PCA for Dimensionality Reduction Before Classification: A Report on the Wine Dataset

Huzaifa Ahmed Khan
Department of Physics
NED University of Engineering & Technology, Karachi, Pakistan
khan4730340@cloud.neduet.edu.pk

***Abstract*—This report demonstrates Principal Component Analysis (PCA) as a dimensionality reduction step prior to supervised classification. Using the Wine dataset, features are standardized and PCA is applied to retain 95% of the variance. A Logistic Regression classifier is trained on the reduced representation. The resulting model achieves 97.78% test accuracy while reducing the feature space to 10 principal components capturing 96.37% of the variance. PCA loadings are further analyzed to identify the most influential original features contributing to the first two principal components.**

***Index Terms*—Principal Component Analysis, Dimensionality Reduction, Logistic Regression, Classification, Feature Loadings, Wine Dataset**

I. Introduction

High-dimensional datasets commonly contain correlated and redundant features, increasing computational cost and potentially reducing generalization. Principal Component Analysis (PCA) addresses this by projecting data into a lower-dimensional orthogonal basis that preserves maximum variance. In this study, PCA is used as a preprocessing step before classification to evaluate (i) performance, (ii) variance retention, and (iii) interpretability via loadings.

II. Methodology

A. Dataset

Experiments use the standard Wine dataset containing three classes and multiple continuous chemical measurements. The objective is to predict the wine class from these features.

B. Preprocessing

Features are standardized to zero mean and unit variance:

$$\tilde{x} = \frac{x - \mu}{\sigma}. \tag{1}$$

Standardization is essential since PCA is scale-sensitive.

C. Dimensionality Reduction via PCA

Given standardized data $\tilde{X} \in \mathbb{R}^{n \times d}$, PCA projects onto a reduced space:

$$Z = \tilde{X} W, \tag{2}$$

where columns of W are eigenvectors of the covariance matrix of $\tilde{X}$, sorted by decreasing eigenvalues. Components are selected to satisfy a variance retention target (95% in this experiment).

TABLE I
PCA Summary

Metric	Value
Components retained	10
Explained variance (sum)	0.9637

TABLE II
Classification Report (Test Set)

Class	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	15
1	0.95	1.00	0.97	18
2	1.00	0.92	0.96	12
Accuracy		0.9778		45
Macro avg	0.98	0.97	0.98	45
Weighted avg	0.98	0.98	0.98	45

D. Classifier

A Logistic Regression model is trained on the PCA-transformed representation. Performance is reported using accuracy, precision, recall, and F1-score on a held-out test set.

III. Results

A. Variance Retention

PCA reduced the feature space to 10 principal components while retaining 96.37% of total variance.

B. Classification Performance

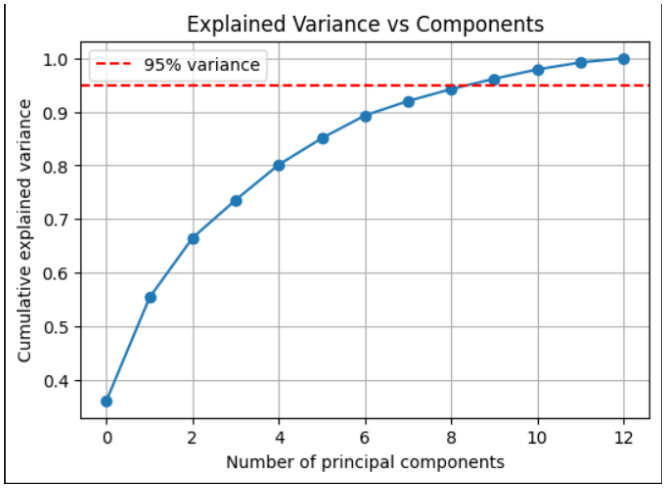
The classifier achieved 97.78% test accuracy on 45 test samples.

C. Feature Loadings (Interpretability)

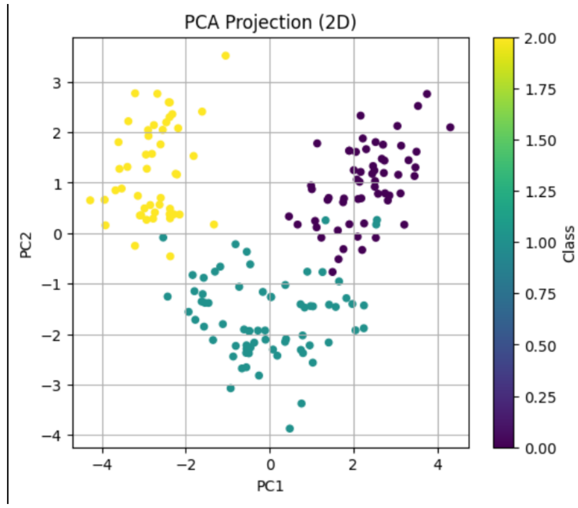
Table III lists the strongest contributors to PC1 and PC2 (from the computed loadings).

IV. Visualizations

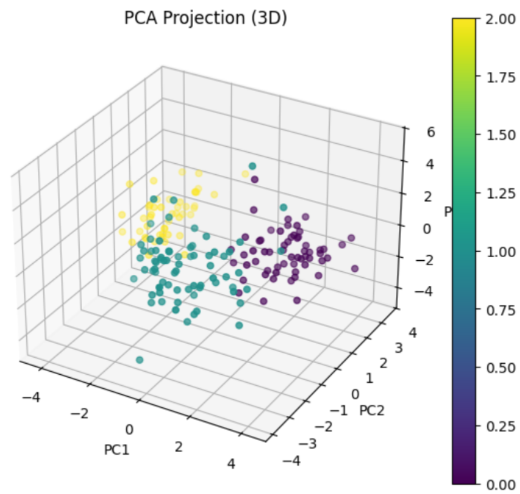
For readability in IEEE two-column format, figures are grouped into two wide panels across both columns. Ensure the image files exist in `images/` with the exact names used below.



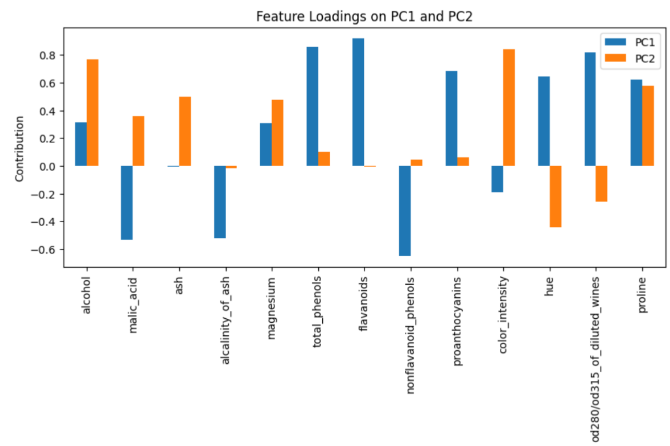
(a) Cumulative explained variance.



(b) 2D PCA projection (PC1 vs PC2).



(c) 3D PCA projection (PC1, PC2, PC3).



(d) Feature loadings on PC1 and PC2.

Fig. 1. PCA visualizations: variance retention, low-dimensional projections, and feature loadings.

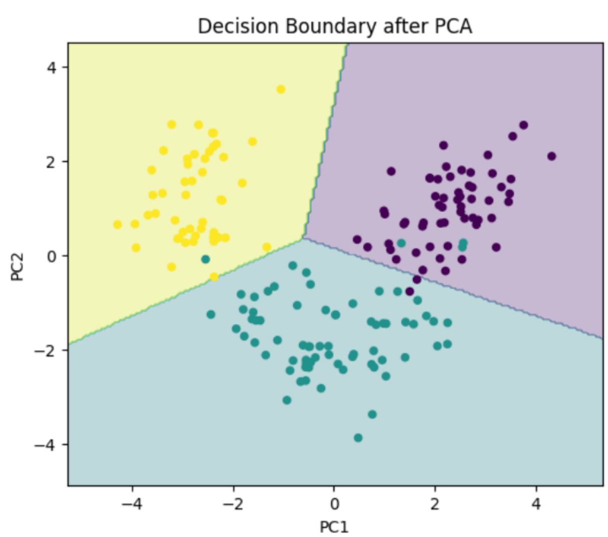


Fig. 2. Logistic Regression decision boundary in PCA space (PC1–PC2).

TABLE III
TOP FEATURE LOADINGS ON PC1 AND PC2

Feature	PC1	PC2
flavanoids	0.9201	-0.0053
total_phenols	0.8586	0.1031
od280/od315_of_diluted_wines	0.8183	-0.2607
proanthocyanins	0.6818	0.0623
hue	0.6455	-0.4425
proline	0.6238	0.5782
alcohol	0.3140	0.7664
magnesium	0.3089	0.4748
ash	-0.0045	0.5009
color_intensity	-0.1928	0.8399

V. DISCUSSION

PCA reduced the original feature space to 10 components while maintaining strong classification performance (97.78% test accuracy). The explained variance (96.37%) indicates that most structure in the dataset lies in a lower-dimensional subspace. Loadings suggest that *flavanoids*, *total_phenols*, and *od280/od315_of_diluted_wines* dominate PC1, while *color_intensity*, *alcohol*, and *proline* strongly influence PC2, improving interpretability of the principal axes.

VI. CONCLUSION

This study demonstrates a practical PCA→ML workflow: standardize features, reduce dimensionality with a variance target, and train a classifier on the compressed representation. On the Wine dataset, PCA retained 96.37% variance using 10 components while achieving 97.78% test accuracy with Logistic Regression. The pipeline provides both performance and interpretability and can be extended to higher-dimensional problems such as imaging and spatiotemporal climate data.

REPRODUCIBILITY NOTE

Results correspond to the pipeline: StandardScaler → PCA (95% variance target) → Logistic Regression, evaluated on a held-out test set.

REFERENCES

[1] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, 2016.

[2] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.