Islamic Azad University

Tabriz Branch

Faculty of Engineering and Technology

Department of Computer Engineering

Project Report Title:

Comparative Classification Analysis of the Iris Dataset using PCA, LDA, QDA, and GNB

Instructor:

Dr. Ahad Esmaeilzadeh

Prepared by:

Shayan Abdollahi Nami, Pouya Esmaeili

January 2026

Table of Contents

Abstract

This project explores the classification of the Iris flower dataset into three species (*Sentosa, Versicolor, Virginica*) using statistical machine learning methods. We applied Principal Component Analysis (PCA) for dimensionality reduction and visualization, followed by training Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Gaussian Naive Bayes (GNB) classifiers. The performance of these models was evaluated using accuracy, precision, recall, and F1-score on a held-out test set.

1. Introduction

The Iris dataset is a classic benchmark in pattern recognition literature. It consists of 150 samples of iris flowers, each described by four features: sepal length, sepal width, petal length, and petal width. The objective of this study is to compare the effectiveness of linear (LDA), quadratic (QDA), and probabilistic (GNB) decision boundaries in distinguishing between the three species, especially after analyzing the data structure through PCA.

2. Methods

Data Source: The standard Iris dataset containing 150 samples (50 per class).

Preprocessing:

- The dataset was split into training (80%, 120 samples) and testing (20%, 30 samples) sets.
- The features were used directly without additional normalization or scaling, as the initial analysis showed the data ranges were compatible with the selected models.

Dimensionality Reduction:

- **PCA:** Applied to project the 4-dimensional into 2D space for visualization purposes.

Models:

- **Linear Discriminant Analysis (LDA):** sklearn.discriminant_analysis.LinearDiscriminantAnalysis
- **Quadratic Discriminant Analysis (QDA):** sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis
- **Gaussian Naive Bayes (GNB):** sklearn.naive_bayes.GaussianNB

- **Evaluation Metrics:** Accuracy, Confusion Matrix, and Classification Report (Precision, Recall, F1-score).

3. Results

3.1. PCA Analysis

Principal Component Analysis was performed to visualize the separability of the classes in a lower-dimensional space.

Explained Variance:

- **Principal Component 1 (PC1):** Explains 72.96% of the variance
- **Principal Component 2 (PC2):** Explains 22.85% of the variance.
- **Total Explained Variance:** The first two components combined explain 95.81% of the total information in the dataset.

Observation:

The PCA scatter plot clearly shows that Iris Setosa is linearly separable from the other two classes along PC1, Iris Versicolor and Iris Virginica shows a slight overlap but are large distinct, suggesting that linear classifiers like LDA should perform well on this dataset.

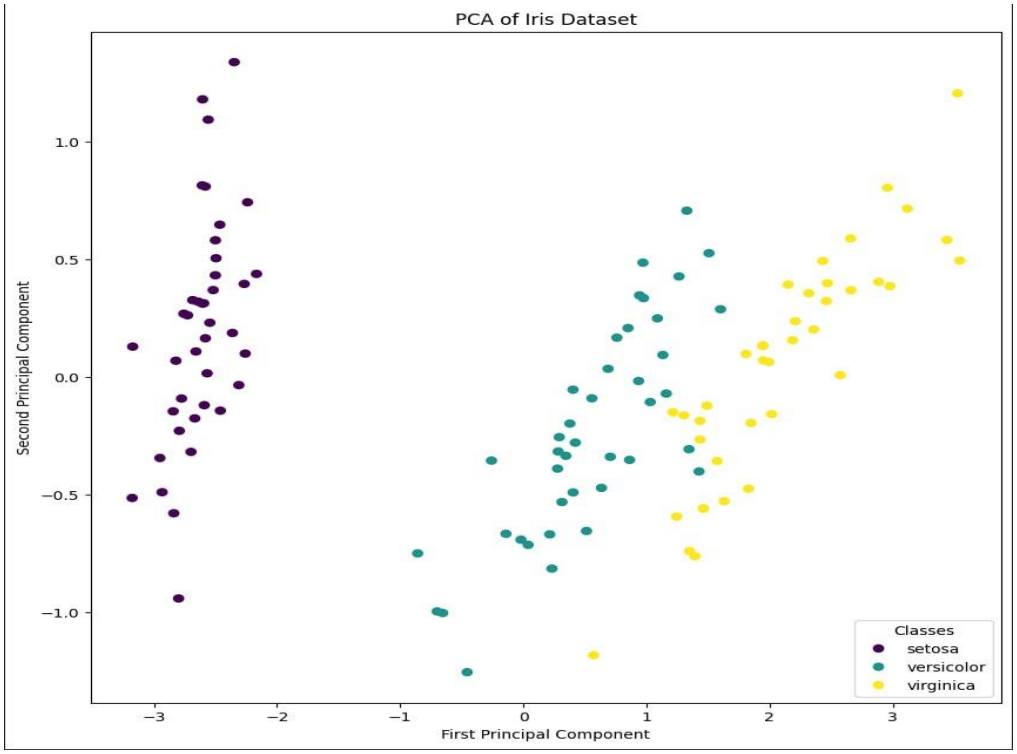


Figure 1: PCA projection of Iris Dataset (PC1 vs PC2)

3.2. Linear Discriminant Analysis (LDA)

The linear Discriminant Analysis model was trained to find the optimal linear boundaries between the classes.

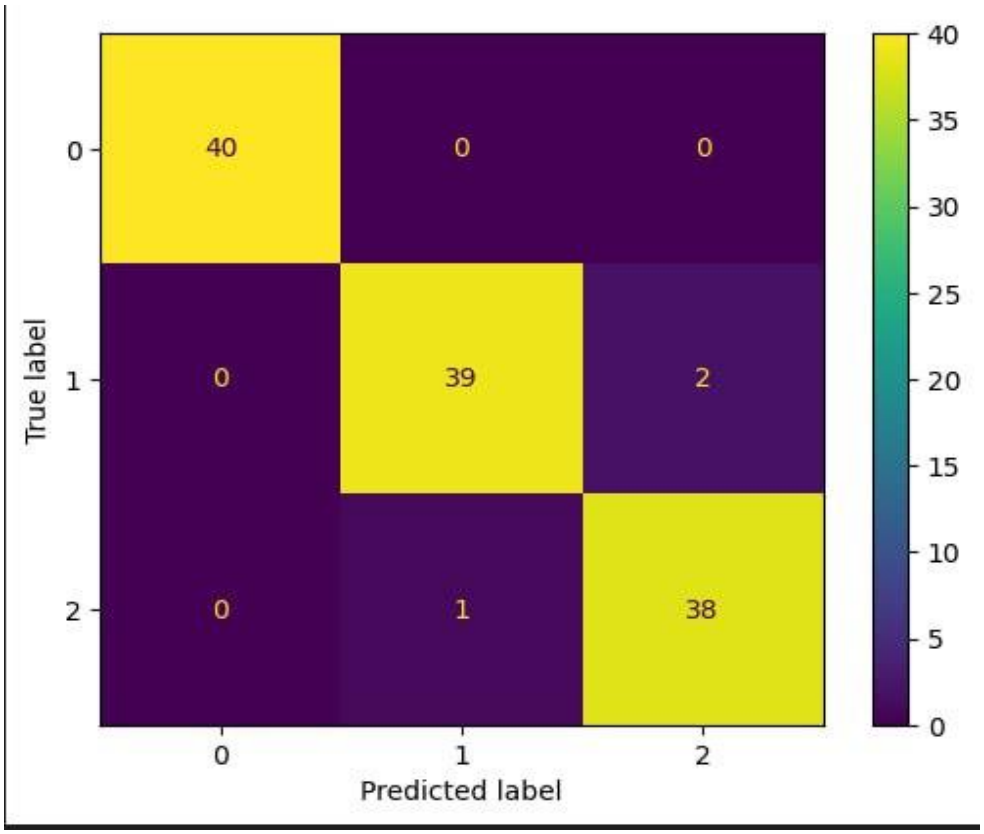
Performance Metrics (Test Set):

- **Accuracy:** 1.00 (100%)
- **Macro Average F1-Score:** 1.00

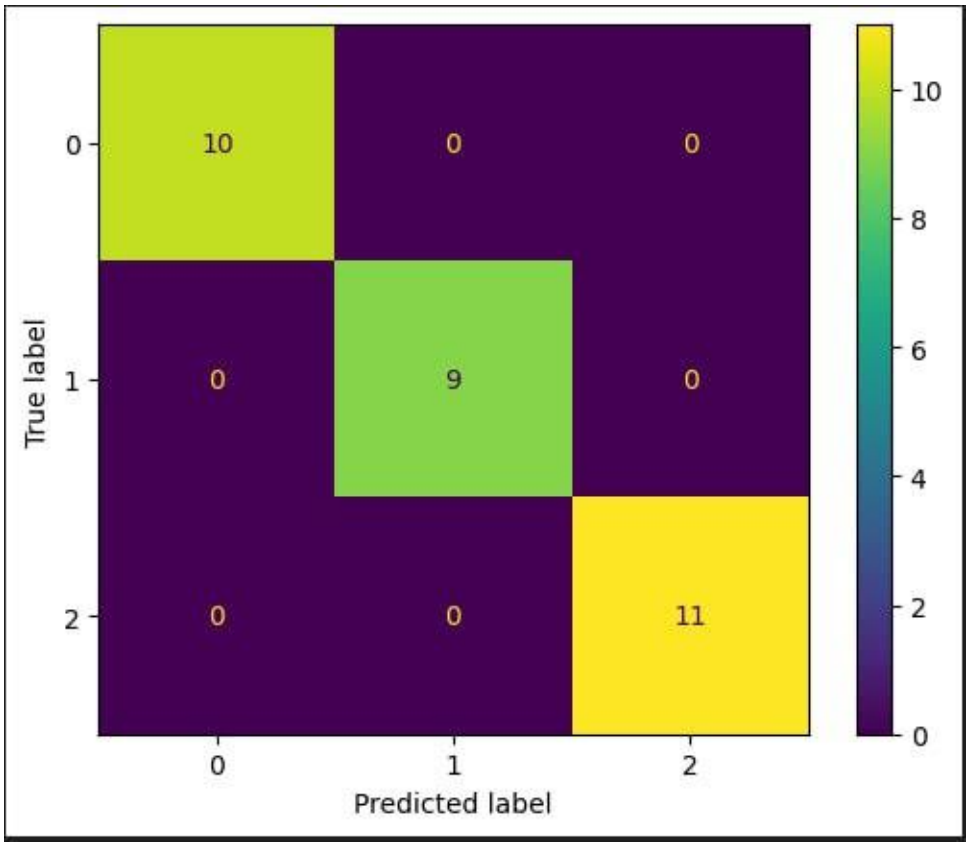
Classification Report: The model achieved perfect precision and recall for all three classes on the test set.

Class	Precision	Recall	F1-Score	Support
Sentosa	1.00	1.00	1.00	10
Versicolor	1.00	1.00	1.00	9
Virginica	1.00	1.00	1.00	11
Accuracy			**1.00**	30

Train



Test



3.3. Quadratic Discriminant Analysis (QDA)

The Quadratic Discriminant Analysis model was trained to allow for non-linear decision boundaries.

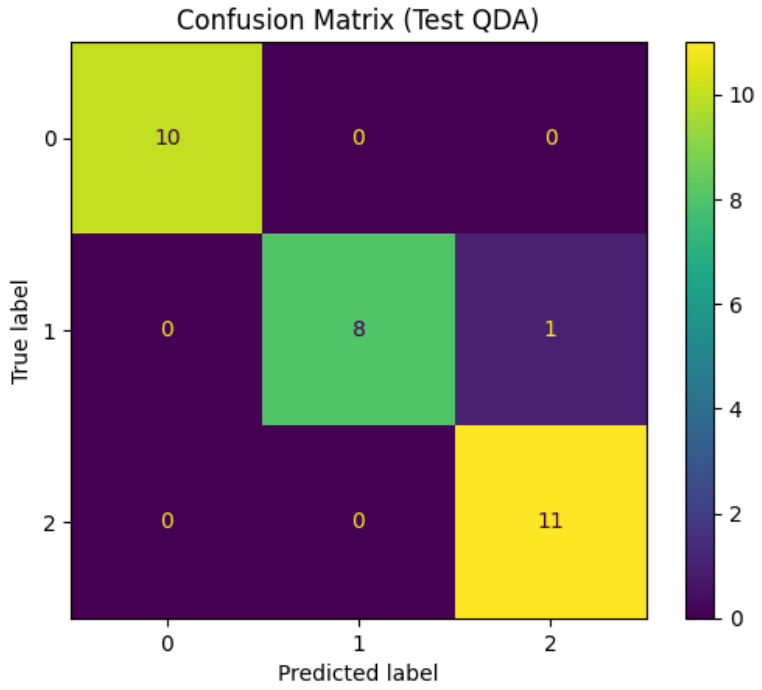
Performance Metrics (Test Set):

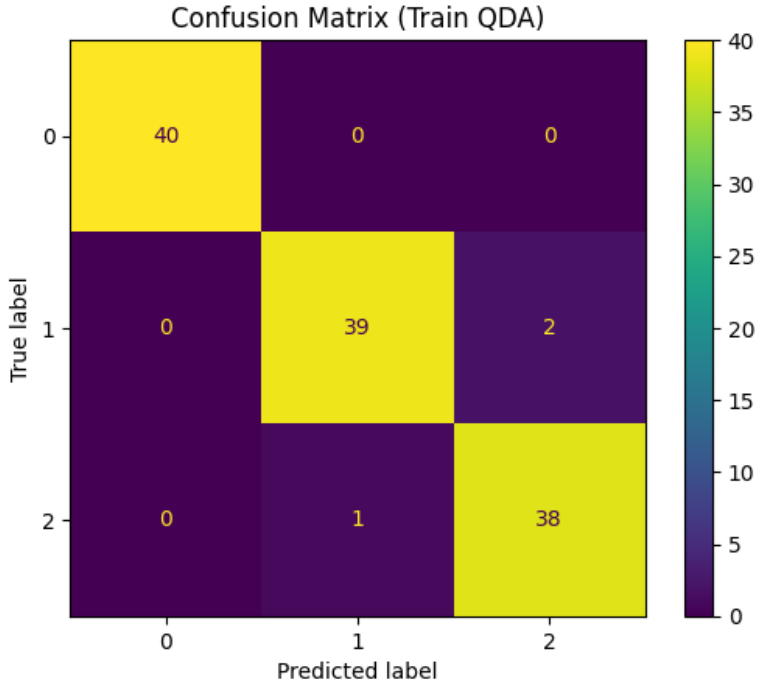
- **Accuracy:** 0.967 (96.7%)
- **Macro Average F1-Score:** 0.97

Comparison with LDA:

While LDA achieved 100% accuracy, QDA achieved approximately 96.7% accuracy with one misclassification on the test set. Although QDA is more flexible and can model different covariance matrices for each class, the linear separability of the dataset meant that the added complexity of QDA did not result in better performance compared to LDA in this specific experiment.

Class	Precision	Recall	F1-Score	Support
Sentosa	1.00	1.00	1.00	10
Versicolor	1.00	0.89	0.94	9
Virginica	0.92	1.00	0.96	11
Accuracy			**0.97**	30





3.4. Gaussian Naive Bayes (GNB) (Without PCA)

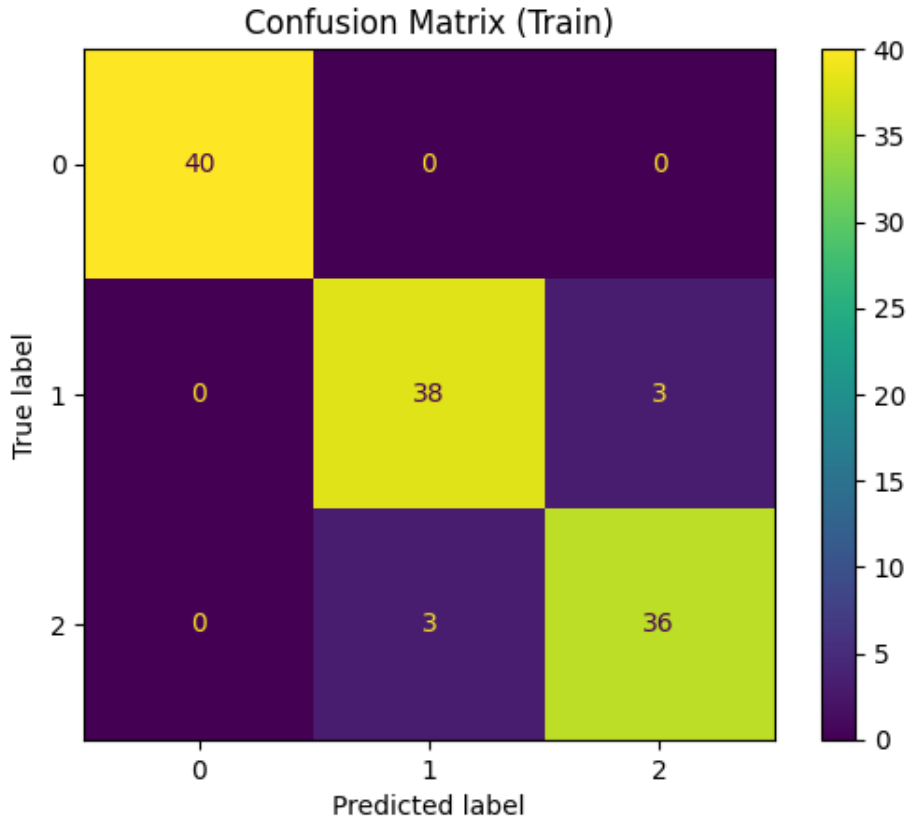
The Gaussian Naive Bayes model was trained assuming that the continuous features associated with each class are distributed according to a Gaussian (Normal) distribution.

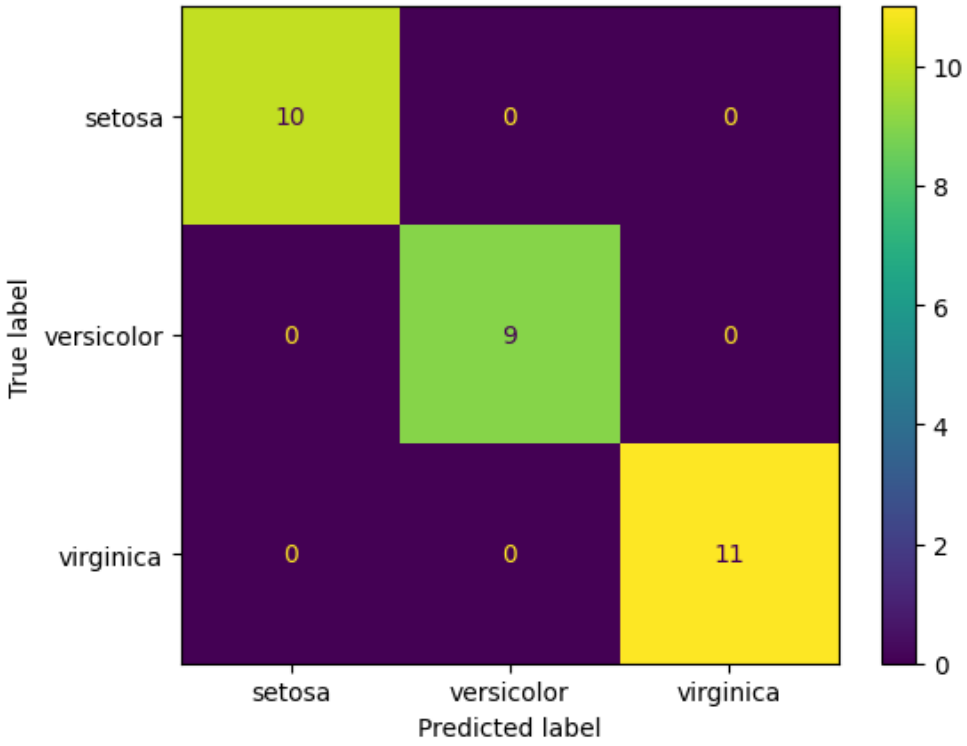
Performance Metrics (Test Set):

- **Accuracy:** 1.00 (100%)
- **Macro Average F1-Score:** 1.00

Classification Report (GNB): The model demonstrated perfect classification performance on the test set. All samples from the three classes (*Setosa*, *Versicolor*, and *Virginica*) were correctly identified with no errors.

Class	Precision	Recall	F1-Score	Support
Sentosa	1.00	1.00	1.00	10
Versicolor	1.00	1.00	1.00	9
Virginica	1.00	1.00	1.00	11
Accuracy			**1.00**	30





Analysis: The GNB model matched the performance of LDA and QDA. This indicates that the assumption of feature independence given the class, while theoretically a simplification, holds well enough for this dataset. Furthermore, it confirms that the feature distributions for each species are approximately normal.

3.5. Gaussian Naive Bayes (GNB) (With PCA)

To evaluate the impact of dimensionality reduction on classification performance, the Gaussian Naive Bayes classifier was also trained using only the first two principal components (PC1 and PC2) derived from the PCA step.

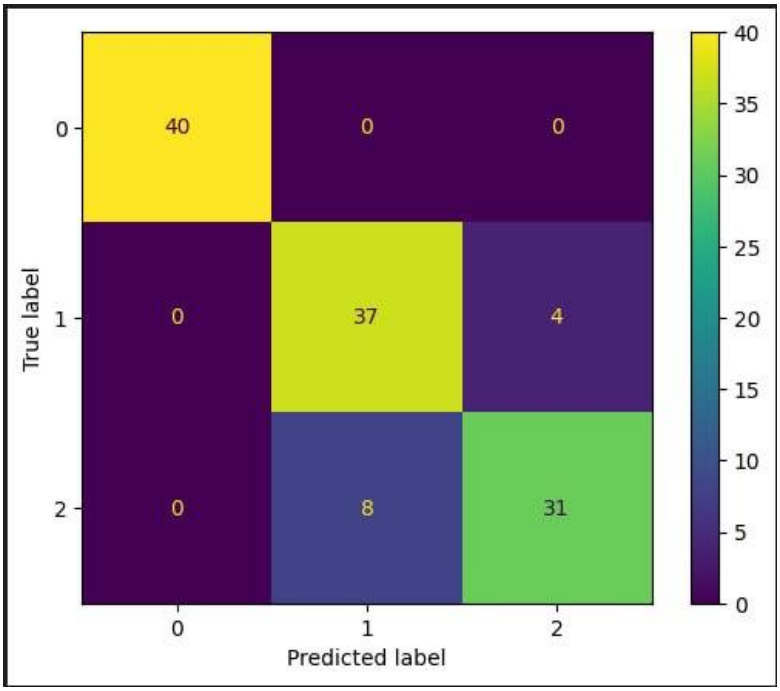
Performance Metrics (Test Set):

- **Accuracy:** 0.933 (93.3%)
- **Macro Average F1-Score:** 0.93

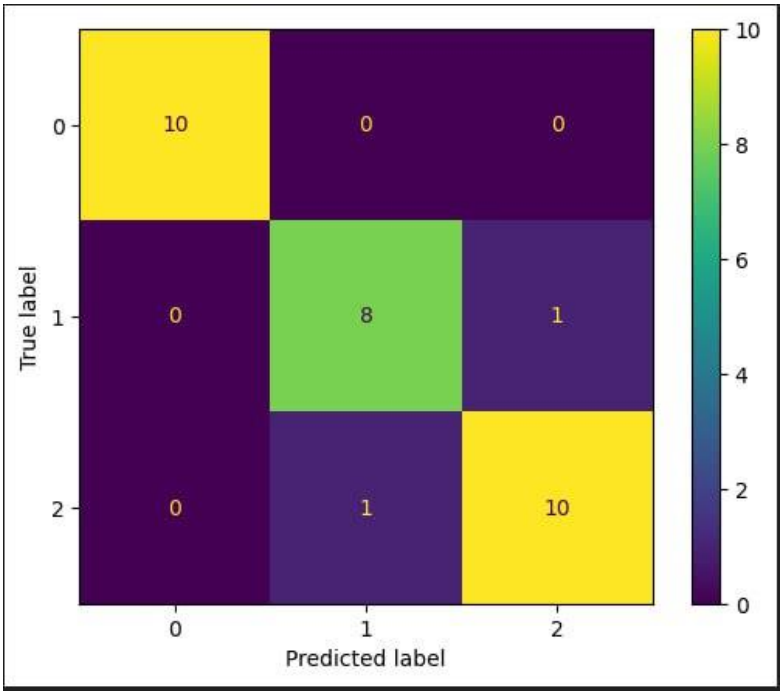
Classification Report: The reduction of features from 4 to 2 resulted in a slight drop in performance compared to the model trained on the full dataset. Specifically, the model struggled slightly to distinguish between Versicolor and Virginica in the reduced space.

Class	Precision	Recall	F1-Score	Support
Sentosa	1.00	1.00	1.00	10
Versicolor	0.89	0.89	0.89	9
Virginica	0.91	0.91	0.91	11
Accuracy			**0.93**	30

Train



Test



Observation & Analysis: While the GNB model on the full dataset achieved 100% accuracy, the accuracy dropped to ~93.3% when using PCA-reduced data. The Confusion Matrix revealed two misclassifications: one *Versicolor* misclassified as *Virginica* and one *Virginica* misclassified as *Versicolor*.

This demonstrates a trade-off: while PCA significantly reduced the complexity and dimensionality of the data (preserving 95.8% of variance), the loss of the remaining ~4% of information contained subtle details necessary for perfectly separating the overlapping classes (*Versicolor* and *Virginica*). However, 93% accuracy on just 2 dimensions is still a robust result.

4. Discussion

Model Comparison: The Linear Discriminant Analysis (LDA) and Gaussian Naive Bayes (GNB) classifiers trained on the full feature set achieved perfect accuracy (100%) on the test set. Quadratic Discriminant Analysis (QDA) followed closely with 96.7% accuracy, misclassifying only one sample. However, when the GNB model was applied to the PCA-reduced data (only 2 dimensions), the accuracy decreased to 93.3%.

Impact of Dimensionality Reduction: The comparison between GNB on full features (100%) and GNB on PCA components (93.3%) highlights a crucial trade-off. While the first two principal components capture approximately 96% of the variance, the remaining ~4% contains subtle information necessary for perfectly distinguishing between the overlapping classes (*Versicolor* and *Virginica*). This suggests that for this specific dataset, retaining all original features is beneficial for maximum accuracy, although the reduced model still performs robustly.

Model Selection: Given the performance analysis:

- LDA is the preferred choice for this dataset as it achieved perfect accuracy with a simple linear model.
- GNB (Full Features) is equally effective but relies on the assumption of feature independence.
- GNB (PCA) offers a good balance if visualization or extreme data compression is required, despite the slight loss in accuracy.

- QDA was slightly less accurate and effectively “overkill” for this problem, as linear boundaries were sufficient.

5. Conclusion

This study successfully classified the Iris species using LDA, QDA, and GNB. The analysis demonstrated that applying these statistical methods to the full feature set is sufficient to achieve perfect classification accuracy (100% for LDA and GNB). Furthermore, the experiment with PCA revealed that while dimensionality reduction is powerful for visualization, it resulted in a minor drop in classification performance (to 93.3%), confirming that the original four features collectively provide the best separability for the Iris species.