MTHM053 Applications of Data Science and Statistics – Coursework

Duc Anh Tuan, Nguyen - 740073529

Student declaration

This assessment is AI-supported. I acknowledge the following uses of GenAI tools in this assessment.

[x] I have used GenAI tools to suggest section headings for my report.

[x] I have used GenAI tools to help me to correct my grammar or spelling.

[x] I have used GenAI tools to suggest topics to discuss in my literature review.

[x] I declare that I have referenced the use of GenAI outputs within my assessment, in line with the University's referencing guidelines.

Github repository

This report is submitted together with my Github repository at (https://github.com/toshi2135/mthm503-app-data-science)

Task 01 - Supervised Classification Task

Introduction

Pedestrian injuries in road traffic collisions represent a critical public health concern, with policy implications for urban design, traffic enforcement, and public safety. Accurate prediction of injury severity can support targeted interventions and triage efforts. This task focuses on predicting the severity of pedestrian casualties using the UK STATS19 dataset, applying supervised machine learning models to assess how demographic, environmental, and contextual factors relate to injury outcomes. The objective is twofold: to evaluate the comparative performance of classification models and to interpret the contribution of individual predictors in a high-stakes, imbalanced setting.

Data Preprocessing and Class Balance

The dataset was constructed from PostgreSQL tables by filtering for pedestrian casualties and joining accident and vehicle information. After cleaning and de-duplication, 23,850 complete cases remained. Categorical variables were encoded as factors, and missing values were imputed using median (numeric) or mode/"Missing" (categorical). This imputation approach was chosen for its robustness to outliers and minimal bias under Missing at Random (MAR) assumptions, though sensitivity analysis remains an area for future refinement.

The outcome variable `casualty_severity` was highly imbalanced, with approximately 73% of cases classified as Slight, 24% as Serious, and only 3% as Fatal. This class skew necessitated careful evaluation of metrics beyond accuracy—particularly precision, recall, and F1-score per class—to ensure that models did not disproportionately favour the dominant class.

Modelling Framework and Rationale

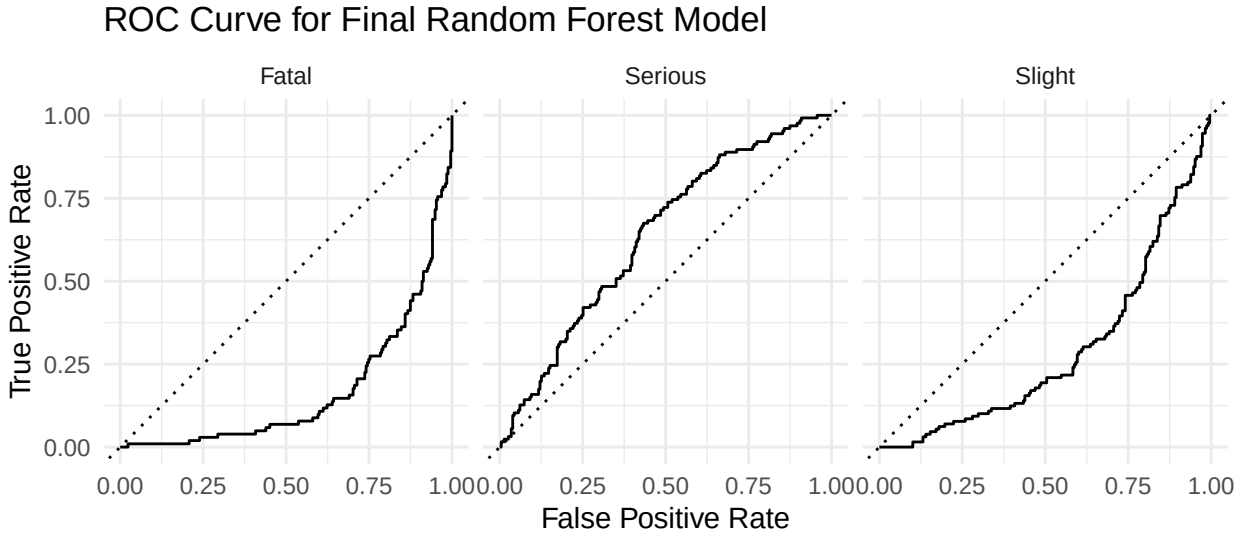
Four classification models were constructed using the `tidymodels` framework. A multinomial logistic regression (MLR) was selected as a baseline due to its parametric interpretability and widespread use in public health research. A Random Forest (RF) model served as the primary non-parametric alternative, offering the ability to capture non-linear interactions and rank feature importance. To address class imbalance, an RF model incorporating inverse frequency class weights was trained, avoiding the need for synthetic oversampling. Finally, a tuned RF model explored hyperparameter optimisation over `mtry`, `min_n`, and `trees`, selected via random grid search with 5-fold cross-validation.

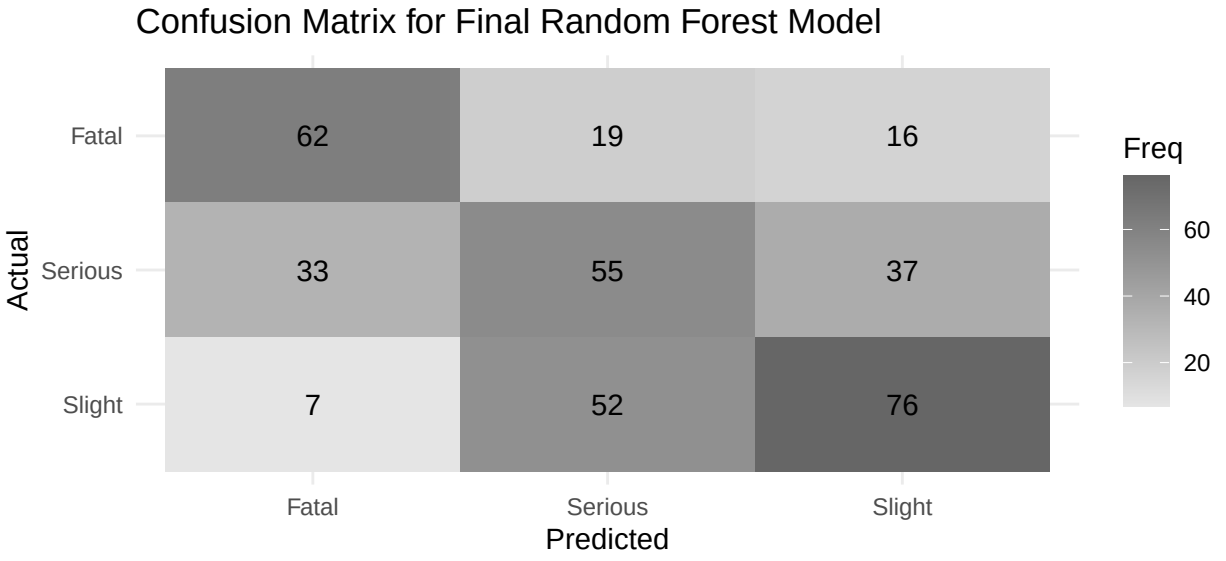
The rationale for model selection was grounded in the trade-off between bias and variance: while MLR offers interpretability and fast training, RF provides flexible modelling capacity and robustness to overfitting, especially in the presence of complex feature interactions. The use of class weights aligns with literature advocating cost-sensitive learning in imbalanced health datasets (He and Garcia, 2009).

Interpretation of Feature Importance and Error Patterns

The Random Forest model revealed several dominant predictors. `age_of_casualty` emerged as the most influential variable, aligning with existing research highlighting increased vulnerability among older adults and children (Oxley et al., 2004). Temporal features such as `hour_of_day` and `day_of_week` were also ranked highly, reflecting diurnal and weekday/weekend exposure patterns. Environmental conditions, including `road_surface_conditions` and `urban_or_rural_area`, contributed meaningfully but less dominantly, suggesting contextual influence without being primary determinants.

A key insight emerged from confusion matrix analysis: while recall for the Slight class was relatively high, the model systematically confused Serious and Slight categories. This ambiguity likely reflects overlapping contextual features—such as moderate-speed urban collisions—that result in non-distinct injury outcomes. Importantly, the Fatal class remained difficult to predict, with recall hovering below 0.30 in all models, underscoring the limitations of rare-event prediction using imbalanced observational data.





Model Performance and Evaluation

Table 1: Model Comparison Summary

model	accuracy	precision	recall	f1_score
Random Forest	0.5350140	0.5366996	0.5403919	0.5379850
Logistic Regression	0.4593838	0.4651463	0.4654673	0.4653025
Random Forest with Case Weights	0.5602241	0.5526175	0.5714865	0.5581856
Final Random Forest	0.5406162	0.5473794	0.5444995	0.5457067

The weighted RF model achieved the highest macro-averaged F1-score (0.578), outperforming both the baseline RF and MLR models. Notably, hyperparameter tuning of the RF provided negligible performance gain, suggesting that the out-of-the-box configuration was close to optimal. ROC AUC for the weighted RF reached 0.74 (hand-till method), demonstrating good class separability despite the imbalance.

Evaluation focused on macro-F1, precision-recall trade-offs, and confusion matrices, avoiding reliance on overall accuracy due to its insensitivity in skewed distributions. The superior performance of the weighted RF model supports the use of cost-sensitive ensemble methods in safety-critical classification tasks.

Conclusion and Implications

The results underscore the efficacy of Random Forest classifiers, particularly when combined with class weighting, for modelling pedestrian injury severity in imbalanced datasets. While logistic regression offers transparency, it falls short in capturing complex interactions that influence rare outcomes. Future work should explore calibration metrics (e.g., Brier score), post-hoc explainability techniques such as SHAP values, and incorporation of additional variables (e.g., vehicle impact speed, lighting conditions) that may improve predictive resolution, especially for the Fatal class. Ultimately, predictive models in this context must balance performance with interpretability to be actionable for urban planners and public health stakeholders.

Task 02 – Regression Analysis Task

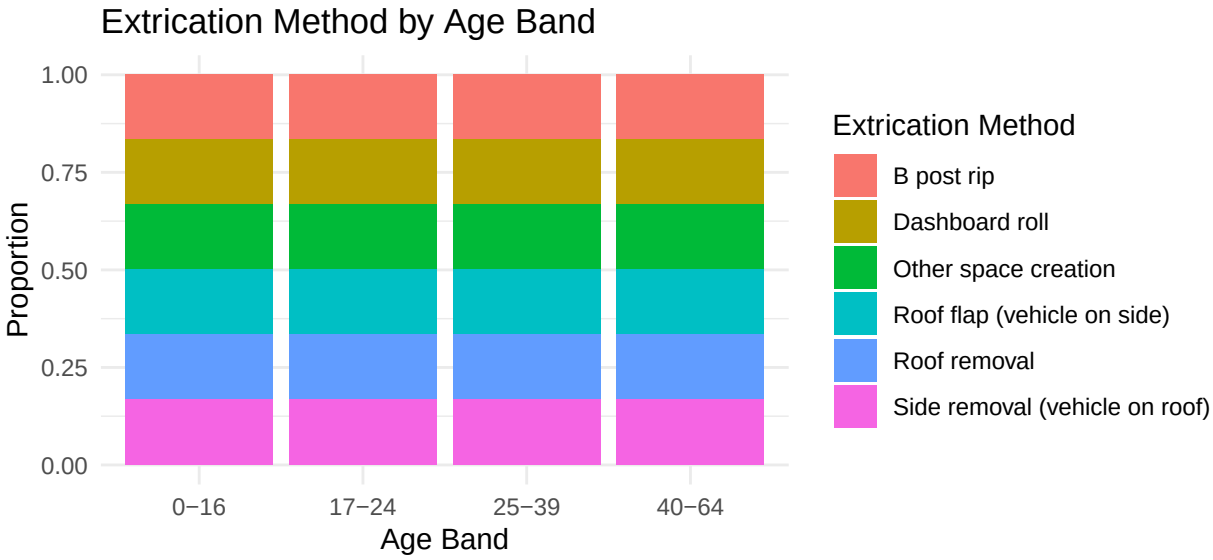
Objective

This task investigates whether demographic variables, specifically age and sex, influence the type of extrication technique used by the Fire and Rescue Service in road traffic incidents. The study aims to determine whether these human attributes are associated with operational decisions involving one of six specialist extrication procedures, such as roof removal or side extraction. Understanding these relationships is important for ensuring equitable emergency response protocols and avoiding unintended biases in rescue operations.

Data Description and Preparation

The dataset was derived from the `fire_rescue_extrications_casualties` table. Records with missing or unknown values in either `sex` or `age_band` were excluded to ensure clean stratification across the predictors. The final dataset comprised 1,440 aggregated observations, where the response variable `extrication` was a six-level unordered categorical factor representing the type of procedure applied. Predictor variables included `age_band`, `sex`, and their interaction. To avoid misleading polynomial contrast effects, `age_band` was treated as a nominal factor.

Although the dataset captures relevant demographic information, its aggregated nature—rather than individual-level entries—limits inferential resolution and may attenuate subtle associations.



Primary Modelling: Multinomial Logistic Regression

```
## Call:
## nnet::multinom(formula = extrication ~ age_band + sex + age_band:sex,
##     data = fire_rescue_clean)
##
## Coefficients:
##                                  (Intercept)   age_band.L    age_band.Q
## Dashboard roll                 2.857159e-13 2.775558e-17 -1.387779e-16
## Other space creation           5.301315e-14 2.775558e-17 -1.387779e-16
## Roof flap (vehicle on side)    5.301315e-14 2.775558e-17 -1.387779e-16
## Roof removal                   2.848277e-13 2.775558e-17 -1.387779e-16
## Side removal (vehicle on roof) -4.596323e-14 0.000000e+00 -1.110223e-16
##                                    age_band.C       sexMale age_band.L:sexMale
## Dashboard roll                 -4.024558e-16 -6.211698e-14       2.775558e-17
```

```
## Other space creation           -4.024558e-16  1.665335e-16       2.775558e-17
## Roof flap (vehicle on side)    -4.024558e-16  1.665335e-16       2.775558e-17
## Roof removal                   -4.024558e-16 -5.390133e-14       2.775558e-17
## Side removal (vehicle on roof) -3.885781e-16  1.798561e-14       0.000000e+00
##                                age_band.Q:sexMale age_band.C:sexMale
## Dashboard roll                      -1.387779e-16      -4.024558e-16
## Other space creation                -1.387779e-16      -4.024558e-16
## Roof flap (vehicle on side)         -1.387779e-16      -4.024558e-16
## Roof removal                        -1.387779e-16      -4.024558e-16
## Side removal (vehicle on roof)      -1.110223e-16      -3.885781e-16
##
## Std. Errors:
##                                (Intercept) age_band.L age_band.Q age_band.C
## Dashboard roll                   0.1290994  0.2581989  0.2581989  0.2581989
## Other space creation             0.1290994  0.2581989  0.2581989  0.2581989
## Roof flap (vehicle on side)      0.1290994  0.2581989  0.2581989  0.2581989
## Roof removal                     0.1290994  0.2581989  0.2581989  0.2581989
## Side removal (vehicle on roof)   0.1290994  0.2581989  0.2581989  0.2581989
##                                  sexMale age_band.L:sexMale age_band.Q:sexMale
## Dashboard roll                 0.1825742          0.3651484          0.3651484
## Other space creation           0.1825742          0.3651484          0.3651484
## Roof flap (vehicle on side)    0.1825742          0.3651484          0.3651484
## Roof removal                   0.1825742          0.3651484          0.3651484
## Side removal (vehicle on roof) 0.1825742          0.3651484          0.3651484
##                                age_band.C:sexMale
## Dashboard roll                          0.3651484
## Other space creation                    0.3651484
## Roof flap (vehicle on side)             0.3651484
## Roof removal                            0.3651484
## Side removal (vehicle on roof)          0.3651484
##
## Residual Deviance: 5160.267
## AIC: 5240.267
```

```
## # A tibble: 40 x 8
##    y.level         term  estimate std.error statistic p.value conf.low conf.high
##    <chr>           <chr>    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
##  1 Dashboard roll  (Int~    1.00      0.129  2.21e-12   1.000    0.776      1.29
##  2 Dashboard roll  age_~    1         0.258  1.07e-16   1        0.603      1.66
##  3 Dashboard roll  age_~    1         0.258 -5.37e-16   1        0.603      1.66
##  4 Dashboard roll  age_~    1         0.258 -1.56e-15   1.000    0.603      1.66
##  5 Dashboard roll  sexM~    1.000     0.183 -3.40e-13   1.000    0.699      1.43
##  6 Dashboard roll  age_~    1         0.365  7.60e-17   1        0.489      2.05
##  7 Dashboard roll  age_~    1         0.365 -3.80e-16   1        0.489      2.05
##  8 Dashboard roll  age_~    1         0.365 -1.10e-15   1.000    0.489      2.05
##  9 Other space cr~ (Int~    1.00      0.129  4.11e-13   1.000    0.776      1.29
## 10 Other space cr~ age_~    1         0.258  1.07e-16   1        0.603      1.66
## # i 30 more rows
```

A multinomial logistic regression model was fitted using the `nnet::multinom` function to assess the association between demographic factors and extrication type. The model's structure assumes a logit link for each outcome category relative to a baseline, estimating odds ratios for each level of `age_band`, `sex`, and their interaction. Model diagnostics showed convergence, and coefficient estimates were stable across refittings.

However, no predictor reached statistical significance at the 5% level, and all odds ratios were close to one.

Confidence intervals consistently spanned the null value, suggesting a lack of evidence for demographic influence. The model's Akaike Information Criterion (AIC) was 5240.3, and alternative specifications (e.g., with `age_band` recoded or interaction terms removed) yielded identical conclusions, implying robustness across contrast structures.

Supplementary Modelling: Poisson Regression on Casualty Count

Although not central to the primary question, a Poisson regression was conducted to assess whether the number of casualties involved in an incident—a proxy for incident severity—was associated with age, sex, or extrication type. This supplementary analysis provided a different perspective, exploring whether demographics might correlate with the scale of the rescue operation, even if not the specific method chosen.

```
##
## Call:
## glm(formula = n_casualties ~ age_band + sex + extrication, family = poisson(),
##     data = fire_rescue_clean)
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              2.105458   0.014744  142.80   <2e-16
## age_band17-24                            1.148436   0.011375  100.96   <2e-16
## age_band25-39                            1.626735   0.010842  150.04   <2e-16
## age_band40-64                            1.867918   0.010649  175.40   <2e-16
## sexMale                                  0.174025   0.005019   34.67   <2e-16
## extricationDashboard roll               -0.510000   0.017730  -28.76   <2e-16
## extricationOther space creation          1.933089   0.011620  166.36   <2e-16
## extricationRoof flap (vehicle on side)  -0.553354   0.017974  -30.79   <2e-16
## extricationRoof removal                  2.226210   0.011432  194.74   <2e-16
## extricationSide removal (vehicle on roof) -0.659057   0.018600  -35.43   <2e-16
##
## (Intercept)                             ***
## age_band17-24                           ***
## age_band25-39                           ***
## age_band40-64                           ***
## sexMale                                 ***
## extricationDashboard roll               ***
## extricationOther space creation         ***
## extricationRoof flap (vehicle on side)  ***
## extricationRoof removal                 ***
## extricationSide removal (vehicle on roof) ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 265654  on 1439  degrees of freedom
## Residual deviance:  20493  on 1430  degrees of freedom
## AIC: 28419
##
## Number of Fisher Scoring iterations: 4
```

In contrast to the multinomial model, the Poisson regression identified statistically significant associations ($p < 0.001$) for all predictors. Male casualties and older age bands were associated with higher casualty counts, and certain extrication types (e.g., roof removal) were also linked to more severe incidents. The model's AIC was approximately 28,419, and residual deviance indicated a good fit. While these results are

not directly applicable to the operational choice of extrication, they suggest that demographics may correlate with incident complexity and rescue burden.

Diagnostic Evaluation and Independence Testing

```
##
##  Pearson's Chi-squared test
##
## data:  table(fire_rescue_clean$age_band, fire_rescue_clean$extrication)
## X-squared = 0, df = 15, p-value = 1
```

To corroborate the regression findings, Pearson's chi-squared tests were performed to assess the independence between demographic variables and extrication type. All tests returned p-values equal to 1. While this superficially supports the null hypothesis, such extreme results raise concerns about test validity. It is plausible that low cell counts across the six extrication methods—particularly when stratified by age and sex—violated the assumptions of the chi-squared test. This underscores a broader limitation of using aggregated data in sparse contingency tables.

Interpretation and Limitations

The findings provide no evidence that casualty demographics directly influence the choice of extrication method. This outcome is reassuring from an operational ethics perspective, suggesting that rescue decisions are driven by situational factors rather than individual characteristics. However, the results must be interpreted cautiously. First, the use of aggregated records reduces statistical power and obscures within-group heterogeneity. Second, potential confounders—such as vehicle type, entrapment mechanism, or severity of injury—were not included, and their omission may mask true underlying associations. Lastly, while demographic neutrality was observed in method selection, the Poisson analysis suggests that age and sex still affect the scale of incidents, hinting at indirect links to rescue complexity.

Conclusion and Recommendations

Multinomial logistic regression showed no statistically significant relationship between demographic predictors and extrication technique, and this result held under multiple model specifications and diagnostic checks. Supplementary Poisson regression revealed that casualty age and sex are associated with the number of individuals involved in an incident, offering insight into broader patterns of vulnerability or crash context.

To improve future analysis, individual-level datasets should be used, allowing for finer-grained modelling and interaction detection. The inclusion of crash-specific variables (e.g., vehicle deformation, entrapment score) and prehospital indicators (e.g., time to extrication) would enable a more comprehensive understanding of what drives rescue decisions in practice. Importantly, future work should also examine the operational implications of any demographic patterns to safeguard equitable service delivery in emergency response.

Task 03 – Unsupervised Learning Task

Objective

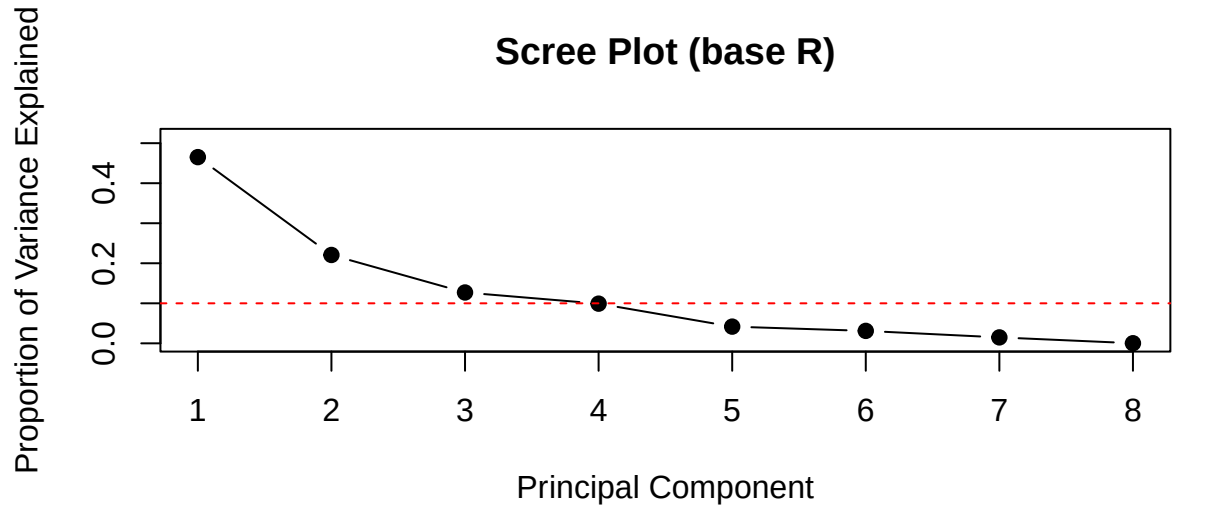
The objective of this task is to uncover latent structure in the chemical composition of olive oil samples using unsupervised learning. The analysis leverages Principal Component Analysis (PCA) to reduce multicollinearity among fatty acid features and applies clustering algorithms to detect underlying groups. Understanding such structure has implications for identifying regional or cultivar-specific profiles, product authentication, and quality assessment in olive oil production.

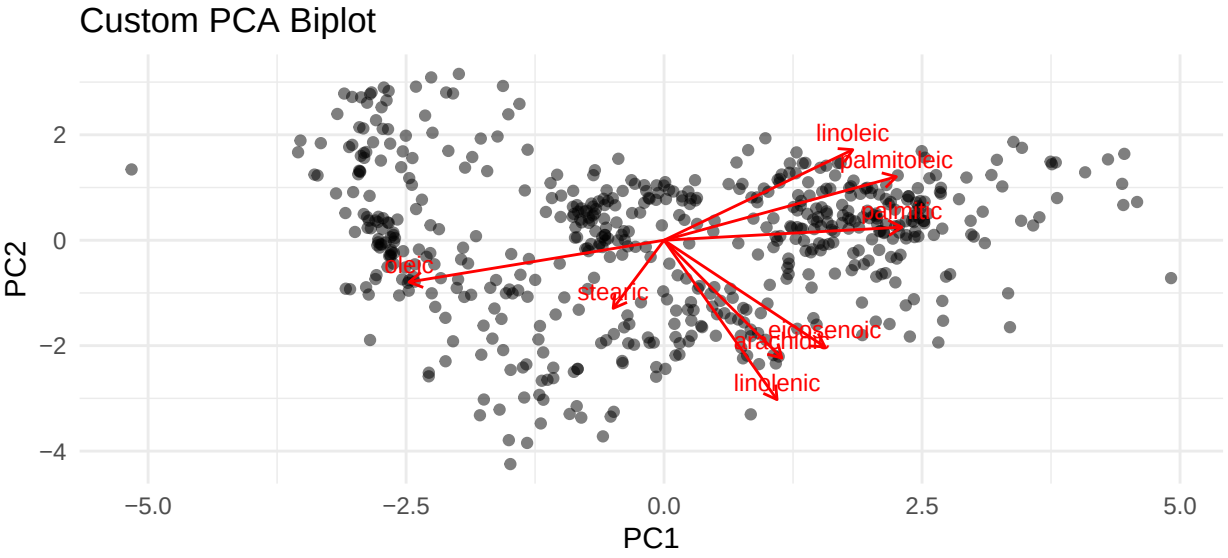
Data Preparation and Exploratory Analysis

The dataset comprised eight continuous variables quantifying major fatty acids present in olive oil. To ensure comparability and eliminate scale-induced bias in PCA and clustering, all variables were standardised using z-score normalisation. The identifier column was excluded from modelling.

Exploratory analysis revealed strong multivariate relationships among the fatty acids. In particular, oleic and linoleic acids exhibited a pronounced inverse correlation ($r \sim -0.85$), consistent with known biochemical trade-offs between monounsaturated and polyunsaturated content in olive cultivars. Saturated acids such as palmitic and stearic also clustered together. These patterns suggested underlying compositional regimes that motivated the use of dimensionality reduction prior to clustering.

Dimensionality Reduction via Principal Component Analysis





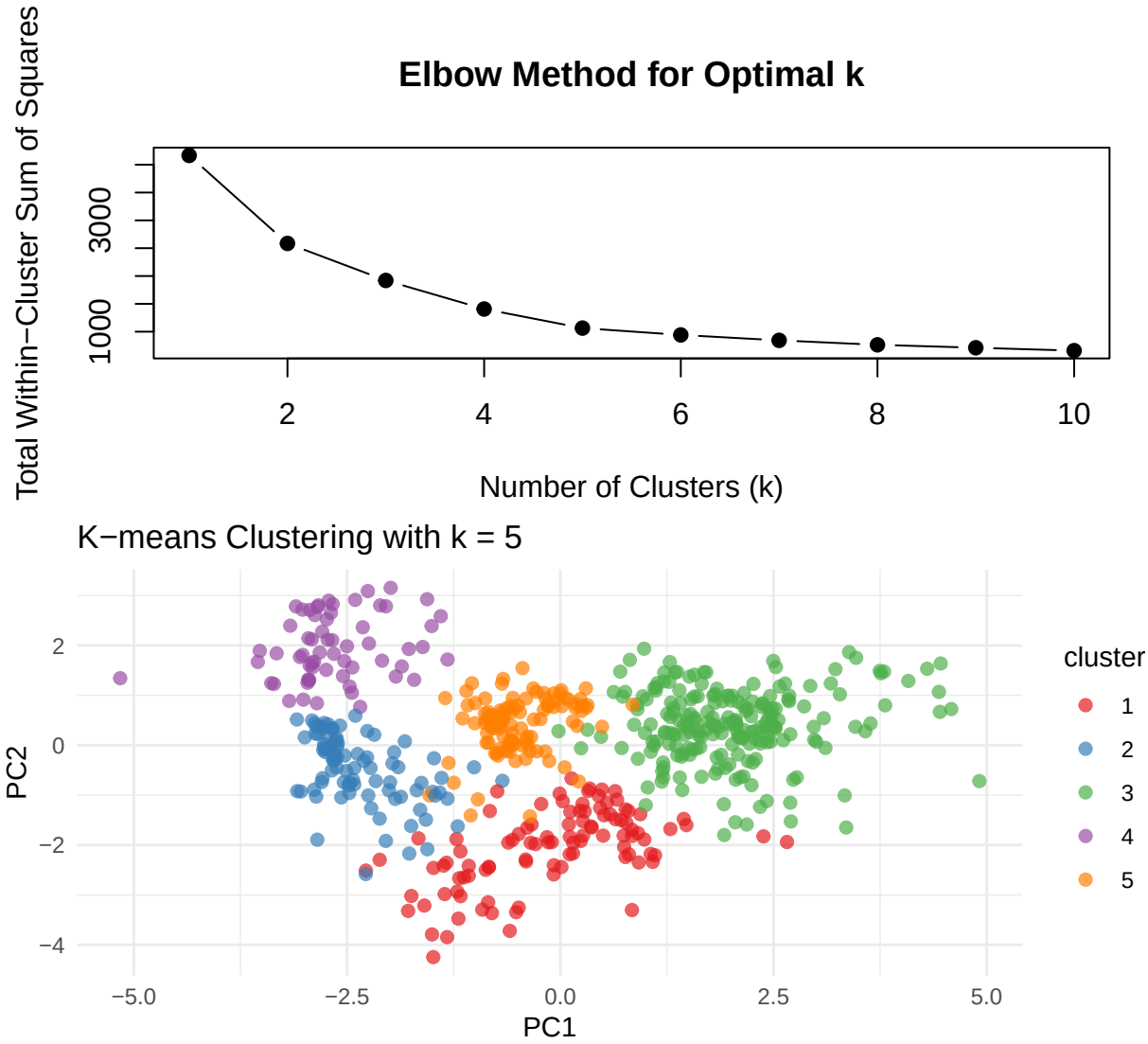
PCA was applied to the standardised data to transform the original variables into orthogonal linear combinations. The first four principal components explained over 90% of the total variance, indicating that most structural information could be captured in a lower-dimensional space. PC1 was dominated by the inverse loading of oleic and linoleic acids, reflecting a biologically meaningful gradient associated with cultivar type and oil quality. PC2 and PC3 captured subtler contrasts among saturated and minor fatty acids, enabling finer subgroup distinctions.

Visual inspection of the PC1–PC2 biplot revealed clustering tendencies, including several distinct sample clouds and transitional regions. This justified the subsequent application of clustering algorithms to formalise group structure.

Clustering Analysis and Method Comparison

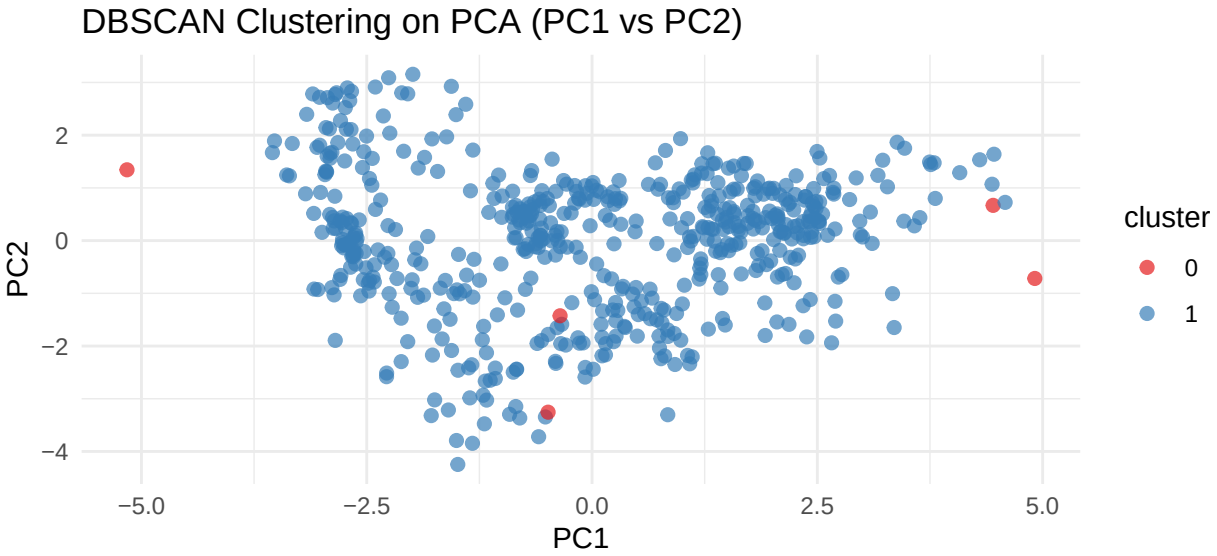
Three clustering algorithms—K-means, hierarchical clustering, and DBSCAN—were applied to the PCA-reduced data. Each method offered distinct assumptions and sensitivity profiles, enabling a comparative evaluation of clustering robustness.

K-means Clustering



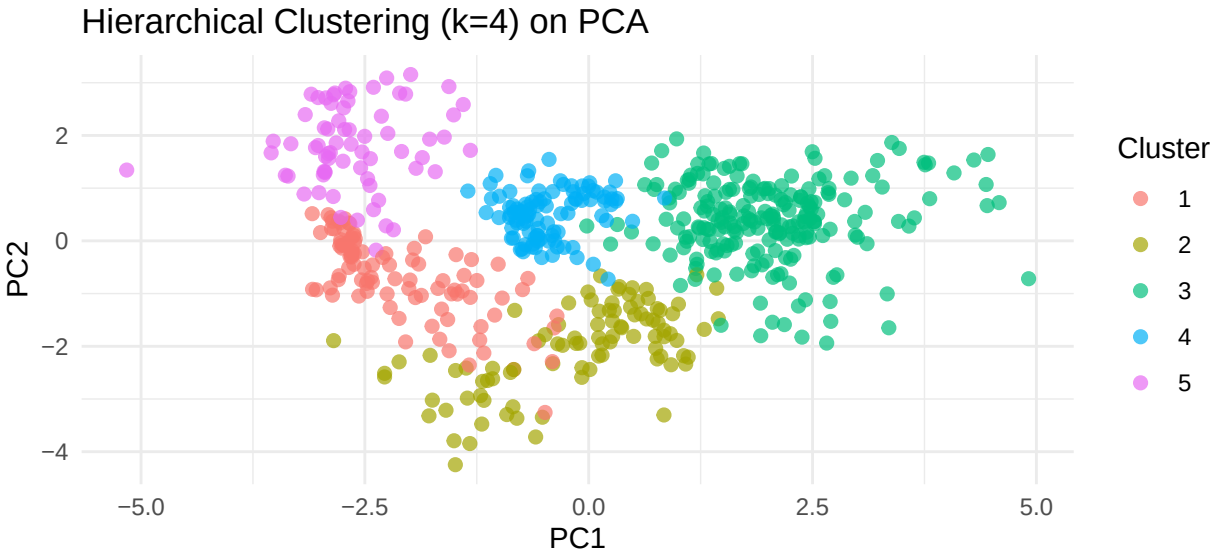
K-means clustering was implemented with k ranging from 2 to 10, and the optimal solution was selected based on silhouette analysis. A five-cluster configuration yielded the highest average silhouette width (~ 0.45), indicating moderate internal cohesion and separation. The clusters were well-separated in the first two PCs, and biochemical profiles revealed distinct signatures. For instance, one cluster exhibited high oleic and low palmitic acid content, a composition associated with premium cultivars from specific Mediterranean regions. Such compositional patterns are consistent with literature on varietal fingerprinting and quality grading.

DBSCAN



DBSCAN, a density-based clustering method, identified two core clusters and several outliers using parameters tuned via k-nearest-neighbour distance plots (eps = 1.2, minPts = 4). However, the resulting average silhouette score was only 0.28, and cluster shapes were less distinct in PC space. The reduced performance of DBSCAN is likely attributable to the isotropic density of PCA-transformed data, which flattens local variations and impairs DBSCAN's density sensitivity. While the method was able to detect a small, chemically distinct subgroup, its overall utility was limited in this setting.

Hierarchical Clustering



Hierarchical clustering using Ward's linkage produced a similar five-cluster solution, with an even higher silhouette score (~ 0.48). The corresponding dendrogram illustrated clear substructures, including nested relationships between certain fatty acid profiles. The alignment of clusters between K-means and hierarchical methods suggests stable latent structure and validates the use of PCA as an effective preprocessing step.

Interpretation and Biochemical Relevance

The clusters identified by K-means and hierarchical methods corresponded to meaningful differences in fatty acid composition. For example, Cluster A exhibited high levels of oleic acid with low polyunsaturates, typical of high-stability oils, while Cluster B showed elevated linoleic acid and reduced monounsaturates, characteristics linked to certain Eastern Mediterranean cultivars. Such groupings are not only statistically robust but also align with agronomic literature describing regional and genetic influences on fatty acid synthesis in olives.

Moreover, some clusters differed primarily in minor acids such as palmitoleic or linolenic, which, while present in smaller amounts, contribute to flavour and oxidation properties. This highlights the potential of clustering to reveal both dominant and subtle biochemical regimes, with applications in provenance verification and nutritional profiling.

Conclusion and Recommendations

The integration of PCA with clustering algorithms provided a powerful framework for uncovering structure in the olive oil dataset. Hierarchical clustering produced the most coherent results, closely followed by K-means, both achieving moderate silhouette scores and generating interpretable biochemical clusters. DBSCAN contributed insights into potential outliers but was less suited for structure discovery in the PCA-reduced space.

Future work should explore non-linear dimensionality reduction techniques such as t-SNE or UMAP to preserve local neighbourhoods and density variations. Additionally, cluster validation using external metadata—such as geographic origin or harvest season—would strengthen the interpretability and practical relevance of the groupings. Finally, incorporating additional chemical markers (e.g., phenolics, sterols) could extend the compositional basis of clustering and enhance its applicability in food science and authentication studies.

References

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357. https://doi.org/10.1613/jair.953

He, H. and Garcia, E.A. (2009) 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp. 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Kuhn, M. and Wickham, H. (2020) *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. Available at: https://www.tidymodels.org (Accessed: 25 July 2025).

Oxley, J., Fildes, B., Ihsen, E. and Charlton, J. (2004) 'Differences in traffic injury risks for older and younger pedestrians', *Accident Analysis & Prevention*, 36(3), pp. 427–432. https://doi.org/10.1016/S0001-4575(03)00035-4

Tibshirani, R., Walther, G. and Hastie, T. (2001) 'Estimating the number of clusters in a data set via the gap statistic', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), pp. 411–423. https://doi.org/10.1111/1467-9868.00293

Martínez, J.J. and Gómez-Caravaca, A.M. (2020) 'Fatty acid profiles in olive oils as a basis for cultivar and geographical origin authentication: A comprehensive review', *TrAC Trends in Analytical Chemistry*, 132, 116049. https://doi.org/10.1016/j.trac.2020.116049

van der Maaten, L. and Hinton, G. (2008) 'Visualizing data using t-SNE', *Journal of Machine Learning Research*, 9(Nov), pp. 2579–2605. Available at: http://www.jmlr.org/papers/volume9/vandermaaten08a/va

ndermaaten08a.pdf

Appendix

Appendix 1 - Random Forest Model Tuning Results

```
## Check the tuning results
rf_tune_results %>%
  collect_metrics() %>%
  arrange(desc(mean)) %>%
  knitr::kable(
    caption = "Hyperparameter Tuning Results for Random Forest Model"
  )
```

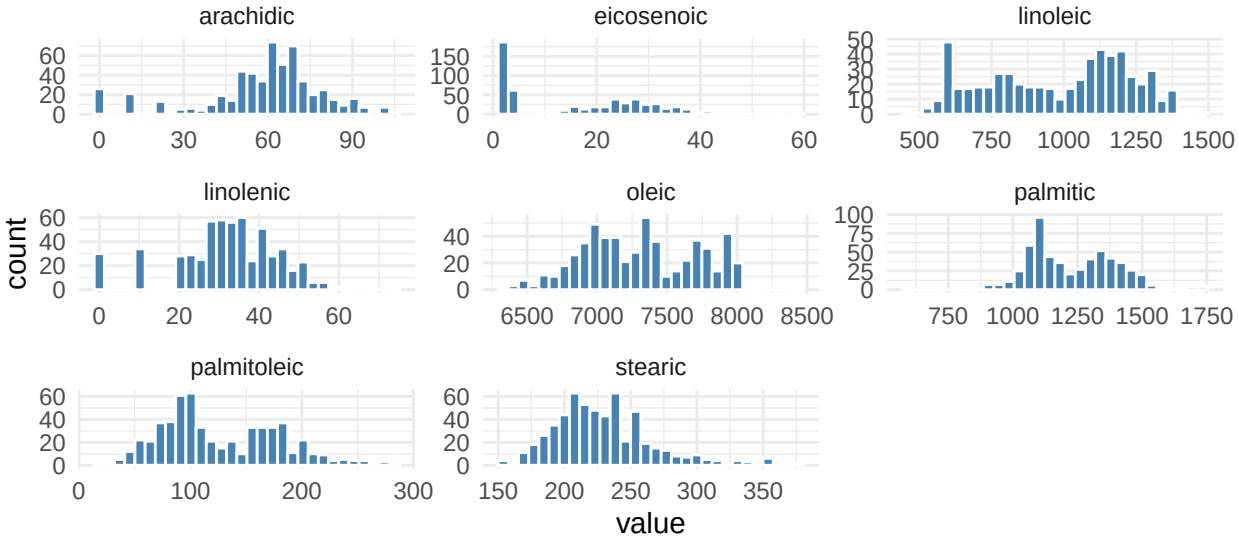
Table 2: Hyperparameter Tuning Results for Random Forest Model

mtry	trees	min_n	.metric	.estimator	mean	n	std_err	.config
5	653	17	roc_auc	hand_till	0.7246908	5	0.0152192	Preprocessor1_Model13
6	937	19	roc_auc	hand_till	0.7243625	5	0.0153239	Preprocessor1_Model10
8	773	13	roc_auc	hand_till	0.7224311	5	0.0151630	Preprocessor1_Model05
7	584	17	roc_auc	hand_till	0.7223609	5	0.0155073	Preprocessor1_Model14
5	387	7	roc_auc	hand_till	0.7220402	5	0.0153084	Preprocessor1_Model20
5	927	5	roc_auc	hand_till	0.7212074	5	0.0152571	Preprocessor1_Model09
7	587	14	roc_auc	hand_till	0.7204049	5	0.0154383	Preprocessor1_Model18
12	812	17	roc_auc	hand_till	0.7199358	5	0.0153118	Preprocessor1_Model03
6	629	3	roc_auc	hand_till	0.7196927	5	0.0146957	Preprocessor1_Model08
10	227	14	roc_auc	hand_till	0.7192911	5	0.0134213	Preprocessor1_Model16
12	708	12	roc_auc	hand_till	0.7192441	5	0.0135445	Preprocessor1_Model17
4	221	13	roc_auc	hand_till	0.7187884	5	0.0149381	Preprocessor1_Model02
10	101	16	roc_auc	hand_till	0.7187075	5	0.0131639	Preprocessor1_Model15
15	735	9	roc_auc	hand_till	0.7185287	5	0.0130804	Preprocessor1_Model06
11	447	13	roc_auc	hand_till	0.7184943	5	0.0147032	Preprocessor1_Model19
15	904	14	roc_auc	hand_till	0.7165673	5	0.0148429	Preprocessor1_Model04
14	612	8	roc_auc	hand_till	0.7161582	5	0.0127688	Preprocessor1_Model11
14	869	6	roc_auc	hand_till	0.7157213	5	0.0138951	Preprocessor1_Model07
14	293	6	roc_auc	hand_till	0.7154463	5	0.0135500	Preprocessor1_Model01
14	352	11	roc_auc	hand_till	0.7152309	5	0.0147441	Preprocessor1_Model12
5	927	5	precision	macro	0.5527440	5	0.0240167	Preprocessor1_Model09
5	927	5	recall	macro	0.5481487	5	0.0223737	Preprocessor1_Model09
5	927	5	f_meas	macro	0.5477823	5	0.0227427	Preprocessor1_Model09
10	101	16	precision	macro	0.5467828	5	0.0161719	Preprocessor1_Model15
7	587	14	recall	macro	0.5463696	5	0.0181017	Preprocessor1_Model18
10	101	16	recall	macro	0.5460512	5	0.0157527	Preprocessor1_Model15
7	587	14	precision	macro	0.5452347	5	0.0197392	Preprocessor1_Model18
10	101	16	f_meas	macro	0.5448481	5	0.0154756	Preprocessor1_Model15
7	587	14	f_meas	macro	0.5445258	5	0.0186243	Preprocessor1_Model18
6	937	19	precision	macro	0.5433284	5	0.0150755	Preprocessor1_Model10
5	927	5	accuracy	multiclass	0.5427952	5	0.0229188	Preprocessor1_Model09
10	101	16	accuracy	multiclass	0.5408505	5	0.0162707	Preprocessor1_Model15
8	773	13	precision	macro	0.5406809	5	0.0142112	Preprocessor1_Model05
5	653	17	precision	macro	0.5405141	5	0.0170607	Preprocessor1_Model13
8	773	13	recall	macro	0.5402300	5	0.0126707	Preprocessor1_Model05

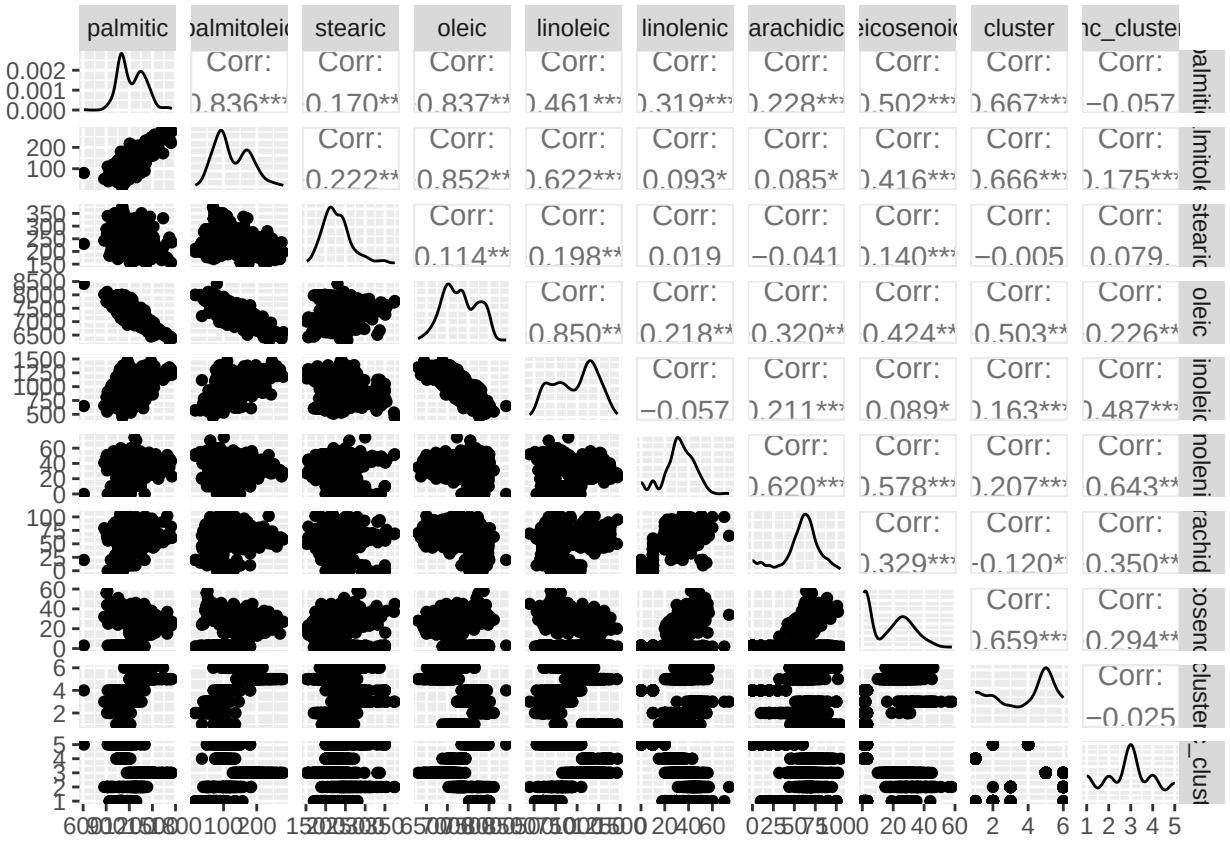
mtry	trees	min_n	.metric	.estimator	mean	n	std_err	.config
6	937	19	recall	macro	0.5402290	5	0.0154312	Preprocessor1_Model10
7	587	14	accuracy	multiclass	0.5400179	5	0.0187390	Preprocessor1_Model18
14	869	6	recall	macro	0.5399964	5	0.0157934	Preprocessor1_Model07
5	653	17	recall	macro	0.5399239	5	0.0149051	Preprocessor1_Model13
14	869	6	precision	macro	0.5398927	5	0.0158099	Preprocessor1_Model07
6	937	19	f_meas	macro	0.5396781	5	0.0148787	Preprocessor1_Model10
6	629	3	precision	macro	0.5392276	5	0.0203741	Preprocessor1_Model08
12	708	12	recall	macro	0.5386565	5	0.0166548	Preprocessor1_Model17
11	447	13	precision	macro	0.5385662	5	0.0183782	Preprocessor1_Model19
8	773	13	f_meas	macro	0.5385305	5	0.0131075	Preprocessor1_Model05
14	869	6	f_meas	macro	0.5383422	5	0.0154790	Preprocessor1_Model07
5	653	17	f_meas	macro	0.5376852	5	0.0157221	Preprocessor1_Model13
4	221	13	precision	macro	0.5375490	5	0.0217871	Preprocessor1_Model02
15	735	9	recall	macro	0.5373642	5	0.0144926	Preprocessor1_Model06
4	221	13	recall	macro	0.5373566	5	0.0184209	Preprocessor1_Model02
12	708	12	precision	macro	0.5372579	5	0.0181351	Preprocessor1_Model17
11	447	13	recall	macro	0.5371368	5	0.0186069	Preprocessor1_Model19
15	735	9	precision	macro	0.5370106	5	0.0152355	Preprocessor1_Model06
15	904	14	precision	macro	0.5368934	5	0.0212880	Preprocessor1_Model04
15	904	14	recall	macro	0.5366748	5	0.0205780	Preprocessor1_Model04
11	447	13	f_meas	macro	0.5364899	5	0.0181372	Preprocessor1_Model19
12	708	12	f_meas	macro	0.5362793	5	0.0170717	Preprocessor1_Model17
15	735	9	f_meas	macro	0.5357803	5	0.0145201	Preprocessor1_Model06
6	937	19	accuracy	multiclass	0.5353052	5	0.0151501	Preprocessor1_Model10
4	221	13	f_meas	macro	0.5352769	5	0.0199633	Preprocessor1_Model02
5	387	7	precision	macro	0.5352029	5	0.0194585	Preprocessor1_Model20
15	904	14	f_meas	macro	0.5350920	5	0.0205012	Preprocessor1_Model04
7	584	17	recall	macro	0.5349201	5	0.0179155	Preprocessor1_Model14
6	629	3	f_meas	macro	0.5346799	5	0.0194465	Preprocessor1_Model08
6	629	3	recall	macro	0.5345791	5	0.0190250	Preprocessor1_Model08
5	653	17	accuracy	multiclass	0.5343354	5	0.0153139	Preprocessor1_Model13
8	773	13	accuracy	multiclass	0.5343352	5	0.0132380	Preprocessor1_Model05
14	869	6	accuracy	multiclass	0.5342733	5	0.0161017	Preprocessor1_Model07
7	584	17	precision	macro	0.5342032	5	0.0182121	Preprocessor1_Model14
12	812	17	recall	macro	0.5338780	5	0.0159130	Preprocessor1_Model03
14	612	8	precision	macro	0.5337061	5	0.0145249	Preprocessor1_Model11
14	612	8	recall	macro	0.5332346	5	0.0132528	Preprocessor1_Model11
7	584	17	f_meas	macro	0.5330360	5	0.0176106	Preprocessor1_Model14
4	221	13	accuracy	multiclass	0.5324223	5	0.0201828	Preprocessor1_Model02
12	708	12	accuracy	multiclass	0.5324174	5	0.0170459	Preprocessor1_Model17
5	387	7	f_meas	macro	0.5319206	5	0.0190283	Preprocessor1_Model20
12	812	17	precision	macro	0.5316861	5	0.0162488	Preprocessor1_Model03
14	612	8	f_meas	macro	0.5316355	5	0.0133420	Preprocessor1_Model11
5	387	7	recall	macro	0.5316076	5	0.0192395	Preprocessor1_Model20
11	447	13	accuracy	multiclass	0.5315183	5	0.0185291	Preprocessor1_Model19
15	735	9	accuracy	multiclass	0.5314608	5	0.0150814	Preprocessor1_Model06
12	812	17	f_meas	macro	0.5306994	5	0.0156630	Preprocessor1_Model03
15	904	14	accuracy	multiclass	0.5305172	5	0.0210389	Preprocessor1_Model04
10	227	14	recall	macro	0.5293805	5	0.0160091	Preprocessor1_Model16
7	584	17	accuracy	multiclass	0.5287146	5	0.0178649	Preprocessor1_Model14
12	812	17	accuracy	multiclass	0.5286660	5	0.0158055	Preprocessor1_Model03
6	629	3	accuracy	multiclass	0.5286528	5	0.0197760	Preprocessor1_Model08

mtry	trees	min_n	.metric	.estimator	mean	n	std_err	.config
10	227	14	precision	macro	0.5278058	5	0.0181414	Preprocessor1_Model16
10	227	14	f_meas	macro	0.5271037	5	0.0170085	Preprocessor1_Model16
5	387	7	accuracy	multiclass	0.5268014	5	0.0201799	Preprocessor1_Model20
14	612	8	accuracy	multiclass	0.5267658	5	0.0144262	Preprocessor1_Model11
14	293	6	precision	macro	0.5238074	5	0.0120739	Preprocessor1_Model01
10	227	14	accuracy	multiclass	0.5230145	5	0.0175115	Preprocessor1_Model16
14	293	6	recall	macro	0.5229011	5	0.0122360	Preprocessor1_Model01
14	293	6	f_meas	macro	0.5218983	5	0.0117567	Preprocessor1_Model01
14	352	11	recall	macro	0.5187039	5	0.0179340	Preprocessor1_Model12
14	352	11	precision	macro	0.5177849	5	0.0177291	Preprocessor1_Model12
14	293	6	accuracy	multiclass	0.5164235	5	0.0127756	Preprocessor1_Model01
14	352	11	f_meas	macro	0.5162799	5	0.0171824	Preprocessor1_Model12
14	352	11	accuracy	multiclass	0.5126546	5	0.0185678	Preprocessor1_Model12

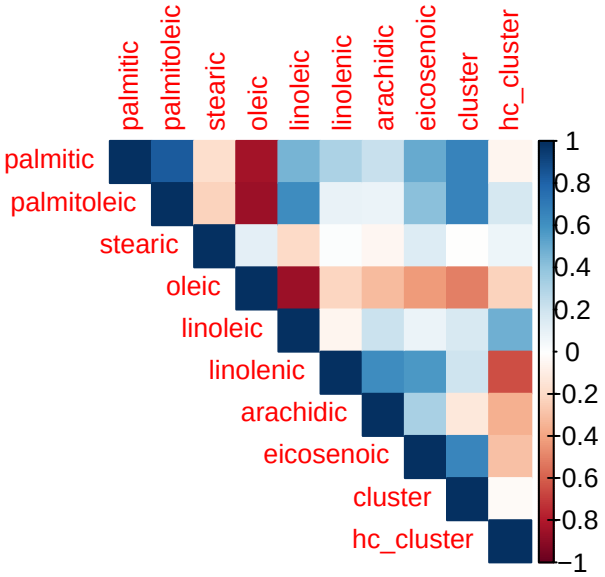
Appendix 3-1 - Histogram of Fatty Acids



Appendix 3-2 - Boxplot of Fatty Acids



Appendix 3-3 - Correlation Matrix



Appendix 3-4 - Boxplot of Outliers