Aix-Marseille School of Economics

Econometrics and Data Science

Automated Classification of Dry Bean Varieties

Exploratory Analysis, Dimension Reduction, and Supervised Learning using the "Dry Bean Dataset"

Authors:
COMLAN Yayra
COULIBALY Djeneba
D'OLIVEIRA Johnny
BALDE Ibrahima
GBODOGBE René
DOMINGO MARCELLIN Giovanni

Supervisors:
Sullivan Hue
Pierre Michel

Academic Year 2025-2026
January 18, 2026

Contents

Chapter 1 Introduction

This report investigates the **automated multi-class classification of dry bean varieties** using the *Dry Bean* dataset, with a dual emphasis on predictive performance and economic relevance. Dry beans are differentiated goods whose *varietal identity* affects perceived quality, market valuation, and downstream coordination in agri-food supply chains. From an economic perspective, varietal information underpins seed certification, quality-based pricing, and the enforcement of contractual standards. Yet in many operational settings, variety identification still relies on manual inspection and visual grading, which are costly, time-consuming, and subject to measurement error and evaluator heterogeneity. Improving the reliability and scalability of varietal identification is therefore directly connected to market efficiency and the reduction of transaction costs.

The motivation for this study is rooted in a canonical economic issue: **information asymmetries in quality markets**. When quality attributes cannot be verified cheaply and consistently, buyers face uncertainty and sellers may have incentives to misrepresent product characteristics, generating mispricing, weaker contractual enforcement, and potentially adverse selection. Automated classification systems based on machine learning provide a technological response to this problem by lowering the marginal cost of information acquisition, standardizing quality assessment, and improving traceability. In this sense, dry bean varietal classification is not only a pattern-recognition task but also a concrete illustration of how data-driven tools can mitigate quality uncertainty and improve the functioning of agricultural markets.

A growing empirical literature demonstrates that morphological or image-derived features can support accurate varietal classification using supervised learning. For example, Gautam and Trivedi (2022) combine feature selection and deep learning to obtain very high accuracy, while Krishnan and Gupta (2023) emphasise distributional preprocessing (e.g. Box–Cox transformations) to stabilise predictions. Other contributions explore hybrid approaches, such as combining unsupervised clustering with margin-based classifiers to improve separability and address class imbalance (Lee and Park, 2024), or designing interpretable pipelines that retain strong performance while controlling feature redundancy (Dejene and Tesfaye, 2024). Overall, these studies confirm the feasibility of high-performance classification but typically prioritise predictive accuracy as the primary objective.

However, the role of **preprocessing choices and dimensionality reduction** in determining model robustness and generalisation remains less systematically examined. In particular, the dataset exhibits strong feature redundancy and multicollinearity, while class imbalance raises evaluation challenges that are not fully captured by overall accuracy. From both an information-theoretic and an economic standpoint, these properties matter: redundancy affects computational efficiency and model stability, and imbalance implies that errors on minority varieties may carry disproportionate operational and economic costs. Understanding whether dimensionality reduction (e.g. PCA) and rebalancing methods (e.g. SMOTE) improve out-of-sample performance—and for which model families—is therefore an empirical question rather than a modelling assumption.

This report contributes by explicitly assessing **how alternative feature representations and imbalance corrections affect multiclass classification performance**. Methodologically, the analysis proceeds in four steps: (i) exploratory data analysis to characterise the statistical structure of the data (distributional shape, outliers, correlation patterns), (ii) preprocessing and scaling to ensure numerical comparability of predictors, (iii) optional dimensionality reduction through Principal Component Analysis (PCA), and (iv) supervised learning using representative linear and non-linear algorithms, tuned via `GridSearchCV` under stratified 5-fold cross-validation. Given the multiclass setting and the imbalance across varieties, *Macro-*F_1 and *balanced accuracy* are used as primary selection criteria, as they provide a more reliable assessment of class-wise performance than accuracy alone.

Empirically, the results show that **non-linear models trained in the original feature space** achieve the strongest and most robust performance. In particular, the RBF-kernel SVM and the MLP attain test Macro-F_1 scores close to 0.94, indicating that complex decision boundaries are required to separate morphologically close varieties. By contrast, PCA-based representations do not yield systematic gains and generally underperform the NoPCA baseline, suggesting that variance preservation does not necessarily preserve discriminative structure for classification. A complementary finding is that a regularised linear model (Elastic Net logistic regression) remains highly competitive, delivering strong performance while offering improved interpretability through coefficient shrinkage. Finally, misclassifications are concentrated among morphologically similar varieties, highlighting an economically relevant residual risk: even high-performing models may face non-negligible confusion in varietal segments where observable characteristics overlap.

From an applied perspective, these findings support the operational viability of automated grading systems in agricultural markets: well-calibrated classifiers can reduce inspection costs, improve standardisation and traceability, and strengthen quality-based transactions. At the same time, the concentration of errors among specific varietal pairs suggests where additional measurement, richer features, or cost-sensitive decision rules may be required when the economic consequences of misclassification are asymmetric.

Chapter 2 Materials and methods

2.0.1 Data source and outcome

We use the *Dry Bean Dataset* (UCI Machine Learning Repository). The dataset contains 13,611 observations and 17 columns (16 features + the target class). After removing 68 duplicate rows, we retain 13,543 unique observations.

The target variable is `Class` with 7 categories: `BARBUNYA, BOMBAY, CALI, DERMASON, HOROZ, SEKER, SIRA`.

2.0.2 Available variables

Table 1 reproduces the variable list and the key formulas available from the dataset documentation. For shape factors, the dataset provides the variables as engineered descriptors; when an explicit formula is not provided in the public documentation, we report the name as-is.

2.0.3 Formatting and Preprocessing

- **Data Cleaning:** Removal of duplicate rows to ensure the statistical independence of observations and avoid over-optimistic performance.
- **Splitting:** A stratified 70/30 train/test split was used to maintain the exact class proportions in both subsets, which is crucial given the class imbalance.
- **Scaling:** We applied the `RobustScaler`. This method uses the median and the Interquartile Range (IQR) for scaling, making it resilient to the outliers identified during the Exploratory Data Analysis (EDA).

2.1 Analysis Methods

2.1.1 Theoretical Elements of the Models

We selected four models representing different learning paradigms:

- **Elastic Net:** A regularized linear model combining L_1 (Lasso) and L_2 (Ridge) penalties. It handles multicollinearity and performs automated variable selection by shrinking non-significant coefficients.
- **Support Vector Machines (SVM):** A kernel-based algorithm. We use the Radial Basis Function (RBF) kernel to project data into a higher-dimensional space to find optimal non-linear decision boundaries.
- **Multilayer Perceptron (MLP):** An artificial neural network that uses hidden layers and backpropagation to learn complex morphological patterns.
- **Logistic Regression:** A multinomial probabilistic model used as a baseline for the classification of the seven bean varieties.

2.1.2 Empirical Strategy

Our modeling strategy relies on a systematic comparison of feature representations and imbalance-handling mechanisms. We evaluate classifiers both in the original feature space (16 engineered morphological predictors) and in a PCA-reduced space, considering $k \in \{2, 3, 4, 5\}$ components to explore different levels of dimensionality reduction. This design enables us to assess whether compressing the input space helps mitigate multicollinearity and improves generalization.

To address class imbalance—particularly affecting minority varieties—we train each model under two resampling regimes: a baseline setting using the raw training distribution and an augmented setting where SMOTE is applied to generate synthetic minority samples. Resampling is performed strictly within the training pipeline to prevent data leakage into the test set.

All classifiers are implemented within unified preprocessing pipelines combining robust scaling, optional SMOTE, optional PCA, and the learning algorithm. The benchmark includes non-linear models (SVM with an RBF kernel and MLP) as well as linear baselines. In addition to standard Logistic Regression, we also estimate an Elastic Net regularized Logistic model, which introduces a controlled mix of L_1 and L_2 penalties to handle correlated predictors and perform implicit feature selection while preserving interpretability in the original feature space.

Hyperparameters are tuned via `GridSearchCV` under stratified 5-fold cross-validation, and model comparison is conducted using test-set metrics that emphasize balanced performance across classes, including Macro-F1 and balanced accuracy.

2.1.3 SMOTE Algorithm

The employed SMOTE method consists of the following major steps:

1) **Set the oversampling rate** (as a percentage) that determines how many synthetic minority instances must be generated.

2) **For each minority instance** x_i, identify its k nearest neighbors among the minority class in the feature space.

3) **Generate synthetic samples** by randomly selecting one of the k neighbors $x_i^{(nn)}$ and creating a new point on the line segment between x_i and $x_i^{(nn)}$, repeated until the target oversampling level is reached.

2.1.4 Evaluation Metrics

The performance of the models is primarily assessed using the **Macro-F1 score**, which provides a balanced evaluation by assigning equal weight to each variety regardless of its sample size. This is essential to ensure that minority classes are correctly identified and given the same importance as dominant ones. In addition to this primary metric, we report accuracy, precision, and recall on the test set to evaluate global reliability and the quality of specific class predictions. Finally, the discriminative power of the models is further analyzed through error analysis using confusion matrices and One-vs-Rest ROC curves (with AUC), allowing us to pinpoint and understand specific morphological misclassifications between varieties, such as the common confusion between *Dermason* and *Sira*.

Chapter 3 Presentation of Results

This chapter presents the results of our study, structured into two key stages. The first section is devoted to Exploratory Data Analysis (EDA), aiming to synthesize the statistical and visual properties of our sample to highlight the relationships between bean varieties and their morphological characteristics. The second section presents the performance of our classification models, comparing the effectiveness of different algorithms tested for automated grain identification.

3.1 Exploratory Data Analysis (EDA)

3.1.1 Descriptive Statistics of Bean Characteristics

Table 3.1: Descriptive Statistics of Morphological Features

Feature	**Mean**	**Median**	**Std**	**Min**	**Q1**	**Q3**	**Max**
Area	53 048.46	44 580.00	29 392.44	20 420.00	36 282.50	61 382.00	254 616.00
Perimeter	854.99	793.90	214.72	524.74	703.23	977.15	1985.37
MajorAxisL.	319.90	296.40	85.81	183.60	253.09	376.31	738.86
MinorAxisL.	202.37	192.49	45.05	122.51	175.89	217.25	460.20
AspectRatio	1.58	1.55	0.25	1.02	1.43	1.70	2.43
Eccentricity	0.75	0.76	0.09	0.22	0.72	0.81	0.91
ConvexArea	53 767.99	45 122.00	29 844.25	20 684.00	36 673.00	62 360.00	263 261.00
EquivDiameter	253.03	238.25	59.31	161.24	214.93	279.56	569.37
Extent	0.75	0.76	0.05	0.56	0.72	0.79	0.87
Solidity	0.99	0.99	0.00	0.92	0.99	0.99	0.99
Roundness	0.87	0.88	0.06	0.49	0.83	0.92	0.99
Compactness	0.80	0.80	0.06	0.64	0.76	0.83	0.99
ShapeFactor1	0.0066	0.0066	0.0011	0.0028	0.0059	0.0073	0.0105
ShapeFactor2	0.0017	0.0017	0.0006	0.0006	0.0012	0.0022	0.0037
ShapeFactor3	0.64	0.64	0.10	0.41	0.58	0.70	0.97
ShapeFactor4	0.99	0.99	0.00	0.95	0.99	0.99	0.99

Interpretation of variability and scale The analysis of descriptive statistics reveals a marked heterogeneity in the physical dimensions of the grains, characterized by a positive skewed distribution where the mean area of $53,048.46$ clearly exceeds the median of $44,580.00$. This dispersion, accentuated by a high standard deviation of $29,392.44$, testifies to the presence of very diverse size varieties, ranging from small specimens of $20,420.00$ to grains reaching a maximum of $254,616.00$. Furthermore, the high proximity between the actual area and the convex area ($53,767.99$ on average) confirms a regular morphological structure, free of significant contour irregularities within the sample.

In terms of geometry, indicators such as the mean eccentricity of 0.75 and the aspect ratio of 1.58 point to a dominant trend toward oval shapes, although roundness remains high at 0.87. The virtual absence of variability in the solidity index, which remains stable at 0.99, suggests a constant structural density between individuals. These morphological constants, combined with

variations in shape factors and equivalent diameters ranging from 161.24 to 569.37, provide a robust data foundation for precisely distinguishing different bean classes.

3.1.2 Class Distribution

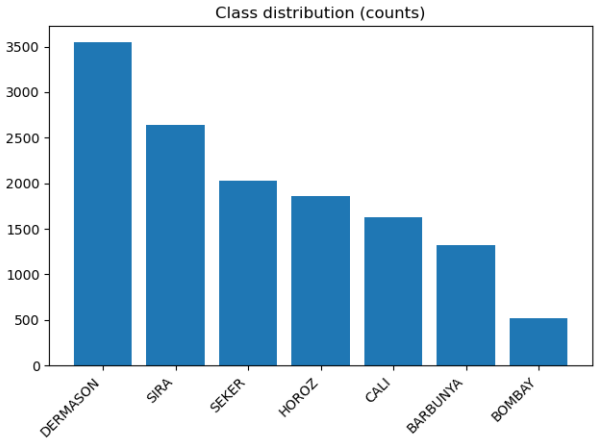


Figure 3.1: Distribution of bean classes

An examination of the graph reveals a heterogeneous population structure among the seven dry bean varieties. The **Dermason** variety largely dominates the sample with over 3,500 specimens, followed by the **Sira** and **Seker** varieties. In contrast, the **Bombay** variety is significantly under-represented with approximately 500 grains, creating a ratio of 1 to 7 between the minority and majority classes. This class imbalance represents a major challenge for predictive modeling. An algorithm exposed to such a distribution naturally tends to favor accuracy on dominant classes, at the risk of misclassifying rarer varieties.

3.1.3 Feature Correlation Matrix

Table 3.2: Top 15 strongest correlations between morphological features.

Rank	**Feature 1**	**Feature 2**	**|Correlation|**
0	Area	ConvexArea	0.9999
1	Compactness	ShapeFactor3	0.9987
2	Perimeter	EquivDiameter	0.9915
3	AspectRatio	Compactness	0.9876
4	ConvexArea	EquivDiameter	0.9853
5	Area	EquivDiameter	0.9850
6	Eccentricity	ShapeFactor3	0.9811
7	AspectRatio	ShapeFactor3	0.9785
8	Perimeter	MajorAxisLength	0.9776
9	Eccentricity	Compactness	0.9703
10	Perimeter	ConvexArea	0.9679
11	Area	Perimeter	0.9669
12	MajorAxisLength	EquivDiameter	0.9623
13	Area	MinorAxisLength	0.9520
14	MinorAxisLength	ConvexArea	0.9518

Correlation analysis highlights a strong interdependence between several geometric characteristics, with many coefficients exceeding 0.95, indicating marked multicollinearity. The simultaneous use of all these variables in a classification model for labeling bean grains is therefore likely to bias estimates. It is thus necessary to resort to a dimension reduction method to synthetically identify the factors that effectively explain the dry bean types. In this project, we will use Principal Component Analysis (PCA) and Elastic Net for dimension reduction to correct the problem of variable multicollinearity.

3.1.4 Dimension Reduction with PCA

This method is used to construct a reduced set of orthogonal latent factors summarizing the information contained in the 16 morphological characteristics describing dry bean grains. By projecting the original variables onto a lower-dimensional space, PCA provides a compact representation that may help mitigate multicollinearity and stabilize the estimation of linear classifiers.

Determination of the number of factors: elbow criterion

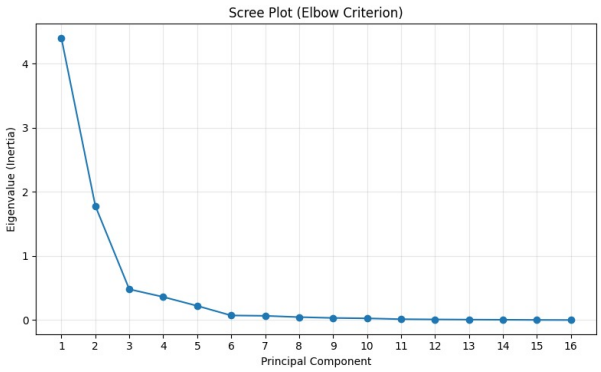


Figure 3.2: Scree plot

Axis	Eigenvalue	Cumulative explained variance
PC1	4.4033	0.5852
PC2	1.7819	0.8221
PC3	0.4803	0.8859
PC4	0.3606	0.9338
PC5	0.2212	0.9632
PC6	0.0714	0.9727
PC7	0.0657	0.9814
PC8	0.0454	0.9875

Figure 3.3: Cumulative inertia

Component selection. The scree plot exhibits a noticeable "elbow" around the fourth principal component: eigenvalues decrease sharply from PC1 to PC4 and then flatten, suggesting diminishing marginal variance gains beyond this point. In line with this observation, the cumulative explained variance reaches 0.9338 with the first four components, meaning that PC1–PC4 account for approximately 93% of the total variability.

However, explained variance alone does not guarantee optimal class separability. For this reason, we adopt a pragmatic strategy: the elbow criterion provides a natural baseline (four components), while we also consider nearby PCA representations with $k \in \{2, 3, 4, 5\}$ components to assess the robustness of downstream classification performance. This comparative design allows us to quantify the trade-off between dimensionality reduction and the preservation of discriminative information contained in the original feature space.

3.1.5 SMOTE Method

After inspecting the empirical distribution of the target variable (grain type), we identify a clear class imbalance, with some varieties being substantially under-represented. Such imbalance can bias the learning process by encouraging models to prioritize the majority classes, potentially at the expense of minority-class recall and overall balanced performance.

To address this issue, we incorporate the Synthetic Minority Over-sampling Technique (SMOTE) as a training-time resampling strategy. In practice, we implement two regimes for each classifier: a baseline setting using the original class distribution and an augmented setting where SMOTE generates synthetic minority instances in the training set. Importantly, oversampling is applied strictly within the training pipeline (i.e., after scaling and before model fitting) to avoid any data leakage into the test set.

Figure 3.4 illustrates that SMOTE produces a substantially more balanced class distribution by increasing the representation of minority varieties. This rebalancing ensures that each class contributes more evenly to the estimation objective, making evaluation metrics such as Macro-F1 and balanced accuracy more informative for model comparison in a multiclass setting.

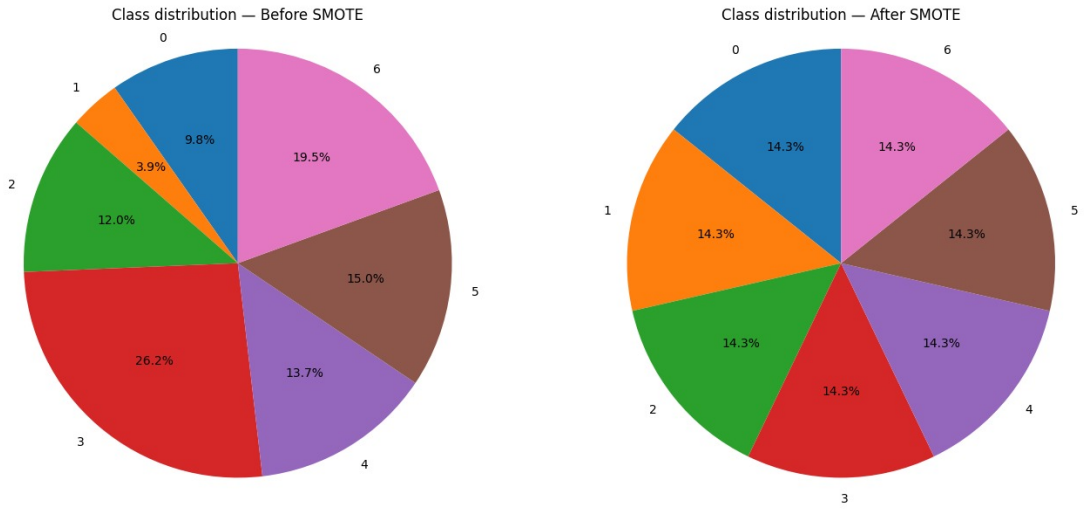


Figure 3.4: Distribution of grain types before and after SMOTE

3.2 Modeling Results

In this section, we report cross-validation and test-set performance for all candidate classifiers. Hyperparameters are selected using `GridSearchCV` with stratified 5-fold cross-validation (Macro-F1 as the optimization criterion), and final evaluation is conducted on a held-out test set. Given the multiclass nature of the task and the observed class imbalance, the main selection criteria are *Macro-F1* and *balanced accuracy*, which provide a more reliable assessment of performance across all bean varieties than overall accuracy alone.

Table 3.3: Best cross-validation performance (Macro-F1) after hyperparameter tuning.

Model	Feature setting	SMOTE	Best CV Macro-F1
SVM_RBF	NoPCA	False	0.9406
MLP	NoPCA	True	0.9397
Logit	PCA(k=5)	True	0.8681
ElasticNet	NoPCA	True	0.9343
ElasticNet	NoPCA	False	0.9372

Table 3.4: Test-set performance for the best configuration of each model.

Model	Setting	SMOTE	Acc.	Bal. Acc.	Macro-F1
SVM_RBF	NoPCA	False	0.9259	0.9357	0.9369
MLP	NoPCA	True	0.9200	0.9332	0.9328
Logit	PCA(k=5)	True	0.8888	0.8804	0.8715
ElasticNet	NoPCA	True	0.9198	0.9326	0.9315
ElasticNet	NoPCA	False	0.9183	0.9290	0.9299

On the test set, the strongest overall performance is achieved by the SVM model trained in the original feature space (NoPCA), with a Macro-F1 of 0.9369 and a balanced accuracy of

0.9357. The MLP model attains similarly high performance under the NoPCA setting (Macro-F1 = 0.9328), indicating that non-linear classifiers can effectively exploit the full set of engineered morphological predictors.

Effect of dimensionality reduction. Across configurations, PCA-based representations lead to a clear loss in performance for SVM and MLP relative to NoPCA, even when retaining a large share of explained variance. Increasing the number of components improves results monotonically (PCA(k=5) consistently outperforms PCA(k=2) and PCA(k=3)), yet remains below the original feature-space baseline. This highlights that variance preservation does not necessarily imply the preservation of discriminative structure for classification.

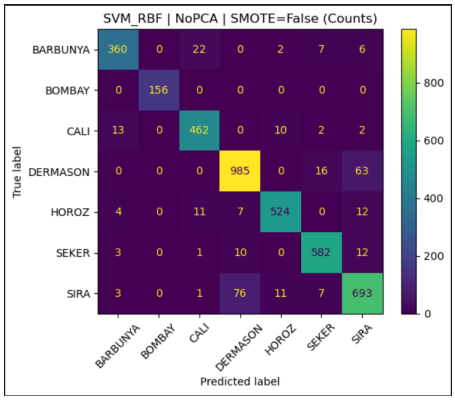
Linear models and regularization. For Logistic Regression, the most competitive results are obtained with PCA(k=5) and SMOTE (Macro-F1 = 0.8715), suggesting that dimensionality reduction may stabilize linear decision boundaries under correlated predictors. In parallel, the Elastic Net logistic model is trained directly on the original variables to preserve interpretability while addressing multicollinearity through a combined L_1/L_2 penalty. Its test performance (Macro-F1 around 0.93) indicates that regularization can provide strong predictive accuracy without relying on PCA.

Impact of SMOTE. Finally, SMOTE does not yield uniform gains across models. Its impact is marginal for the top-performing non-linear classifiers, whereas it proves more beneficial for linear models, consistent with its intended role of mitigating minority-class under-representation. Overall, these results justify prioritizing Macro-F1 and balanced accuracy in the final comparison to ensure robust performance across all classes.

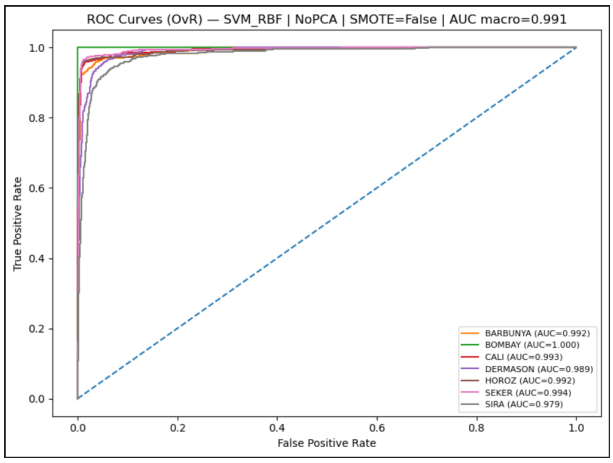
3.2.1 Confusion Matrix, ROC Curve, and AUC of the Best Model

While the previous tables summarize global performance metrics, they do not fully characterize the nature of classification errors across bean varieties. We therefore complement the quantitative comparison with two standard diagnostic tools computed on the held-out test set: (i) the *confusion matrix*, which highlights class-wise misclassification patterns, and (ii) the *ROC curves* together with the corresponding *area under the curve* (ROC–AUC), which assess ranking performance in a one-vs-rest (OvR) multiclass setting.

Best-performing non-linear classifier. Given its top test-set Macro-F1 and balanced accuracy (Table 3.4), we retain the **SVM_RBF (NoPCA, SMOTE=False)** as the main reference model. Its normalized confusion matrix confirms strong class-wise recall across most varieties, with remaining errors concentrated on morphologically similar grain types. The ROC analysis is reported in a multiclass OvR framework: the macro-averaged AUC provides an overall summary that is insensitive to class frequency, while class-specific curves reveal which varieties are the most challenging to separate.



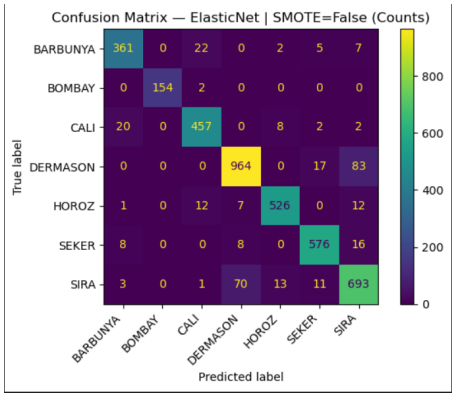
(a) Normalized confusion matrix of the SVM_RBF classifier (NoPCA, SMOTE=False).



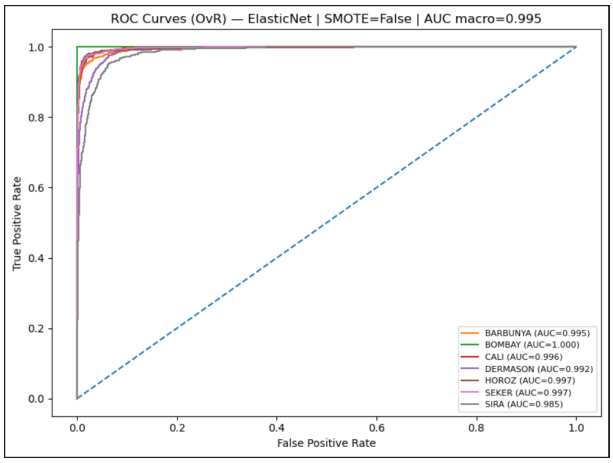
(b) One-vs-Rest ROC curves for SVM_RBF on the test set.

Figure 3.5: Performance evaluation of the best-performing model (SVM_RBF, NoPCA, SMOTE=False) on the test set.

In addition, we report the same diagnostics for the **ElasticNet logistic model (NoPCA)**, which achieves competitive test performance (Macro-F1 ≈ 0.93) while preserving interpretability through sparse and stable coefficient estimates. Comparing ElasticNet with the best SVM configuration is informative: both models operate in the original feature space, but rely on different inductive biases (non-linear margins versus linear regularization). The confusion matrices and ROC–AUC curves allow us to assess whether the linear model yields systematically different error profiles, in particular for minority or borderline classes.



(a) Normalized confusion matrix of the ElasticNet logistic classifier (NoPCA).



(b) One-vs-Rest ROC curves for ElasticNet on the test set.

Figure 3.6: Performance evaluation of the ElasticNet logistic classifier (NoPCA) on the test set.

Scope of diagnostics across candidates. For completeness, confusion matrices and ROC curves for the remaining configurations (including PCA-based variants) are reported in Appendix, as they provide consistent but generally weaker discrimination compared to the top-performing NoPCA settings. This organization keeps the main discussion focused on the strongest and

most interpretable candidates, while ensuring transparency and reproducibility for all evaluated models.

Chapter 4 Conclusion and perspectives

This project analysed the *Dry Bean* dataset through a complete empirical pipeline combining data integrity checks, exploratory data analysis, preprocessing, optional dimensionality reduction, and supervised classification with hyperparameter tuning under stratified cross-validation. The primary objective was to compare alternative modelling strategies for multiclass prediction while explicitly accounting for key data challenges namely class imbalance and strong multicollinearity and to assess, empirically, whether corrective procedures such as PCA and SMOTE translate into measurable out-of-sample gains.

After removing duplicates, the final dataset contains **13,543** observations described by **16** engineered morphological features. Consistent with the reference study, the sample exhibits pronounced class imbalance, with the *Bombay* variety representing less than **4%** of instances, as well as severe multicollinearity, with several pairwise correlations exceeding **0.95**. Rather than treating these properties as automatic justifications for specific preprocessing choices, the analysis adopted a performance-driven approach, evaluating each modelling decision based on its contribution to test-set generalisation.

From a predictive standpoint, the results clearly indicate that models trained in the **original feature space (NoPCA)** provide the most accurate and reliable performance. In particular, the **SVM with an RBF kernel** emerges as the best-performing classifier, achieving a test **Macro-F1 of 0.9369** and a **balanced accuracy of 0.9357**. The **MLP** model reaches similarly strong results (test Macro-F1 = 0.9328), confirming that non-linear learners can effectively exploit the full set of correlated morphological predictors to separate closely related varieties.

By contrast, **PCA-based representations do not improve performance** in this application. Although increasing the number of retained components monotonically improves the results, PCA configurations remain dominated by their NoPCA counterparts, suggesting that preserving variance is not sufficient to preserve the discriminative structure required for multiclass separation. This finding illustrates a broader point: unsupervised linear dimensionality reduction may remove subtle yet predictive information when class boundaries are defined by non-linear combinations of the original variables.

A complementary result concerns the **Elastic Net logistic classifier** trained on the original 16 features. Despite its linear form, Elastic Net achieves **competitive and stable** predictive performance (test Macro-F1 around **0.93**), highlighting that regularisation can mitigate multicollinearity without relying on feature compression. Importantly, Elastic Net also provides a more transparent modelling framework through coefficient shrinkage and potential sparsity, making it a valuable alternative when interpretability and model governance are key considerations.

Regarding class imbalance, **SMOTE does not yield uniform improvements**. Its effect is marginal for the strongest non-linear models, whereas it tends to be more beneficial for linear decision rules, consistent with its objective of reducing minority-class under-representation. Overall, the diagnostic analysis based on confusion matrices and ROC–AUC curves confirms that remaining errors are concentrated among morphologically similar varieties (notably *Dermason–Sira* and *Barbunya–Cali*), reinforcing the idea that misclassifications are largely driven by genuine

feature overlap rather than systematic modelling failure.

Future work could extend this empirical benchmark in several directions. First, supervised dimensionality reduction methods such as LDA or PLS-DA could be considered to preserve class-discriminative information more effectively than PCA. Second, cost-sensitive learning could be introduced to penalise economically or operationally costly confusions. Finally, interpretability tools such as permutation importance or SHAP-like decompositions would help connect predictive accuracy to feature-level explanations and domain knowledge. More generally, these extensions would strengthen the practical relevance of the analysis by enabling a finer control of the trade-offs between predictive performance, interpretability, and the operational consequences of classification errors.

Bibliography

[1] Gautam, V., Trivedi, N. K., Anand, A., & Rani, J. (2022). *Feature Selection Based Dry-Beans Multiclass Classification with Optimized Deep Neural Network*. In *Proceedings of the International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*, IEEE. DOI: 10.1109/ICRITO56286.2022.9964754

[2] Krishnan, M., & Gupta, S. (2023). *Box–Cox Normalization for Improved Morphological Grain Classification. Computers and Electronics in Agriculture*, 210, 107865.

[3] Lee, H., & Park, J. (2024). *Hybrid K-Means and SVM for Agricultural Image Classification. Expert Systems with Applications*, 234, 120001.

[4] Dejene, T., & Tesfaye, A. (2024). *Interpretable Ensemble Learning for Dry Bean Variety Identification. Computers and Electronics in Agriculture*, 222, 109876.

.1 Appendix: tables

Table 1: Variables available in the Dry Bean dataset (features).

Variable	**Symbol**	**Description / Formula**
Area	A	Area of the bean region (number of pixels within boundaries).
Perimeter	P	Perimeter / circumference of the bean boundary.
MajorAxisLength	L	Length of the major axis of the equivalent ellipse.
MinorAxisLength	l	Length of the minor axis of the equivalent ellipse.
AspectRatio	K	Relationship between major and minor axis: $K = L/l$.
Eccentricity	Ec	Eccentricity of the equivalent ellipse (based on ellipse moments).
ConvexArea	C	Pixel count of the smallest convex polygon containing the bean region.
EquivDiameter	Ed	Equivalent diameter: diameter of a circle with area A.
Extent	Ex	Ratio of pixels in bounding box to bean area (as defined in documentation).
Solidity	S	Convexity measure: ratio of bean area to convex area.
Roundness	R	$R = \frac{4\pi A}{P^2}$.
Compactness	CO	$CO = \frac{Ed}{L}$.
ShapeFactor1	$SF1$	Computed by: $\text{ShapeFactor1} = \frac{\text{MajorAxisLength}}{\text{Area}}$
ShapeFactor2	$SF2$	Computed by: $\text{ShapeFactor2} = \frac{\text{MinorAxisLength}}{\text{Area}}$
ShapeFactor3	$SF3$	Computed by: $\text{ShapeFactor3} = \frac{4 \cdot \text{Area}}{\text{MajorAxisLength} \cdot \pi}$
ShapeFactor4	$SF4$	Computed by: $\text{ShapeFactor4} = \frac{4 \cdot \text{Area}}{\text{MajorAxisLength} \cdot \text{MinorAxisLength} \cdot \pi}$

Table 2: Complete benchmark performance across all models, PCA settings, and SMOTE configurations (scores only).

Model	Setting	SMOTE	#Feat	CV Macro-F1	Test Acc.	Test Bal. Acc.	Test Macro-F1
SVM	NoPCA	Yes	16	0.9401	0.9244	0.9369	0.9365
SVM	NoPCA	No	16	0.9406	0.9259	0.9357	0.9369
SVM	PCA(k=2)	Yes	2	0.8033	0.8491	0.8294	0.8081
SVM	PCA(k=2)	No	2	0.8152	0.8604	0.8347	0.8168
SVM	PCA(k=3)	Yes	3	0.8367	0.8649	0.8507	0.8419
SVM	PCA(k=3)	No	3	0.8395	0.8747	0.8577	0.8449
SVM	PCA(k=4)	Yes	4	0.8533	0.8831	0.8780	0.8722
SVM	PCA(k=4)	No	4	0.8449	0.8828	0.8665	0.8644
SVM	PCA(k=5)	Yes	5	0.8874	0.9025	0.9031	0.8928
SVM	PCA(k=5)	No	5	0.8894	0.9030	0.8893	0.8900
MLP	NoPCA	Yes	16	0.9397	0.9200	0.9332	0.9328
MLP	NoPCA	No	16	0.9392	0.9190	0.9320	0.9319
MLP	PCA(k=2)	Yes	2	0.7732	0.8159	0.7986	0.7736
MLP	PCA(k=2)	No	2	0.7908	0.8307	0.8179	0.7935
MLP	PCA(k=3)	Yes	3	0.8098	0.8398	0.8275	0.8072
MLP	PCA(k=3)	No	3	0.8151	0.8410	0.8294	0.8152
MLP	PCA(k=4)	Yes	4	0.8363	0.8627	0.8475	0.8360
MLP	PCA(k=4)	No	4	0.8305	0.8597	0.8400	0.8306
MLP	PCA(k=5)	Yes	5	0.8668	0.8757	0.8654	0.8668
MLP	PCA(k=5)	No	5	0.8734	0.8843	0.8763	0.8734
Logit	PCA(k=2)	Yes	2	0.7811	0.8184	0.7749	0.7829
Logit	PCA(k=2)	No	2	0.7871	0.8253	0.7454	0.7728
Logit	PCA(k=3)	Yes	3	0.8226	0.8592	0.8341	0.8233
Logit	PCA(k=3)	No	3	0.8100	0.8508	0.7932	0.8034
Logit	PCA(k=4)	Yes	4	0.8390	0.8681	0.8509	0.8395
Logit	PCA(k=4)	No	4	0.8083	0.8656	0.8096	0.8172
Logit	PCA(k=5)	Yes	5	0.8681	0.8888	0.8804	0.8715
Logit	PCA(k=5)	No	5	0.8534	0.8900	0.8546	0.8627
ElasticNet	NoPCA	Yes	16	0.9343	0.9198	0.9326	0.9315
ElasticNet	NoPCA	No	16	0.9372	0.9183	0.9290	0.9299

Note: CV Macro-F1 refers to the best score obtained under stratified 5-fold cross-validation during hyperparameter tuning. Test metrics are computed on the held-out test set using the selected estimator.

Table 3: Best hyperparameters selected by GridSearchCV for each model configuration.

Model	Setting	SMOTE	Best hyperparameters
SVM	NoPCA	Yes	C: 10, gamma: scale
SVM	NoPCA	No	C: 10, gamma: scale
SVM	PCA(k=2)	Yes	C: 10, gamma: 1
SVM	PCA(k=2)	No	C: 100, gamma: 1
SVM	PCA(k=3)	Yes	C: 1, gamma: 1
SVM	PCA(k=3)	No	C: 100, gamma: scale
SVM	PCA(k=4)	Yes	C: 100, gamma: 0.1
SVM	PCA(k=4)	No	C: 100, gamma: scale
SVM	PCA(k=5)	Yes	C: 10, gamma: 0.1
SVM	PCA(k=5)	No	C: 10, gamma: 0.1
MLP	NoPCA	Yes	alpha: 0.0001, hidden_layer_sizes: (100,), learning_rate_init: 0.01
MLP	NoPCA	No	alpha: 0.0001, hidden_layer_sizes: (100,), learning_rate_init: 0.01
MLP	PCA(k=2)	Yes	alpha: 0.01, hidden_layer_sizes: (50,), learning_rate_init: 0.01
MLP	PCA(k=2)	No	alpha: 0.01, hidden_layer_sizes: (50,), learning_rate_init: 0.01
MLP	PCA(k=3)	Yes	alpha: 0.0001, hidden_layer_sizes: (100,), learning_rate_init: 0.01
MLP	PCA(k=3)	No	alpha: 0.0001, hidden_layer_sizes: (100,), learning_rate_init: 0.01
MLP	PCA(k=4)	Yes	alpha: 0.0001, hidden_layer_sizes: (100,), learning_rate_init: 0.01
MLP	PCA(k=4)	No	alpha: 0.0001, hidden_layer_sizes: (100,), learning_rate_init: 0.01
MLP	PCA(k=5)	Yes	alpha: 0.0001, hidden_layer_sizes: (100,), learning_rate_init: 0.01
MLP	PCA(k=5)	No	alpha: 0.0001, hidden_layer_sizes: (100,), learning_rate_init: 0.01
Logit	PCA(k=2)	Yes	C: 10
Logit	PCA(k=2)	No	C: 100
Logit	PCA(k=3)	Yes	C: 100
Logit	PCA(k=3)	No	C: 100
Logit	PCA(k=4)	Yes	C: 10
Logit	PCA(k=4)	No	C: 100
Logit	PCA(k=5)	Yes	C: 1
Logit	PCA(k=5)	No	C: 100
ElasticNet	NoPCA	Yes	C: 10, l1_ratio: 0.9
ElasticNet	NoPCA	No	C: 10, l1_ratio: 0.9

Note: CV Macro-F1 refers to the best score obtained under stratified 5-fold cross-validation during hyperparameter tuning. Test metrics are computed on the held-out test set using the selected estimator.