SOUTH DAKOTA STATE UNIVERSITY

STAT 602: Modern Applied Statistics
Group Project

Spring 2023

Classification and price prediction of dry beans using Different methods

Theophilus Anim Bediako
Bridget Bafowaa
Asif Mahmud Chowdhury
Emmanuel Boah

Problem Statement

In a recent paper, Koklu et al. explored the possibility of using morphometric measurements on seven commonly cultivated white beans to develop an automated (or at least machine assisted) method for separating the white beans when presented at market. Using certified dry bean seeds in Turkey is around 10% (Bolat et al., 2017). Dry bean cultivation in Turkey and Asian countries usually in the form of populations containing mixed species of seeds. Also, there is not much certified seed planting area (Varankaya and Ceyhan, 2012). Since different populations which contain different genotypes are cultivated, the final products contain different species of seeds. Thus, when the dry bean seeds obtained from population cultivation are released to the market without being separated by species, the market value decreases immensely (Varankaya and Ceyhan, 2012). -Murat Koklu, Ilker Ali Ozkan, Multiclass classification of dry beans using computer vision and machine learning techniques, Computers and Electronics in Agriculture, Volume 174, 2020
For this project you have a slightly different task- you are expected to develop an automated method that predicts the value of a harvest from a 'population cultivation' from a single farm that has been presented at market. For each of the different varieties of white beans, you will be given the price per pound and the average number of seeds per pound. You will be expected to take into account the differing costs of a mistake when you classify the seed varieties.

Data Exploration

To properly understand the data, we explored some possible plots and statistical summaries of the data. We identified some variables to have larger scales as compared to others. For instance, whereas the Eccentricity variable is in decimals, the Area variable was in the tenths of thousands. The choice of classification method will determine whether we have to transform the variables or not. Figure 1[we show only four variables because of plot size] indicates that some predictor variables are highly correlated within classes. For instance, the correlation coefficient between Area and MajorAxisLength within ”BOMBAY” class is 0.86. Several other variables are correlated within classes[refer to the full correlation plot in the RMD file]. Similarly, a boxplot of each predictor against the response class(CLASS variable), shows significant variability in the seed varieties. We show three of such plots [see Figures 2, 3, 4].

Classification Methods

Classification involves assigning observations to known classes based on some common features. In this project, we consider multiple classification methods for the purpose of grouping commonly cultivated white beans into six varieties using morphometric measurements. The classification results are applied to predict the value of a harvest presented at the market. We also examine the cost of misclassification when each classification method is used. The rest of the paper discusses the classification methods we considered:

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) assumes the class-specific samples are drawn from a multivariate normal(MVN) distribution and that the covariance matrices are equal across classes. A test of the multivariate normal assumption reveals that the observations in each class do not appear to have come from a MVN. This raises suspicion regarding the appropriateness of LDA as a classifier for this problem. To verify that the results from the LDA model are not due to random, we attempt 10-fold cross-validation. This gives an average accuracy of 85.7%. The consistency in the cross-validation results could be attributed to the idea that the departure from normality may not be severe [2]. We go ahead to fit an LDA model by splitting the data into 70% training and 30% testing. On the unseen data, we obtain a confusion matrix shown in Table 1. The prediction accuracy for the LDA model is 84.7%.

The total value of seeds presented at the market based on our LDA prediction is $6.51. For the varieties of seeds that are incorrectly predicted, the cost of misclassification is $1.56.

Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) also assumes that the observations within each class follow a multivariate normal distribution. QDA slightly differs from LDA by allowing the covariance matrix of each class to be unique. The absence of enough information to conclude the within-class normality assumption calls for an investigation into the appropriateness of using QDA for this problem. The 10-fold cross-validation on a QDA model results in an average accuracy of 90.7%. Similarly, the consistent accuracy in the cross-validation could be a rare case where the normality assumption is not satisfied yet QDA performs reasonably well [2]. Table 2 shows the confusion matrix when the QDA model is used on the unseen data. This gives a 90.1% prediction accuracy. The total value of seeds presented at the market based on our QDA prediction is \$6.48 and the cost of misclassification is \$1.45.Comparing both models, LDA has a higher error rate which explains its higher cost of misclassification.

K Nearest Neigbors

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm for classification. KNN classifies a given observation to the class with the highest estimated probability [1]. The performance of the KNN model is very sensitive to the choice of K. Hence, we fit the KNN with different values of k and choose the value of K that has the highest(lowest) accuracy rate(error rate). The KNN is a non-parametric method and hence it is distribution-free.

We suspect the KNN model using the original correlated predictors may be significantly impacted since the KNN function in **R** employs the Euclidean distance as a distance metric. To solve this issue, we attempt a dimension reduction technique and compare the KNN model built using the original dataset to the model built after applying the dimension reduction technique.

Principal component analysis (PCA) is a technique used to reduce the dimensionality of large data sets while retaining as much of the original information as possible. It achieves this by transforming the data into a new coordinate system, where the new axes are linear combinations of the original variables. PCA transforms data with correlated predictors to principal components that are uncorrelated.[2].

From Figure 5, The first principal component explains about 68% of variances in the data. The second, third, and fourth principal components explain about 19%, 10% and 2% of the variation in the data. Overall, the first four principal components explain about 99% of the variance. Here, we consider the first four components to build our KNN model. In subsequent sections, we consider KNN using the original datasets and KNN using the principal components

To find the optimum K for the KNN model, we try different K values on the original data and the principal components while maintaining the same testing and training ratio. The errors are extracted for each value of K and plotted. The plot helps to detect the value of K which minimizes the error in each case. Figure 6 and Figure 7 show K=23 and K=23 are needed for a KNN model in the case of using the original data and the principal components respectively.

Using K = 23 on the original data, the KNN model gives 72.0% prediction accuracy[see Table 3] whereas the KNN model on the principal components gives 87% prediction accuracy[see Table 4]. We focus on the PCA-KNN-produced model(has higher prediction accuracy) to calculate the value of our prediction and the cost of misclassification. In that case, the total value of seeds presented at the market is \$6.52 and the cost of misclassification is \$0.47.

Based on these results, we conclude by choosing the QDA model as the best model (90.1% prediction accuracy). Choosing the best model is solely determined by prediction accuracy. The high total value and and low cost of misclassification associated with a particular model do not provide enough information for determining the best model, particularly when the market price per pound is different across the seed varieties.

References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[2] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, NJ, 2002.

[3] Murat Koklu and Ilker Ali Ozkan. “Multiclass classification of dry beans using computer vision and machine learning techniques.” In: *Computers and Electronics in Agriculture* 174 (2020), p. 105507.

[4] *Lecture Notes and Slides*.

[5] KARA Mazhar, Bahadır Sayinci, Erdal Elkoca, İsmail ÖZTÜRK, and Talha ÖZMEN. “Seed size and shape analysis of registered common bean (Phaseolus vulgaris L.) cultivars in Turkey using digital photography.” In: *Journal of Agricultural Sciences* 19.3 (2013), pp. 219–234.

Internet Resources

1. https://stackoverflow.com/questions/68638725/how-to-address-overplotting-in-ggallyggpairs
2. https://www.rdocumentation.org/packages/MASS/versions/7.3-58.3/topics/lda
3. https://www.rdocumentation.org/packages/MASS/versions/7.3-58.3/topics/qda.html
4. https://www.rdocumentation.org/packages/class/versions/7.3-21/topics/knn
5. https://www.rdocumentation.org/packages/FactoMineR/versions/2.4/topics/PCA
6. https://www.rdocumentation.org/packages/GGally/versions/1.5.0/topics/ggpairs

Appendix

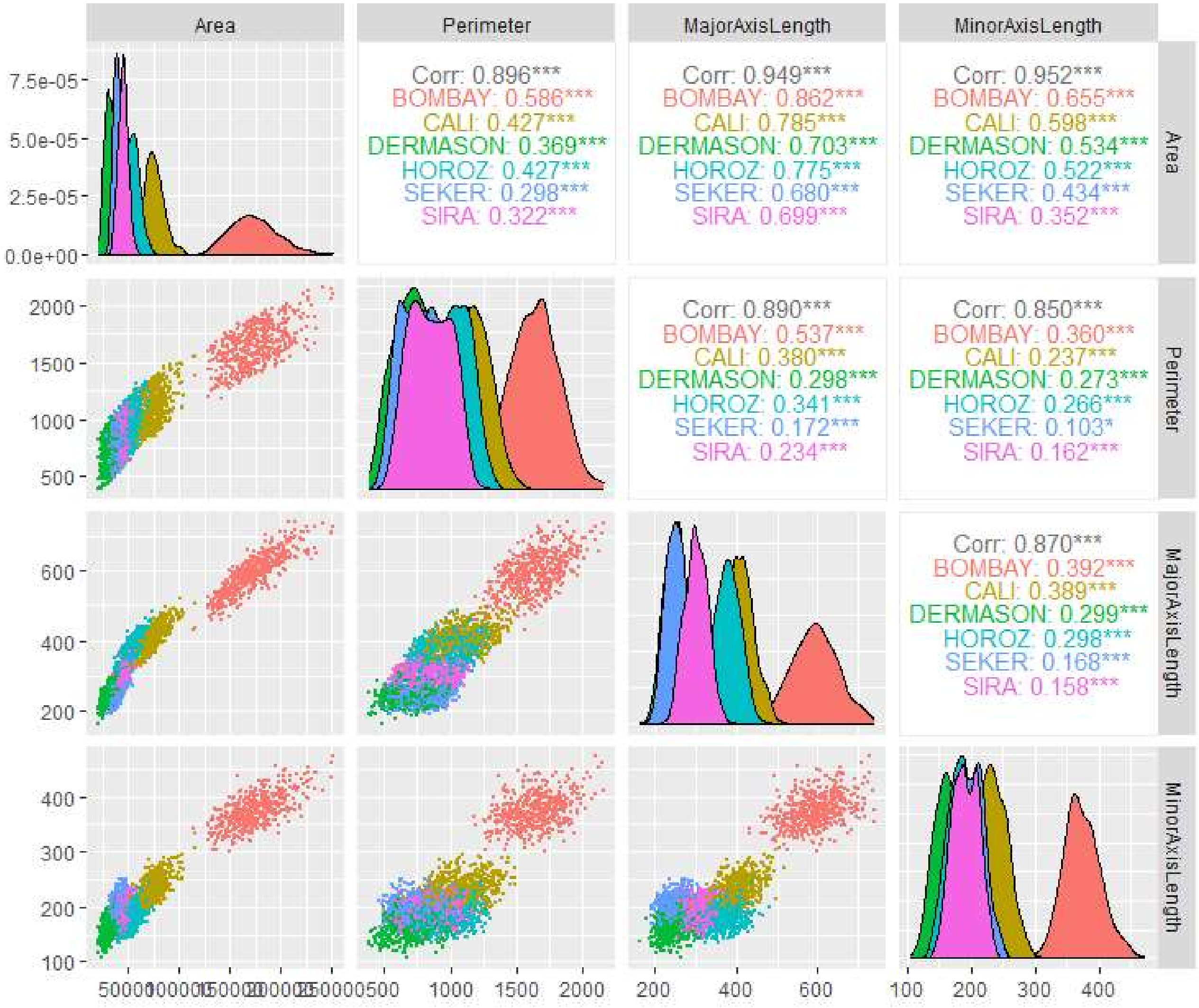

Figure 1: A scatterplot showing how variables are correlated within each Class

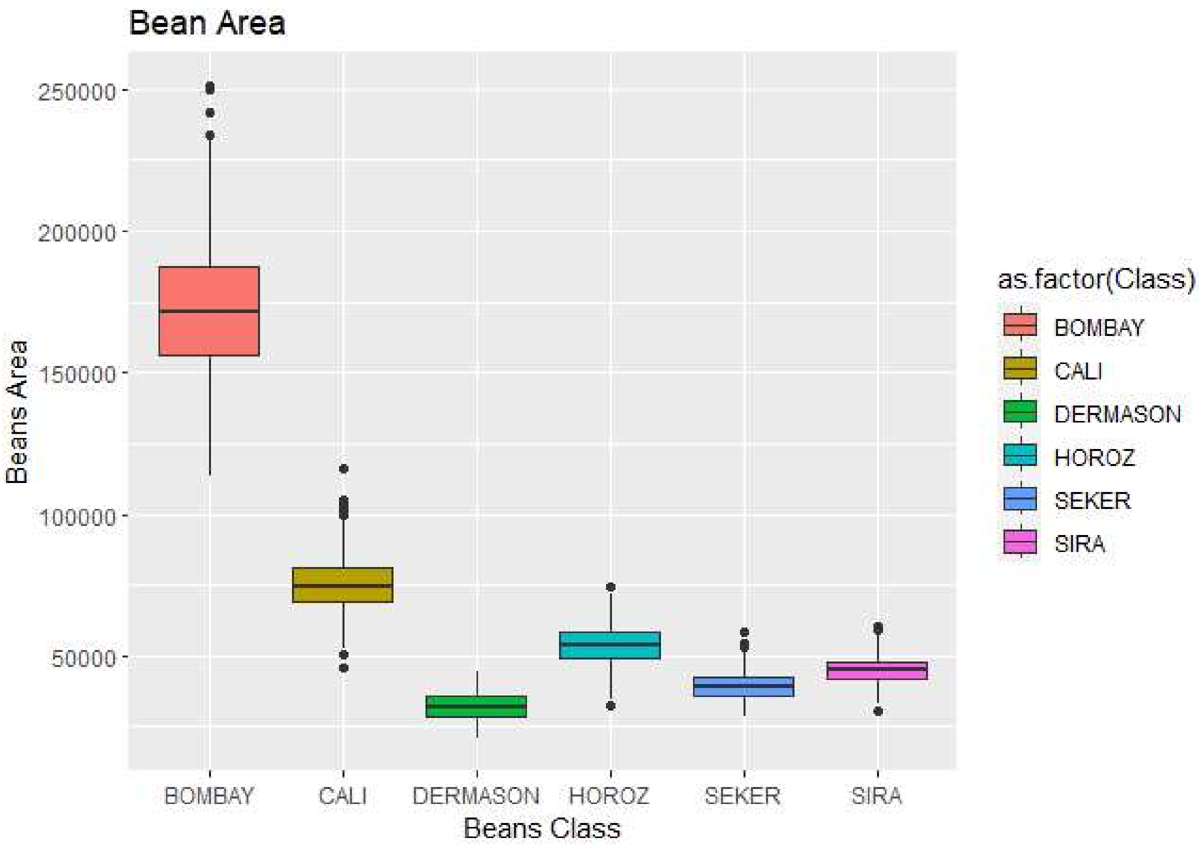

Figure 2: Boxplot of Bean Area by Class

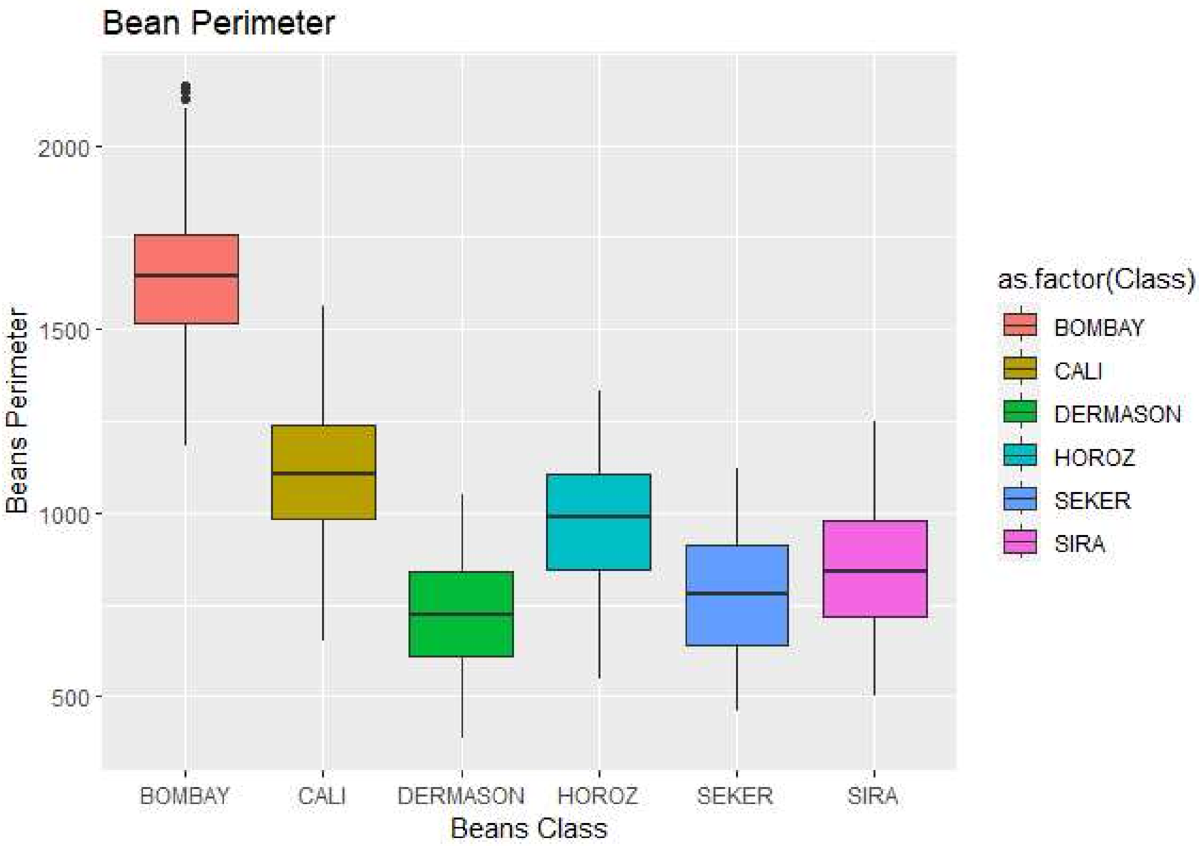

Figure 3: Boxplot of Bean Perimeter by Class

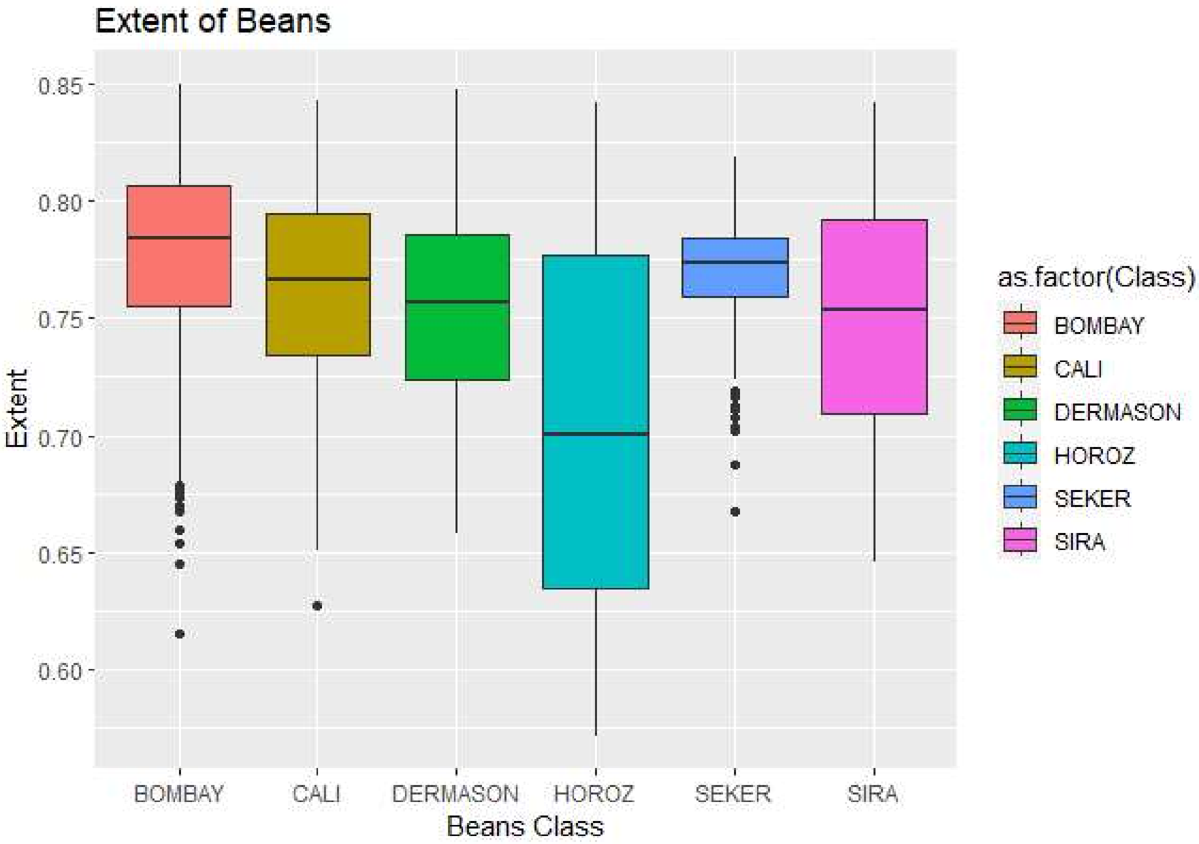

Figure 4: Boxplot of Extent by Class

Table 1: Confusion matrix of the LDA model

			True class			
Predicted class	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	150	0	0	0	0	0
CALI	0	134	0	14	0	1
DERMASON	0	0	115	1	7	25
HOROZ	0	5	0	123	0	10
SEKER	0	1	8	0	128	2
SIRA	0	10	27	12	15	112

Table 2: Confusion matrix of the QDA model

			True class			
Predicted class	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	150	1	0	0	0	0
CALI	0	142	0	5	0	0
DERMASON	0	0	129	2	8	14
HOROZ	0	5	0	135	0	6
SEKER	0	2	5	0	132	7
SIRA	0	0	16	8	10	123

Table 3: Confusion matrix of the KNN model on the original data

			True class			
Predicted class	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	150	0	0	0	0	0
CALI	1	130	0	18	0	1
DERMASON	0	0	113	0	35	2
HOROZ	0	15	1	93	7	34
SEKER	0	0	33	3	66	48
SIRA	0	0	4	14	33	99

Table 4: Confusion matrix of the KNN model after PCA

			True class			
Predicted class	BOMBAY	CALI	DERMASON	HOROZ	SEKER	SIRA
BOMBAY	150	0	0	0	0	0
CALI	0	139	0	13	0	2
DERMASON	0	0	120	1	6	21
HOROZ	0	3	0	122	0	5
SEKER	0	2	10	0	132	6
SIRA	0	6	20	14	12	116

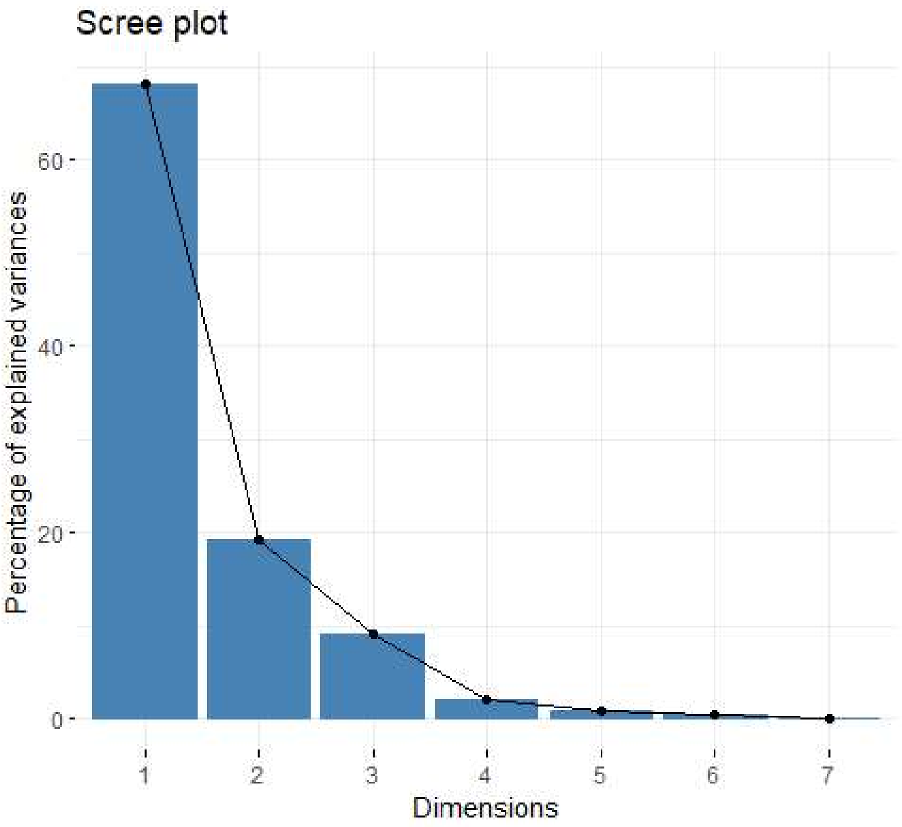

Figure 5: Scree plot showing the percentage of variability explained by components

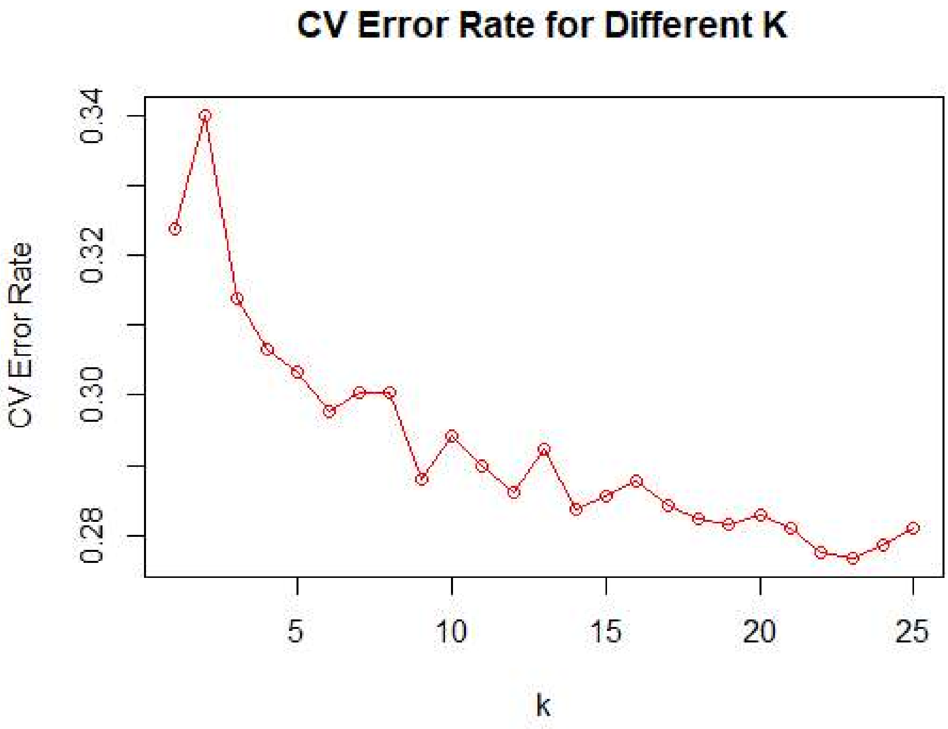

Figure 6: Optimal K using the original Data

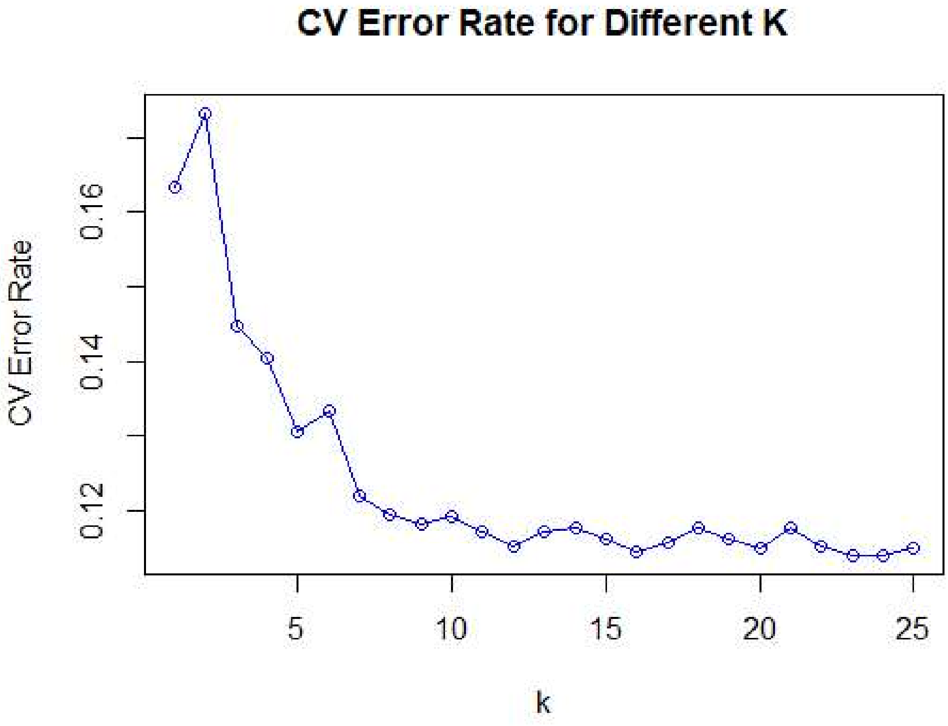

Figure 7: Optimal K using the Principal Components