**Introduction** The motivation behind this project was to learn more about the quick application of ML algorithms by experimenting with various classification models and the tuning parameters associated with each model. Caret is a powerful R library which is used for Classifaction And REgression Training, and I wanted to get familiar with this package and different classification algorithms.

**Data** This dataset consists of the:

* Refractive Index
* Measures of various elements in the glass
* The Glass Type (1-7) eg. *for headlamps, containers, building windows, vehicle windows etc. \*\** (Target Variable)\*\*

**Aim**

* The aim was to identify patterns between variables in the dataset through Principle Component Analysis (PCA)
* Apply and evaluate different classification models and see how each one performs based on the 'Accuracy' and 'Kappa' metrics

*I expected the Random Forest model to perform the best, so just felt lke exploring different parameter tuning methods (randomSearch and gridSearch through caret)*

Context

This is a Glass Identification Data Set from UCI. It contains 10 attributes including id. The response is glass type(discrete 7 values)

Content

Attribute Information:

1. Id number: 1 to 214 (removed from CSV file)
2. RI: refractive index
3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
4. Mg: Magnesium
5. Al: Aluminum
6. Si: Silicon
7. K: Potassium
8. Ca: Calcium
9. Ba: Barium
10. Fe: Iron
11. Type of glass: (class attribute) -- 1 building\_windows\_float\_processed -- 2 building\_windows\_non\_float\_processed -- 3 vehicle\_windows\_float\_processed -- 4 vehicle\_windows\_non\_float\_processed (none in this database) -- 5 containers -- 6 tableware -- 7 headlamps

Acknowledgements

<https://archive.ics.uci.edu/ml/datasets/Glass+Identification> Source:

Creator: B. German Central Research Establishment Home Office Forensic Science Service Aldermaston, Reading, Berkshire RG7 4PN

Donor: Vina Spiehler, Ph.D., DABFT Diagnostic Products Corporation (213) 776-0180 (ext 3014)

Inspiration

Data exploration of this dataset reveals two important characteristics : 1) The variables are highly corelated with each other including the response variables: So which kind of ML algorithm is most suitable for this dataset Random Forest , KNN or other? Also since dataset is too small is there any chance of applying PCA or it should be completely avoided?

2) Highly Skewed Data: Is scaling sufficient or are there any other techniques which should be applied to normalize data? Like BOX-COX Power transformation?