Unofficial summary report of Banana Data Analysis using ML models

Dr. Subaru Muroi

IDEA Bio

Contents

[Summary 4](#_Toc202794228)

[Banana Ripeness Classification Using PLS-DA 6](#_Toc202794229)

[Overview 6](#_Toc202794230)

[Workflow Summary 6](#_Toc202794231)

[1. Data Preparation 6](#_Toc202794232)

[2. Label Encoding 6](#_Toc202794233)

[3. Model Evaluation 6](#_Toc202794234)

[4. Final Model 7](#_Toc202794235)

[5. Sanity Check 7](#_Toc202794236)

[6. Visualisation 7](#_Toc202794237)

[Key Outputs & Interpretation 7](#_Toc202794238)

[Cross-Validated Accuracy 7](#_Toc202794239)

[Classification Report 7](#_Toc202794240)

[Sanity Check (Shuffled Labels) 8](#_Toc202794241)

[PLS-DA Score Plot 8](#_Toc202794242)

[VIP Scores 8](#_Toc202794243)

[Conclusion 9](#_Toc202794244)

[Random Forest Classification of Banana Ripeness Using Aroma Compounds 10](#_Toc202794245)

[Overview 10](#_Toc202794246)

[Workflow Summary 10](#_Toc202794247)

[1. Data Preparation 10](#_Toc202794248)

[2. Feature Selection 10](#_Toc202794249)

[3. Model Training 10](#_Toc202794250)

[4. Feature Importance 11](#_Toc202794251)

[5. Distribution Analysis 11](#_Toc202794252)

[6. SHAP Interpretation 11](#_Toc202794253)

[7. Permutation Importance 11](#_Toc202794254)

[8. Dimensionality Reduction 11](#_Toc202794255)

[Key Outputs & Interpretation 11](#_Toc202794256)

[Cross-Validated Accuracy 11](#_Toc202794257)

[Feature Importance 12](#_Toc202794258)

[Boxplots 12](#_Toc202794259)

[Permutation Importance 16](#_Toc202794260)

[t-SNE & UMAP Plots 16](#_Toc202794261)

[Conclusion 18](#_Toc202794262)

[Ordinal Regression of Banana Ripeness Using Aroma Compounds 19](#_Toc202794263)

[Overview 19](#_Toc202794264)

[Workflow Summary 19](#_Toc202794265)

[1. Data Preparation 19](#_Toc202794266)

[2. Target Encoding 19](#_Toc202794267)

[3. Model Training 19](#_Toc202794268)

[4. Feature Importance 20](#_Toc202794269)

[5. Trend Analysis 20](#_Toc202794270)

[Key Outputs & Interpretation 20](#_Toc202794271)

[Feature Coefficients 20](#_Toc202794272)

[Line plots of top aroma compounds(Linear & Log Scale) 21](#_Toc202794273)

[Conclusion 22](#_Toc202794274)

# Summary

The models and pipelines presented in this analysis are exploratory and designed to evaluate the potential of aroma compound data for ripeness classification. While current sample sizes (n=12) are limited and will lead to overfitting, the goal is to establish a robust framework that can be scaled as more data becomes available. These pipelines enable early identification of promising features and latent patterns, which can inform future experimental design and data collection. As sample sizes grow, the same workflows can be retrained and validated to yield generalisable, interpretable models suitable for scientific and operational use.

The original dataset contained 119 features (chemical compounds), however, 35 were dropped. By default, we use 84 features by dropping any columns with NaNs to prevent model’s from interpreting an effect that does not exist.

This report contains 3 models: Partial Least Squares Discriminant Analysis (PLS-DA), Random Forest, and Ordinal regression.

The first 2 were chosen for their robustness and interpretability, which will allow for these model types to be foundational in fast analysis of a given dataset when we have many features, and we’d like to understand which ones are most consequential.

The 3rd model was specifically chosen for this dataset as the class ‘green’, ‘ripe’, and ‘overripe’ are ordinal in nature, making it a useful tool for similarly ordered classes, however, this means it cannot be applied as generally as the first two. Considering the classes were categorical, a one-hot encode method was used to encode data for the PLS-DA model as well, however, this is standard operating procedure when the target variable of interest for pls-da.

4 compounds appear across the 3 analyses indicating their potential importance:

* Ethyl 3-methylbutanoate Area
* 2-Heptanol Area
* (E)-2-Octen-1-ol Area
* 1-Butanol Area

While this list is inclusive by using the top 35 features using VIP (3/4 of these compounds have VIP<1, while the 4th is just above VIP=1). It suggests some interesting interpretation that would not be possible with a standard pls-da alone. (Reminder VIP score less than 1 means the variable is less influential in explaining class across all components).

As the random forest can capture non-linear relations, and ordinal regression monotonic (ordered) trends, the fact that these relatively low VIP scores were high (almost all within top 10) in feature importance for random forest and ordinal regression highlights their ability to capture more complex relations that may be missed by discriminant analysis.

The modelling here suggests that a latent or subset of data could be informative to develop a predictive model that can identify the ripeness of bananas (I.e., predict banana ripeness with only 10-20 out of 119 targeted features) based on t-sne and umap using top 15 features identified via ANOVA, and that the methods in this report could be used to refine the most important features necessary to develop such a model.

On a final note, even if a feature ranks low on any of these analyses, they may still be biologically relevant, integration of domain knowledge is key in interpretation. Though it could be suggested that compounds that are top features across multiple models signal robustness, because multiple compounds may correlate together with the target variable of interest, even advanced models will struggle to tease out these differences and identify which compound is actually biologically meaningful.

Furthermore, with small datasets, predictive models like random forests may latch onto spurious correlations and rank them highly. The models themselves are sensitive to the data so having a greater number of samples, would not only improve stability of rankings, but allow for better splitting (between training and testing sets) to develop a predictive classification model.

TL: DR

    - This practice is useful for hypothesis generation of important compounds relation with an objective outcome, which requires biological and a biochemical interpretation. With only 12 samples and 50+ to 100+ features any ML model will overfit. This just means the model will not generalise (I.e. for prediction/ decision making tools).

    - The purpose of this practice is to develop models with early indicators of useful features and latent patterns to assist in model development, experimental design, and data collection (therefore overfitting is not an issue, but care is required when interpreting).

    - 4 compounds were identified across 3 models, suggesting potentially robust biomarkers. However, further work is required to determine which subset of features can be used in a generalisable model that can robustly predict the ripeness of a banana given this type of data.

# Banana Ripeness Classification Using PLS-DA

## Overview

This notebook applies **\*\*Partial Least Squares Discriminant Analysis (PLS-DA)\*\*** to classify banana samples into three ripeness categories:

**\*\*Green\*\***, **\*\*Ripe\*\***, and **\*\*Overripe\*\***.

The goal is to:

- Evaluate how well biochemical features can distinguish ripeness stages.

- Identify which features are most important for classification.

---

## Workflow Summary

### 1. Data Preparation

   - Loaded imputed dataset

   - Removed columns with missing values (generally whole categories for a given feature could be NaN: rather than assuming they were below detection threshold and using pseudo count, we remove the entire column).

   - Applied log transformation to normalize feature ranges.

   - Standardized features to ensure equal weighting.

### 2. Label Encoding

   - Converted ripeness categories into numerical format using one-hot encoding.

### 3. Model Evaluation

   - Tested PLS-DA with 1 to 6 components using cross-validation.

   - Selected **\*\*2 components\*\*** based on accuracy performance.

### 4. Final Model

   - Trained PLS-DA with 2 components.

   - Evaluated classification accuracy and generated a detailed report.

### 5. Sanity Check

   - Shuffled labels to confirm the model isn’t learning random noise.

### 6. Visualisation

   - Plotted PLS-DA scores to show group separation.

   - Calculated **\*\*VIP (Variable Importance in Projection)\*\*** scores to rank feature importance.

---

## Key Outputs & Interpretation

### Cross-Validated Accuracy

- Accuracy maxes to 100% at 2 components. This was unnecessary for this, because the samples are only 12 and features (even with removal of columns with NaNs) are 84 (total was 119 with no removal). This will almost certainly overfit. This just means the model will not generalise, but for this given dataset I just want to identify important features that could be useful indicators for aromatic compounds.

- However, we keep this section as it is good practice for larger datasets.

### Classification Report

- Shows precision, recall, and F1-score for each ripeness group.

- **\*\*Interpretation\*\***: Again, because this is an overfitted, small sample model, accuracy is 100% across all metrics (only displays 2 component model)

### Sanity Check (Shuffled Labels)

- Accuracy drops significantly when labels are randomized.

- **\*\*Interpretation\*\***: Confirms the model is learning meaningful patterns, not noise.

### PLS-DA Score Plot

- Each point represents a banana sample projected onto two PLS components.

- Colours indicate ripeness groups.

- **\*\*Interpretation\*\***: Clear separation between groups suggests strong discriminatory power (all 84 features used).

A graph with red dots and numbers

AI-generated content may be incorrect.

### VIP Scores

- Features with **\*\*VIP > 1\*\*** are considered important for classification (I arbitrarily chose top 35 compounds to include VIP values <1).

- Top features are visualized.

- **\*\*Interpretation\*\***: These features are key biochemical markers for ripeness.

A graph with a bar and text

AI-generated content may be incorrect.

---

## Conclusion

This analysis demonstrates that PLS-DA can effectively classify banana ripeness based on biochemical data. The model is accurate, interpretable, and highlights the most influential features — making it a valuable tool for quality control or further research.

# Random Forest Classification of Banana Ripeness Using Aroma Compounds

## Overview

This notebook applies a **\*\*Random Forest classifier\*\*** to predict banana ripeness stages — **\*\*Green\*\***, **\*\*Ripe\*\***, and **\*\*Overripe\*\*** — using aroma compound data.

The focus is on identifying the most important compounds contributing to ripeness classification and interpreting their influence using multiple techniques.

---

## Workflow Summary

### 1. Data Preparation

   - Loaded imputed dataset.

   - Dropped columns with missing values.

   - Encoded ripeness categories numerically.

### 2. Feature Selection

   - Applied ANOVA F-test to select the top 15 aroma compounds most associated with ripeness.

   - This step improves interpretability and reduces dimension (and noise by proxy).

### 3. Model Training

   - Trained a **\*\*Random Forest classifier\*\*** with cross-validation.

   - Used **\*\*Repeated Stratified K-Fold\*\*** to ensure robust accuracy estimates.

### 4. Feature Importance

   - Ranked features by tree-based importance.

   - Visualized top compounds contributing to classification.

### 5. Distribution Analysis

   - Plotted boxplots showing how top compounds vary across ripeness stages.

### 6. SHAP Interpretation

   - Used SHAP values to explain how each compound influences predictions for each ripeness class.

### 7. Permutation Importance

   - Validated feature importance by measuring how prediction accuracy changes when feature values are randomly shuffled (not valid: model overfit; kept in place as could be useful robustness check for predictive model dev).

### 8. Dimensionality Reduction

   - Applied **\*\*t-SNE\*\*** and **\*\*UMAP\*\*** to visualize sample clustering based on selected features.

---

## Key Outputs & Interpretation

### Cross-Validated Accuracy

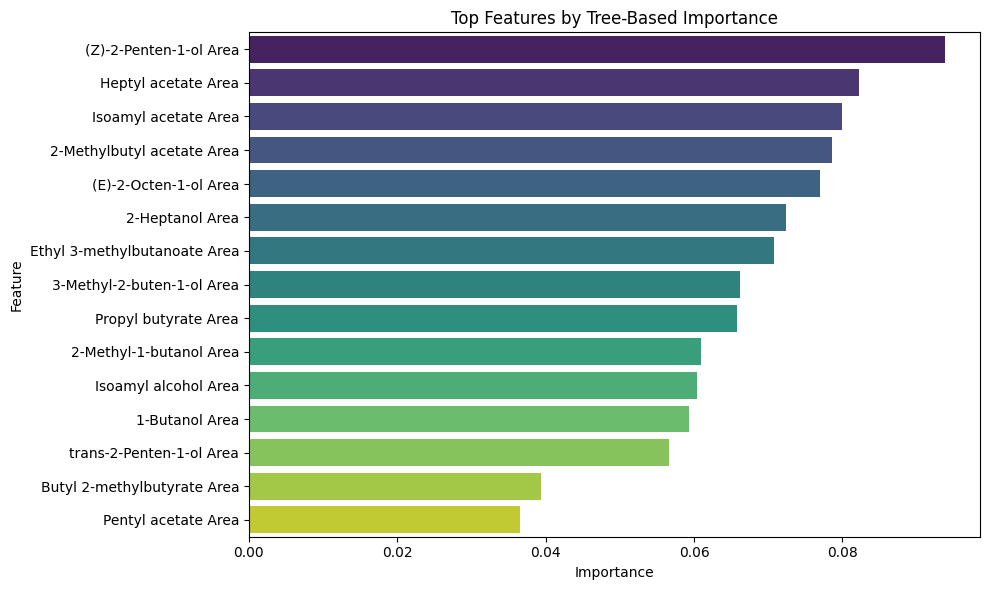
- Shows how well the model predicts ripeness across multiple folds (only using a subset of 15 features).

- **\*\*Interpretation\*\***: High accuracy indicates reliable classification, though generalization should be tested on unseen data (this was again, a given (model accuracy is 100%), considering the overfit model, and can be interpreted as “reliable classification for this (subset of) dataset”). It’s important to note that this is not a valid method for traditional cross-validation to prove model generalisability.

### Feature Importance

- Tree-based and permutation importance highlight the most influential aroma compounds (top-15).

- **\*\*Interpretation\*\***: These compounds are strong candidates for ripeness biomarkers.



### Boxplots

- Show how top 10 compound levels differ across ripeness stages.

- **\*\*Interpretation\*\***: Helps visualize which compounds increase or decrease with ripeness.

A graph with different colored squares

AI-generated content may be incorrect.SHAP Summary Plots

- Explain how each compound contributes to predictions for each ripeness class.

- **\*\*Interpretation\*\***: Offers a transparent view of model decision-making.

A chart of different types of area

AI-generated content may be incorrect.Green (SHAP1)

A chart with text and numbers

AI-generated content may be incorrect.Ripe (SHAP2)

A chart with text and numbers

AI-generated content may be incorrect.Overripe (SHAP3)

### Permutation Importance

- Invalid for this data and model but could be a good robustness check with more samples or reduced dimension model. Kept in code as it will be important in future implementation with more samples (to prove we’re picking robust features that generalise)

### t-SNE & UMAP Plots

- Visualize how samples cluster based on selected features. Top 15 features (using ANOVA) are used here (again, for modelling the random forest, as well as to generate clusters in t-sne and umap), showing that even with a smaller subset, clusters can be generated using various techniques. Clusters were sensitive when I adjusted the number of samples featured (while keeping hyperparameters to their default), and what we show here is that it’s possible to retain differentiable information with a significantly reduced set. This possibly suggests that a latent or subset of data could be informative to develop a predictive model that can identify the ripeness of bananas (I.e., predict banana ripeness with only 10-20 out of 119 targeted features), and that the methods in this report could be used to refine the most important features necessary to develop such a model.

- **\*\*Interpretation\*\***: Clear separation between groups suggests strong discriminatory power (only using top 15 features and default tsne/umap params). In UMAP the clusters are a bit hard to distinguish without labels, however this can be tightened via hyperparameters.

A graph with red blue and green dots

AI-generated content may be incorrect.

A graph with red blue and green dots

AI-generated content may be incorrect.

---

## Conclusion

This analysis demonstrates that Random Forests, combined with feature selection and interpretation tools, can effectively classify banana ripeness and identify key aroma compounds by using a subset of features to predict its ripeness. The results support the use of these compounds as potential indicators for ripeness monitoring or product development.

# Ordinal Regression of Banana Ripeness Using Aroma Compounds

## Overview

This notebook applies **\*\*ordinal logistic regression\*\*** to model banana ripeness stages — **\*\*Green\*\***, **\*\*Ripe\*\***, and **\*\*Overripe\*\*** — based on aroma compound measurements.

Unlike standard classification, ordinal regression respects the **\*\*natural order\*\*** of ripeness stages, making it ideal for this type of progression.

---

## Workflow Summary

### 1. Data Preparation

   - Loaded imputed dataset.

   - Dropped columns with missing values.

   - Log-transformed features to normalize large value ranges.

   - Scaled features using MinMaxScaler for model compatibility.

### 2. Target Encoding

   - Mapped ripeness categories to ordinal values:

     `Green = 0`, `Ripe = 1`, `Overripe = 2`.

### 3. Model Training

   - Fitted an **\*\*ordinal logistic regression model\*\*** using `LogisticAT`.

   - Extracted feature coefficients to identify which compounds most influence ripeness prediction.

### 4. Feature Importance

   - Ranked features by absolute coefficient values.

   - Visualized the top 10 aroma compounds contributing to ripeness classification.

### 5. Trend Analysis

   - Plotted how the top compounds change across ripeness stages:

     - **\*\*Linear scale\*\*** for raw abundance.

     - **\*\*Log scale\*\*** to highlight low-abundance compounds.

---

## Key Outputs & Interpretation

### Feature Coefficients

- Coefficients indicate the **\*\*direction and strength\*\*** of each compound's influence on ripeness.

- **\*\*Positive values\*\*** suggest increasing abundance with ripeness.

- **\*\*Negative values\*\*** suggest decreasing abundance.

- As coefficients range from approx. -0.04 to +0.03, the individual effect is very small, here's an interpretation for the top negative and top positive coefficient and their compound:

“For trans-2-Hexenal Area, the odds of being in a higher ripeness class decrease by about 4% over the scaled range (range in raw (approx. 100,000 ~ 5,000,000)).”

“For Ethyl 3-methylbutanoate Area increases from its minimum to maximum observed value, the odds of being in a higher ripeness class increase by approximately 3% (raw range 3,500 ~ 1,500,000).”

A graph with different colored bars

AI-generated content may be incorrect.

### Line plots of top aroma compounds (Linear & Log Scale)

- Show how compound levels change from Green → Ripe → Overripe.

- **\*\*Interpretation\*\***: These trends help identify biomarkers that consistently increase or decrease with ripeness and are potential candidates for ripeness indicators in quality control or product development.

A graph of different colored lines

AI-generated content may be incorrect. A graph of different colored lines

AI-generated content may be incorrect.

---

## Conclusion

This analysis demonstrates that ordinal regression is a powerful tool for modelling ripeness progression. It identifies key aroma compounds that change predictably across ripeness stages, offering insights into biochemical markers of banana maturity.