Math 218: Final Report

Gauss-Seidel

Dan & DJ

December 2022

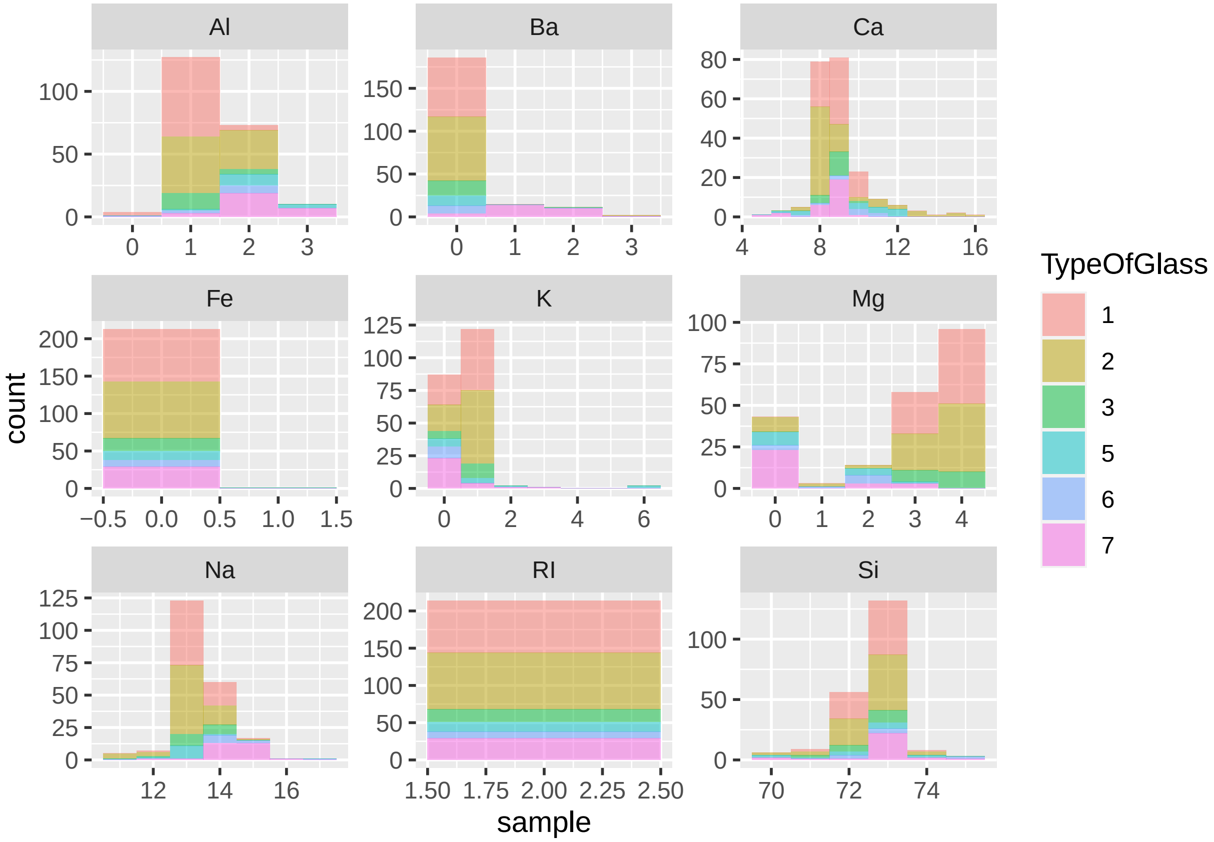
Introduction

Identifying the type of glass from a fragment has important forensic use. To this end, the U.K. Home Office Forensic Science Service conducted chemical analyses in 1987 on 214 samples of glass from 7 broad categories: float-processed building windows (1), non-float-processed building windows (2), float-processed vehicle windows (3), non-float-processed vehicle windows (4), containers (5), tableware (6), and headlamps (6). Float processing refers to the method of manufacturing where the glass is floated on a bath of molten tin after being poured to ensure flatness (a video describing the process is available at https://www.pilkington.com/en/global/knowledge-base/glass-technology/the-float-process/the-float-process#). The chemical analysis included percent composition by mass of Na, Mg, Al, Si, K, Ca, Ba, and Fe, as well as the refractive index. The dataset was downloaded from Kaggle (at https://www.kaggle.com/datasets/danushkumarv/glass-identification-data-set) and the Home Office's paper on the data set is available in the GitHub repository titled "evett_1987_reifs.pdf". Note that there are no observations in the data set corresponding to non-float-processed vehicle windows (4). This is because almost all car windows are float-processed.

We use supervised learning methods KNN and decision trees to determine the type of glass from the elemental composition and refractive index. Additionally we use unsupervised k-means clustering on the elemental composition and refractive index to determine if we would indeed expect 6 types of glass to be present in this database. Specifically, we will used our supervised methods and our unsupervised method to answer the follow questions, respectively: 1.) Given the measure of each element, and the refractive index, can we correctly classify the type of glass? 2.) Given only the measure of each element, will clusters reflect the actual classifications?

EDA

Before we begin analyses let us first consider the distribution of our predictors: elemental composition and refractive index. Glass is primarily silicon with small amounts of other elements added depending on the desired properties. Due to their small quantity these elements have small ranges of values which will lead to distance sensitive methods like KNN not giving them equal influence. Additionally, refractive index is in different units than percent mass. Thus, to ensure equal weighting of all predictors, we see we must standardize the data.



Methodology

Note a random seed of 1 was used throughout.

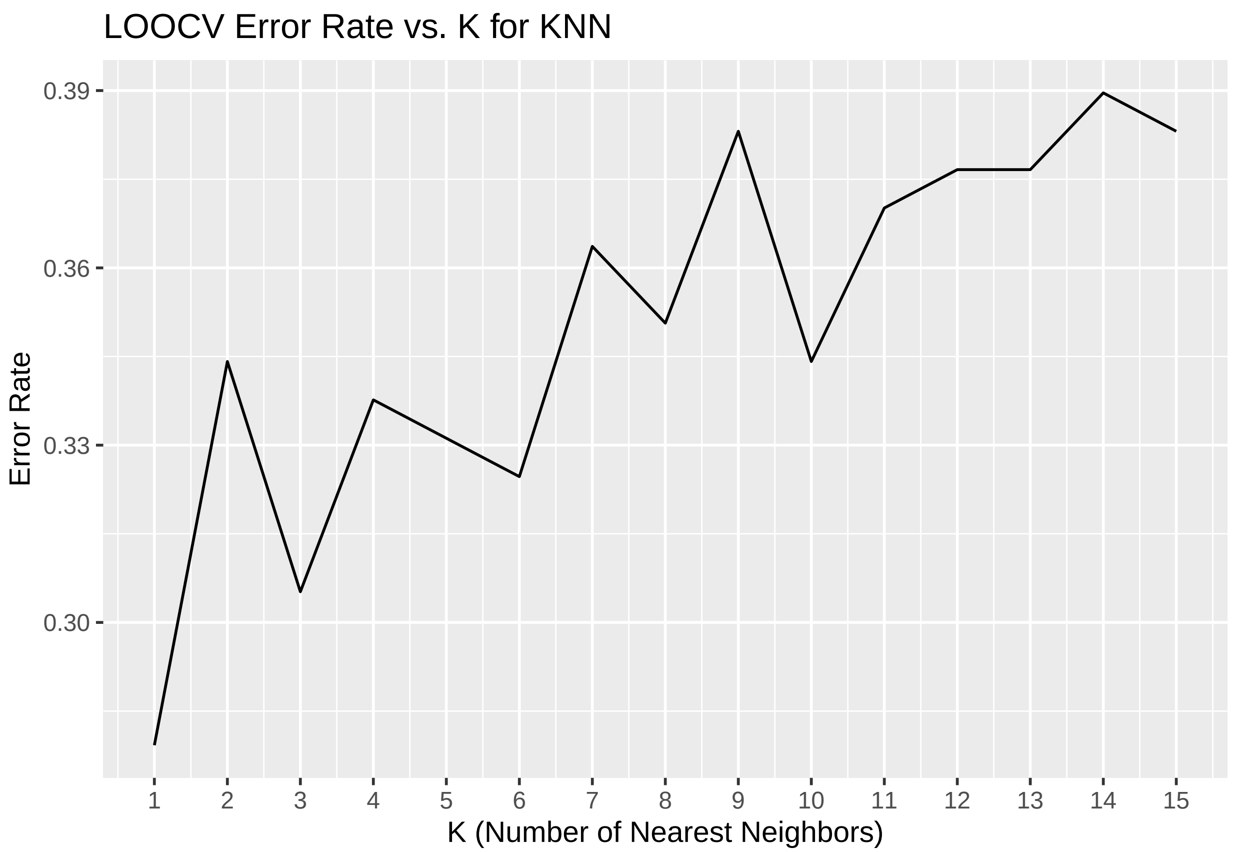
We start by standardizing the data, such that we have standard deviation of 1 and mean of 0 for each predictor.

Next, we split the data proportionally, to ensure each type of glass has proportional representation in our training and testing sets. We use roughly 70% of the data for training, and the remaining data for testing.

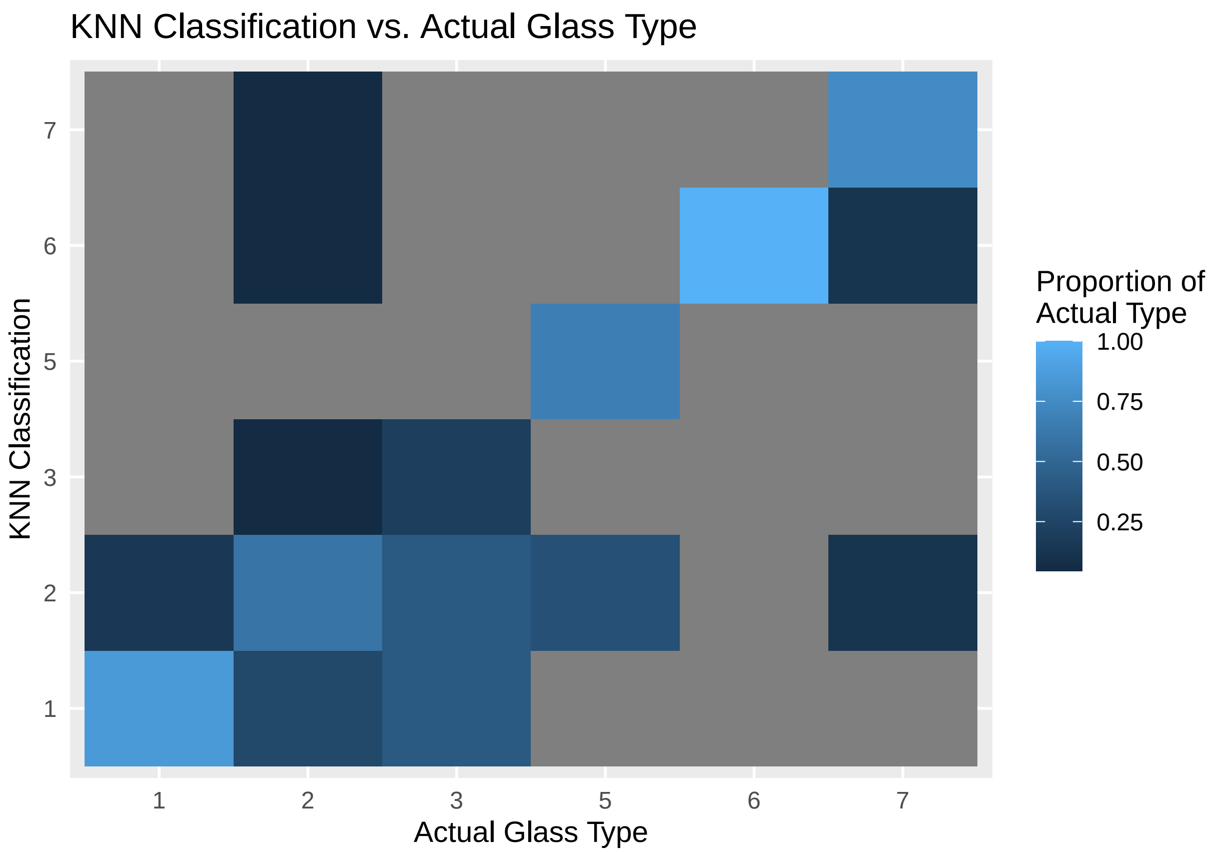
KNN

We start with our supervised learning models. First is KNN. KNN makes predictions by looking at the K nearest neighbors to a given point, and assigning that given point the most common value of these neighbors. This method makese sense in our data, since we expect each type of glass to fit some elemental profile. That is, we expect one piece of tableware glass to have similar properties to another piece of tableware glass.

We use LOOCV on our training data to determine the optimal K (number of neighbors).



Based on our graph, we see K=1 nearest neighbor is the optimal value. We now use this K to build a model, which we then use to make predictions for our test data.



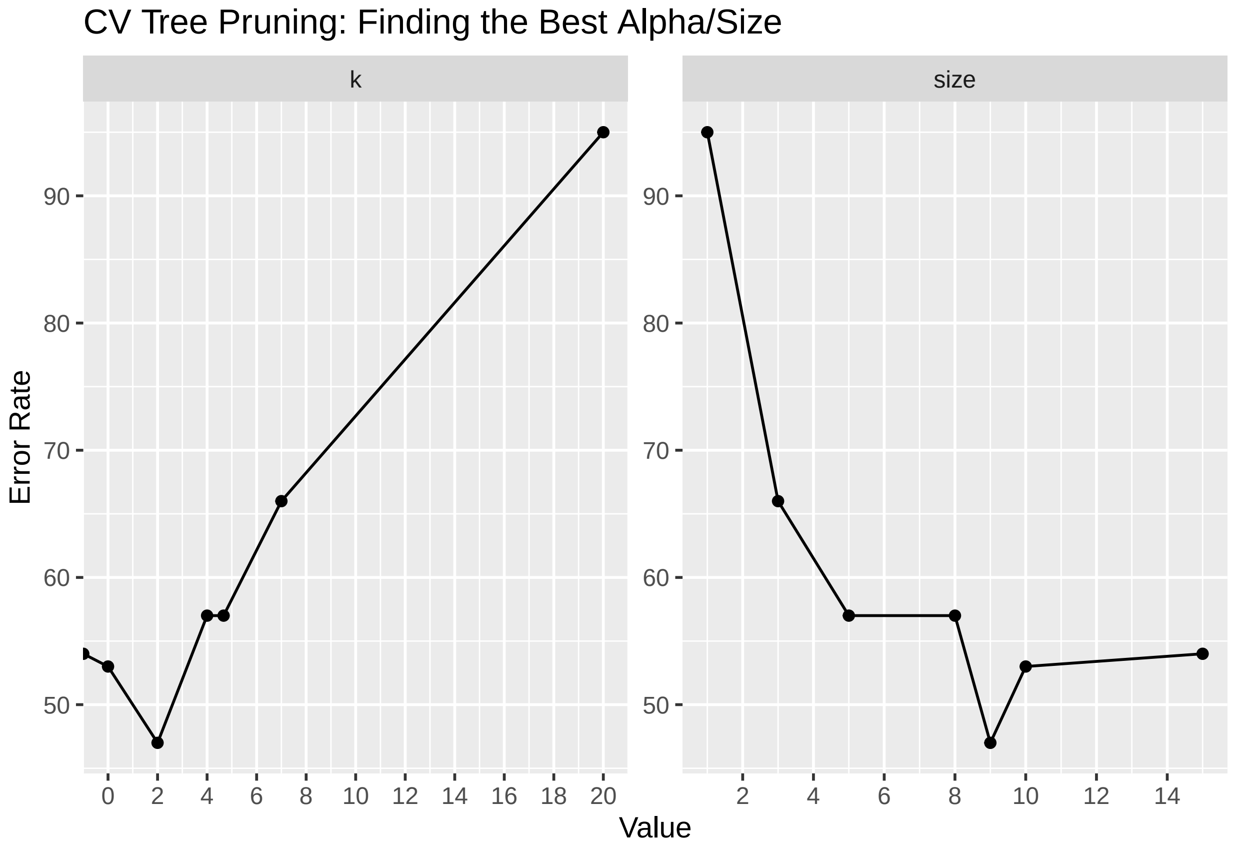
We can see that the test error rate for our KNN model is about 31.7%.

We also include above a heat map showing the proportion of our method's classification of each type of glass. That is, the x-axis represents the actual glass type and the y-axis represents the type KNN assigned to points. Recall that we have no samples of glass type 4, and thus type 4 is left off our axes. Note that each column is proportional to the number of points of the corresponding true type, and thus the colors are based on the proportion of the points of each actual type assigned each predicted type. For each column, the values associated with each color thus sum to 1. Looking at this heat map, KNN has done a reasonably good job predicting the type of our points. The mostly light blue anti-diagonal shows this, as it represents a relatively high proportion of accurate classification. We do note worse performance in classifying points of type 3, as well as a fair number of misclassified points of type 2.

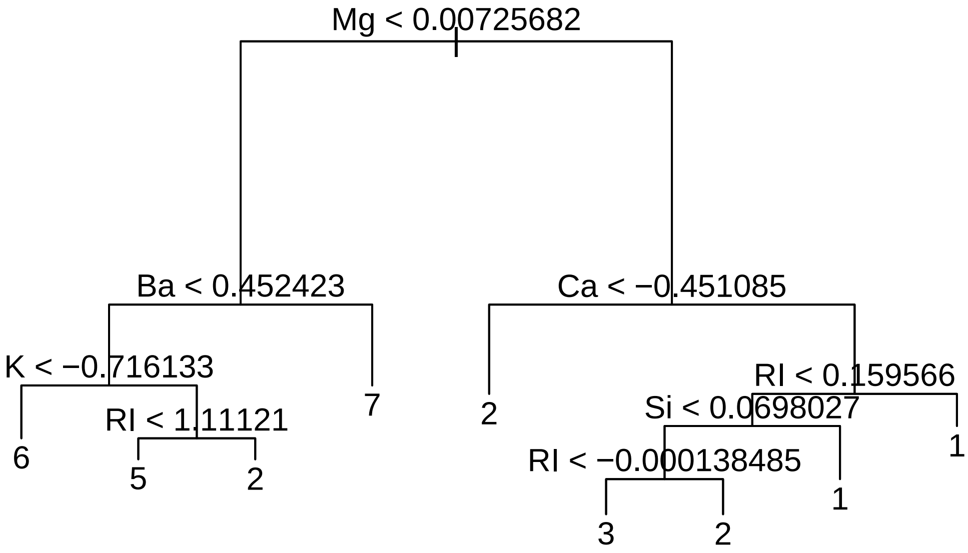
Decision Tree

We now turn to making a decision tree. A decision tree is formed by greedily choosing the pivot points that maximize node purity. A decision tree is a good choice for our data set because, again, we expect observations of the same type of glass to be relatively similar in our predictors, and thus to be near each other in predictor space. Since our decision tree essentially creates a partition of predictor, we expect it might be able to accurately assign types of glass to our points. We will now make our initial tree using our training data.

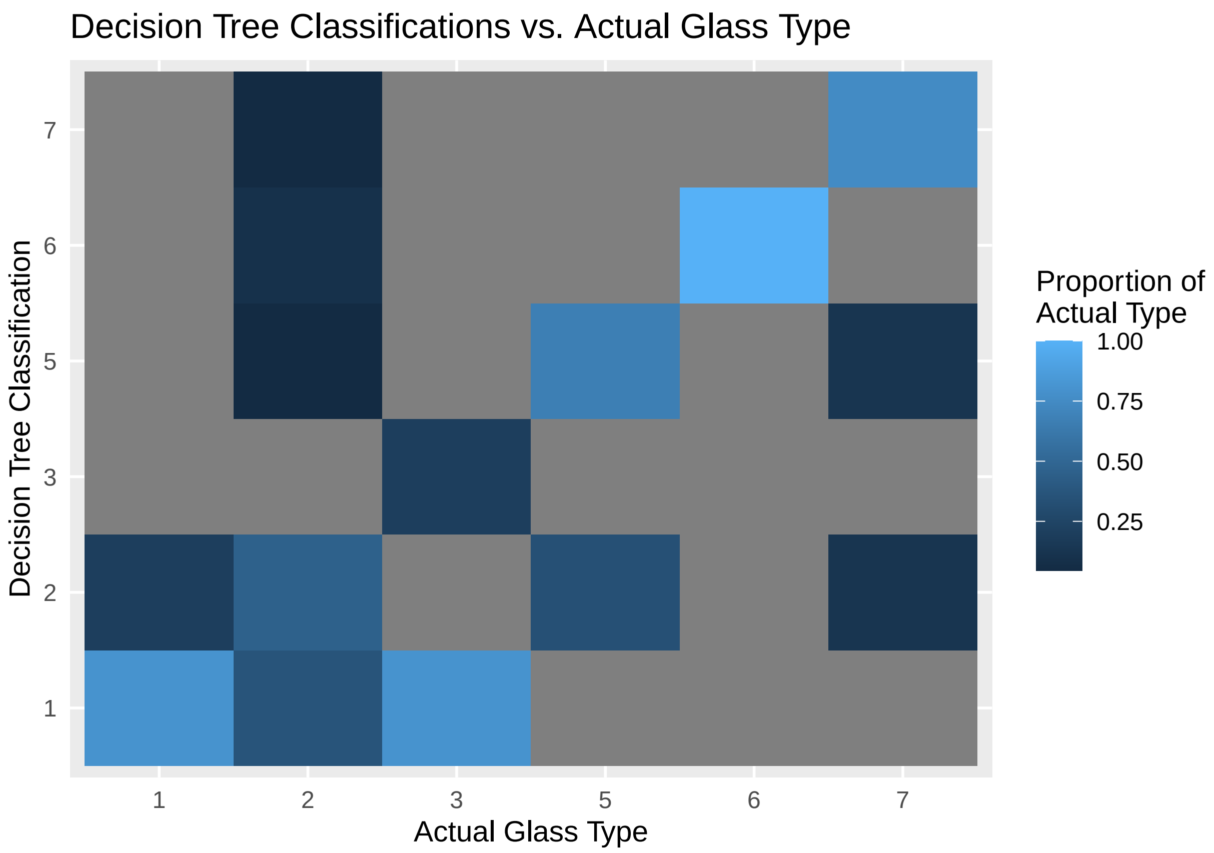
Having made our initial tree, we will prune it. First, we use K-fold cross validation, with K = 10, guided by the misclassification rate, to find the optimal tree size to prune to.



Looking at our graph, we choose a tree size of 9 leaves, as this is the size minimizing the error rate. We now prune our tree to this size.



Above, we have our pruned tree. It seems as though magnesium is the most important element in our tree's classification. We now use our tree to make glass type predictions on for test data.



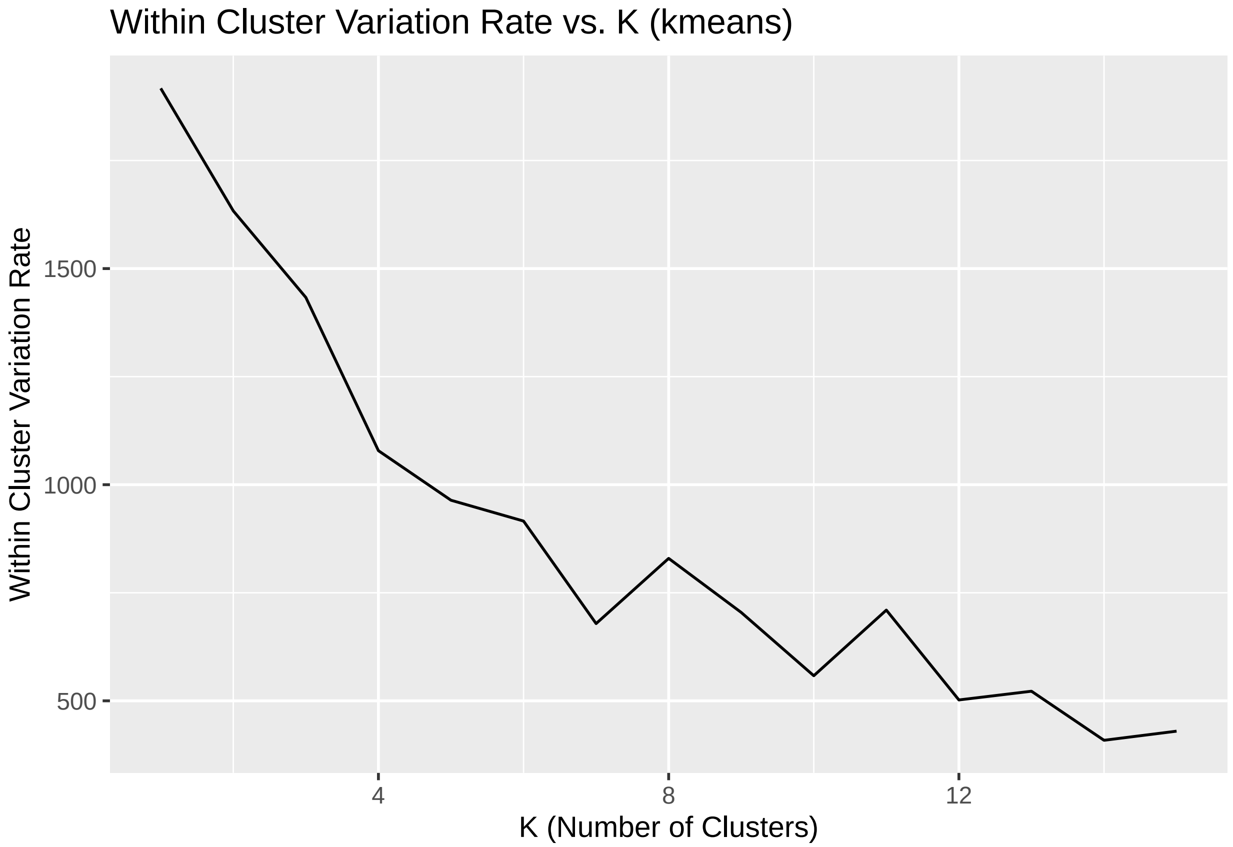
```
## [1] 0.3833333
```

We can see that our decision tree's misclassificaiton rate is about 38.3%. Looking at our heat map, we can see again by the light blue anti-diagonal that our decision tree does a fair job. Again, points of type 3 are mostly misclassified, seemingly moreso than by KNN. Points of type 2 are also again broadly misclassified, also moreso than they were by KNN.

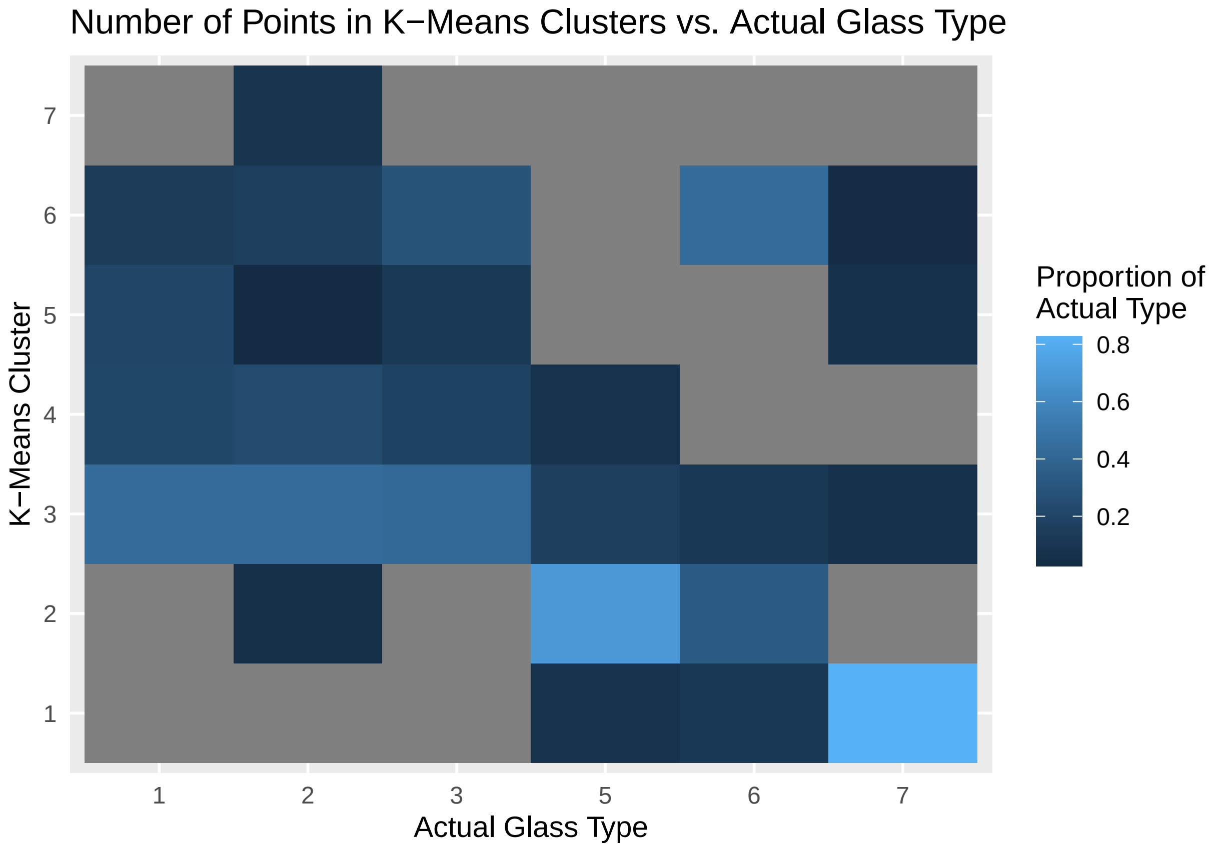
K-Means

Finally, we move to unsupervised learning. Specifically, we run K-means clustering on our entire data set (there is no need to use the train/test data, since we aren't actually predicting anything.) K-means clustering starts with K centroids, and K clusters formed by assigning points to the closest centroid. The algorithm then iteratively computes new centroids based on the preceding clusters, new clusters based on these new centroids, and so on. K-means clustering is a good choice for our data because we expect each type of glass to represent a predictor profile, i.e., we expects the natural clusters inherent in our data to each represent, at least loosely, a type of glass.

First, we find the ideal K (number of clusters).



Looking at our above graph, we choose K=7, since this represents a local min, and any increase of K past this yields only small reductions in within cluster variation. With our choice K, we can now again run the Lloyd clustering algorithm.



We need to take care working with this heat map as, unlike previously, the predicted cluster numbers has no correspondence to the actual type. Even if all points of glass type 1 are clustered together, that cluster might be labelled with any number. Because of this, we no longer give special consideration to the anti-diagonal. If K-Means clustered points by type perfectly, we would see a single bright square in each row and column. Looking at our graph, this is not the case. 5 types of glass are represented in clusters 5 and 4. Types 1, 2, and 3 each have a majority of their points in cluster 5. Only types 5, 6, and 7 are consistently classified into a cluster that no other types are consistently classified as (i.e., only columns 5, 6, and 7 have a bright rectangle in rows 2, 4, and 7, respectively).

Results

Looking at the misclassification errors KNN appears to be the method of choice over decision trees. The decision tree is significantly computationally easier, though, so if computation was truly limited or there was a very large amount of glass to be classified, a decision tree would be a feasible choice.

Method	Error Rate
KNN	31.67%
Tree	38.33%

Although accuracy in the 65% range over 6 categories is rather good, it is probably not good enough to be used in a forensic context where it would need to hold up in court.

For K-means we found that there exists 7 natural groupings within the dataset compared to the true 6 categories. This means that each natural profile does not correspond to a glass profile or the given categories are not specific enough to fit the natural profiles.

Discussion

With our analysis we have found we can accurately predict which of the six categories of glass a sample belongs with accuracy around 65%, with KNN being the best method explored with a 31.67% misclassification. We do note glasses of type 1 and 2 are often misclassified as each other, which makes some sense, as they are both building windows, albeit processed differently. We also found that the 6 categories given did not align with k-means clustering.

Going forward we would to improve accuracy of the results, as forensic work requires high accuracy. One possible approach would be to weight certain predictors more heavily than others, since some elements (or the refractive index) may be more significant than others in classification.

An extension of a more fundamental nature would be to predict refractive index from the elemental composition.