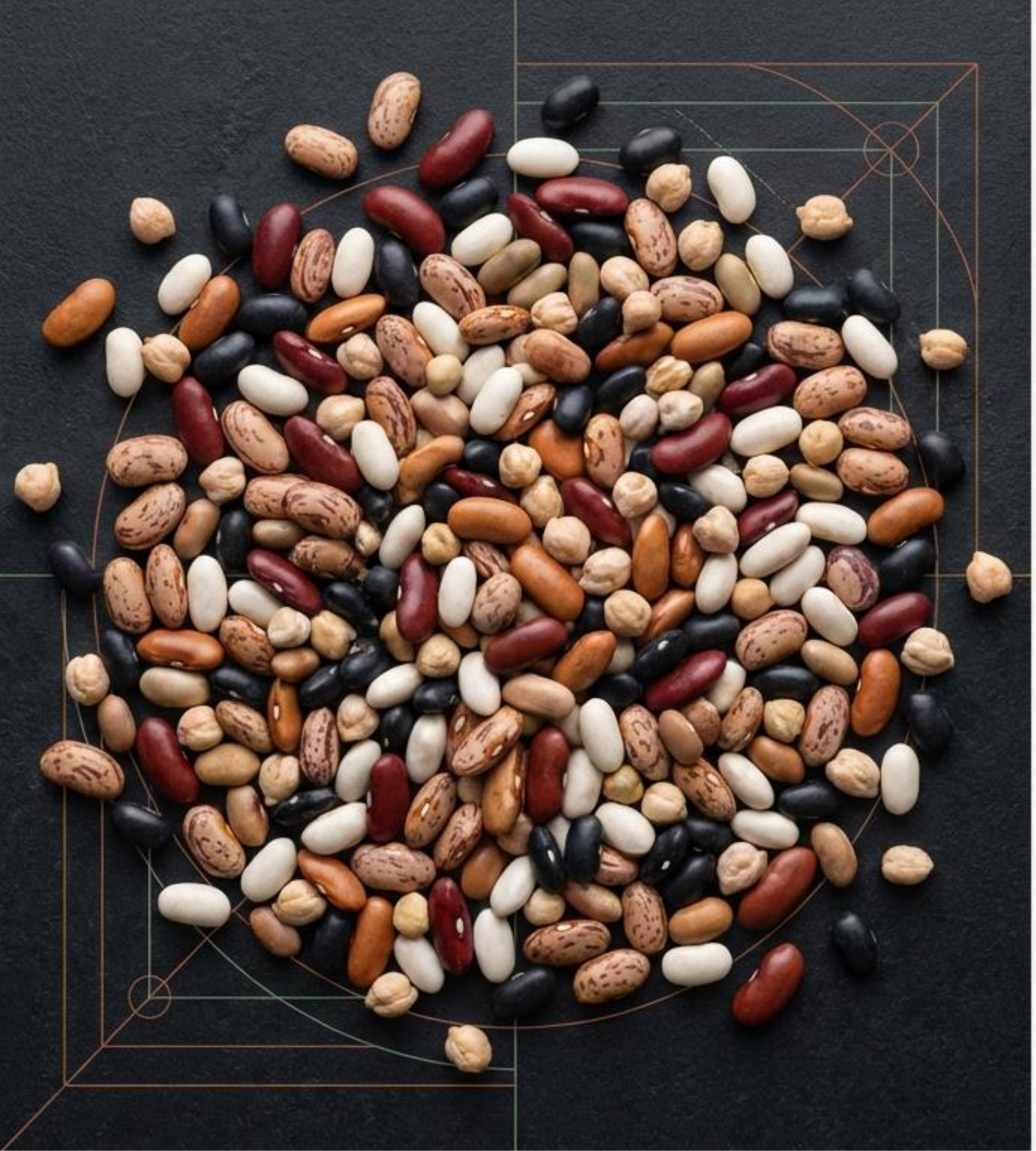


Automated Classification of Dry Bean Varieties

Mitigating Information Asymmetries through Exploratory Analysis, Dimension Reduction, and Supervised Learning

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

Supervisors: Sullivan Hue, Pierre Michel.

Institution: Aix-Marseille School of Economics (Econometrics and Data Science).

Date: January 18, 2026.

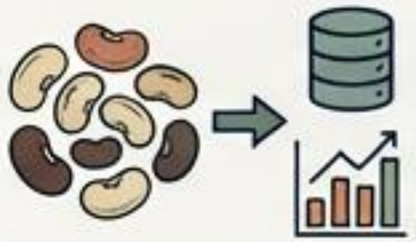
Executive Summary

The Economic Problem

Manual grading creates transaction costs and information asymmetry.

Automated classification is necessary for market efficiency.

The Data

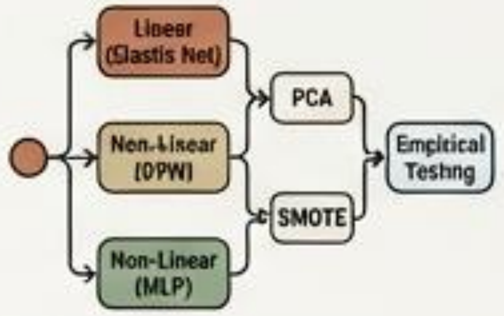


UCI Dry Bean Dataset.

13,543 unique observations.

7 varieties classified by 16 morphological features.

Methodology



Comparative analysis of Linear (Elastic Net) vs. Non-Linear (SVM, MLP) models.

Empirical testing of PCA for dimension reduction and SMOTE for class imbalance.

The Verdict

Best Model: SVM with RBF Kernel (NoPCA).

Macro-F1 Score: 0.9369

Key Insight: Dimensionality reduction (PCA) actively reduced performance; complex non-linear boundaries are required.

The Economics of Varietal Identification

Core Argument:

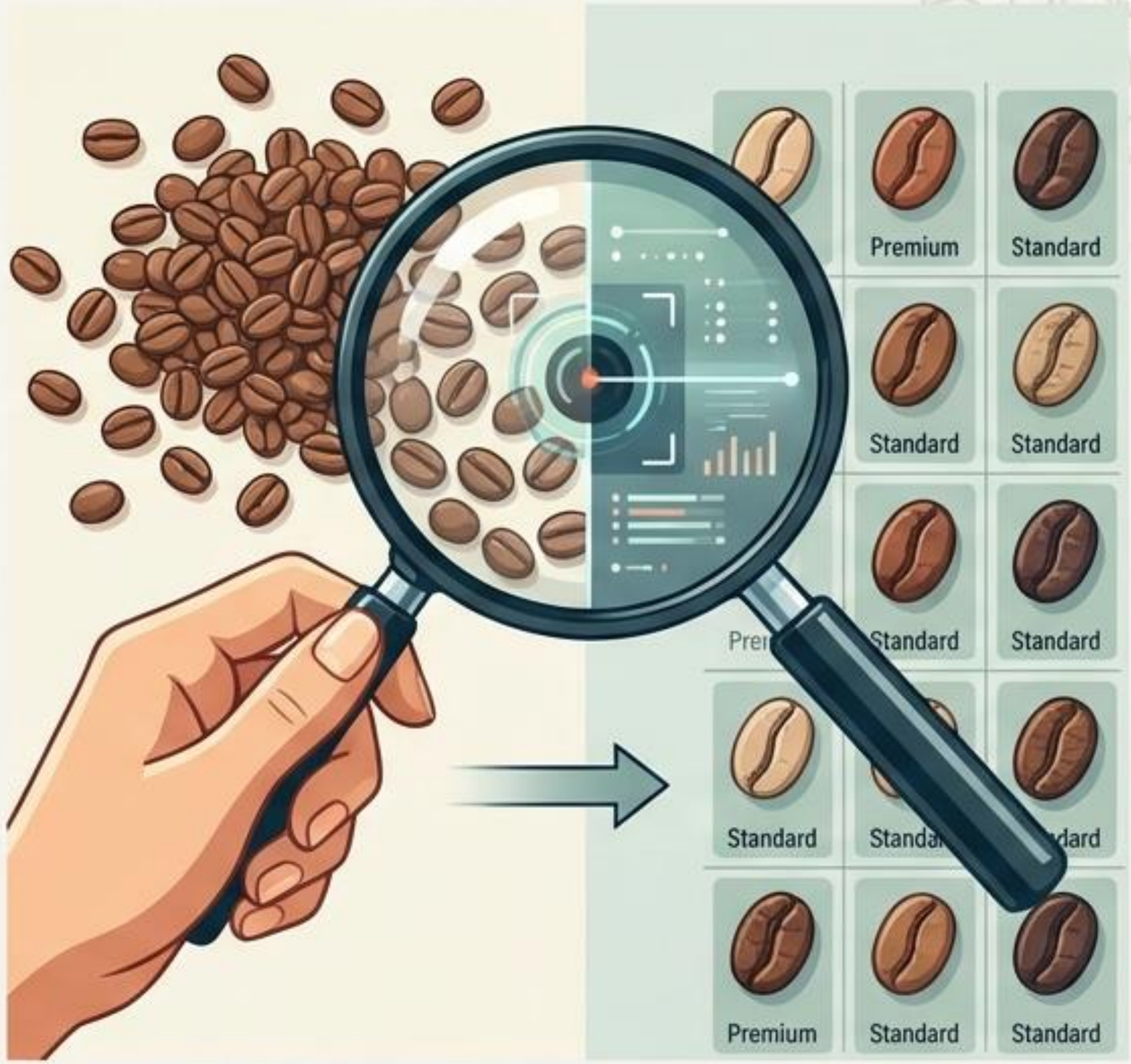
Varietal identity determines perceived quality, pricing, and supply chain coordination.

The Problem: Manual Inspection.

- Costly and time-consuming.
- Subject to human measurement error.
- Creates Information Asymmetry (Adverse Selection risk).

The Solution: Automated Classification.

- Lowers marginal cost of information.
- Standardizes quality assessment.
- Strengthens contract enforcement.



Literature Review & Research Objectives

Gautam & Trivedi (2022)

High accuracy via feature selection & deep learning.

Krishnan & Gupta (2023)

Distributional preprocessing (Box-Cox).

Lee & Park (2024)

Hybrid clustering/SVM approaches.

Our Contribution: The Robustness Gap

- 1. Empirical assessment of feature representation (Original vs. PCA).
- 2. Evaluation of imbalance handling (SMOTE) across model families.
- 3. Focus on Macro-F1 to account for minority variety transaction costs.

Methodological Roadmap

Phase 1: EDA

Distribution, Outliers, Correlation.

Phase 2: Preprocessing

Cleaning, Robust Scaling, SMOTE.

Phase 3: Dim Reduction

PCA (k = 2, 2, 3, 4, 5).

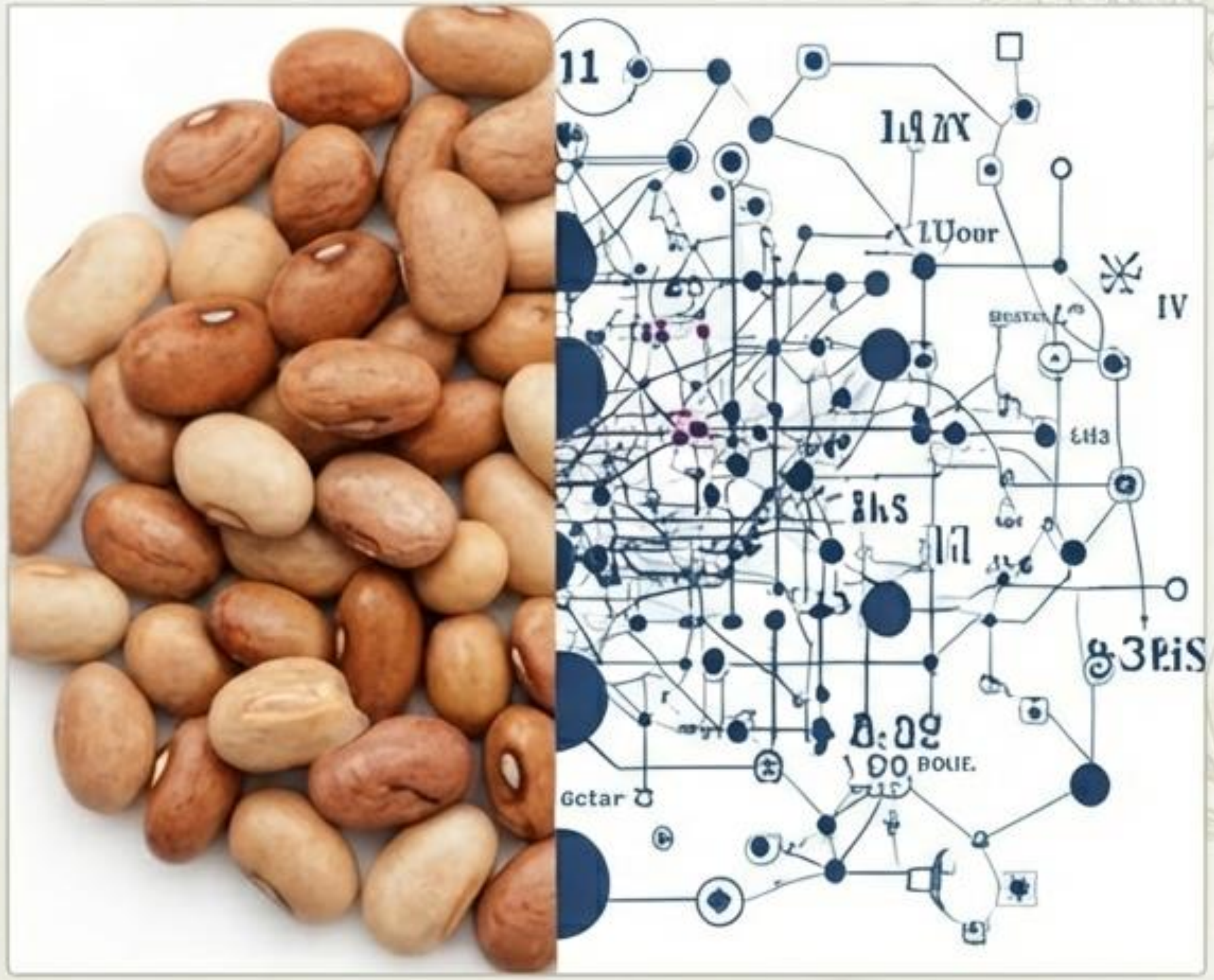
Phase 4: Learning

Elastic Net, SVM, MLP, Logit.

Validation: Stratified 5-Fold Cross-Validation.

Data Source & Preprocessing Protocols

- **Source**: UCI Machine Learning Repository ‘Dry Bean Dataset’.
- **Observation Count**: 13,543 unique samples (68 duplicates removed).
- **Features**: 16 Morphological Descriptors (Area, Perimeter, Shape Factors).
- **Classes**: 7 Varieties (Barbunya, Bombay, Cali, Dermason, Horoz, Seker, Sira).
- **Split**: Stratified 70/30 Train/Test.
- **Scaling Strategy**: RobustScaler.
- **Rationale**: Uses Median and IQR instead of Mean/Variance to remain resilient against significant physical outliers.



Transformation of Physical Matter to Data

Model Selection & Theoretical Framework

Linear Baselines (Interpretable)

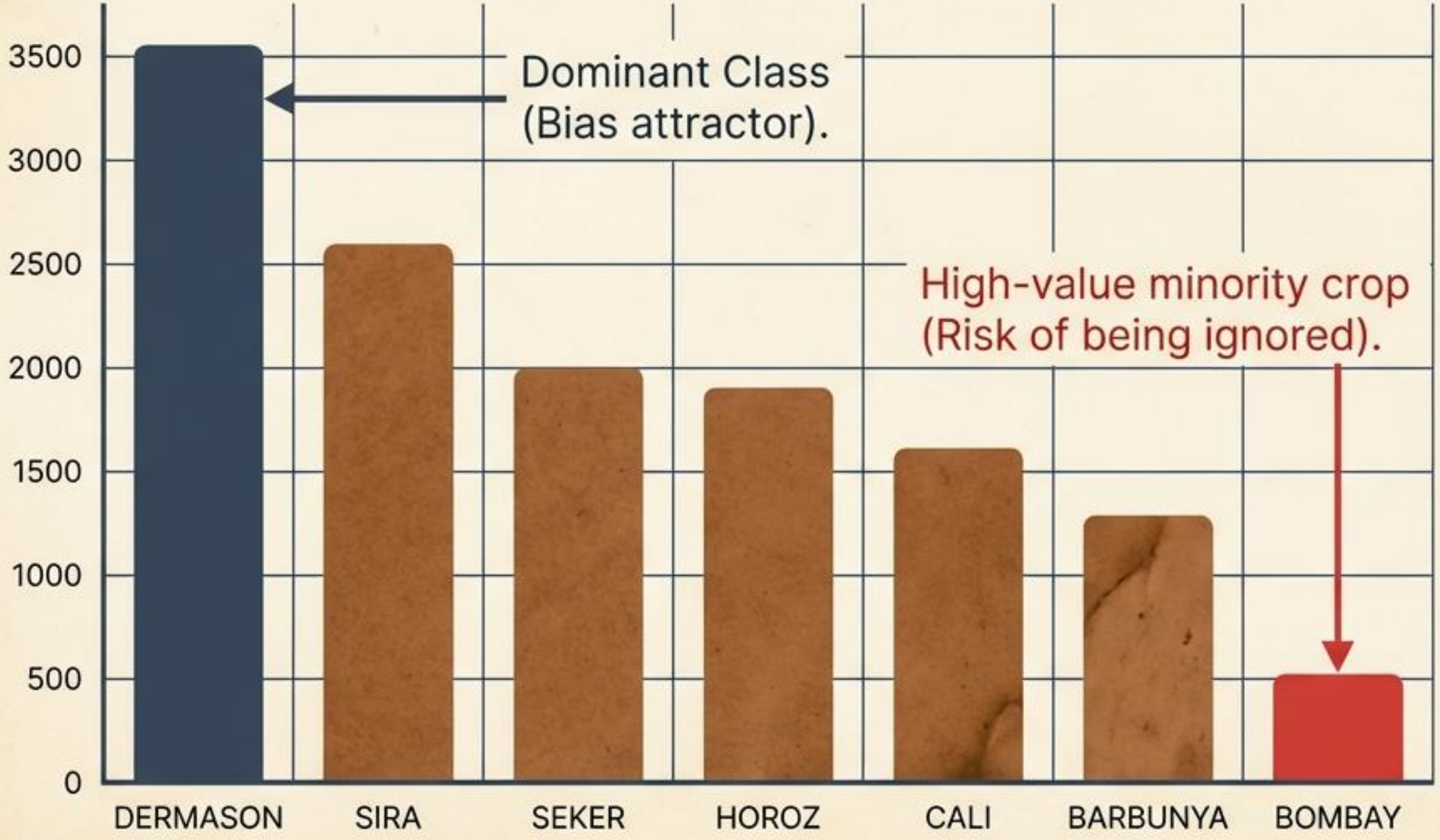
- **Logistic Regression:**
  Multinomial probabilistic baseline.
- **Elastic Net:**
  Combines L1/L2 penalties. Handles multicollinearity and performs variable selection.

Non-Linear Powerhouses (Complex)

- **Support Vector Machines (SVM):**
  RBF Kernel projects to high-dimensional space. Optimal for non-linear boundaries.
- **Multilayer Perceptron (MLP):**
  Neural network with hidden layers to capture complex morphology.

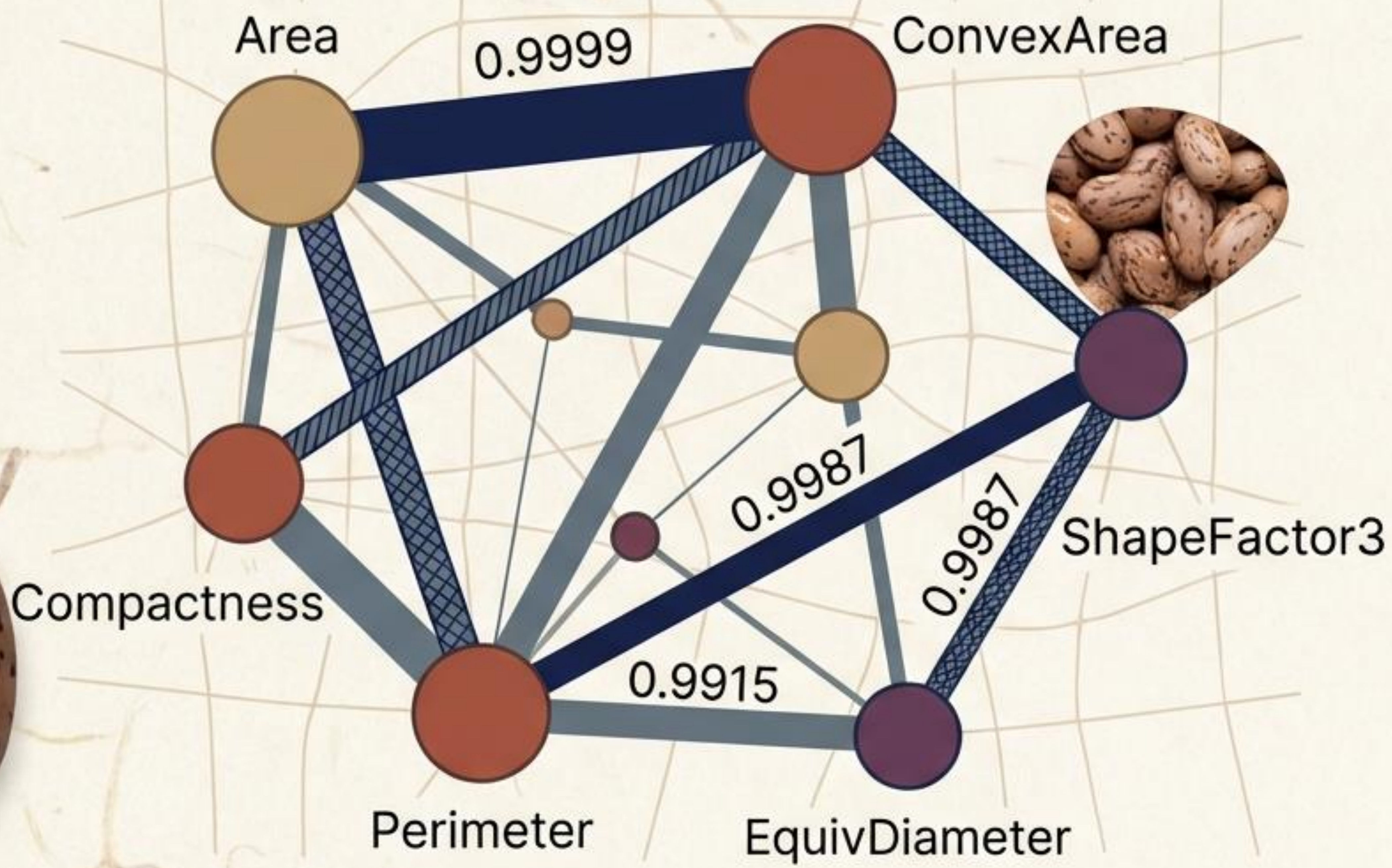
Experimental Grid: All models tested [With/Without PCA] and [With/Without SMOTE].

The Challenge of Heterogeneity and Imbalance



Consequence: A 1:7 imbalance ratio creates a bias toward majority classes, risking revenue loss for specialized farmers.

EDA: The Multicollinearity Problem



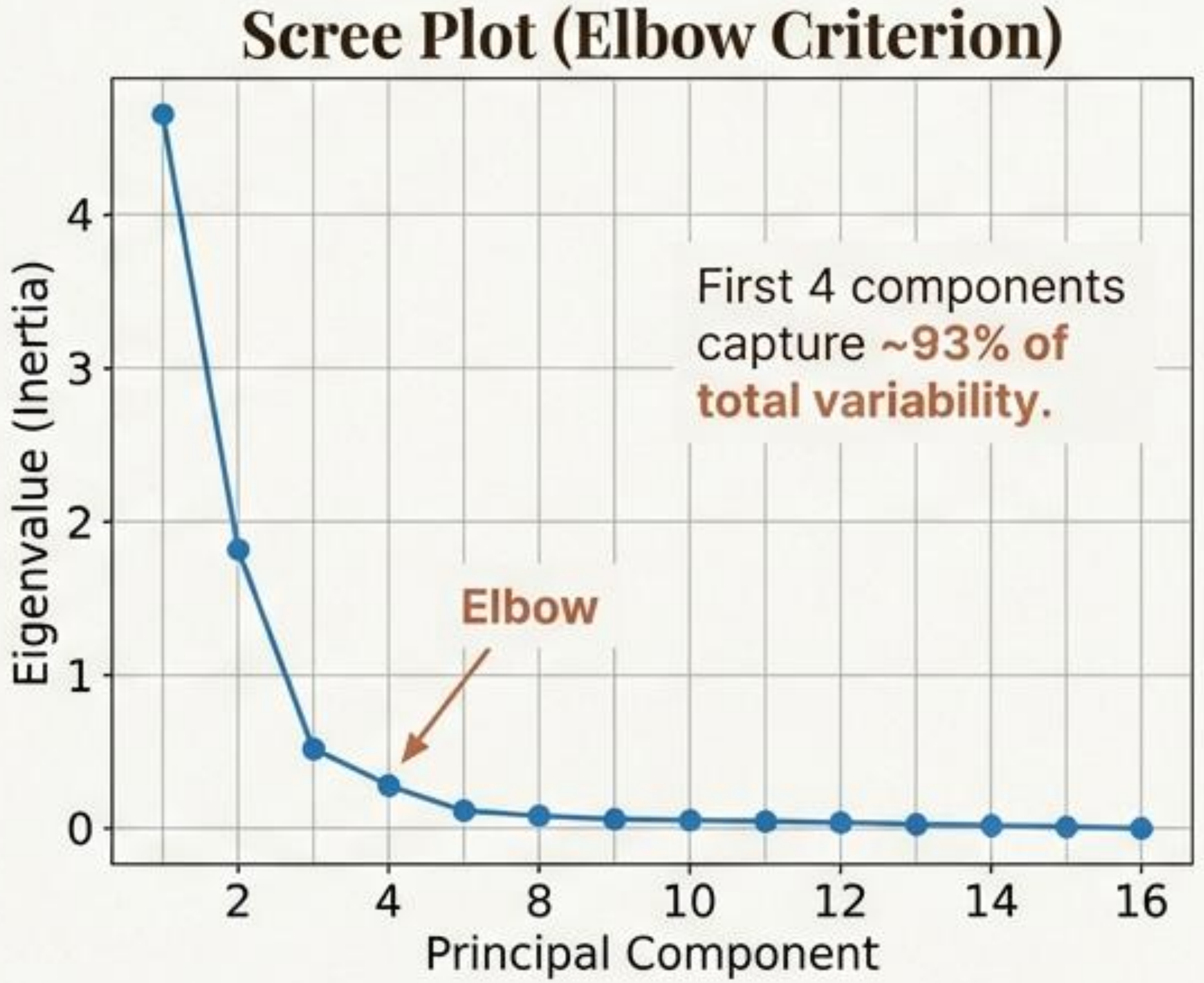
Insight: Geometric features are highly interdependent.

Hypothesis: Since variables overlap, can we compress them? (Motivation for PCA).

EDA III: Multicollinearity & Dimensionality Reduction

The Problem: Extreme Feature Correlation.

- Area vs. ConvexArea: 0.9999
- Compactness vs. ShapeFactor3: 0.9987
- Risk: Unstable estimates in linear models.



Empirical Strategy & Model Selection

Experimental Variables

Feature Space

- Original (16 features) → PCA (k=2, 3, 4, 5)

Imbalance Handling

- Raw Distribution → SMOTE

The Models

Group 1: Linear

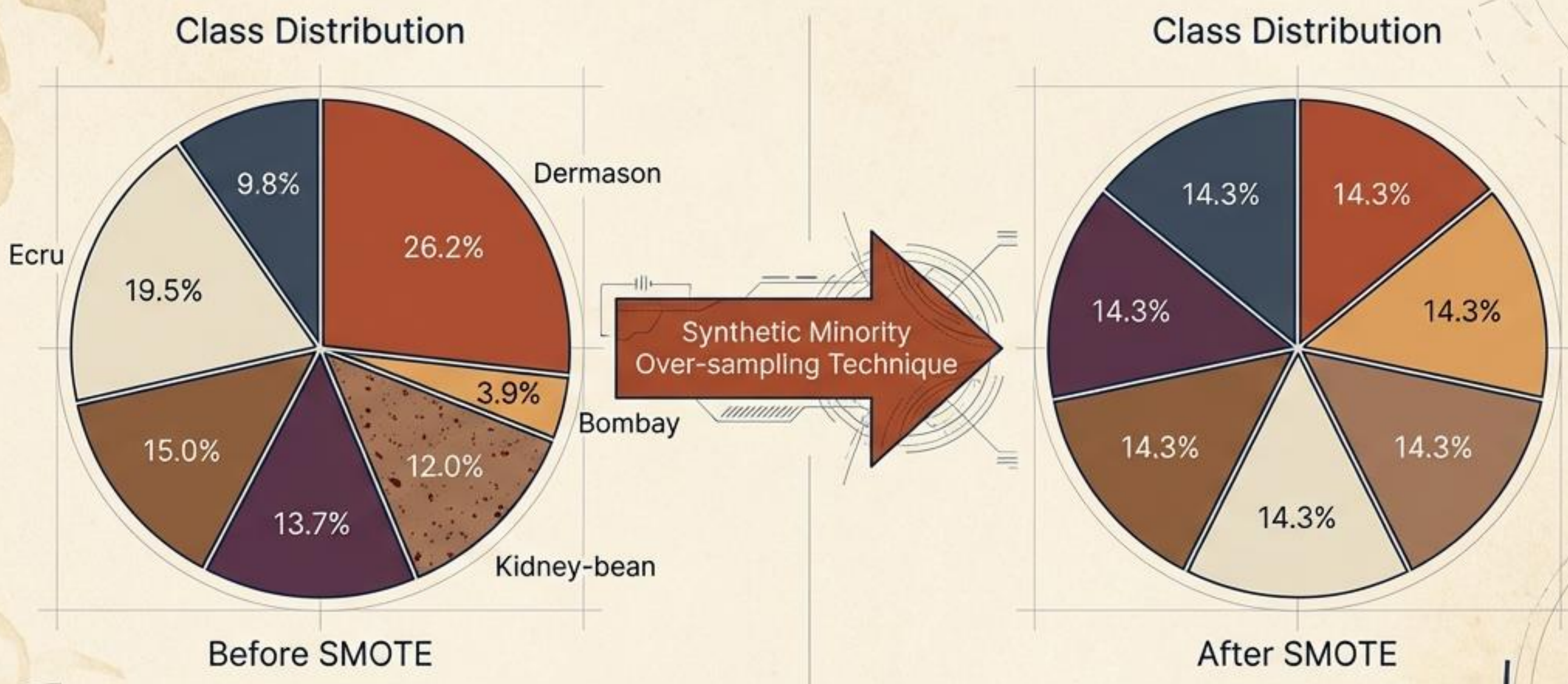
- Logistic Regression
- Elastic Net (L1/L2 Regularization)

Group 2: Non-Linear

- SVM (RBF Kernel)
- MLP (Neural Network)

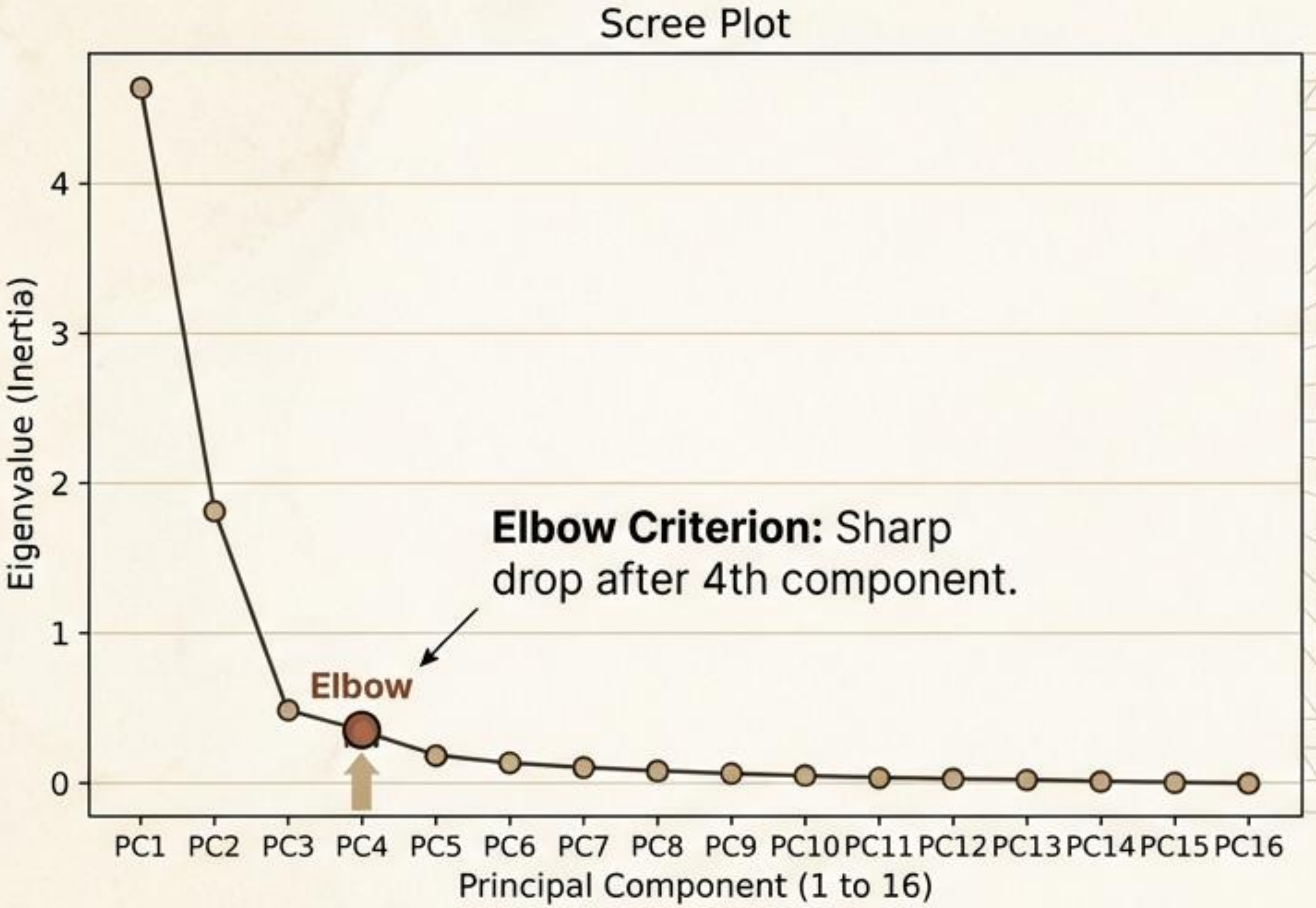
Validation: GridSearchCV with Stratified 5-Fold Cross-Validation.

Correcting Market Bias: The SMOTE Algorithm



The Problem: Algorithms ignore minority crops.
The Fix: Generating synthetic examples by interpolating between nearest neighbors in feature space.

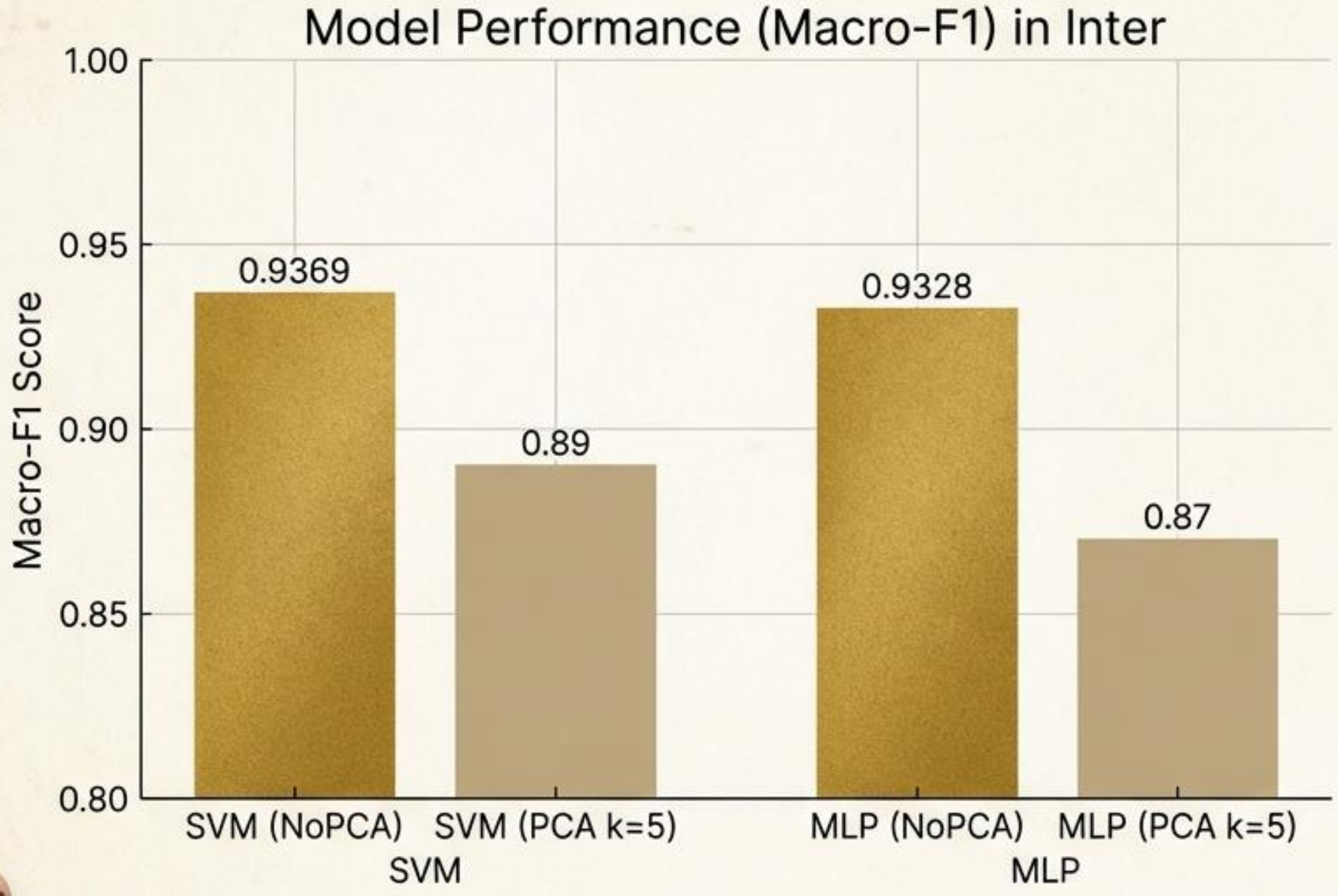
Dimensionality Reduction: The PCA Hypothesis



Explained Variance: First 4 components capture ~93% of variability.

Expectation:
Compressing 16 features into 4 should retain signal and remove noise.

The Verdict: Complexity Triumphs Over Simplification



Key Insight:
Variance ≠ Separability.
PCA captured 93% of variance but discarded the subtle signals needed to distinguish similar varieties.

Result: Original feature space consistently outperforms PCA.

Champion Model: SVM (RBF Kernel)

The optimal technical solution for the economic problem.

Macro-F1: 0.9369

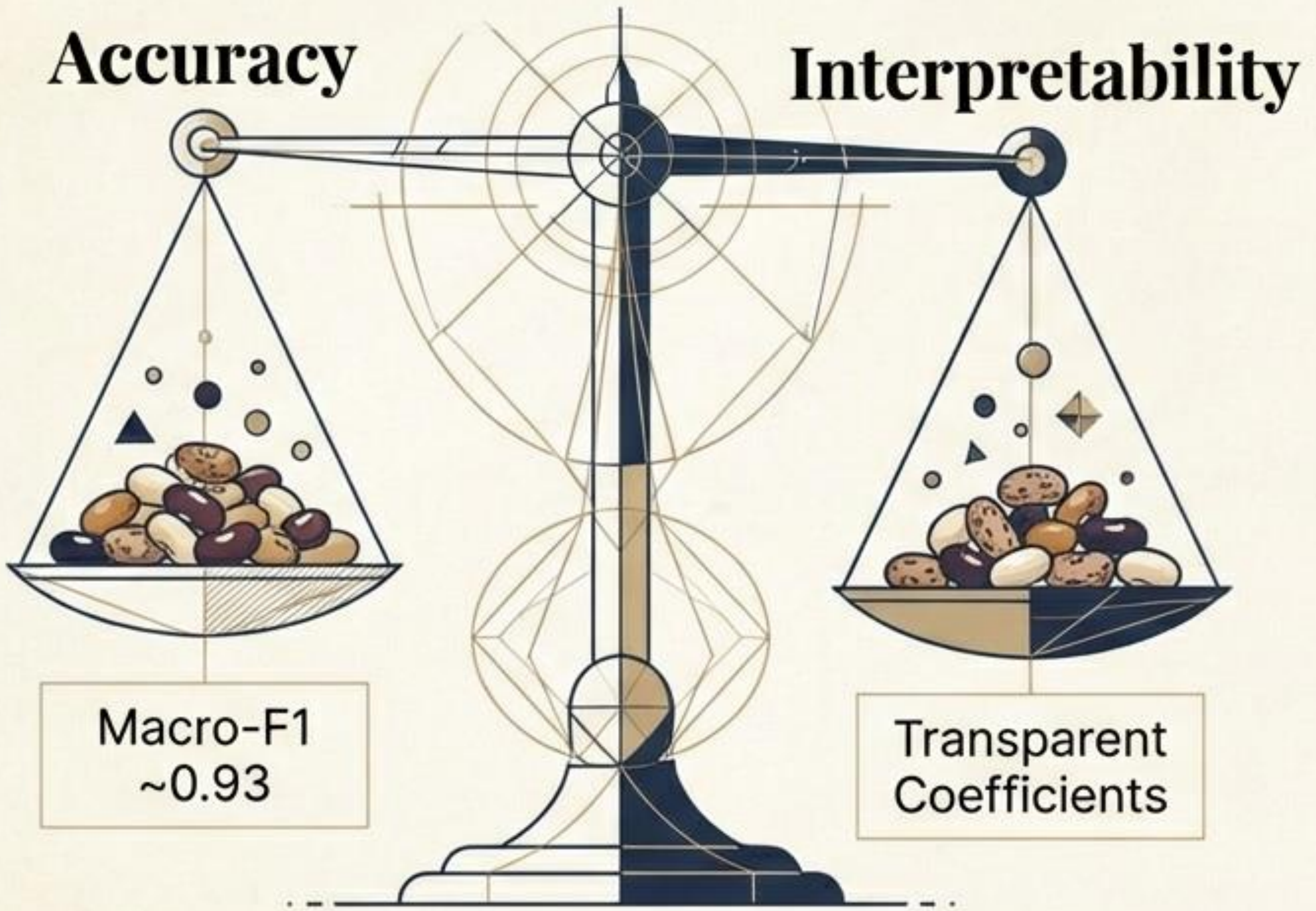
Balanced Accuracy: 0.9357

- Feature Space: Original (NoPCA)
- Imbalance Handling: Raw (NoSMOTE)
- Kernel: Radial Basis Function (RBF)

Why it won:

The RBF kernel projects data into higher dimensions, creating complex decision boundaries that linear models miss.

The Linear Alternative: Elastic Net

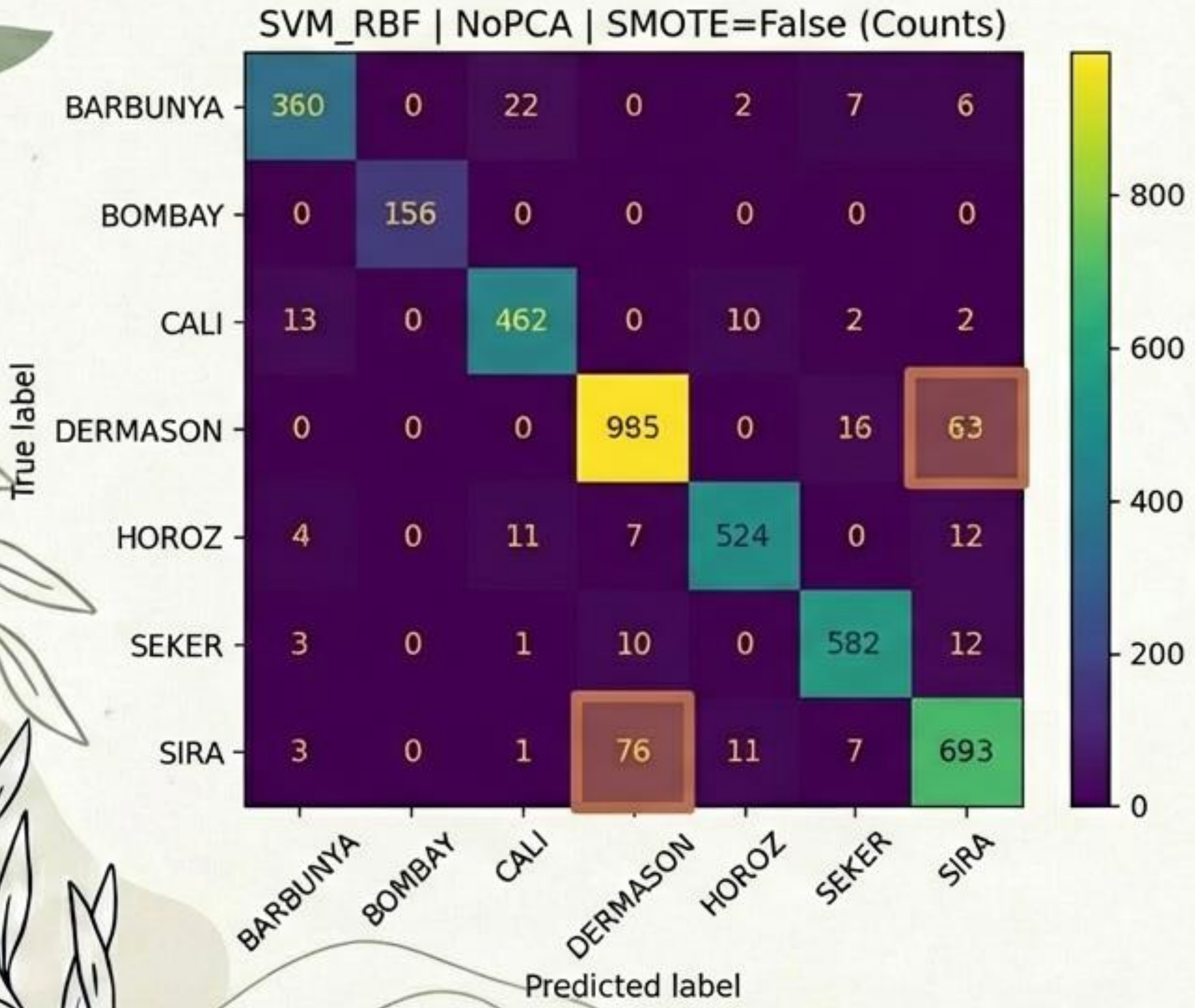


Why it matters: Elastic Net handles multicollinearity via L1/L2 regularization *without* losing feature meaning.

Economic Advantage: A "Glass Box" solution. We can explain to a farmer exactly which physical trait determined the grade.

Competitive performance to SVM, but with superior explainability.

Error Analysis & Confusion Matrix



The Problem Pairs: Dermason vs. Sira.

- Root Cause: Extreme morphological similarity.
- Economic Implication: These specific transaction pairs represent the highest risk of misgrading. Automated systems may require human supervision or higher thresholds for these specific varieties.

Conclusion: Key Findings

Best Performance:
SVM-RBF on Raw Data. **Macro-F1: 0.9369.** Complexity wins over reduction.

PCA Limitation:
Information loss in **projection** outweighed benefits of dimensionality reduction.

Economic Viability:
High accuracy validates the use of **automation to reduce inspection costs and standardize quality.**

Future Directions

- **Supervised Dimensionality Reduction:** Investigate LDA or PLS-DA instead of PCA to preserve class separability.
- **Cost-Sensitive Learning:** Implement loss functions that penalize errors based on the actual market price difference between varieties.
- **Interpretability:** Apply SHAP values to explain individual grading decisions for trust and transparency.