Classification Tools for Chemical Analysis Data

Siva Karteek Bolisetti - siva2.bolisetti@live.uwe.ac.uk

09/05/2023

Table of contents

[1 Introduction 2](#_Toc134892208)

[2 Exploratory Data Analysis 2](#_Toc134892209)

[2.1 Data Summary 2](#_Toc134892210)

[2.2 Missing Data 3](#_Toc134892211)

[2.3 Outliers 5](#_Toc134892212)

[2.4 Data Summarisation 8](#_Toc134892213)

[2.5 Splitting training and test sets 10](#_Toc134892214)

[2.6 Data Normalisation 11](#_Toc134892215)

[2.7 Correlation Analysis 11](#_Toc134892216)

[3 Dimensionality Reduction and Feature Selection 12](#_Toc134892217)

[3.1 Principal Component Analysis (PCA) 12](#_Toc134892218)

[3.1.1 Parallel Analysis 12](#_Toc134892219)

[3.1.2 PCA Observations 13](#_Toc134892220)

[3.2 Decision Tree and Random Forest for Feature Selection 16](#_Toc134892221)

[4 Classification 18](#_Toc134892222)

[4.1 Random Forest 21](#_Toc134892223)

[4.2 MclustDA Discriminant Analysis 21](#_Toc134892224)

[5 Recommendation and Validation 22](#_Toc134892225)

[6 Conclusion 23](#_Toc134892226)

[References 24](#_Toc134892227)

## 1 Introduction

In this report, R programming language is used to explore the chemical analysis dataset and recommend suitable classification tools. Exploratory Data Analysis (EDA) is performed to understand the underlying statistical distributions of the data. Various dimensionality reduction techniques and classification tools are investigated. Finally, classification performances and recommendations are discussed.

## 2 Exploratory Data Analysis

In this section, EDA is performed on the chemical analysis dataset. Some of the objectives of EDA include:

* Understand the structure of the data.
* Analyse various summary statistics for the data.
* Identify missing entries and other anomalies such as outliers.
* Perform data visualisation to visually explore the data.
* Data normalisation to adjust scales.

### 2.1 Data Summary

The features in the dataset include X1, X2, X3, X4, etc. Features and corresponding data types are summarised in Table 1.

| **Variable** | **Data Type** | **Missing Values (%)** |
| --- | --- | --- |
| **X1** | double | 0.16 |
| **X2** | double | 0.08 |
| **X3** | double | 0.16 |
| **X4** | double | 0.04 |
| **X5** | double | 0.12 |
| **X6** | double | 0.08 |
| **X7** | double | 0.16 |
| **X8** | double | 0.00 |
| **X9** | double | 0.08 |
| **X10** | double | 0.12 |
| **X11** | double | 0.16 |
| **X12** | double | 0.08 |
| **X13** | double | 0.12 |
| **X14** | double | 0.12 |
| **X15** | double | 0.24 |
| **X16** | double | 0.00 |
| **X17** | double | 0.00 |
| **X18** | double | 0.08 |
| **X19** | double | 0.16 |
| **X20** | double | 0.04 |
| **label** | character | 0.00 |

Table 1**:** Chemical Analysis Data Structure

The dataset consists of 21 features and 2500 samples. The unique categorical values in the label feature and their distribution are summarised in *Table 2***.**

| **A** | **B** | **C** | **D** | **E** |
| --- | --- | --- | --- | --- |
| **498** | **487** | **499** | **525** | **491** |

Table 2**:** Distribution of Labels

### 2.2 Missing Data

Presence of missing data may interfere with statistical analysis procedures which usually require complete cases. The missing data can either be systematic or completely random. If it is systematic or “not random”, the data needs to be evaluated to identify the reasons.

The dataset consists of 50 missing values and 50 incomplete entries. Ten features with most number of missing values are plotted in [Figure 1](#fig-missing-values-0).

|  |
| --- |
| Figure 1: Missing Value Contributions by Features (Top 10) |

The number of missing values by label is summarised in Table 3.

| **A** | **B** | **C** | **D** | **E** |
| --- | --- | --- | --- | --- |
| **50** | **50** | **50** | **50** | **50** |

Table 3**:** Distribution of Missing Values by Labels

Table 1suggests that the proportion of missing values by features is relatively low. Missing value proportions by feature and label are plotted in [Figure 2](#fig-missing-values-1).

|  |
| --- |
| Figure 2: Missing Values in each Feature by label |

Based on the limited available information and available evidence, it is assumed that the data is missing completely at random. Low number of missing entries and relatively similar distributions of missing values across labels suggest no risk of inducing bias in the data due to removal of the missing entries.

### 2.3 Outliers

Outliers are the data points that deviate significantly from the group mean values. Data points which deviate significantly from their group means can be observed in [Figure 3](#fig-outliers-0).

|  |
| --- |
| Figure 3: Outliers by Features |

In the absence of further information, the data points which deviate from the group median by more than 3 median absolute deviations are referred to as outliers.

| **Features** | **Outlier Count** | **Outlier Proportion (%)** |
| --- | --- | --- |
| **X1** | **5** | **0.20408163** |
| **X2** | **3** | **0.12244898** |
| **X3** | **3** | **0.12244898** |
| **X4** | **7** | **0.28571429** |
| **X5** | **3** | **0.12244898** |
| **X6** | **7** | **0.28571429** |
| **X7** | **1** | **0.04081633** |
| **X8** | **96** | **3.91836735** |
| **X9** | **1** | **0.04081633** |
| **X10** | **5** | **0.20408163** |
| **X11** | **11** | **0.44897959** |
| **X12** | **12** | **0.48979592** |
| **X13** | **25** | **1.02040816** |
| **X14** | **5** | **0.20408163** |
| **X15** | **9** | **0.36734694** |
| **X16** | **6** | **0.24489796** |
| **X17** | **6** | **0.24489796** |
| **X18** | **14** | **0.57142857** |
| **X19** | **6** | **0.24489796** |
| **X20** | **8** | **0.32653061** |

Table 4**:** Distribution of Outliers by Features

The outlier count and respective proportions by features is summarised in Table 4. For most of the features, the proportion of outliers is low. However, X8 is of concern as 3.92% of the data points in this feature are outliers.

Presence of outliers can lead to poor classification performances due to issues such as bias, increased error variance, etc. in the model. Although it is possible for the outliers to contain useful information in specific circumstances, given the limited information regarding the nature of the dataset, it is not possible to classify if a given data point is an outlier.

|  |
| --- |
| Figure 4: Mahalobus Distance Distrbution |

Another approach to identify outliers is to consider multivariate distributions of individual samples (rows). This allows us to identify if a particular data point deviated significantly from the dominant trend. [Figure 4](#fig-mahalobus-distance) shows the distribution of mahalobus distances and few data points can be seen which have significantly higher distances compared to the rest. They are of concern, and it is recommended to verify these observations at the following row indexes:

256, 462, 1879, 2009, 2026, 2076, 2231, 2284, 2302

### 2.4 Data Summarisation

Some of the key data summary statistics are summarised in Table 5. Mean is used as the measure of central tendency. Two sets of features are visible based on mean values which can also be observed in [Figure 3](#fig-outliers-0) and [Figure 5](#fig-mean-std-distribution).

| **Variable** | **Mean** | **Std Deviation** | **Min** | **Max** | **Std Err** | **Q0.25** | **Q0.5** | **Q0.75** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X1 | **9.949** | **0.747** | 7.3219 | 12.26 | 0.01510 | 9.449 | 9.962 | 10.448 |
| X2 | **10.054** | **1.023** | 6.6307 | 14.02 | 0.02066 | 9.346 | 10.051 | 10.749 |
| X3 | **8.923** | **0.879** | 5.8410 | 12.05 | 0.01776 | 8.298 | 8.933 | 9.525 |
| X4 | **9.025** | **1.265** | 4.3643 | 13.11 | 0.02556 | 8.149 | 9.013 | 9.890 |
| X5 | **13.749** | **0.907** | 10.5128 | 17.02 | 0.01832 | 13.145 | 13.766 | 14.375 |
| X6 | **7.892** | **1.158** | 3.5931 | 11.94 | 0.02340 | 7.115 | 7.903 | 8.662 |
| X7 | **0.422** | **0.263** | -0.4702 | 1.31 | 0.00532 | 0.215 | 0.410 | 0.623 |
| X8 | **0.251** | **0.220** | -0.6464 | 1.13 | 0.00444 | 0.111 | 0.194 | 0.380 |
| X9 | **0.590** | **0.324** | -0.2345 | 1.68 | 0.00654 | 0.339 | 0.638 | 0.847 |
| X10 | **0.445** | **0.200** | -0.1624 | 1.20 | 0.00404 | 0.305 | 0.417 | 0.584 |
| X11 | **9.132** | **1.098** | 5.2898 | 12.70 | 0.02218 | 8.359 | 9.070 | 9.829 |
| X12 | **11.804** | **1.002** | 8.0456 | 15.48 | 0.02023 | 11.147 | 11.789 | 12.430 |
| X13 | **8.369** | **0.838** | 5.2614 | 11.23 | 0.01694 | 7.833 | 8.404 | 8.919 |
| X14 | **7.727** | **1.233** | 3.3328 | 11.66 | 0.02492 | 6.889 | 7.734 | 8.565 |
| X15 | **10.695** | **0.979** | 7.6032 | 14.61 | 0.01977 | 10.048 | 10.701 | 11.373 |
| X16 | **7.655** | **1.079** | 3.6035 | 11.06 | 0.02180 | 6.954 | 7.653 | 8.378 |
| X17 | **0.497** | **0.224** | -0.2380 | 1.32 | 0.00452 | 0.346 | 0.495 | 0.646 |
| X18 | **0.675** | **0.208** | -0.0472 | 1.43 | 0.00421 | 0.533 | 0.686 | 0.813 |
| X19 | **0.528** | **0.257** | -0.3473 | 1.52 | 0.00520 | 0.358 | 0.530 | 0.699 |
| X20 | **0.600** | **0.229** | -0.2072 | 1.39 | 0.00462 | 0.444 | 0.601 | 0.749 |

Table 5**:** Summary Statistics

|  |
| --- |
| Figure 5: Distribution of Feature Means and Standard Deviations |

Some features having significantly higher means and standard deviations compared to others could lead to biased classification models. Therefore, the data needs to be normalised.

### 2.5 Splitting training and test sets

To prevent information leak, the data is split into training and test samples (Kuhn and Max, 2008) prior to normalisation. This allows us to train the classification models on the training data and measure their performances on the test data which is completely unknown to the model.

|  | **A** | **B** | **C** | **D** | **E** |
| --- | --- | --- | --- | --- | --- |
| **Training Set** | 366 | 359 | 366 | 385 | 363 |
| **Test Set** | 73 | 72 | 74 | 77 | 73 |
| **Validation Set** | 48 | 47 | 48 | 51 | 48 |
| **Total** | 487 | 478 | 488 | 513 | 484 |

Table 6**:** Distribution of Labels in Training and Test Sets

Label distributions after training and test splits are summarised in Table 6.

### 2.6 Data Normalisation

Data normalisation improves the interpretability of the data by changing the values of numeric features to a common scale. The data is normalised using z-score. z-score is measured by subtracting the feature mean from each value which is divided by the feature standard deviation as shown in [Equation 1](#eq-z-score-measurement).

### 2.7 Correlation Analysis

Correlation is a statistical measure which explains the relationship between two variables. Measure of correlation ranges between -1 and 1. Correlation values closer to 0 indicate weak correlation with 0 being no correlation and correlation values at the extremes indicate strong negative or positive correlations. [Figure 6](#fig-corr-plot) shows the correlations between features in the dataset.

|  |
| --- |
| Figure 6: Correlation between Features |

Highly correlated features may contain redundant information and could lead to overfitting the classification model and effect its performance. In [Section 3](#sec-feature-selection), data is explored for dimensionality reduction and feature selection to investigate if any of the correlated features could be removed.

## 3 Dimensionality Reduction and Feature Selection

Dimensionality of a dataset increases with the number of features. Dimensionality reduction reduces the complexity of the classification model and in certain cases, improves the overall classification performance. Exploring the contribution of each feature to explain the variance in the dataset allows us to select important features to retain.

### 3.1 Principal Component Analysis (PCA)

PCA is one of the commonly used dimensionality reduction techniques (Abdi and Williams, 2010). It functions by projecting the n-dimensional feature space to a relatively smaller dimensional space called principal components. Each principal component is a linear combination of the original features at different proportions. The principal components are ordered with the first principal component explaining the most variance in the dataset followed by the second principal component and so on. A decision regarding the number of principal components to retain is made based on the amount of variance to retain.

#### 3.1.1 Parallel Analysis

Statistical methods such as parallel analysis can be used to estimate the number of principal components to retain. Parallel analysis is performed using paran library (Dinno, 2009) which recommends 4 principal components to be retained.

*Please note the difference between dimensions in* Table 7 *and features. Each dimension is a combination of multiple features at different proportions.*

| **Dimensions** | **Eigenvalue** | **Variance (%)** | **Cumulative Variance (%)** |
| --- | --- | --- | --- |
| **Dim.1** | **2.9668982** | **14.8344910** | **14.83449** |
| **Dim.2** | **2.8283799** | **14.1418997** | **28.97639** |
| **Dim.3** | **2.2024043** | **11.0120214** | **39.98841** |
| **Dim.4** | **1.1603140** | **5.8015702** | **45.78998** |
| **Dim.5** | 1.0346035 | 5.1730174 | 50.96300 |
| **Dim.6** | 0.9968554 | 4.9842771 | 55.94728 |
| **Dim.7** | 0.9687413 | 4.8437065 | 60.79098 |
| **Dim.8** | 0.9343126 | 4.6715631 | 65.46255 |
| **Dim.9** | 0.9081312 | 4.5406561 | 70.00320 |
| **Dim.10** | 0.8766685 | 4.3833425 | 74.38654 |
| **Dim.11** | 0.8549703 | 4.2748517 | 78.66140 |
| **Dim.12** | 0.8001130 | 4.0005651 | 82.66196 |
| **Dim.13** | 0.7644286 | 3.8221428 | 86.48410 |
| **Dim.14** | 0.7272942 | 3.6364712 | 90.12058 |
| **Dim.15** | 0.6155401 | 3.0777004 | 93.19828 |
| **Dim.16** | 0.4861481 | 2.4307407 | 95.62902 |
| **Dim.17** | 0.3174128 | 1.5870642 | 97.21608 |
| **Dim.18** | 0.2404162 | 1.2020809 | 98.41816 |
| **Dim.19** | 0.1680152 | 0.8400758 | 99.25824 |
| **Dim.20** | 0.1483524 | 0.7417622 | 100.00000 |

Table 7**:** Distribution of Labels in Training and Test Sets

Eigen values indicate the amount of variance captured by a principal component. The cumulative variance explained by the principal components recommended by parallel analysis is 45.7899823. This is of concern as significant amount of variance is not retained.

#### 3.1.2 PCA Observations

Feature contributions to each of the principal components is summarised in Table 8 with the most dominant features for each principal component highlighted.

| **Features** | **Comp1** | **Comp2** | **Comp3** | **Comp4** |
| --- | --- | --- | --- | --- |
| **X1** | -0.019 | -0.165 | 0.194 | 0.567 |
| **X2** | 0.012 | -0.092 | -0.491 | 0.151 |
| **X3** | -0.081 | 0.232 | 0.646 | -0.160 |
| **X4** | -0.021 | -0.193 | -0.039 | -0.491 |
| **X5** | -0.064 | 0.102 | -0.188 | 0.249 |
| **X6** | -0.032 | 0.501 | -0.220 | -0.128 |
| **X7** | -0.082 | 0.449 | 0.692 | 0.147 |
| **X8** | 0.072 | -0.761 | -0.251 | -0.300 |
| **X9** | -0.045 | 0.927 | -0.045 | 0.082 |
| **X10** | 0.009 | -0.645 | 0.455 | 0.136 |
| **X11** | -0.009 | 0.257 | 0.395 | -0.355 |
| **X12** | 0.042 | 0.226 | -0.366 | 0.159 |
| **X13** | -0.062 | -0.408 | 0.312 | 0.184 |
| **X14** | 0.080 | 0.110 | -0.401 | 0.029 |
| **X15** | 0.068 | -0.183 | 0.023 | 0.360 |
| **X16** | -0.039 | 0.173 | -0.335 | 0.078 |
| **X17** | 0.815 | 0.050 | 0.045 | -0.015 |
| **X18** | -0.890 | 0.063 | -0.072 | -0.097 |
| **X19** | 0.903 | 0.100 | 0.037 | -0.043 |
| **X20** | -0.806 | -0.097 | -0.041 | 0.030 |

Table 8**:** Feature Contributions to Principal Components

Dominant feature associations with each of the principal components are as follows:

Principal Component 1: X17, X18, X19 and X20, Principal Component 2: X6, X8, X9 and X10, Principal Component 3: X3 and X7, Principal Component 4: X1

[Figure 7](#fig-pca-feature-interpretation) visualises feature loadings across the principal components. The length of an arrow corresponding to each feature along a principal component indicates the strength of its association with it.

|  |
| --- |
| Figure 7: Feature Association with Principal Components |

[Figure 8](#fig-pca-lable-interpretation) visualises (Kassambara, 2016) the distribution of samples corresponding to each label when projected along different principal axes. There is substantial overlap for the samples corresponding to labels B and E. This suggests that PCA may not be a suitable option for dimensionality reduction.

|  |
| --- |
| Figure 8: Label Distribution With Principal Components |

### 3.2 Decision Tree and Random Forest for Feature Selection

Decision Trees (DT) provide an alternative method for feature selection. DT estimates importance of each feature using metrics such as Gini impurity or information gain. Feature selection can subsequently be performed by selecting features with most importance. R rpart library (Therneau and Atkinson, 2022) is used to generate a DT model.

|  |
| --- |
| Figure 9: Node Selection Using Decision Tree |

The series of decisions used by a DT model are plotted in [Figure 9](#fig-decision-tree). It can be observed that only a subset of the original features is used in the decision making process.

Based on DT feature importance, the selected features are: X9, X7, X8, X10, X11, X3 and X6.

Random forest (RF) can be seen as an extension to decision trees. RF generates multiple random subsets of samples and features, and DT models are trained on them. These multiple DT models are aggregated for a final optimised estimation. R randomForest library (Liaw and Wiener, 2002a) is used to generate a RF model and perform feature selection.

|  |
| --- |
| Figure 10: Random Forest Feature Importance |

Based on random forest feature importance, the selected features are: X9, X7, X8, X10, X3, X11 and X2.

Since RFs generate multiple DT models for estimations, their execution times are longer compared to DT. RF decision making process is also not as interpretable as DTs. However, the RF models are not tightly constrained around the training data and RF models usually tend to have better classification accuracy on the test data.

## 4 Classification

Classification is a supervised learning technique where predictions are made based on models trained using labelled inputs. Prior to classifying the data, a classification model is trained to identify relationships and patterns between the features. The knowledge of the statistical properties of the training data is used to classify new unknown data samples.

[Figure 11](#fig-pairs-panels) shows the correlations and distributions of the features selected using DT and RF models. The distribution of the features appears to be approximately normal and most of the features are not highly correlated. These are some of the essential assumptions regarding the statistical properties of the data.

|  |  |  |
| --- | --- | --- |
| |  | | --- | | (a) Distribution of DT Selected Features | | (b) Distribution of RF Selected Features |   Figure 11: Histograms and Correlations for the Training Data |

[Figure 12](#fig-multivariate-normal-dist) shows the presence of multivariate normal distributions in the data.

|  |
| --- |
| Figure 12: Multivariate Normal Distributions |

Based on the performances on training and test data, RF and MclustDA discriminant analysis are selected to be the potential classification models suitable for this dataset. Random Forest is non-parametric and makes no explicit assumptions regarding the distribution of the data. MclustDA discriminant analysis is suitable when the data follows multivariate normal distribution. Performances of RF and MclustDA are compared using confusion matrices and F1 scores. F1 score is the measure of model’s accuracy, and it is calculated as shown in [Equation 2](#eq-F1-score-measurement):

### 4.1 Random Forest

As discussed in [Section 3.2](#sec-dt-rf-feature-selection), RF is an extension to DT which solves the overfitting issue. randomForest::randomForest function (Liaw and Wiener, 2002b; Breiman, 2001) is used to train the RF model.

|  | **A** | **B** | **C** | **D** | **E** |
| --- | --- | --- | --- | --- | --- |
| **A** | 72 | 0 | 0 | 0 | 3 |
| **B** | 0 | 63 | 2 | 2 | 9 |
| **C** | 0 | 2 | 70 | 0 | 5 |
| **D** | 1 | 2 | 0 | 74 | 7 |
| **E** | 0 | 5 | 2 | 1 | 49 |

Table 9**:** Random Forest Confusion Matrix with All Features

|  | **A** | **B** | **C** | **D** | **E** |
| --- | --- | --- | --- | --- | --- |
| **A** | 73 | 0 | 0 | 0 | 1 |
| **B** | 0 | 64 | 1 | 2 | 11 |
| **C** | 0 | 1 | 70 | 0 | 4 |
| **D** | 0 | 1 | 0 | 73 | 3 |
| **E** | 0 | 6 | 3 | 2 | 54 |

Table 10**:** Random Forest Confusion Matrix with Features Selected Using RF

An accuracy of 0.89 is observed using all the features with F1 score of 0.97. An accuracy of 0.91 is observed using RF recommended features and the corresponding F1 score is 0.99. Corresponding confusion matrices are summarised in Table 9 and Table 10.

### 4.2 MclustDA Discriminant Analysis

Discriminant analysis based on Gaussian finite mixture modelling is a classification model which is suitable when the data follows multivariate normal distributions. mclust::MclustDA function (Scrucca *et al.*, 2016; Fraley *et al.*, 2012) is used for MclustDA Discriminant Analysis.

|  | **A** | **B** | **C** | **D** | **E** |
| --- | --- | --- | --- | --- | --- |
| **A** | 72 | 0 | 0 | 0 | 1 |
| **B** | 0 | 67 | 0 | 0 | 9 |
| **C** | 0 | 1 | 74 | 0 | 1 |
| **D** | 0 | 0 | 0 | 75 | 0 |
| **E** | 1 | 4 | 0 | 2 | 62 |

Table 11**:** MclustDA Confusion Matrix with Features Selected Using DT

|  | **A** | **B** | **C** | **D** | **E** |
| --- | --- | --- | --- | --- | --- |
| **A** | 72 | 0 | 0 | 0 | 1 |
| **B** | 0 | 68 | 0 | 0 | 6 |
| **C** | 0 | 1 | 74 | 0 | 2 |
| **D** | 0 | 0 | 0 | 75 | 1 |
| **E** | 1 | 3 | 0 | 2 | 63 |

Table 12**:** MclustDA Confusion Matrix with Features Selected Using RF

When features recommended by DT and RF are used, accuracies of 0.95 and 0.95 respectively are observed with the corresponding F1 scores being 0.99 and 0.99 respectively. Corresponding confusion matrices are summarised in Table 11and Table 12.

## 5 Recommendation and Validation

Based on the performances on the training data, RF and MclustDA classification models are recommended. For the case of RF, mtry hyperparameter is set to 3. Their performances are evaluated using the validation data. Features recommended by RF are used in both the models.

For the case of RF, an accuracy of 0.91 is observed and the corresponding F1 score is 0.98. For the case of MclustDA, an accuracy of 0.95 is observed and the corresponding F1 score is 0.98. The corresponding confusion matrices are summarised in Table 13 and Table 14.

|  | **A** | **B** | **C** | **D** | **E** |
| --- | --- | --- | --- | --- | --- |
| **A** | 48 | 0 | 0 | 0 | 2 |
| **B** | 0 | 42 | 3 | 0 | 3 |
| **C** | 0 | 0 | 42 | 0 | 4 |
| **D** | 0 | 0 | 0 | 51 | 2 |
| **E** | 0 | 5 | 3 | 0 | 37 |

Table 13**:** RF Validation Results

|  | **A** | **B** | **C** | **D** | **E** |
| --- | --- | --- | --- | --- | --- |
| **A** | 47 | 0 | 0 | 0 | 1 |
| **B** | 0 | 45 | 0 | 0 | 2 |
| **C** | 0 | 0 | 48 | 0 | 3 |
| **D** | 0 | 0 | 0 | 49 | 2 |
| **E** | 1 | 2 | 0 | 2 | 40 |

Table 14**:** MclustDA Validation Results

When tested on validation samples, accuracy of RF model marginally increased compared to accuracy on test samples while accuracy of MclustDA is similar in both the cases.

## 6 Conclusion

MclustDA marginally outperformed RF in classification accuracy while computational complexity of RF is low compared to MclustDA. Optimum overall accuracy is achieved when the features are selected based on feature importance metrics obtained from the RF model. If computational complexity is a concern, RF classification model is recommended. However, classification performance can be optimised by using MclustDA.

## References

Abdi, H. and Williams, L.J. (2010) Principal component analysis: Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*. [online]. 2 (4), pp.433–459. Available from: <https://onlinelibrary.wiley.com/doi/10.1002/wics.101> [Accessed 9 May 2023].

Breiman, L. (2001) Random forests. *Machine learning*. 45, Springer, pp.5–32.

Dinno, A. (2009) Exploring the Sensitivity of Horn’s Parallel Analysis to the Distributional Form of Random Data. *Multivariate Behavioral Research*. [online]. 44 (3), pp.362–388. Available from: <https://www.tandfonline.com/doi/full/10.1080/00273170902938969> [Accessed 9 May 2023].

Fraley, C., Raftery, A.E., Murphy, T.B. and Scrucca, L. (2012) *mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation* Technical report.

Kassambara, A. (2016) Factoextra: extract and visualize the results of multivariate data analyses. *R package version*. 1.

Kuhn and Max (2008) Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*. [online]. 28 (5), pp.1–26. Available from: <https://www.jstatsoft.org/index.php/jss/article/view/v028i05>.

Liaw, A. and Wiener, M. (2002a) Classification and Regression by randomForest. *R News*. [online]. 2 (3), pp.18–22. Available from: <https://CRAN.R-project.org/doc/Rnews/>.

Liaw, A. and Wiener, M. (2002b) Classification and Regression by randomForest. *R News*. [online]. 2 (3), pp.18–22. Available from: <https://CRAN.R-project.org/doc/Rnews/>.

Scrucca, L., Fop, M., Murphy, T.B. and Raftery, A.E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R journal*. 8 (1), NIH, p.289.

Therneau, T. and Atkinson, B. (2022) *rpart: Recursive Partitioning and Regression Trees* [online]. Available from: <https://CRAN.R-project.org/package=rpart>.