Mushroom Classification

Winfield Lai

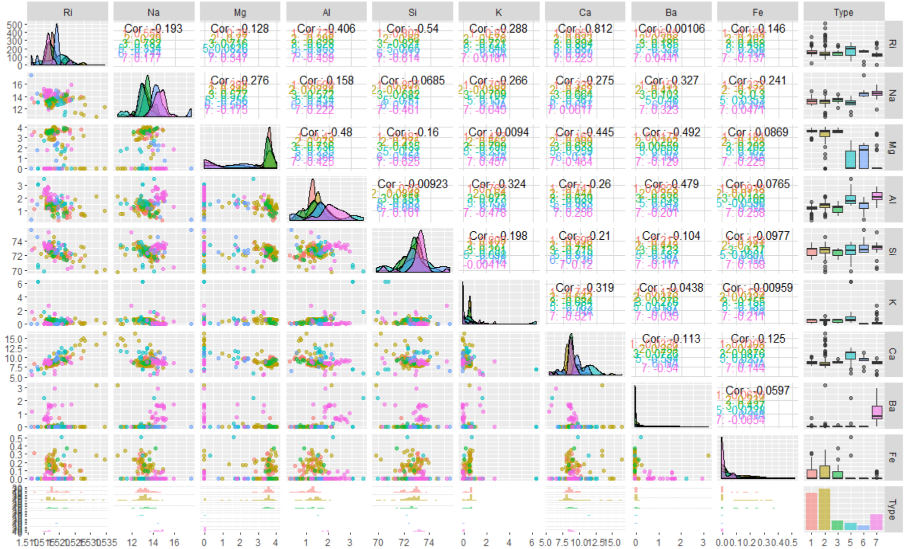
March 31, 2019

Glass Data

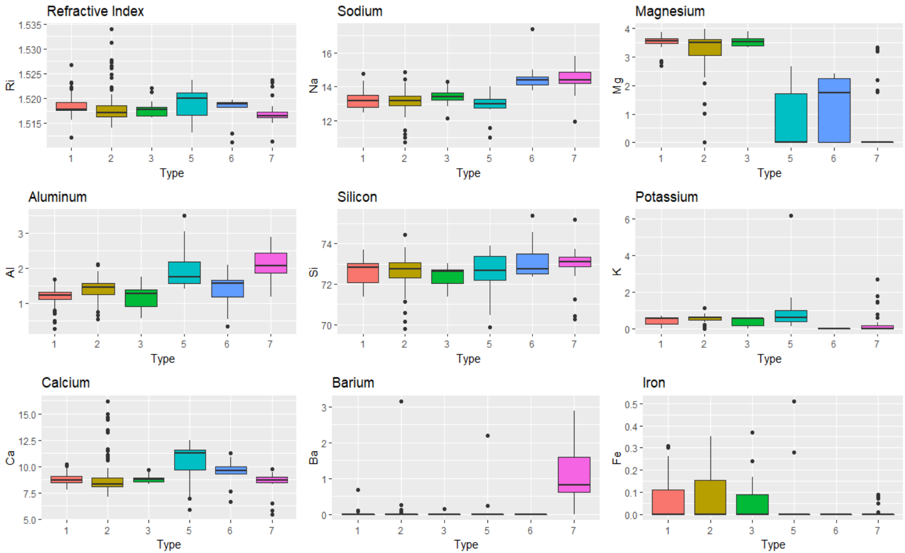
- There are 7 types of glass we are interested in
  - 1 → Building Window Float Processed
  - 2 → Building Window Non-Float Processed
  - 3 → Vehicle Window Float Processed
  - 4 → Vehicle Window Non-Float Processed
  - 5 → Container, 6 → Tableware, 7 → Headlamp
- There are no observation of glass type 4, Vehicle Window Non-Float Processed
- We have the percent composition of 8 different elements and the refractive index(Ri) for each observation

Id	Ri	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type
1	1.52	13.89	3.60	1.36	72.73	0.48	7.83	0.00	0.00	1
2	1.52	13.53	3.55	1.54	72.99	0.39	7.78	0.00	0.00	1
3	1.52	13.21	3.69	1.29	72.61	0.57	8.22	0.00	0.00	1
4	1.52	13.27	3.62	1.24	73.08	0.55	8.07	0.00	0.00	1
5	1.52	12.79	3.61	1.62	72.97	0.64	8.07	0.00	0.26	1
6	1.52	13.30	3.60	1.14	73.09	0.58	8.17	0.00	0.00	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Paris Plot of Glass Data



Box Plots of Glass Data



Classification Methods

- We will be the following Classification methods
  - K-Nearest Neighbours
  - Classification Trees
- We will be using stratified sampling due to unbalanced number of glass types

Type of Glass	1	2	3	4	5	6	7
Number of Observations	69	76	17	0	13	9	29

K Nearest Neighbours

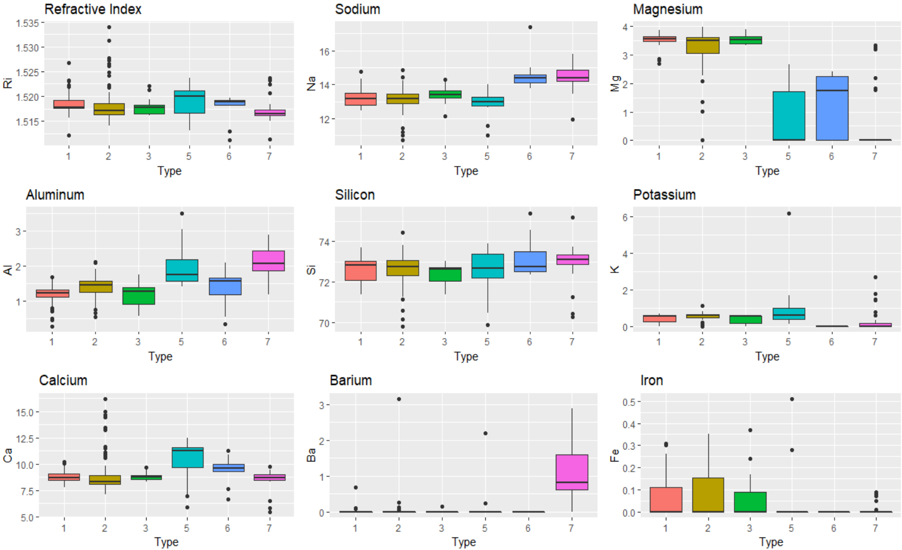
- Testing the KNN algorithm for 1 to 10 Neighbours(N). Performance indicated by the basic misclassification rate and adjusted rand index(ARI)

Neighbours	Misclassification Rate	ARI
1.00	0.68	0.35
2.00	0.61	0.31
3.00	0.59	0.26
4.00	0.58	0.28
5.00	0.64	0.34
6.00	0.54	0.16
7.00	0.63	0.26
8.00	0.56	0.19
9.00	0.61	0.24
10.00	0.63	0.28

	1	2	3	5	6	7
1	14	2	2	0	0	0
2	5	12	0	2	0	0
3	4	0	1	0	0	0
5	0	0	0	4	0	0
6	0	1	0	0	4	0
7	0	1	1	1	0	5

K Nearest Neighbours: Overlap

- Clearly from the box plots, we see overlap in many of the variables



K Nearest Neighbours: Removal

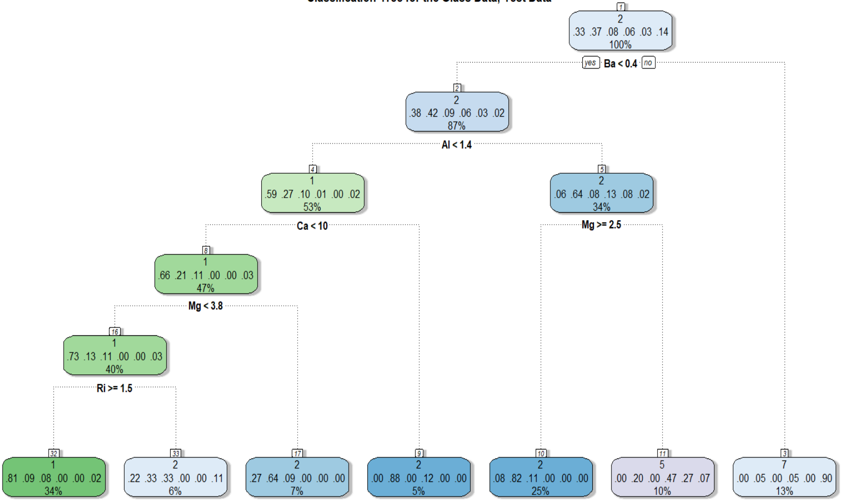
- If we remove some of the overlapping variables, we might increase the ARI. Ri, Na, Si, K, and Ca were chosen to be removed because they had overlapping boxes in the box plots and densities in the density plots (slide 3) in the Pairs plot. Some variables were left in as to not over remove every variable.

Neighbours	Misclassification Rate	ARI
1.00	0.98	0.98
2.00	0.93	0.94
3.00	0.95	0.96
4.00	0.93	0.95
5.00	0.93	0.95
6.00	0.92	0.97
7.00	0.95	0.97
8.00	0.92	0.93
9.00	0.90	0.91
10.00	0.90	0.93

	1	2	3	5	6	7
1	18	0	0	0	0	0
2	0	19	0	0	0	0
3	0	0	5	0	0	0
5	0	0	0	4	0	0
6	0	0	0	0	5	0
7	0	0	0	0	1	7

Classification Tree

Classification Tree for the Glass Data, Test Data



Rattle 2019-Mar-31 16:40:15 Winfield Lai

Classification Tree

- ARI of 0.3088442
- Misclassification Rate of 0.6440678
- Root Node Error of $137/213 = 0.64319$

CP	nsplit	rel error	xerror	xstd
0.22	0.00	1.00	1.06	0.06
0.07	2.00	0.56	0.58	0.06
0.04	3.00	0.48	0.55	0.06
0.01	5.00	0.40	0.52	0.06
0.01	6.00	0.39	0.49	0.06

	1	2	3	5	6	7
1	12	4	1	0	1	0
2	5	14	4	0	2	0
3	0	0	0	0	0	0
5	0	1	0	4	2	0
6	0	0	0	0	0	0
7	1	0	0	0	0	8

Conclusion

- The classification tree is poor at predicting the correct classifications of data.
- The KNN algorithm performed about the same as classification trees when accounting for all variables
- The KNN algorithm performed extremely well, better than classification trees, after the removal of some variables