MATH 642 Project
Zijing Gao

Introduction:
Glass identification and classification is applied when searching for the evidence at the scene of the crime. The glass left can be used as the evidence if the glass type is correctly identified. Correct identification can be helpful to improve the efficiency for cracking a criminal in a lot of cases. If you are interested in this, please click here to check.

Problem Description:
The objective of this project is to identify and classify the type of the glass. There are several types of the glass which can be hard to identify by professional approaches since the glass in the criminal scene is fragment or covered by other materials. My goal is to use the amount of materials that consists of the fragments of glass, such as Fe, Al, Mg and so on to identify the type of glass. Since there are many kinds of glasses to take into account, this task can be challenging.

I assume the customer will be the criminological investigation since correct classification is greatly helpful in glass analysis. After correct identification of glass type, it can be helpful when determining direction and sequence of force.

This type of determination compares a known sample to a glass fragment to see if the two samples came from the same source. Glass can be made from a variety of different materials that differ from batch to batch. The presence of the different materials in the glass makes it easier to distinguish one sample from another. Also, the properties of glass can vary depending upon the temperature the glass is exposed to during manufacturing. Basic properties, such as color, thickness, and curvature, can also help to identify different samples of glass just by looking at them. Optical properties, such as refractive index (RI), are defined by various manufacturing methods. RI is the manner in which light passes through the glass. This can be measured easily even on small fragments of glass. These properties help to indicate that two samples of glass could be from the same source.

Data Description:
- Title : Glass Identification Database
- Sources:

Here is the data source

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

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

(c) Date: September, 1987

There is no missing/ambiguous data in this dataset.

- Challenges:

  There are 7 types of glasses for classification

  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

  And we have attribute information

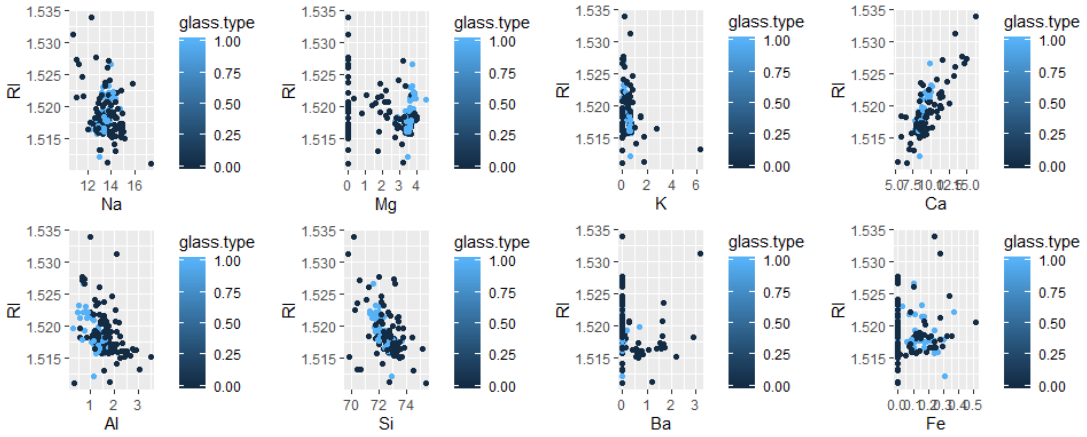
  Attribute Information:

  1. Id number: 1 to 214
  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

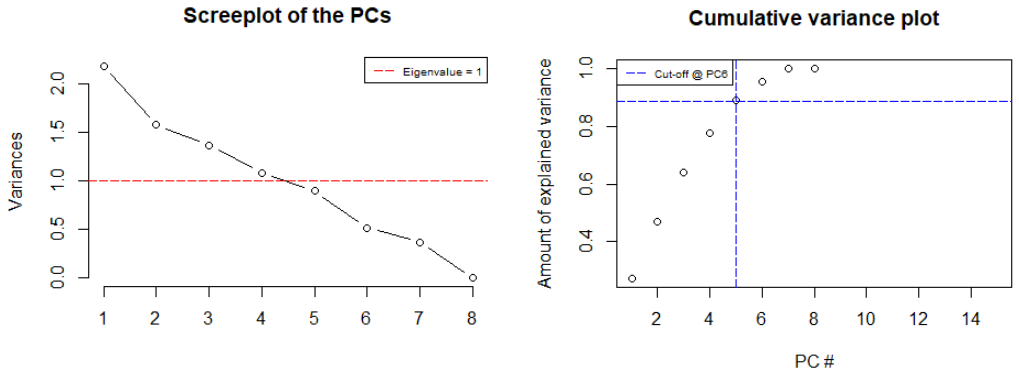
  It is hard to reduce the accuracy of the classification error since there are 7 types of glasses. Model selection and error testing are both challenging since I need to test the validity and accuracy of the model and select the best model by comparison.

Unsupervised Learning Results:

At first, I try to simplify the target to classify the float and non-float glass. So I classify the Type 1 and Type 3 as “float” and others as “non-float” so that I can apply the model to a binomial classification problem. Then, I plot the response (refractive index) with respect to each material in the glass because I think refractive index can be a good indicator of classification.



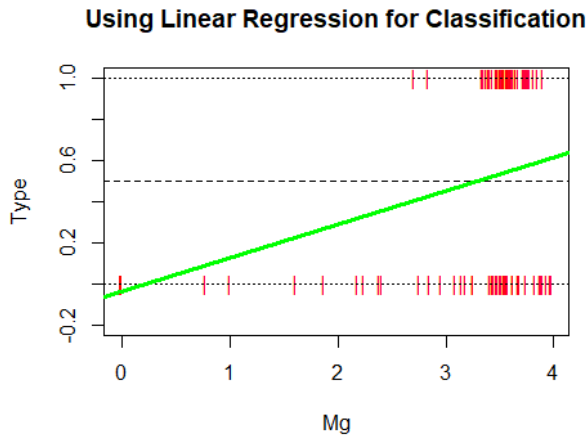
Then, I apply PCA on the attributes, hoping to reduce the feature columns to reduce the dimension. Here is the PVE figure.



I find that the first 4 pricinpal components out of 8 can explain 80% variance so we can reduce the dimension to 4.

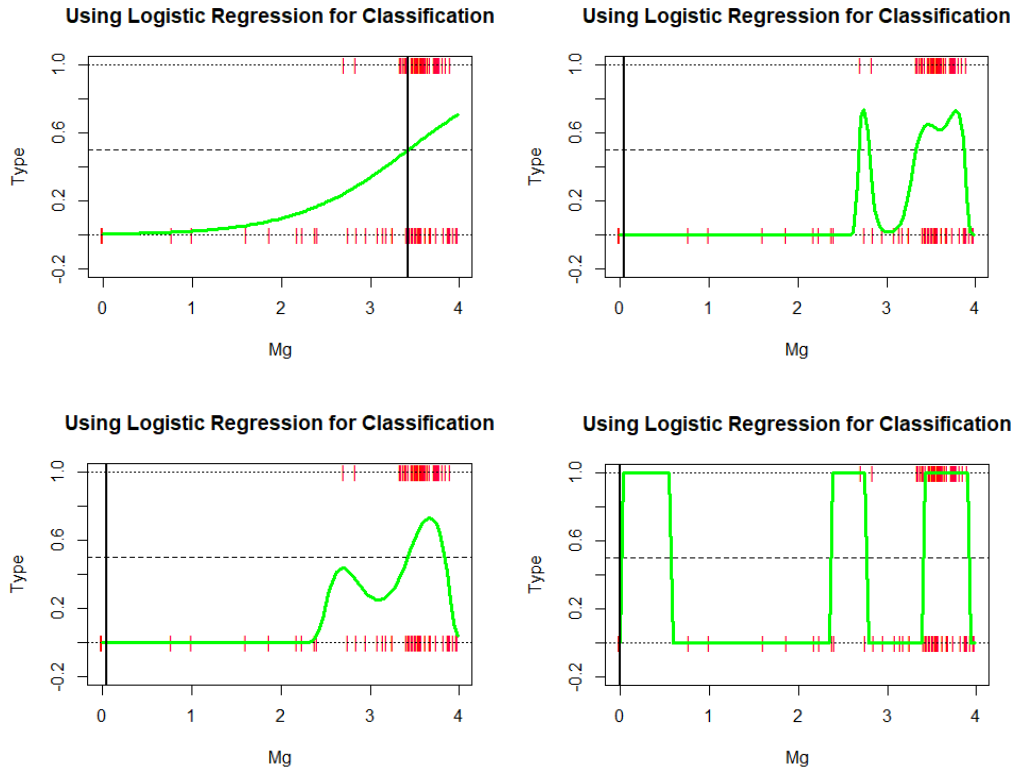
Then, I apply LDA on the data after PCA and I split the data after LDA into train set and test set equally. By using additive LDA model, I predict the classification by the new data which is the test set. After evaluating the performance of the model, I plot the AUC - ROC plot which is a good evaluation metrics for checking the model's performance. ROC curve is a probability curve and AUC represents degree of separability, telling me that the model accuracy is almost 82%, which stilled needs improving.

Secondly, I use Linear Regression for classification. Still, I split the set into train set and test set first. Then, I apply LR on the glass type with respect to each predictor and test the MSE for each case. Then, we find the amount of the metal Mg is closely related to the type of glass. Here is the plot for visualization.

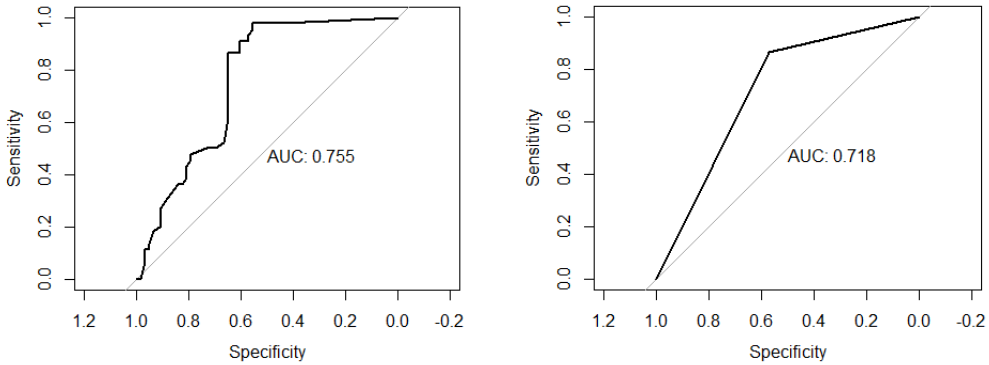


The model is not good since there are cases that the probabilities are below 0.

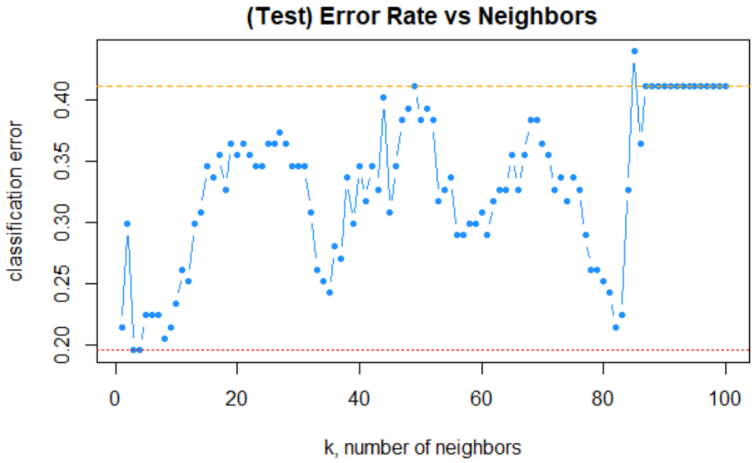
Then, I apply LOOCV and K-fold cross validation on the train set using glm function and test the error for each polynomial. And I find the polynomial to be 4 which gives me the lowest cross validation (cv) error = 0.1788766. Also, I find the polynomial to be 4 or 9 when cv error = is 0.1788766 which is the lowest. The I calculate the MSE for each model and derive the relative plot with different polynomial.



Here is the ROC-AUC plot,



At last, I apply K-Nearest Neighbors on the train set. After scaling and set k = 3 for a trial, I calculate the classification error which is 0.2336449, Then, I try to choose the best k and plot error vs choice of k.



Here, we choose k = 4 since the largest one is the least varibable and has the least chance of overfitting.

Analysis plan:

Here, I set 1 to represent float glass and 0 to represent non-float glasses. I try to use Linear Regression, Logistic Regreesion, PCA, LDA and KNN approaches to model the glass type distribution to identify glass type more accuately. By comparing the MSE, classification error and constructing ROC-AUC plot, parameter and model selection are done prelimanorily.

In the future, I will further investigate the dataset in these ways:

- Detect more features of the dataset by computing covariance matrix and mutual information.
- Exception handling if needed.

For analytical approaches, I will improve in these ways:

- Upgrade the binomial model into multilevel models by adding more classification of the glass type, not just float glass and non-float glass.

- Plan to apply Mixture models and Hierarchical clustering which are both unsupervised learning on the data.
- Improve the parameter and model selection techniques from the respect of accuracy, sensitivity and specificity for different classifiers.
- Perform the more explicit visualization for the data.

Preliminary Results

1. Basic investigation of the data

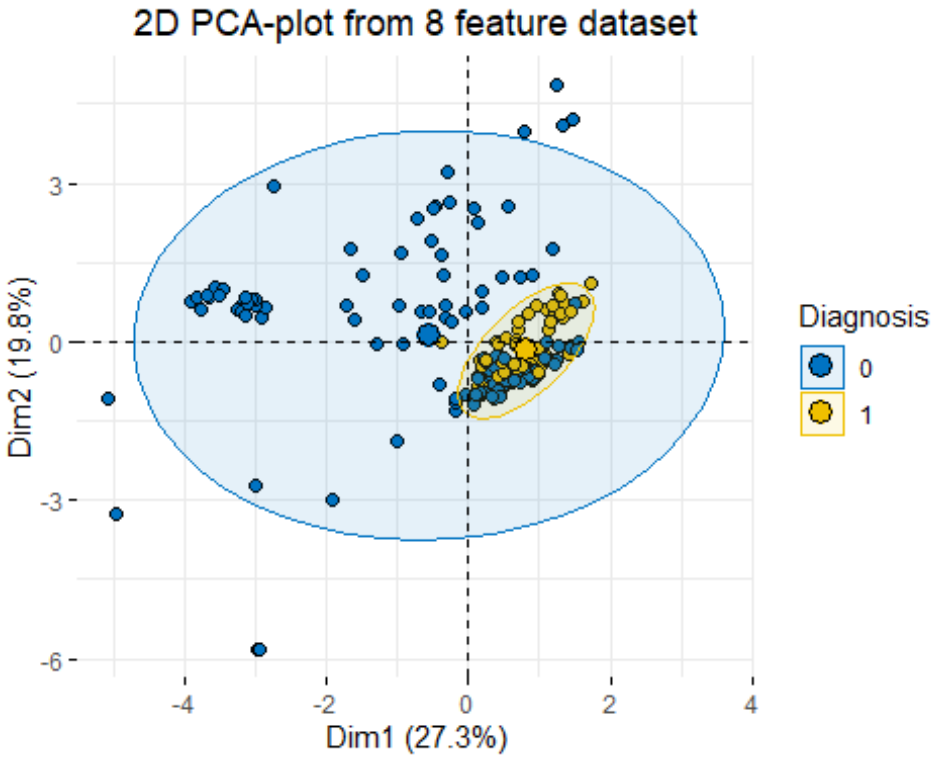
```
head(glass)

##        RI    Na   Mg   Al    Si    K   Ca Ba   Fe Type
## 1 1.52101 13.64 4.49 1.10 71.78 0.06 8.75  0 0.00    1
## 2 1.51761 13.89 3.60 1.36 72.73 0.48 7.83  0 0.00    1
## 3 1.51618 13.53 3.55 1.54 72.99 0.39 7.78  0 0.00    1
## 4 1.51766 13.21 3.69 1.29 72.61 0.57 8.22  0 0.00    1
## 5 1.51742 13.27 3.62 1.24 73.08 0.55 8.07  0 0.00    1
## 6 1.51596 12.79 3.61 1.62 72.97 0.64 8.07  0 0.26    1
```

2. PCA results

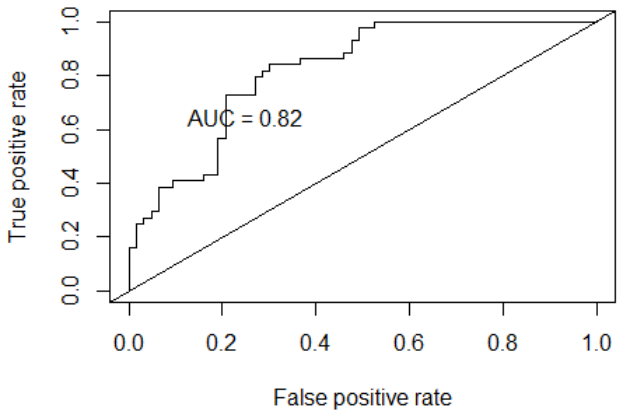
   Here, I plot the 2D PCA plot from 8 features after searching for 4 principal components which expains almost 80% of the variance. The target is the rotate the data in a new axis to have a better visulaization of the classification.

   From the plot, it is clear that the data of the float glass type is clustering within the yellow ellipse.

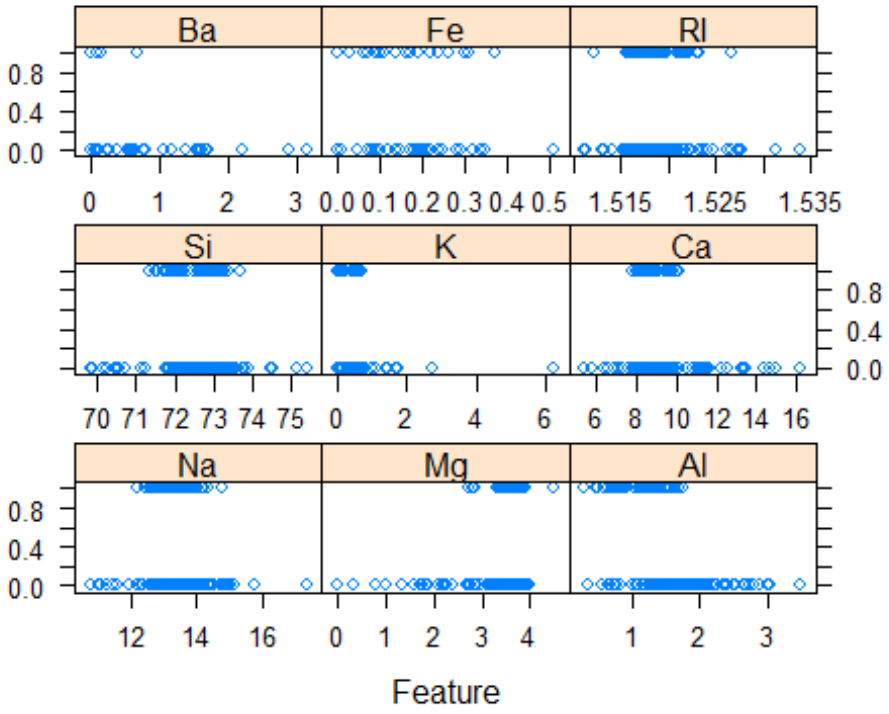


3. LDA results

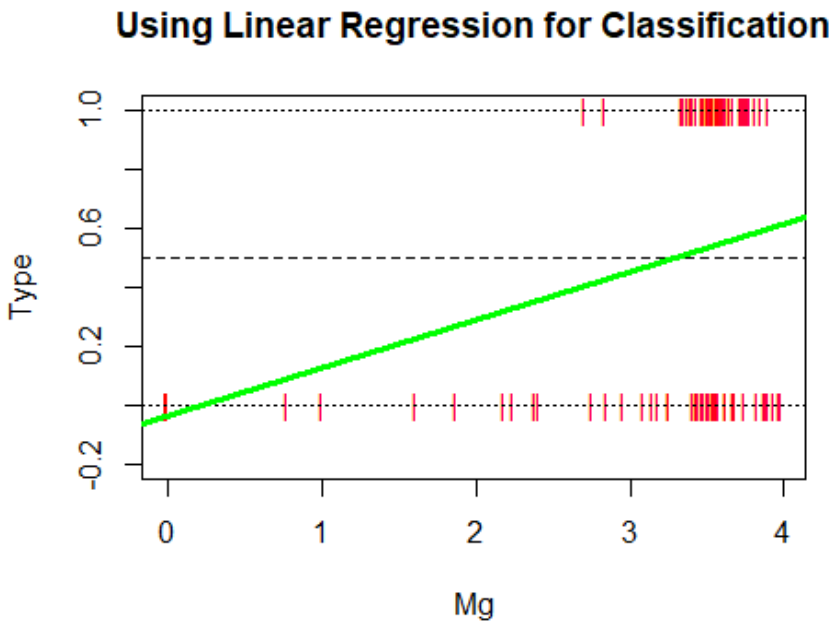
After reducing dimension to 4 by PCA, I apply LDA on the transformed glass data. Also, I use additive LDA model to predict by using test set and construct the ROC-AUC plot and received the AUC = 0.82.



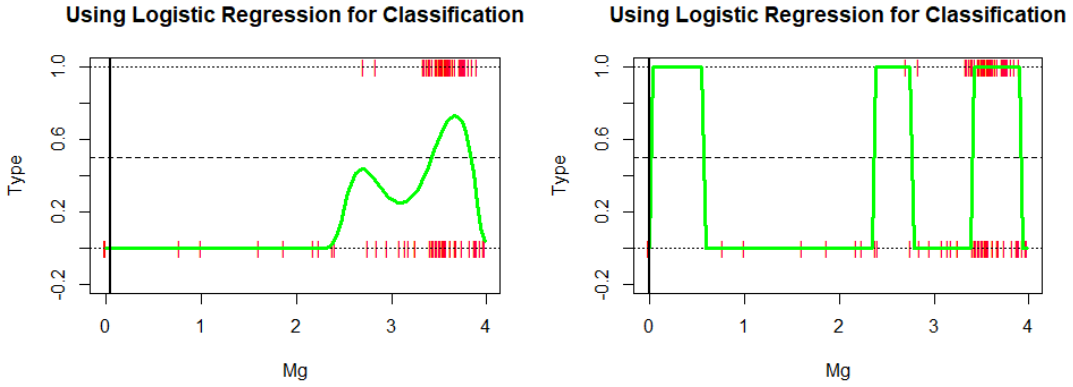
4. Linear Regression results
   Before I apply LR, I use plot the response with respect to each predictor.



Then, I use additive linear model to regress on the train set and the predictor Mg gives me the lowest MSE. But the result is not that good since there are probabilities that are below zero.

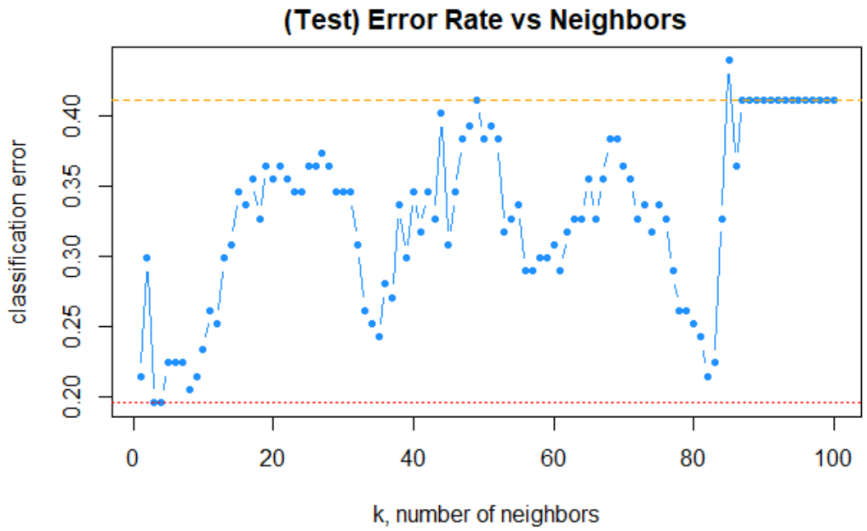


5. Logistic Regression results
   Here, I use LOOCV and K-fold cross validation. Each gives me different polynomial parameters for the model.



   For each case, the AUC is 0.755 and 0.718 which means the model needs improving.

6. K-Nearest Neighbor results
   After scaling the train set and test set, I apply knn on the train set with k = 3 for a trial. After a specific approach to choose k, I plot the test error rate vs k, the number of neighbors.



Here, we choose K = 4 since the largest one is the least variable and has the least chance of overfitting.