Feature scaling in Dry Bean classification

Standardizing 16 seed measurements within training folds raised mean balanced accuracy from 71.62% to 93.46% for a fixed 7-nearest-neighbor classifier. The gain was 21.84 percentage points across five paired Dry Bean splits.

Research question

Koklu and Özkan (2020) used geometry from segmented seed images to compare four classifiers. Area and shape ratios have different scales. We test how rescaling this feature representation changes a fixed distance-based classifier.

Methods

We removed 68 exact duplicate feature-and-label rows before splitting and retained 13,543 of 13,611 rows. A shuffled stratified five-fold split (seed 42) balanced class proportions. In each fold, z-scoring used means and standard deviations estimated only from training rows, then applied them unchanged to held-out rows. Both pipelines used those same splits, all 16 features, Euclidean distance and uniform 7-neighbor votes. Balanced accuracy was mean recall across seven varieties.

Results

Standardization improved balanced accuracy in each fold by 21.45 to 22.23 points. Barbunya recall increased from 49.1% to 90.1%. Barbunya predictions classified as Cali fell from 460 to 86. Sira remained difficult, with 87.4% recall after scaling.

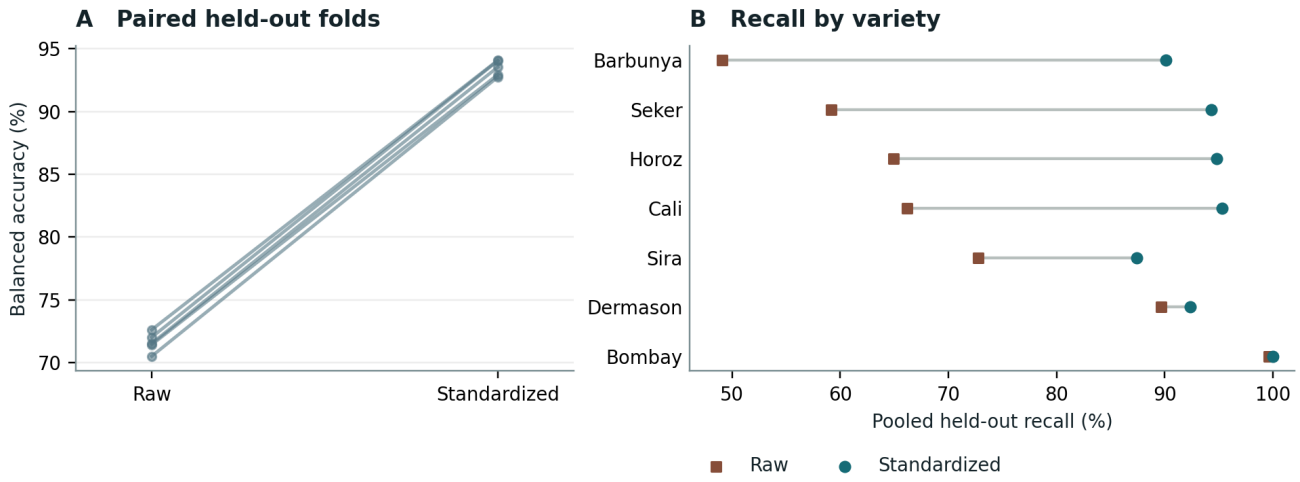


Figure 1. (A) Each line joins the two balanced accuracy scores for one held-out fold. (B) Out-of-fold recall pools predictions for each variety across five folds.

Limitations

Overlapping training folds make the reported range descriptive within one dataset. An independently acquired image batch would test whether the advantage survives changes in lighting, camera calibration, segmentation or lot. Such a test should fit the scaler and classifier on the original data and apply both unchanged to the new batch.

Sources

Koklu M and Özkan İA (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. Computers and Electronics in Agriculture 174, 105507. doi.org/10.1016/j.compag.2020.105507.

Dry Bean (2020). UCI Machine Learning Repository. doi.org/10.24432/C50S4B.