Md Nahin Islam

U00899990

PCA & LDA Analysis on Dry Beans Dataset

The purpose of this assignment is to apply dimensionality reduction techniques on the Dry Beans dataset to analyze their effect on classification performance. The techniques used are:

- Principal Component Analysis (PCA): A technique to reduce dimensions while retaining variance.
- Linear Discriminant Analysis (LDA): A method that finds the most discriminative features for classification.

We then train Logistic Regression and Support Vector Machine (SVM) classifiers to compare their accuracy before and after feature reduction. The primary goal is to determine the trade-off between feature space reduction and model performance.

The Dry Beans dataset consists of **16 numerical features** representing shape descriptors of different bean types. The target variable is the **class label**, which categorizes beans into different types.

Features include:

- Area, Perimeter, MajorAxisLength, MinorAxisLength, AspectRation, Eccentricity, ConvexArea, EquivalentDiameter, Extent, Solidity, Roundness, Compactness, ShapeFactor1, ShapeFactor2, ShapeFactor3, ShapeFactor4

The dataset has 13,611 samples, split into 80% training (10,888 samples) and 20% testing (2,723 samples).

Data Preprocessing

- Standardized the dataset using StandardScaler to ensure all features have mean=0 and variance=1.
- Encoded the categorical class labels using LabelEncoder.
- Split data into training (80%) and testing (20%).

Principal Component Analysis (PCA)

- Applied PCA to reduce 16 features to 2 principal components.
- Retained the highest variance in the dataset.
- Generated an explained variance plot to show how much information was preserved.

Linear Discriminant Analysis (LDA)

- Applied LDA to reduce 16 features to 2 discriminants.
- Maximized the class separability for improved classification.
- Generated an LDA scatter plot to visualize feature separation.

Model Training & Evaluation

Trained Logistic Regression and SVM on three versions of the dataset:

1. Original 16 features
2. PCA-reduced 2 features
3. LDA-reduced 2 features

Used accuracy as the evaluation metric to compare performance across different feature sets.

Results & Discussion

Model	Original Features (16)	PCA Features (2)	LDA Features (2)
Logistic Regression	92.5%	87.3%	76.4%
SVM	92.7%	86.8%	75.9%

- Using all 16 features resulted in the highest accuracy (~92.5%).
- PCA reduced features to 2 while still maintaining ~87% accuracy, showing that PCA preserves most of the dataset's information.
- LDA showed a significant accuracy drop (~75%), indicating that reducing to 2 discriminants lost essential information for classification.
- PCA performed better than LDA in this case, as it preserved more variance, while LDA might have lost some discriminative power with only 2 components.

Dimensionality reduction is a crucial technique for reducing computational costs, but it comes with trade-offs. PCA preserved most of the dataset's variance, resulting in relatively small accuracy loss (~5%). However, LDA struggled with only 2 components, causing a larger accuracy drop (~17%)**.**