Robert Shepherd
2431907S

Data Mining and Machine Learning I – Olive Oil Classification

Executive Summary

This report compares a variety of modelling methods to find the most predictive model to classify olive oil samples.

The dataset has a high number of predictors but a small number of observations making it difficult to build a robust model. The data was partitioned further to create training, validation and test datasets.

The exploratory analysis focused on understanding the training dataset. This identified that the data had a varying scale and a high amount of covariance between predictors. Principle components were created to reduce the number of dimensions while accounting for as much variance as possible.

Potential outliers were observed but were kept in the dataset as their performance was in line with other observations of the same class and removing them would have meant a significant loss of information.

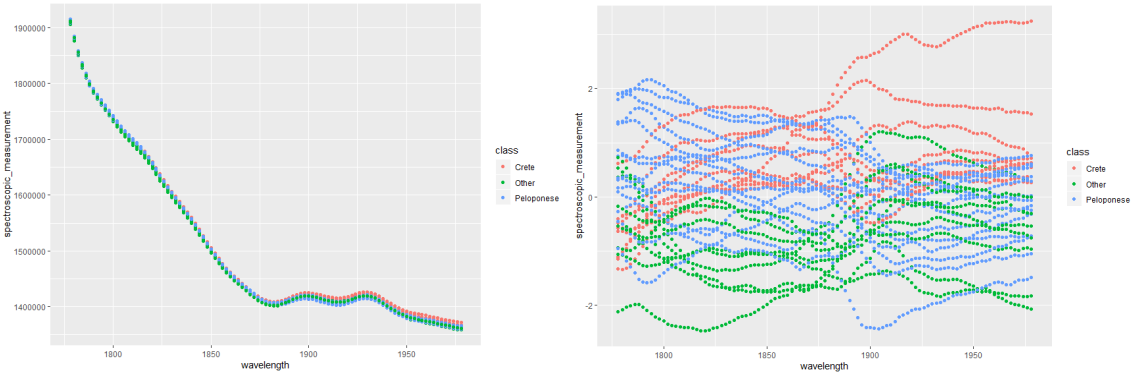
K-Nearest Neighbour (KNN), Decision trees and Support Vector Machine models were built using both the principle components and the original variables. Models were tuned and pruned using the validation dataset before being assessed against their classification accuracy over the test dataset.

The final model selected was a random forest decision tree model using the original variables which saw a correct classification rate of 94% over the test dataset.

Exploratory Analysis

The data has been partitioned in to three datasets; training (50%), validation (25%) and test (25%) Exploratory analysis has been created on the training dataset only. The intention is to inform how the data should be treated as well as to detect any outliers or compounding factors that would bias the model build.

Visualising the data



Visualising the data shows a big drop in spectroscopic measurement values as the wavelength increases before It levels off around 1,875. The varying scale makes it difficult to parse how classes vary across different wavelengths, however we can start to see a divergence in classes from when the data levels off.

Standardising the scale across different wavelengths shows some patterns that we can look out for in our model build process as below:

- 'Crete' oils tend to have **higher** spectroscopic measurements at higher wavelengths (> 1,875)
- 'Other' oils tend to have **lower** spectroscopic measurements at lower wavelengths (< 1,875)

Detecting outliers

Running a boxplot on the scaled data highlights a couple of potential outliers (appendix – figure 1). However, the visualisations show that these are generally consistent with other points of the same class. Given that the number of datapoints is quite low (32 in the training set), these have been kept in for the model as removing them would lose a lot of information.

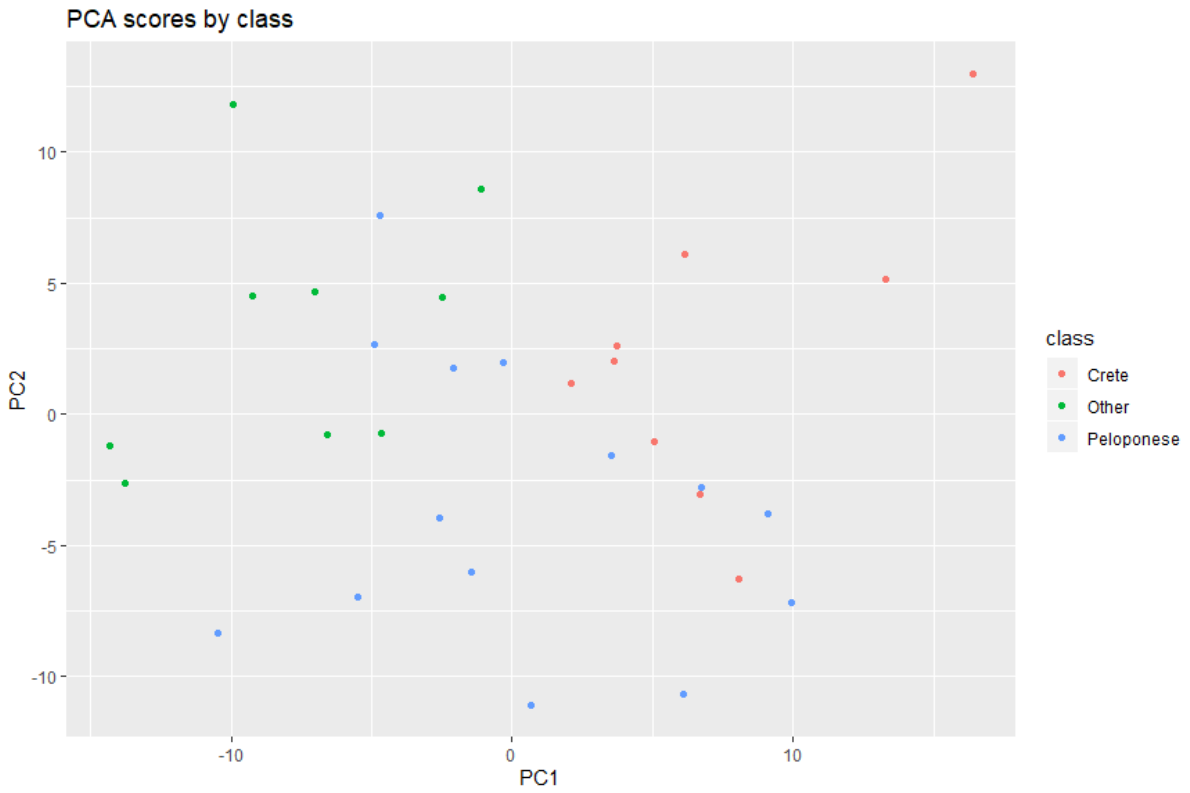
Building Principle Components

Plotting correlation shows a large volume of correlation between variables (appendix – figure 2), particularly when their wavelength is close. Principle components analysis (PCA) will be used to remove much of the correlation between predictors to avoid model bias.

PCA was applied using singular value decomposition due to the low number of observations. The observations were scaled and centred to avoid any bias caused by the varying scaled observed when visualising the data.

Plotting the total variance by the components shows a big drop in variance after the first two components (appendix – figure 3). Together these explained 94% of total variance. As such, only the first two variables will be used when building models on the components.

The two selected components had a high number of dimensions with a strong contribution (appendix – figure 4).



Plotting classes by principle component scores shows groups starting to develop:

- Crete oils tend to have positive score for PC1 and mixed scores for PC2
- Peloponese oils tend to have mixed scores for PC1 and negative scores for PC2
- Other oils tend to have negative scores for PC1 and positive scores for PC2

Model building

Three model types will be built on the training dataset and optimised based on their performance on the test dataset. These are **KNN**, **Support Vector Machines** and **Decision Trees**.

Within each model type, variations will be built using the **original variables** and the **principle components** to see which return the most predictive models. Variants will be used within each model type where applicable to find the best performing model.

A final model will be selected from each class and a final 'winner' will be chosen based on their ability to correctly classify the validation dataset.

KNN

Models were built for a series of values for K and selected based on which had the highest overall classification rate.

Original predictors

The KNN model build with the original predictors was found to perform best on the test dataset where **K was equal to 3** (appendix – figure 5).

Total correct classification rate	**82%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	2	0	2
	Other	0	6	0
	Peloponese	1	0	6
Correct classification rate by class		67%	100%	75%

The selected model accurately predicted all of the 'Other' oils but misclassified one 'Crete' oil as 'Peloponese' and two 'Peleoponese' oils as 'Crete.

PCA predictors

The KNN model build with the PCA predictors was found to perform best on the test dataset where **K was equal to 3** (appendix – figure 6).

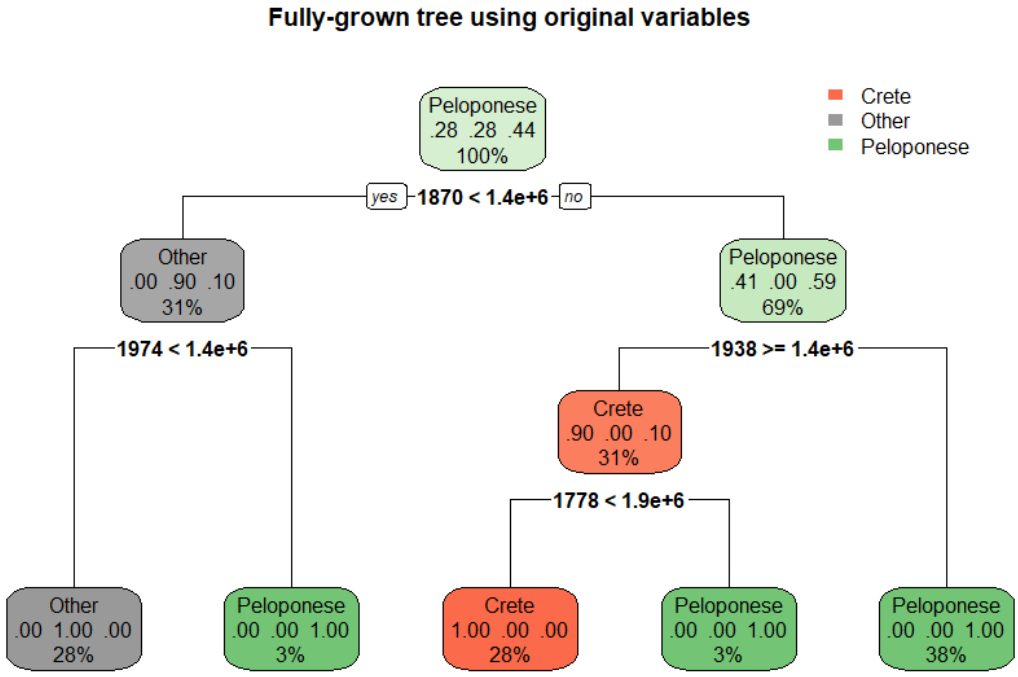
Total correct classification rate	**71%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	2	0	3
	Other	0	6	1
	Peloponese	1	0	4
Correct classification rate by class		67%	100%	50%

Similar to the original model, the PCA model correctly classified all 'Other' oils but had a higher misclassification rate for 'Peloponese' oils.

Decision trees

Decision trees were created by building a fully-grown tree and then pruning the tree based on the complexity and error values. Two models were tested for the original variables (2 split and 4 split) and the PCA variables (2 split and 3 split). A random forest model was also built for both the original and PCA variables.

Original predictors



Pruning values

CP	nsplit	rel error	xerror	xstd	selected?
0.444444	0	1	1	0.1559	*Not selected*
0.055556	2	0.11111	0.44444	0.13608	*Two split model*
-1	4	0	0.44444	0.13608	*Four split model*

Two split model

Total correct classification rate	**82%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	3	0	3
	Other	0	6	0
	Peloponese	0	0	5
Correct classification rate by class		100%	100%	63%

The model accurately predicted all of the 'Crete' and 'Other' oils but misclassified some 'Peloponese' oils as 'Crete'.

Four split model

Total correct classification rate	**94%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	3	0	1
	Other	0	6	0
	Peloponese	0	0	7
Correct classification rate by class		100%	100%	88%

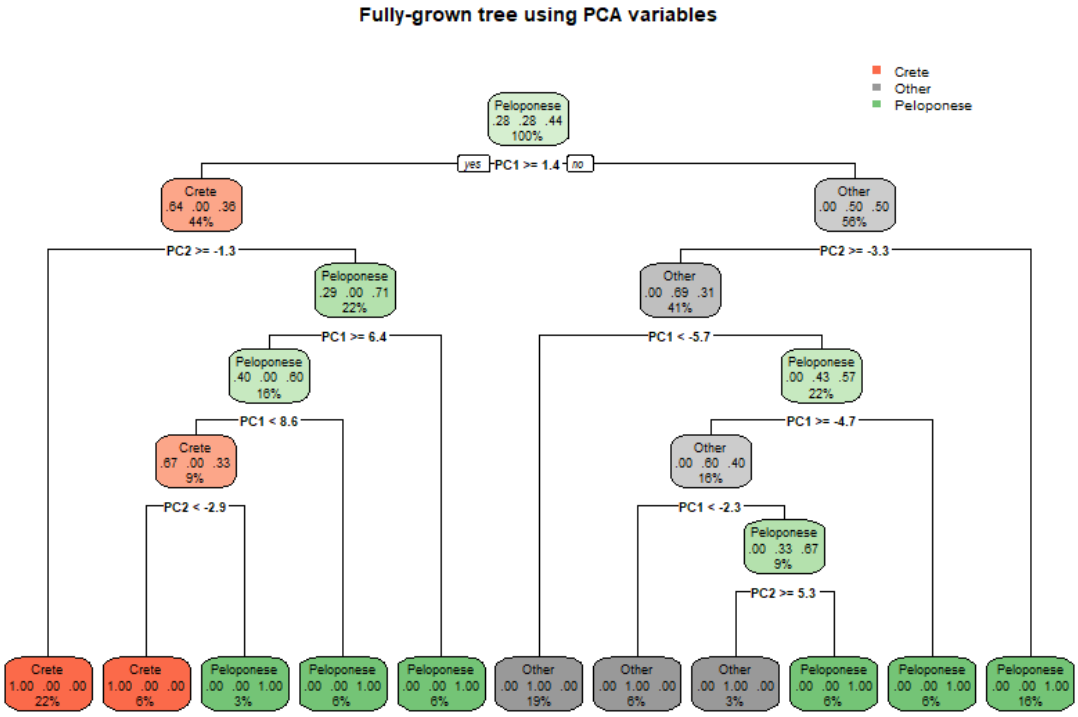
The model performed very well, only misclassifying one 'Peloponese' oil as 'Crete'.

Random forest

Total correct classification rate	**94%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	3	0	1
	Other	0	6	0
	Peloponese	0	0	7
Correct classification rate by class		100%	100%	88%

The random forest model performed the same as the 'Four split model'.

PCA predictors



Pruning values

CP	nsplit	rel error	xerror	xstd	selected?
0.25	0	1	1	0.1559	*Not selected*
0.166667	2	0.5	1.33333	0.13608	*Two split model*
0.055556	3	0.33333	0.83333	0.15683	*Three split model*
0.037037	7	0.11111	0.88889	0.15713	*Not selected*
-1	10	0	0.88889	0.15713	*Not selected*

Two split model

Total correct classification rate	**65%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	2	0	3
	Other	0	6	2
	Peloponese	1	0	3
Correct classification rate by class		67%	100%	38%

The two split model struggled to identify ‘Peloponese’ oils, often misclassifying them as ‘Crete’ or ‘Other’.

Three split model

Total correct classification rate	**71%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	2	0	3
	Other	0	6	1
	Peloponese	1	0	4
Correct classification rate by class		67%	100%	50%

The three split model performed slightly better but still struggled to identify 'Peloponese' oils.

Random forest

Total correct classification rate	**71%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	2	0	3
	Other	0	6	1
	Peloponese	1	0	4
Correct classification rate by class		67%	100%	50%

The random forest model performed the same as the 'Three split model'.

SVM

Both linear and radial models were built for each of the sets of variables (original and PCA). Each model was tested over a series of cost parameters and the models were chosen based on the classification error on the test data.

Original predictors

Linear model (cost parameter 0.1)

Total correct classification rate	**76%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	3	0	3
	Other	0	5	0
	Peloponese	0	1	5
Correct classification rate by class		100%	83%	63%

The linear model correctly classified all 'Crete' oils and most 'Other' oils but misclassified some 'Peloponese' oils as 'Crete'.

Radial model (cost parameter 2)

Total correct classification rate	**88%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	2	0	1
	Other	0	6	0
	Peloponese	1	0	7
Correct classification rate by class		67%	100%	88%

The radial model performed better than the linear model, correctly classifying all 'Other' oils and most 'Peloponese' oils but, unlike the linear model, incorrectly classified one 'Crete' oil.

PCA predictors

Linear model (cost parameter 5)

Total correct classification rate	**71%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	3	0	4
	Other	0	5	0
	Peloponese	0	1	4
Correct classification rate by class		100%	83%	50%

The PCA linear model performed similarly to the original variable equivalent but had a higher misclassification rate for 'Peloponese' oils.

Radial model (cost parameter 0.5)

Total correct classification rate	**82%**	Observed		
		Crete	Other	Peloponese
Predicted	Crete	2	0	2
	Other	0	6	0
	Peloponese	1	0	6
Correct classification rate by class		67%	100%	75%

Again, the PCA radial model performed similarly to the original variable equivalent but had a higher misclassification rate for 'Peloponese' oils.

Model selection

Model type	Variables	Model description	Correct classification rate	Rank
Decision tree	Original	Random forest	94%	1
Decision tree	Original	Split = 4	94%	1
SVM	Original	Radial	88%	3
Decision tree	Original	Split = 2	82%	4
KNN	Original	K = 3	82%	4
SVM	PCA	Radial	82%	4
SVM	Original	Linear	76%	7
Decision tree	PCA	Random forest	71%	8
Decision tree	PCA	Split = 3	71%	8
KNN	PCA	K = 3	71%	8
SVM	PCA	Linear	71%	8
Decision tree	PCA	Split = 2	65%	12

All of the models using the 'original' variables outperformed their PCA equivalents. The strongest models were the **decision tree with 4 splits** and the **random forest**. The **radial SVM** also performed quite well with only a slightly higher misclassification rate.

As only one model is required, the **random forest model** has been selected as the final classification model as the technique is generally more robust than classical decision trees.

Appendix

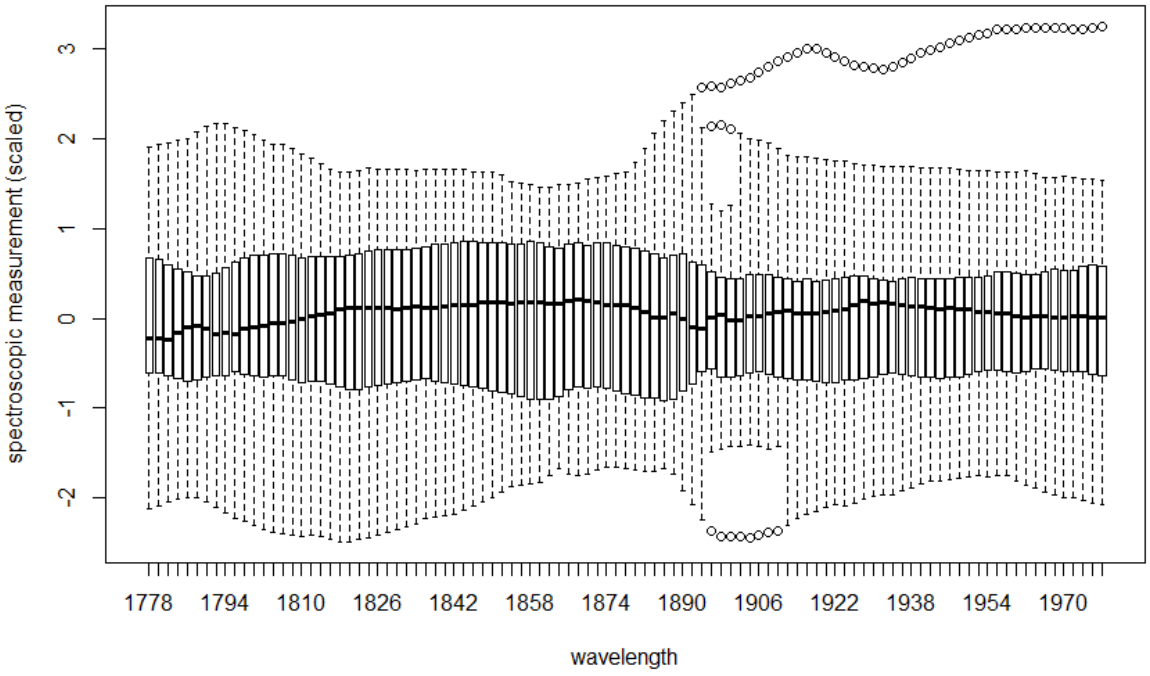


Figure 1- Boxplot used to identify outliers

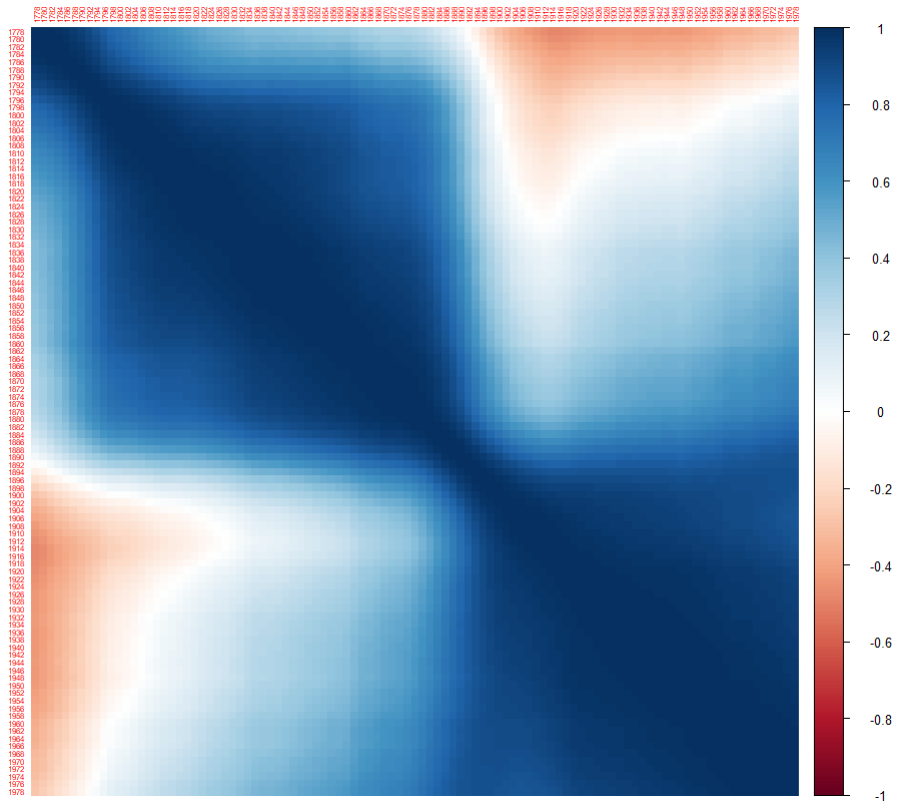


Figure 2- Correlation plot

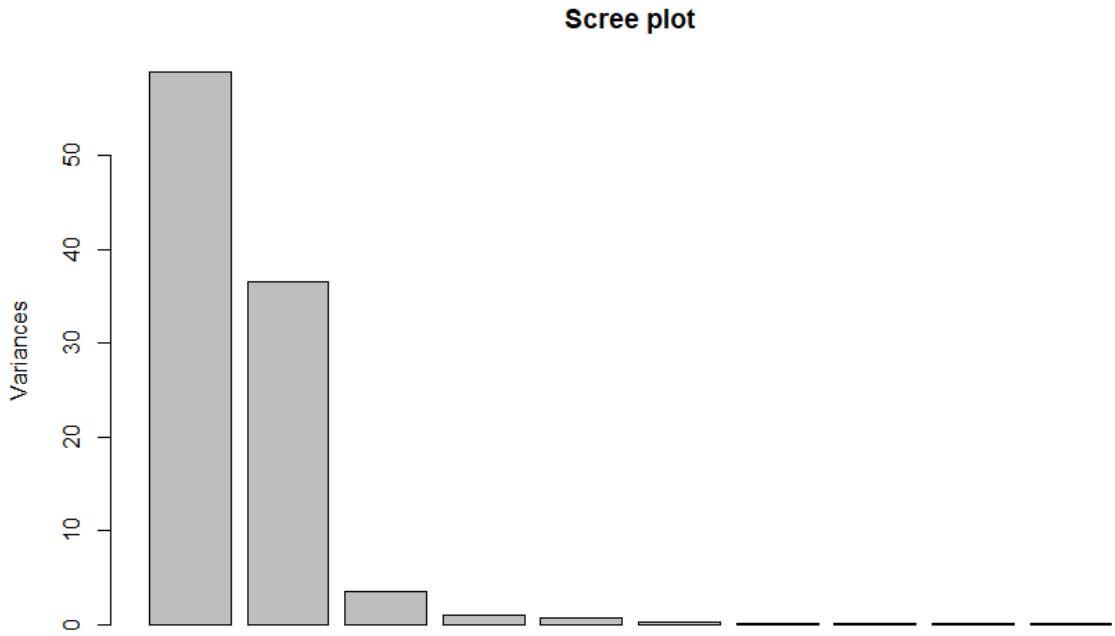


Figure 3 - Scree plot showing total variance by component

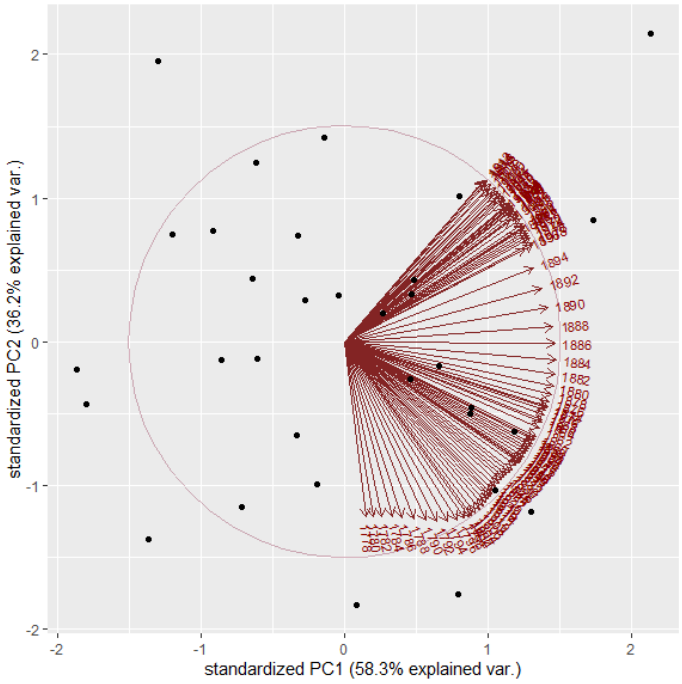


Figure 4 - Biplot showing variable contribution to selected principle components

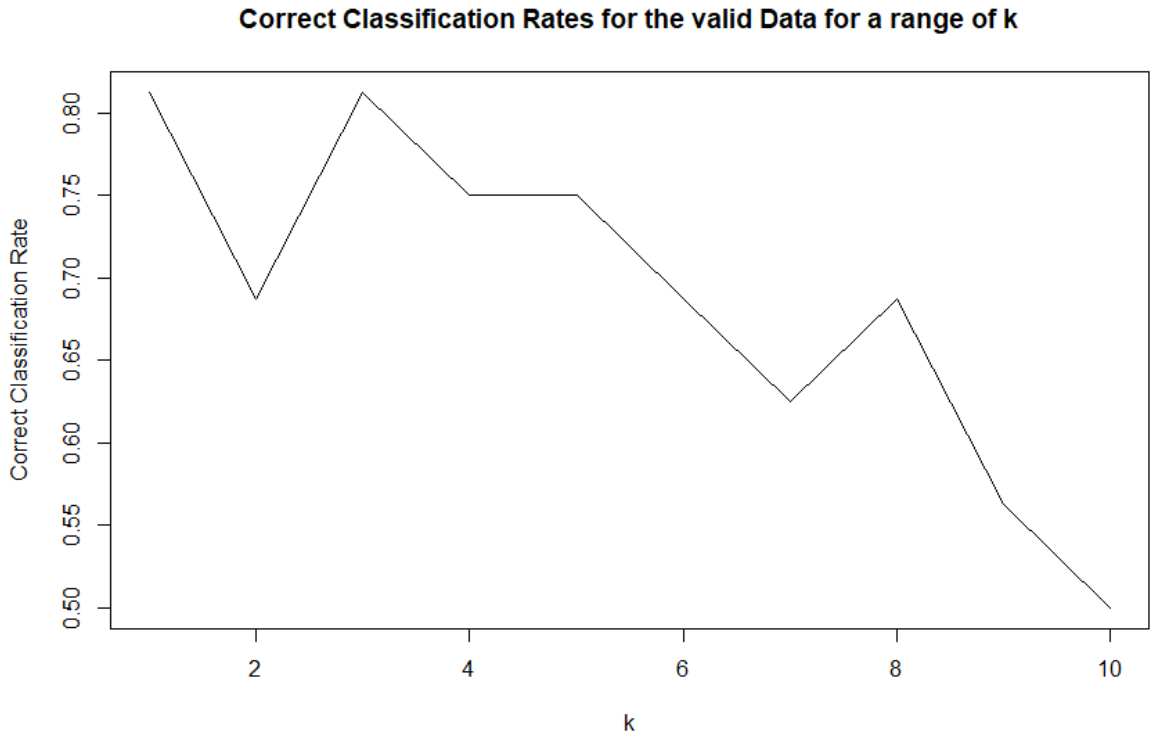


Figure 5 - Classification rate for different values of K using the original variables

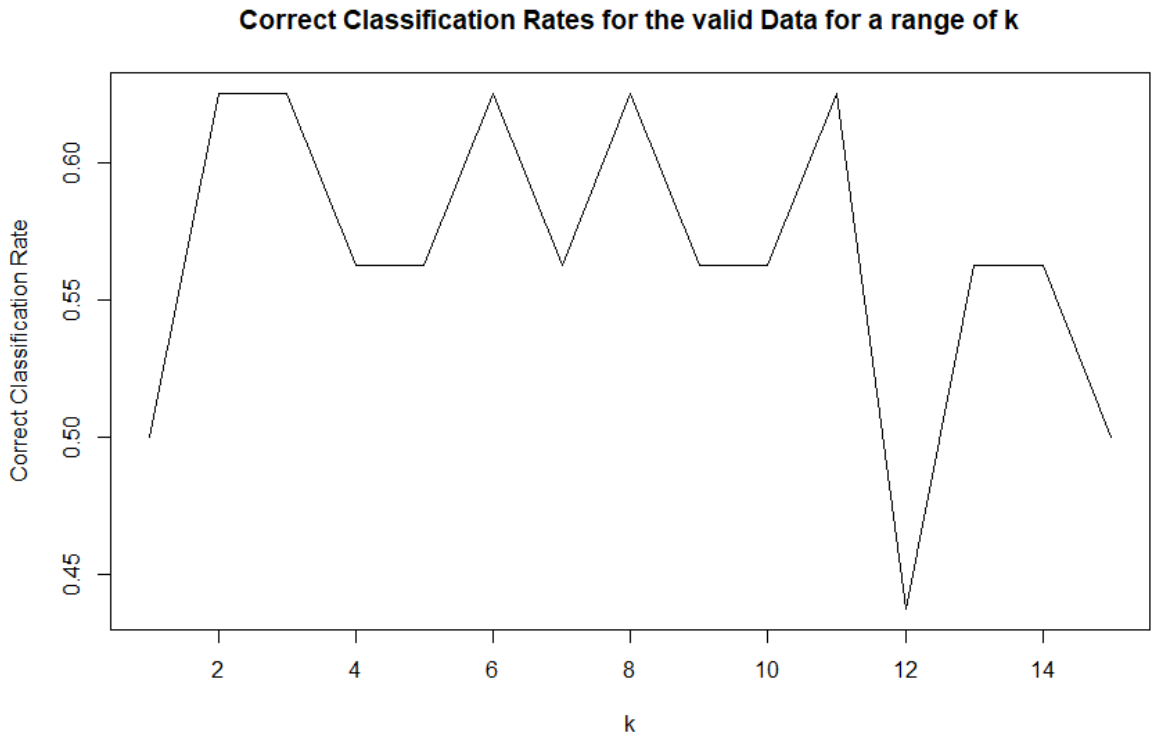


Figure 6 - Classification rate for different values of K using the PCA variables