	0.61	0.50	0.49	0.45	0.62	0.61	0.57	0.68	0.42	0.53	0.49	0.54
A31	1.00	0.56	0.34	0.54	0.40	0.38	0.60	0.44	0.42	0.33	1.00	0.50	0.41	0.62	0.42	0.49	0.39	0.34	0.32	0.29
A32	0.69	0.84	0.75	0.84	0.77	0.69	0.89	0.83	0.80	0.76	0.69	0.84	0.82	0.83	0.79	0.72	0.81	0.79	0.76	0.76
A33	0.49	0.63	0.78	0.71	0.77	0.72	0.68	0.78	0.78	0.79	0.49	0.60	0.79	0.72	0.74	0.80	0.61	0.72	0.69	0.73
A34	0.68	0.85	0.75	0.85	0.78	0.72	0.90	0.84	0.81	0.75	0.68	0.82	0.80	0.85	0.80	0.74	0.80	0.78	0.77	0.75
A35	0.51	0.72	0.82	0.79	0.83	0.78	0.77	0.85	0.85	0.83	0.51	0.68	0.82	0.79	0.80	0.83	0.69	0.77	0.75	0.78
A36	0.49	0.59	0.74	0.66	0.74	0.71	0.63	0.74	0.76	0.75	0.49	0.52	0.72	0.70	0.71	0.83	0.54	0.65	0.64	0.67
A37	0.61	0.85	0.80	0.86	0.83	0.75	0.87	0.87	0.84	0.79	0.61	0.83	0.85	0.85	0.82	0.76	0.81	0.82	0.79	0.80
A38	0.47	0.74	0.83	0.80	0.84	0.78	0.77	0.86	0.86	0.83	0.47	0.71	0.84	0.79	0.81	0.81	0.71	0.80	0.77	0.80
A39	0.43	0.76	0.86	0.81	0.87	0.81	0.78	0.88	0.88	0.86	0.43	0.72	0.86	0.79	0.84	0.83	0.74	0.82	0.81	0.83
A40	0.39	0.69	0.83	0.76	0.83	0.77	0.71	0.83	0.83	0.81	0.39	0.66	0.83	0.74	0.79	0.81	0.67	0.78	0.75	0.79

	A21	A22	A23	A24	A25	A26	A27	A28	A29	A30	A31	A32	A33	A34	A35	A36	A37	A38	A39	A40
A21	1.00																			
A22	0.68	1.00																		
A23	0.48	0.54	1.00																	
A24	0.60	0.97	0.61	1.00																
A25	0.54	0.88	0.81	0.94	1.00															
A26	0.49	0.90	0.65	0.95	0.96	1.00														
A27	0.67	0.98	0.51	0.98	0.87	0.91	1.00													
A28	0.58	0.87	0.85	0.93	0.99	0.92	0.87	1.00												
A29	0.55	0.87	0.81	0.94	0.99	0.96	0.88	0.99	1.00											
A30	0.49	0.55	0.98	0.63	0.82	0.67	0.54	0.86	0.83	1.00										
A31	1.00	0.68	0.48	0.60	0.54	0.49	0.67	0.58	0.55	0.49	1.00									
A32	0.69	0.86	0.64	0.85	0.83	0.82	0.86	0.83	0.83	0.65	0.69	1.00								
A33	0.49	0.69	0.82	0.76	0.87	0.79	0.68	0.88	0.88	0.83	0.49	0.78	1.00							
A34	0.68	0.84	0.58	0.83	0.80	0.81	0.84	0.78	0.80	0.59	0.68	0.97	0.75	1.00						
A35	0.51	0.72	0.74	0.77	0.86	0.82	0.72	0.85	0.87	0.75	0.51	0.84	0.97	0.84	1.00					
A36	0.49	0.63	0.71	0.68	0.78	0.73	0.63	0.78	0.80	0.73	0.49	0.69	0.94	0.72	0.95	1.00				
A37	0.61	0.82	0.71	0.85	0.87	0.85	0.82	0.86	0.87	0.73	0.61	0.96	0.84	0.95	0.89	0.75	1.00			
A38	0.47	0.73	0.78	0.79	0.88	0.83	0.72	0.88	0.89	0.79	0.47	0.86	0.97	0.84	0.99	0.91	0.92	1.00		
A39	0.43	0.71	0.73	0.77	0.86	0.83	0.71	0.84	0.87	0.74	0.43	0.84	0.95	0.83	0.98	0.91	0.90	0.99	1.00	
A40	0.39	0.67	0.81	0.75	0.87	0.81	0.66	0.87	0.88	0.82	0.39	0.79	0.98	0.75	0.96	0.91	0.87	0.98	0.98	1.00

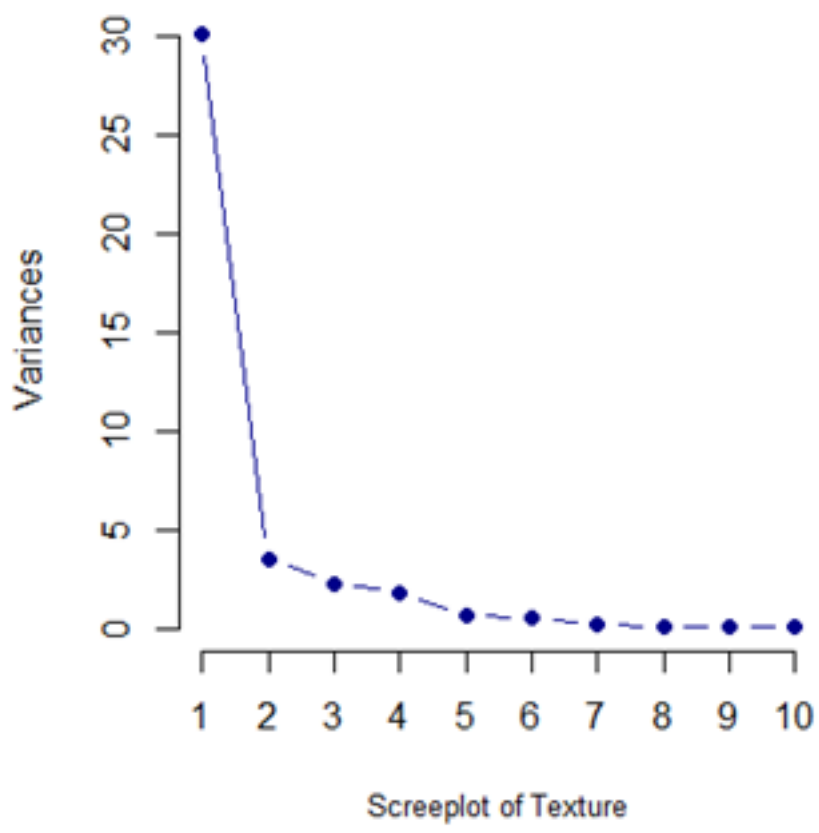
CD.2: Principal Components Analysis

R codes for Principal Component Analysis

With such a high degree of dependency it is recommended that a PCA is done on the data and

only the top few components are used for classification.

	SD	Prop of Variation	Cumulative Prop
PC1	*5.48*	*0.75*	*0.75*
PC2	*1.89*	*0.09*	*0.84*
PC3	*1.52*	*0.06*	*0.90*
PC4	*1.37*	*0.05*	*0.95*
PC5	*0.82*	*0.02*	*0.96*
PC6	*0.75*	*0.01*	*0.98*
PC7	*0.44*	*0.00*	*0.98*
PC8	*0.41*	*0.00*	*0.99*
PC9	0.33	0.00	0.99
PC10	0.30	0.00	0.99
PC11	0.26	0.00	0.99
PC12	0.23	0.00	0.99
PC13	0.21	0.00	0.99
PC14	0.19	0.00	1.00
PC15	0.17	0.00	1.00
PC16	0.16	0.00	1.00
PC17	0.13	0.00	1.00
PC18	0.13	0.00	1.00
PC19	0.11	0.00	1.00
PC20	0.11	0.00	1.00
PC21	0.09	0.00	1.00
PC22	0.08	0.00	1.00
PC23	0.08	0.00	1.00
PC24	0.07	0.00	1.00
PC25	0.07	0.00	1.00
PC26	0.07	0.00	1.00
PC27	0.06	0.00	1.00
PC28	0.06	0.00	1.00
PC29	0.06	0.00	1.00
PC30	0.05	0.00	1.00
PC31	0.05	0.00	1.00
PC32	0.05	0.00	1.00
PC33	0.04	0.00	1.00
PC34	0.03	0.00	1.00
PC35	0.03	0.00	1.00
PC36	0.03	0.00	1.00
PC37	0.02	0.00	1.00
PC38	0.00	0.00	1.00
PC39	0.00	0.00	1.00
PC40	0.00	0.00	1.00



From the table and the screeplot above it is clear that it is sufficient to consider only the first 8 PCs. They are given below.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
A1	0.11	0.40	0.11	-0.08	-0.07	-0.13	0.10	-0.04
A2	0.16	-0.02	0.21	-0.23	0.14	0.12	-0.06	-0.17
A3	0.16	-0.17	0.05	0.03	-0.19	-0.14	0.33	-0.03
A4	0.16	-0.03	0.11	-0.28	0.17	0.09	-0.06	0.04
A5	0.16	-0.13	0.04	-0.24	0.06	-0.09	0.03	0.00
A6	0.15	-0.11	0.00	-0.36	0.09	-0.12	-0.07	0.19
A7	0.16	0.00	0.19	-0.20	0.05	0.17	0.03	0.08
A8	0.17	-0.12	0.08	-0.15	-0.05	-0.02	0.16	0.06
A9	0.17	-0.12	0.04	-0.21	-0.02	-0.04	0.08	0.14
A10	0.16	-0.16	0.04	0.06	-0.29	-0.04	0.45	0.08
A11	0.11	0.40	0.11	-0.08	-0.07	-0.13	0.10	-0.04
A12	0.16	-0.05	0.25	0.10	0.16	0.08	0.04	-0.37
A13	0.17	-0.12	0.05	0.09	0.05	-0.13	0.12	0.11
A14	0.17	0.02	0.13	-0.04	0.13	-0.14	-0.35	-0.43
A15	0.17	-0.12	0.10	0.04	0.05	-0.22	-0.13	-0.01
A16	0.16	-0.05	-0.05	-0.04	-0.05	-0.48	-0.29	0.36
A17	0.16	-0.11	0.22	0.15	0.08	0.10	0.04	-0.21
A18	0.17	-0.16	0.11	0.11	0.07	-0.07	0.03	-0.11
A19	0.16	-0.18	0.13	0.09	0.04	-0.14	-0.06	0.03
A20	0.16	-0.18	0.07	0.13	0.03	-0.07	0.13	0.08
A21	0.11	0.40	0.11	-0.08	-0.07	-0.13	0.10	-0.04
A22	0.16	0.13	0.09	0.28	-0.06	0.05	-0.05	0.05
A23	0.13	0.10	-0.36	-0.07	0.42	-0.03	0.15	-0.01
A24	0.16	0.08	0.01	0.30	-0.01	0.03	-0.03	0.09
A25	0.17	0.04	-0.15	0.19	0.14	-0.01	-0.05	-0.01
A26	0.17	-0.01	-0.04	0.25	0.05	-0.03	-0.21	0.07
A27	0.16	0.12	0.10	0.29	-0.09	0.04	0.01	0.15
A28	0.16	0.09	-0.18	0.18	0.15	0.00	0.02	-0.02
A29	0.17	0.05	-0.15	0.18	0.12	-0.03	-0.05	0.04
A30	0.13	0.10	-0.36	-0.06	0.39	-0.05	0.17	0.07
A31	0.11	0.40	0.11	-0.08	-0.07	-0.13	0.10	-0.04
A32	0.17	0.10	0.06	-0.01	-0.01	0.37	0.04	0.25
A33	0.16	0.01	-0.27	-0.04	-0.19	0.04	0.14	-0.19
A34	0.17	0.09	0.10	-0.04	-0.09	0.32	-0.32	0.33
A35	0.17	-0.01	-0.18	-0.07	-0.27	0.10	-0.13	-0.10
A36	0.15	0.01	-0.25	-0.09	-0.39	-0.15	-0.28	-0.17
A37	0.17	0.04	-0.01	-0.03	0.04	0.35	-0.02	0.10
A38	0.17	-0.03	-0.19	-0.06	-0.13	0.20	-0.02	-0.10
A39	0.17	-0.07	-0.16	-0.07	-0.18	0.15	-0.09	-0.10
A40	0.16	-0.06	-0.25	-0.05	-0.10	0.10	0.08	-0.19

For classification, therefore, only the first 8 PCs will be used, instead of all the 40 attributes.

CD.3: 10-Fold Cross-validation

R codes for Cross-Validation

Since data set is large enough, 10-fold cross-validation is applied to evaluate model performance. After removing the outliers 5372 observations are included in the master data and the first 8 principal components are used for prediction. For each observation (row) a score corresponding to each PC is computed and this is the value of the predictors (PCs) used to

evaluate model performance. Hence the master data used has 5372 rows (observations) and 8 predictors and one response (total of 9 columns) indicating the classes each observation belongs to.

Data is divided into 10 sets randomly of which 9 sets have 537 observations and the last set has 539 observations. Training data is formed by taking 9 sets at a time and leave one set out as the Test data. Hence 10 different combinations of Training and Test sets are formed. On each of Training and Test pair a technique is applied and evaluated. Final evaluation of the technique is determined by the average mis-classification probability over the 10 Test sets.

Following table shows classification proportion in the Master data as well as in the Training data sets. Distribution of different classes is almost identical over the data sets. Moreover all categories have almost uniform representation

	Texture Classification										
	2	3	4	6	7	8	9	10	12	13	14
Overall	9.3%	9.3%	9.3%	9.3%	9.3%	8.4%	9.1%	8.4%	9.3%	9.0%	9.3%
Training Set 1	9.4%	9.3%	9.1%	9.2%	9.2%	8.5%	9.1%	8.5%	9.5%	9.0%	9.3%
Training Set 2	9.2%	9.3%	9.4%	9.4%	9.3%	8.5%	9.1%	8.4%	9.5%	8.9%	9.0%
Training Set 3	9.3%	9.5%	9.4%	9.2%	9.5%	8.3%	9.0%	8.2%	9.4%	9.0%	9.4%
Training Set 4	9.2%	9.1%	9.2%	9.4%	9.2%	8.5%	9.3%	8.2%	9.3%	9.1%	9.5%
Training Set 5	9.5%	9.5%	9.1%	9.5%	9.0%	8.4%	9.2%	8.3%	9.4%	9.0%	9.3%
Training Set 6	9.2%	9.3%	9.3%	9.3%	9.2%	8.3%	9.3%	8.5%	9.2%	9.1%	9.4%
Training Set 7	9.5%	9.4%	9.2%	9.3%	9.6%	8.1%	9.2%	8.4%	9.2%	8.9%	9.2%
Training Set 8	9.2%	9.3%	9.4%	9.2%	9.5%	8.5%	9.2%	8.3%	9.2%	8.9%	9.2%
Training Set 9	9.4%	9.1%	9.5%	9.3%	9.2%	8.5%	9.1%	8.4%	9.1%	9.0%	9.4%
Training Set 10	9.2%	9.4%	9.5%	9.2%	9.4%	8.1%	9.0%	8.4%	9.3%	9.2%	9.2%

Linear Discriminant Analysis

R codes for Discriminant Analysis

Since PCs are linear combinations of original variables, they may also be assumed to follow multivariate normal distribution. For each Training set a linear discriminant function is developed using all 8 PCs. Prior probability distribution for each Training set is very similar as given in the table above.

Details are given for the Training Data 1:

Group Means: Training Set 1								
Classes	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
2	-3.106	0.739	-0.749	-0.276	-1.054	0.393	-0.302	0.088
3	-2.281	-1.343	0.318	-1.211	0.270	-0.963	-0.052	-0.102
4	-8.502	1.385	-0.488	2.091	-0.093	0.107	0.056	-0.407
6	3.294	-2.068	-0.290	-0.148	-0.953	0.219	0.344	-0.482
7	2.072	-1.818	3.288	1.990	0.078	0.413	0.053	0.296
8	3.730	1.203	-0.163	-0.693	-0.457	0.182	-0.075	0.561
9	-1.131	1.228	0.057	0.124	-0.300	-1.273	-0.078	0.205
10	4.679	0.245	0.167	-0.093	-0.336	-0.222	-0.389	-0.218
12	-7.622	0.761	0.437	-1.580	0.864	0.758	0.182	0.230
13	8.338	0.343	-2.584	1.175	0.997	0.210	0.299	0.154
14	1.579	-0.552	-0.135	-1.430	0.904	0.160	-0.101	-0.262

Coefficients of Linear Discriminants								
	LD1	LD2	LD3	LD4	LD5	LD6	LD7	LD8
PC1	-0.108	-0.665	0.023	0.124	-0.039	0.025	-0.041	0.050
PC2	-0.738	0.534	-0.239	-0.339	-0.433	-0.004	-0.152	0.355
PC3	2.393	0.590	0.661	0.417	-0.030	-0.095	-0.117	0.265
PC4	2.246	-1.078	-0.437	-0.872	-0.191	-0.309	-0.051	-0.117
PC5	-2.244	-0.404	1.716	-0.736	0.307	-0.955	-0.338	-0.068
PC6	0.521	-0.343	1.098	-1.221	0.347	1.461	-0.300	-0.030
PC7	-0.236	-1.618	0.701	-1.090	0.848	-0.122	1.974	1.077
PC8	-0.200	-0.145	2.134	0.208	-3.133	0.268	0.930	-0.777

Proportion of Trace							
LD1	LD2	LD3	LD4	LD5	LD6	LD7	LD8
0.501	0.310	0.076	0.057	0.032	0.019	0.004	0.001

Results from other Training sets are also very similar and are not shown here. In the following table misclassification probabilities in Training and Test sets created for the 10-fold cross-validation are shown.

	Misclassification Probability: LDA	
	Training Set	Test Set
Set 1	2.5%	2.6%
Set 2	2.5%	3.3%
Set 3	2.5%	3.3%
Set 4	2.5%	2.8%
Set 5	2.6%	2.0%
Set 6	2.5%	3.0%
Set 7	2.6%	1.7%
Set 8	2.7%	1.3%
Set 9	2.6%	2.0%
Set 10	2.4%	3.5%

Therefore overall misclassification probability of the 10-fold cross-validation is 2.55%, which is the mean misclassification probability of the Test sets.

Note that for Sets 5, 7, 8 and 9 mis-classification probability in Test set is less than that in the corresponding Training set. This may seem fallacious; however, several points to be noted here. Training set size is much larger compared to Test set. With 11 classes in Test sets, each class

has sometimes even fewer than 40 representations. This might lead to the standard error of probability of misclassification to be relatively higher, in turn leading to apparent counter-intuitive results. Average error of Training set is 2.54%.

Overall results indicate accurate and stable classification rules.

CD.4: K Nearest Neighbour

For this method $k = 7$ is used, i.e. 7 nearest neighbours are used to predict class membership of each observation in the Test set. Following is the misclassification error rate for Test sets.

	Misclassification Rate for Test Set
Set 1	3.0%
Set 2	4.1%
Set 3	4.8%
Set 4	2.2%
Set 5	1.9%
Set 6	2.6%
Set 7	2.0%
Set 8	2.4%
Set 9	2.0%
Set 10	2.6%

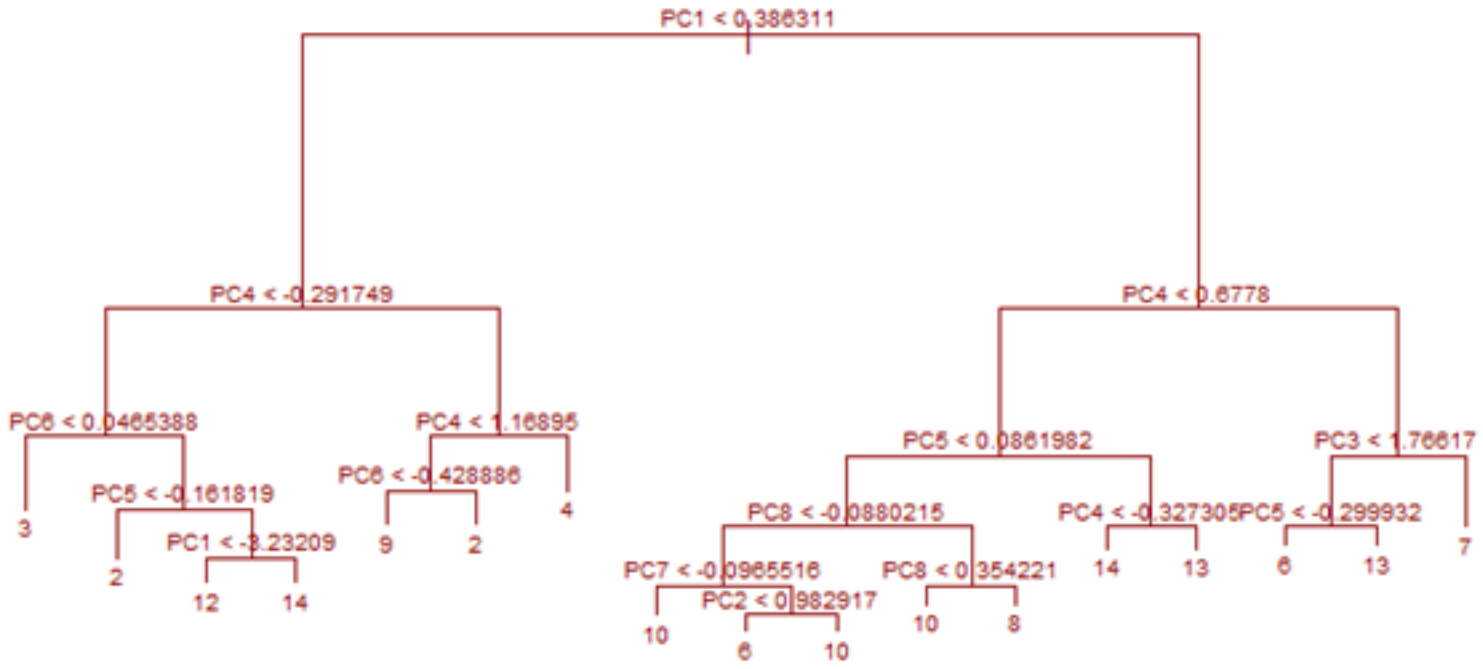
Overall error rate is 3.0%. This value is comparable to the LDA error rate.

CD.5: Decision Tree

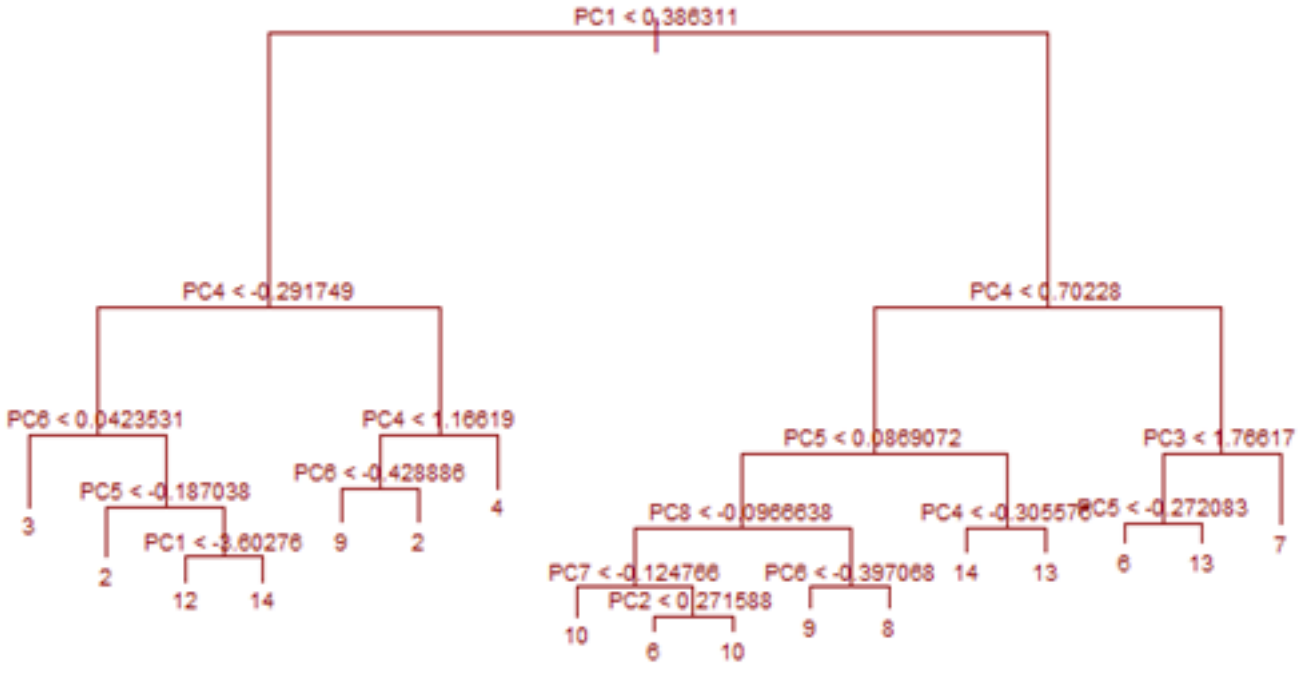
R codes for Tree Based Algorithms

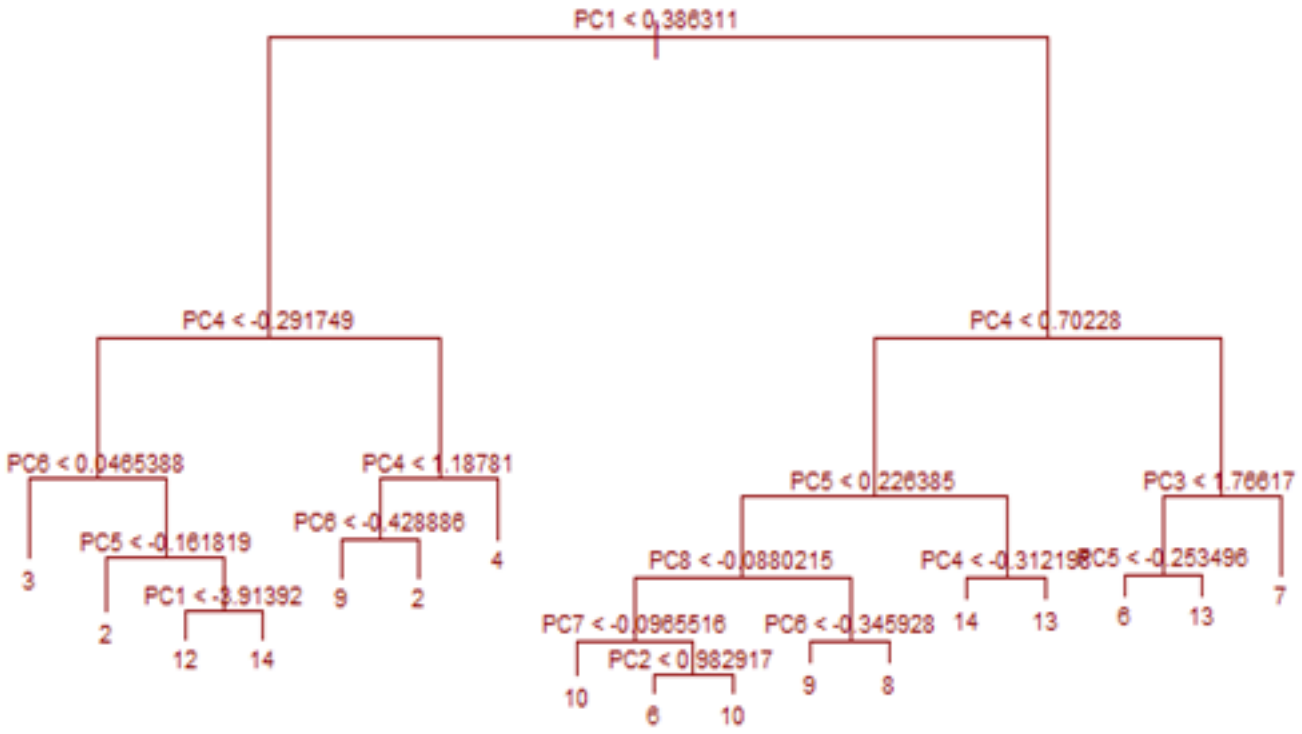
Unsupervised tree algorithm is applied to all Training sets and misclassification probability was calculated for both the Training and Test sets. All the Training Sets give rise to very similar decision trees. Three representative trees are shown below as examples.

Decision Tree (Unpruned) for Training Set 1



Decision Tree (Unpruned) for Training Set 5





Following table summarizes the misclassification probabilities for Tree classification

	Misclassification Probability: Tree	
	Training Set	Test Set
Set 1	17.1%	18.8%
Set 2	16.0%	17.3%
Set 3	17.1%	22.3%
Set 4	16.6%	18.4%
Set 5	16.5%	15.3%
Set 6	17.5%	15.6%
Set 7	17.3%	18.2%
Set 8	16.0%	18.6%
Set 9	17.7%	15.9%
Set 10	15.9%	18.9%

Therefore overall mis-classification probability of the 10-fold cross-validation is 17.9%, which is the mean mis-classification probability of the Test sets.

Pruning was tried for this decision tree, but it did not improve the result.

At the first glance this high error rate compared to *k*-NN and LDA looks surprising. LDA uses only linear classifier all over the sample space, but Tree procedure recursively partitions the sample space to reduce mis-classification error. It is therefore expected that Tree procedure will always give better results than LDA.

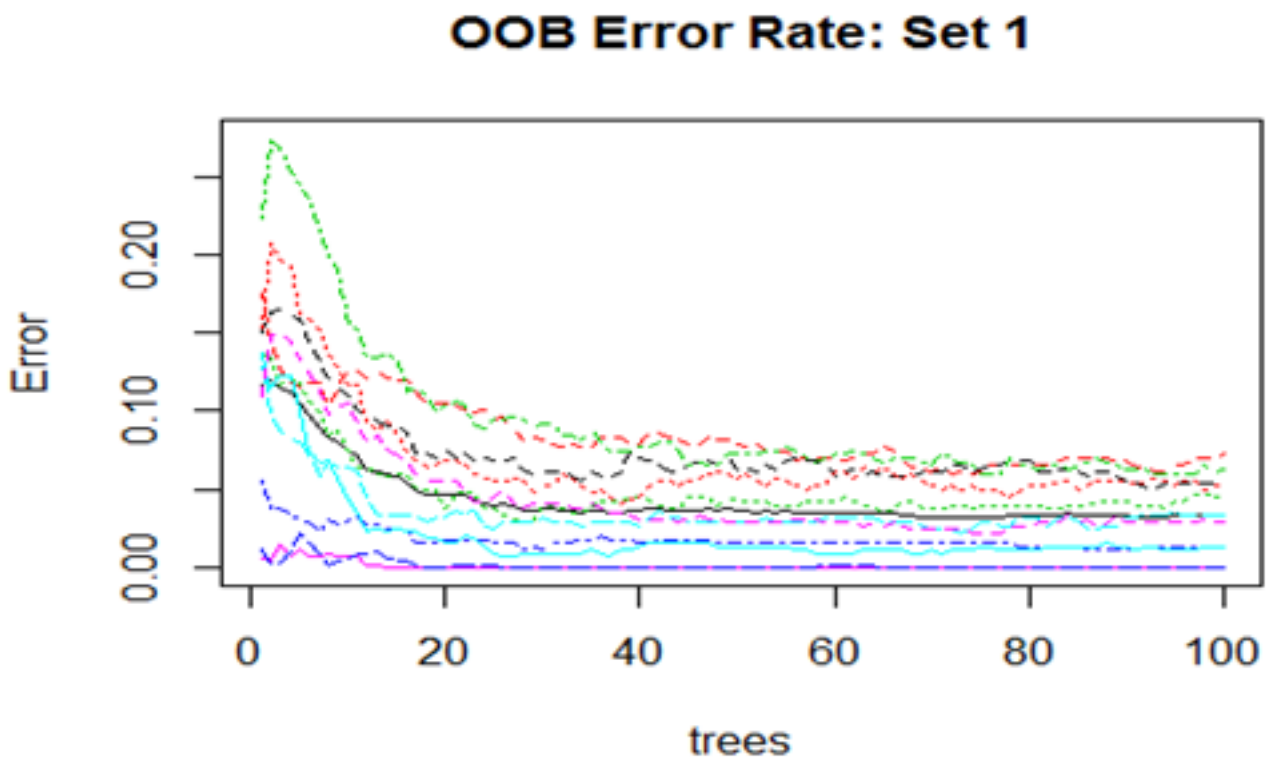
However, it is to be noted that LDA takes into account linear combinations of the predictors, whereas Tree always divides the sample space into splits parallel to the axes. If separation is along any other line, Tree wil not be able to capture that. This is exactly what is happening here.

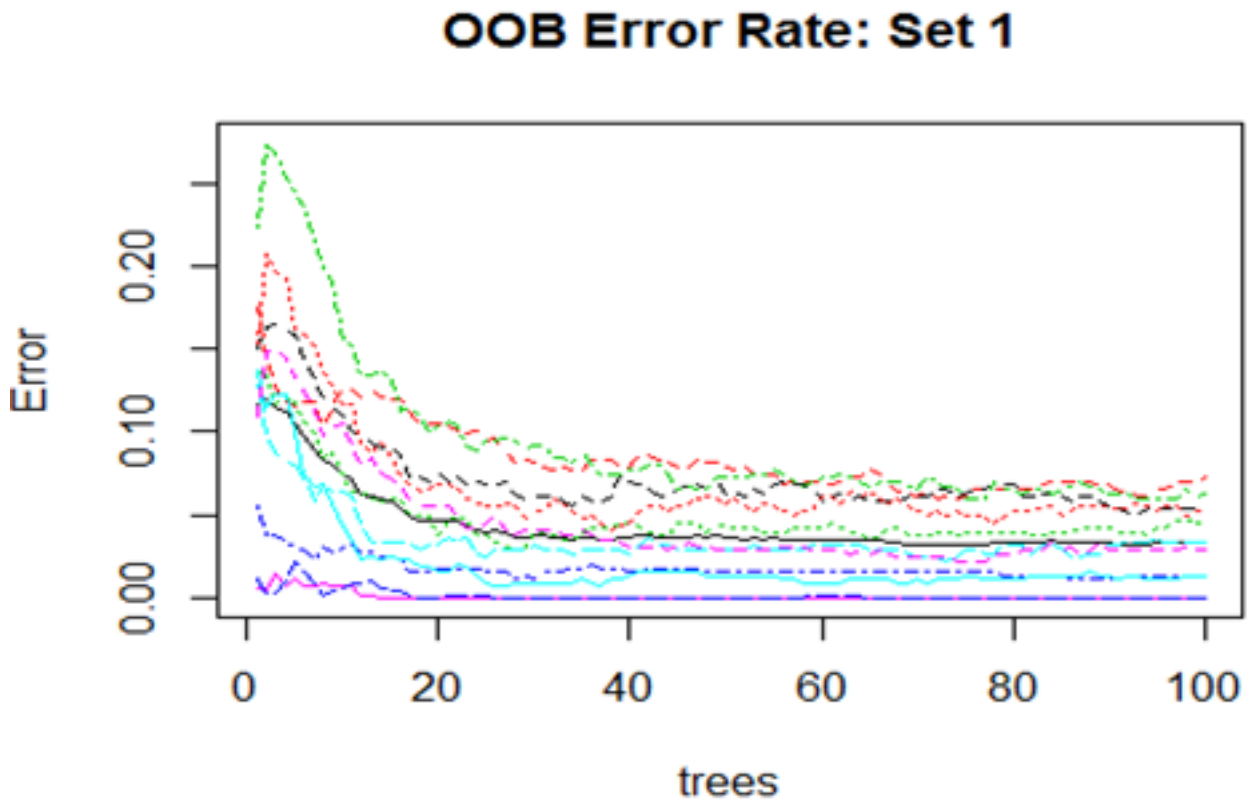
CD.6: Random Forest

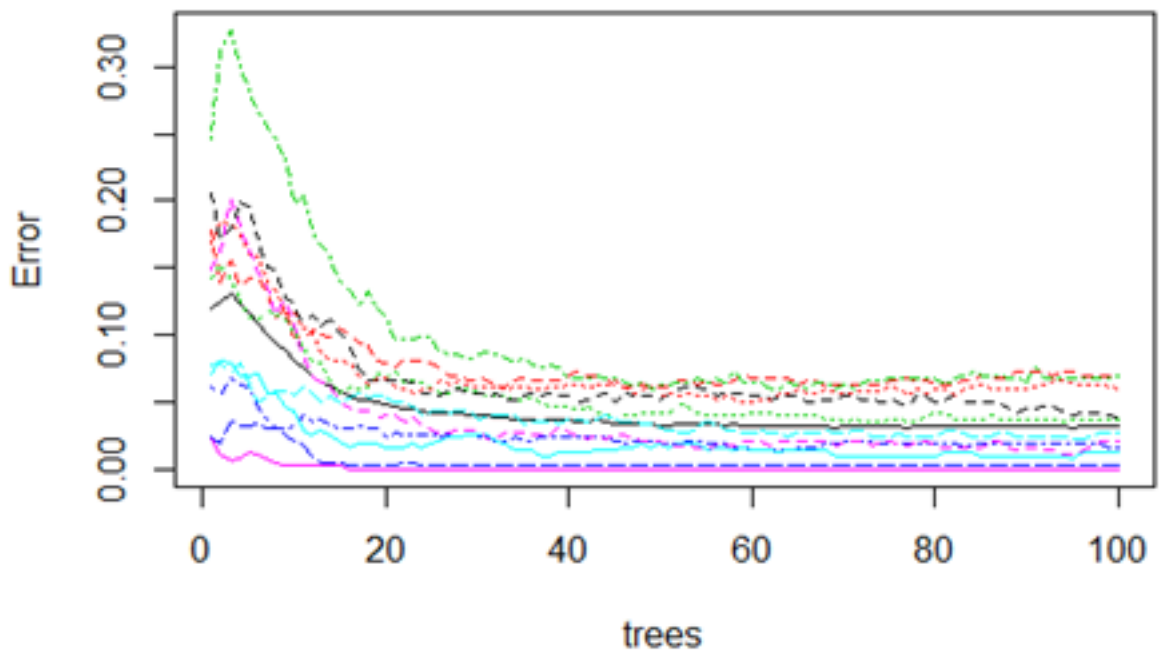
High mis-classification error rate is corrected to a large extent by using Random Forest. Unsupervised random forest method is applied to each Training set and both Out-of-Bag error rate and Test error rate are calculated for error or mis-classification corresponding to each of the 11 categories. Variable importance plots are also shown. All results are for *ntree* = 100. Number of variables tried at each split is 2.

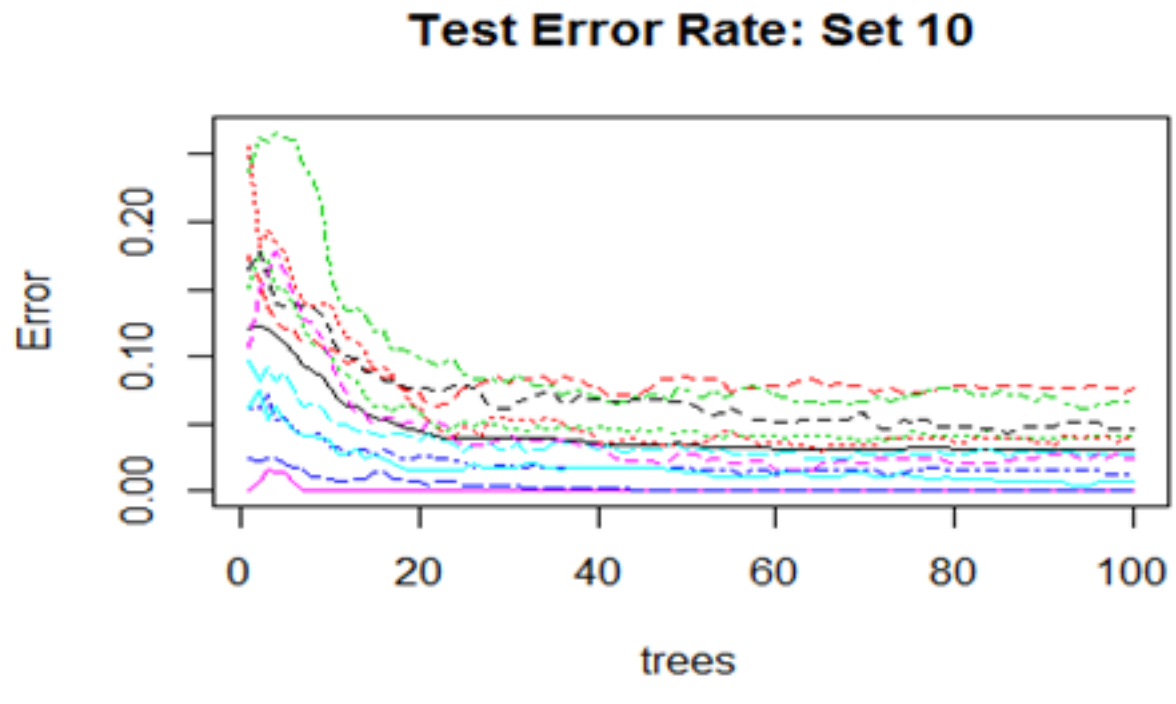
	Random Forest Error Rate	
	Out-of-Bag Error	Test Error
Set 1	3.27%	2.98%
Set 2	3.14%	3.17%
Set 3	3.21%	3.35%
Set 4	3.23%	2.42%
Set 5	3.14%	2.42%
Set 6	3.02%	3.17%
Set 7	3.19%	3.54%
Set 8	2.81%	2.79%
Set 9	3.25%	2.61%
Set 10	3.04%	3.53%

Convergence of the errors are shown for two Sets only since behaviour of the error is very similar for all the 10 Sets random forest technique was applied. For Set 1 both the Out-of-Bag error rate and Test error rate are shown.

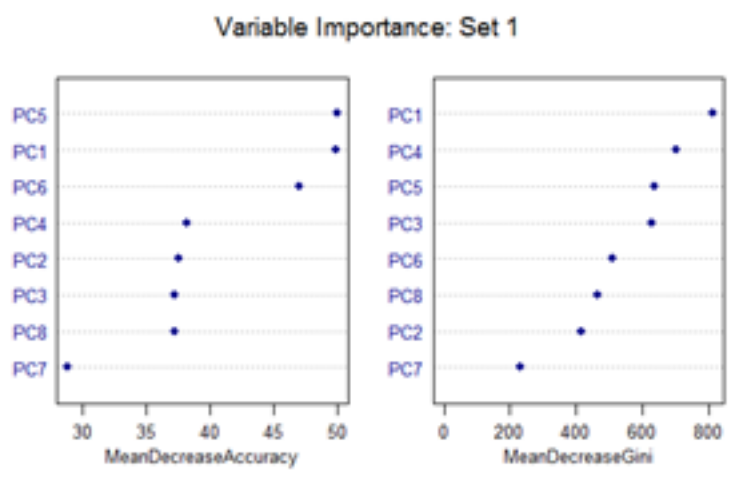


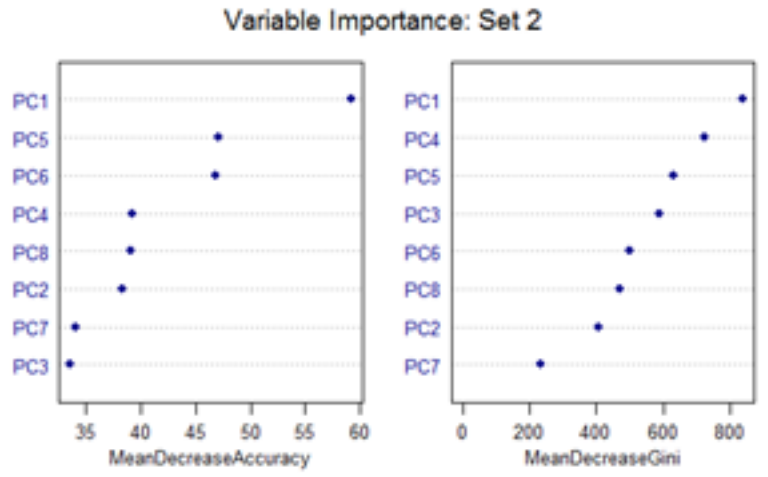

OOB Error Rate: Set 1
Error
0.00
0.10
0.20
0
20
40
60
80
100
trees

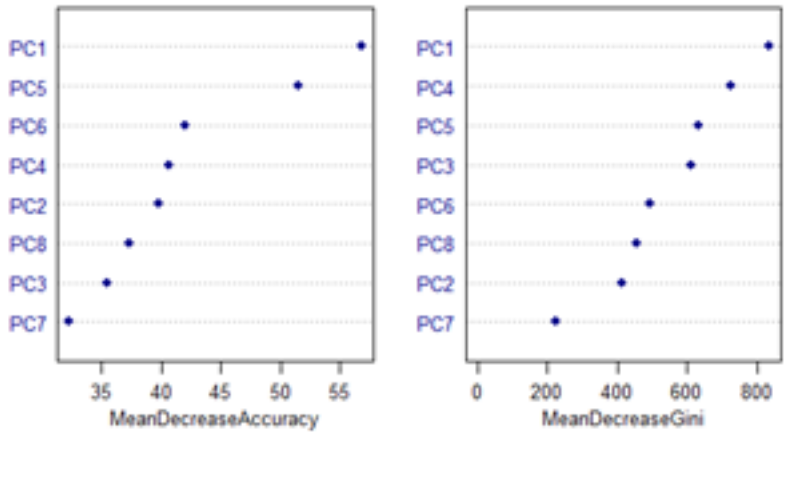

Test Error Rate: Set 5
Error
0.00
0.10
0.20
0.30
0
20
40
60
80
100
trees

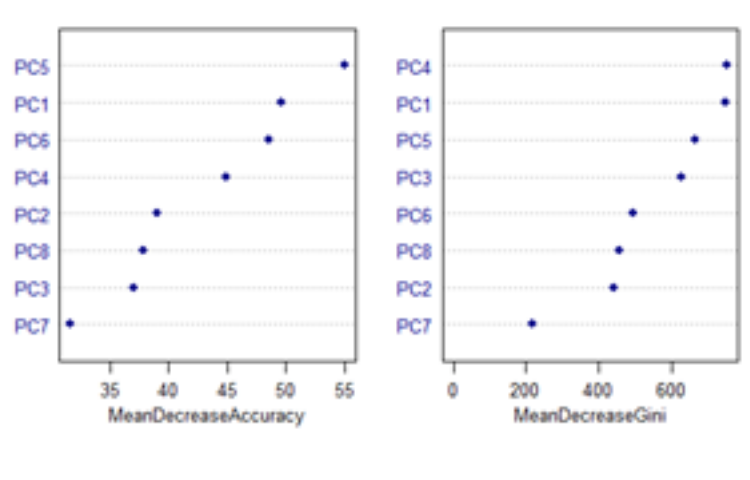

Test Error Rate: Set 10
Error
0.00
0.10
0.20
0
20
40
60
80
100
trees

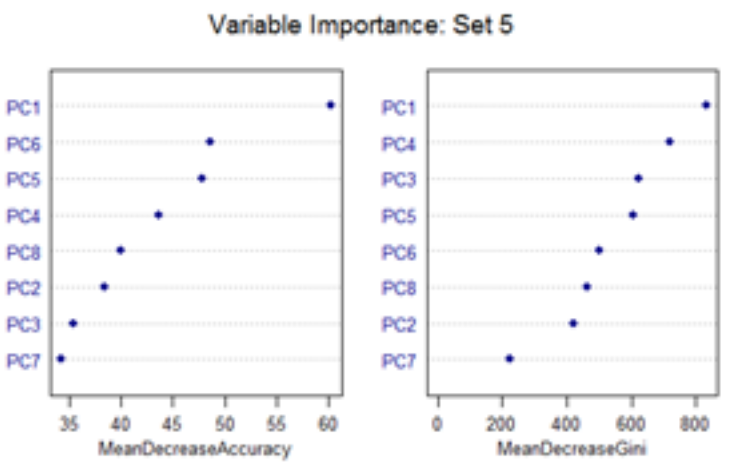
However, there are slight differences in the variable importance plots. All the plots are shown below. It is clear from the plots below that PC1, PC5 and PC6 are the primary influential variables, but they appear in different order in different cross-validation sets.

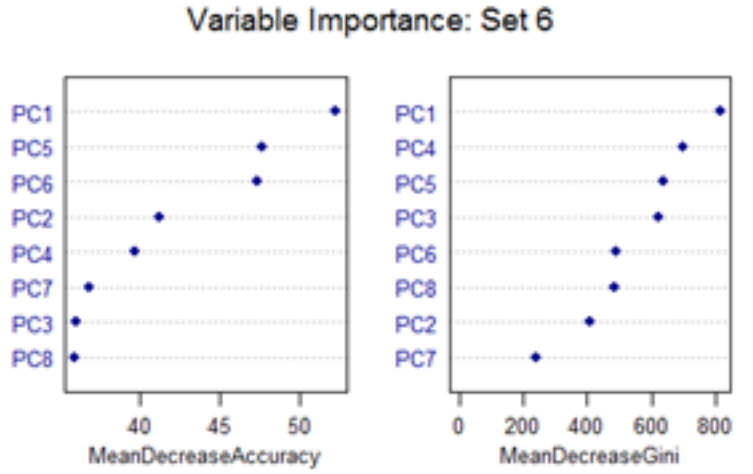


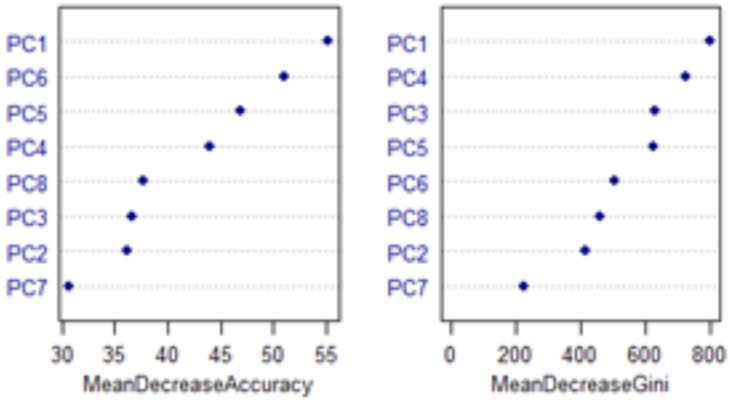


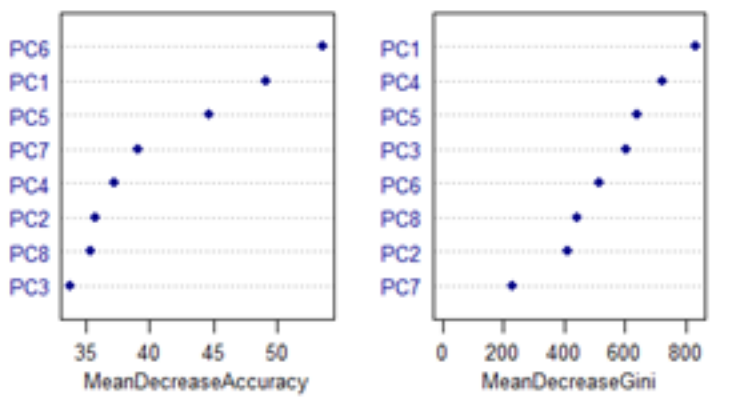


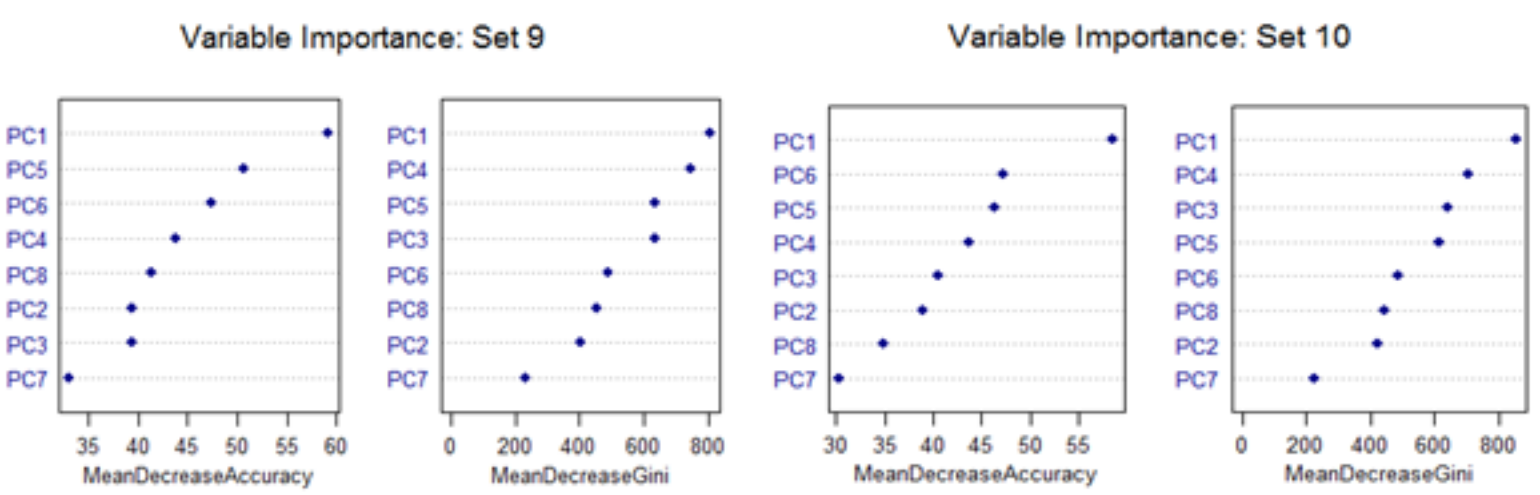












CD.7: Conclusion

All the 11 different categories can be separated by various classification methods with almost similar misclassification error rates. Classification Tree-based method shows less than optimal performance. But this can be improved by application of Random Forest or Oblique Tree method.

Source URL: https://onlinecourses.science.psu.edu/stat857/node/231

Links:
[1] http://sci2s.ugr.es/keel/dataset.php?cod=72#sub2
[2] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/Texture.csv
[3] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTrain1.csv
[4] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTrain2.csv
[5] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTrain3.csv
[6] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTrain4.csv
[7] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTrain5.csv
[8] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTrain6.csv
[9] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTrain7.csv
[10] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTrain8.csv
[11] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTrain9.csv
[12] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTrain10.csv
[13] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTest1.csv
[14] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTest2.csv
[15] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTest3.csv
[16] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTest4.csv
[17] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTest5.csv
[18] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTest6.csv
[19] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTest7.csv
[20] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTest8.csv
[21] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTest9.csv
[22] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/TXTest10.csv
[23] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/Texture.zip