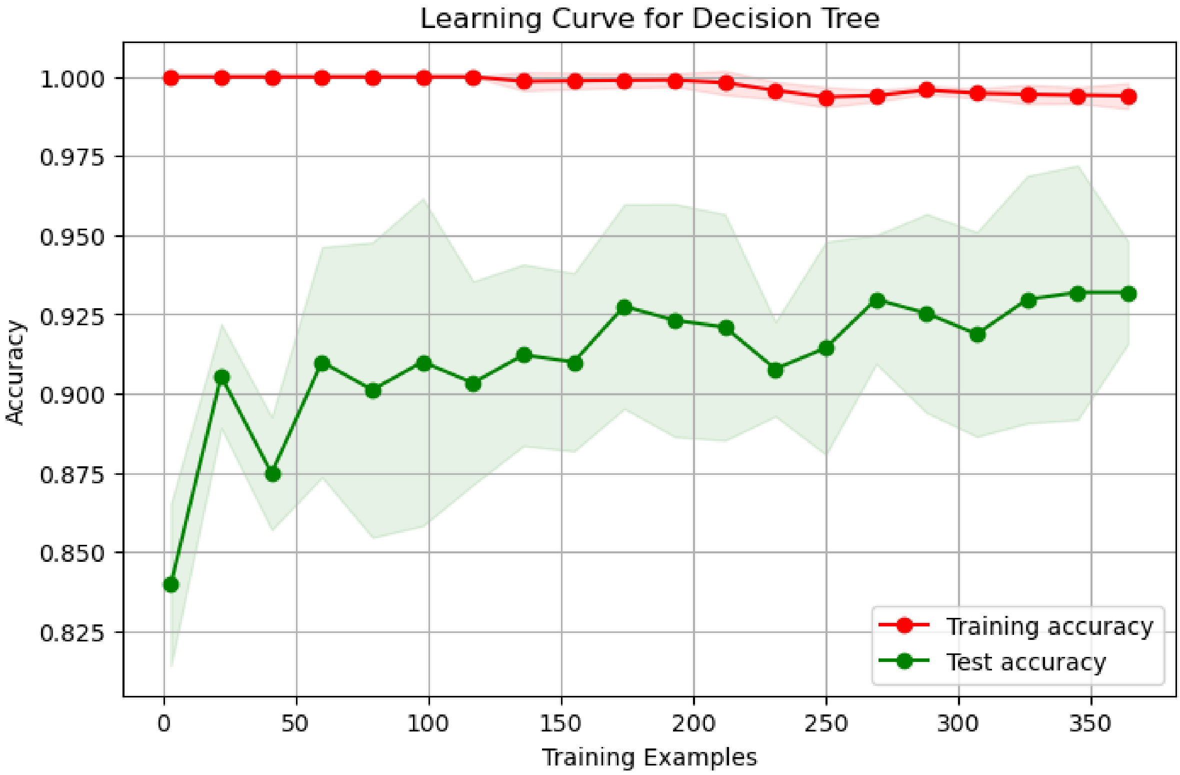
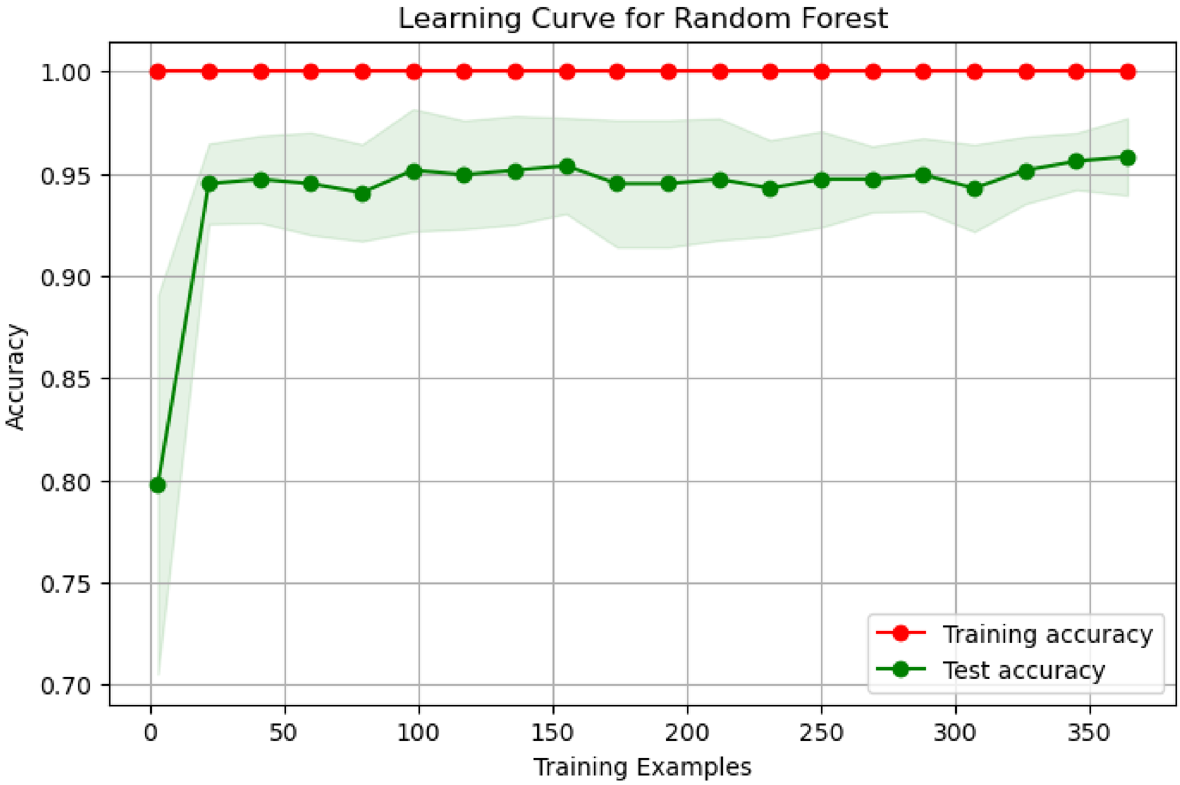
Report

2023148056 이서현

1.1

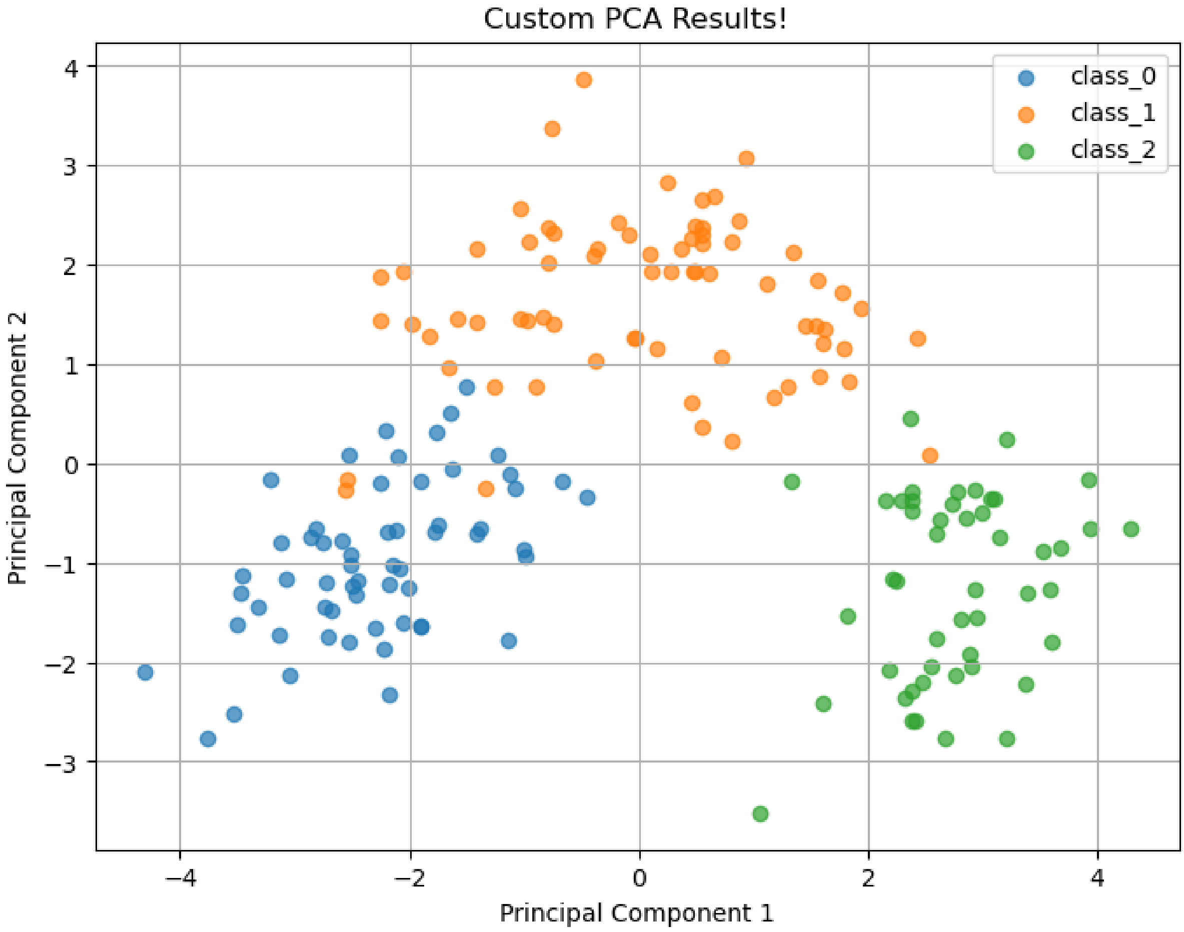


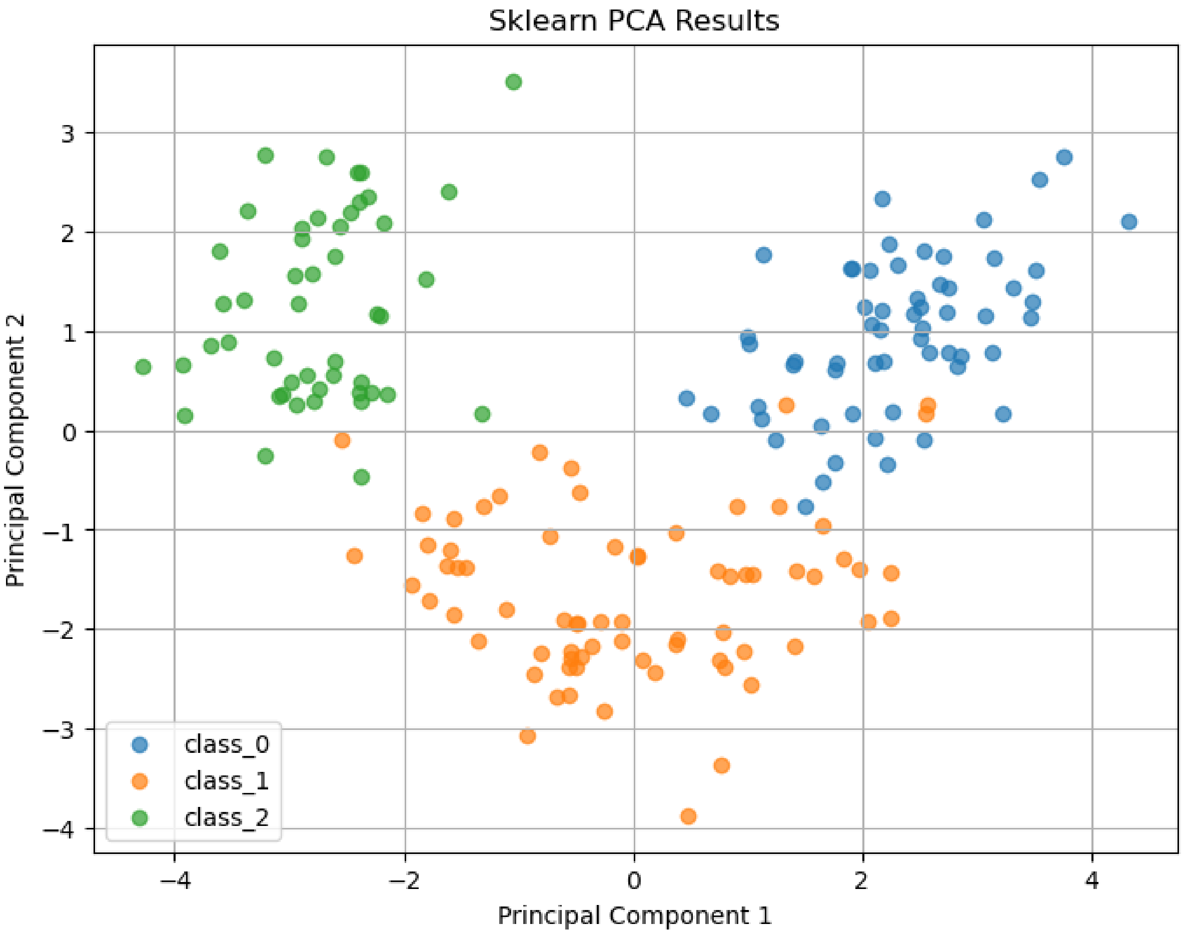


1.2

The Random Forest model outperformed the Decision Tree in both training and test accuracy. While the Decision Tree demonstrated strong performance, its slightly lower test accuracy indicates that it is more prone to overfitting compared to the Random Forest model.

2.1





2.2

1. Benefits of PCA

Dimensionality Reduction: PCA reduces the number of features while retaining the most important information, making data analysis more efficient.

Noise Reduction: It helps remove noise from data by focusing on principal components that explain the majority of variance.

Improved Model Performance: By reducing redundancy, PCA can improve the performance of machine learning models and reduce overfitting.

2. Disadvantages of PCA

Loss of Interpretability: The transformed components are linear combinations of original features, making them less interpretable.

Linear Assumption: PCA assumes linear relationships between features, which may not capture complex patterns in the data.

Sensitive to Scaling: PCA is sensitive to the scale of the data, requiring proper standardization beforehand.

3. Alternative Dimensionality Reduction Method

t-SNE (t-Distributed Stochastic Neighbor Embedding):

Benefits: t-SNE is highly effective for visualizing high-dimensional data in 2D or 3D space by preserving local data structure.

Drawbacks: It is computationally expensive and not ideal for feature selection or large datasets.

3.1

Hard Margin SVM:

Assumes that the data is perfectly linearly separable.

Finds the maximum margin hyperplane with no tolerance for misclassification.

Advantages:

Simple to implement when data is perfectly separable.

Guarantees a unique solution if separability is maintained.

Disadvantages:

Fails if the data is noisy or not linearly separable.

Highly sensitive to outliers.

Soft Margin SVM:

Allows for some misclassification by introducing a slack variable to handle noisy or non-linearly separable data.

Finds the best balance between maximizing the margin and minimizing misclassification errors.

Advantages:

Can handle noisy and non-linearly separable data.

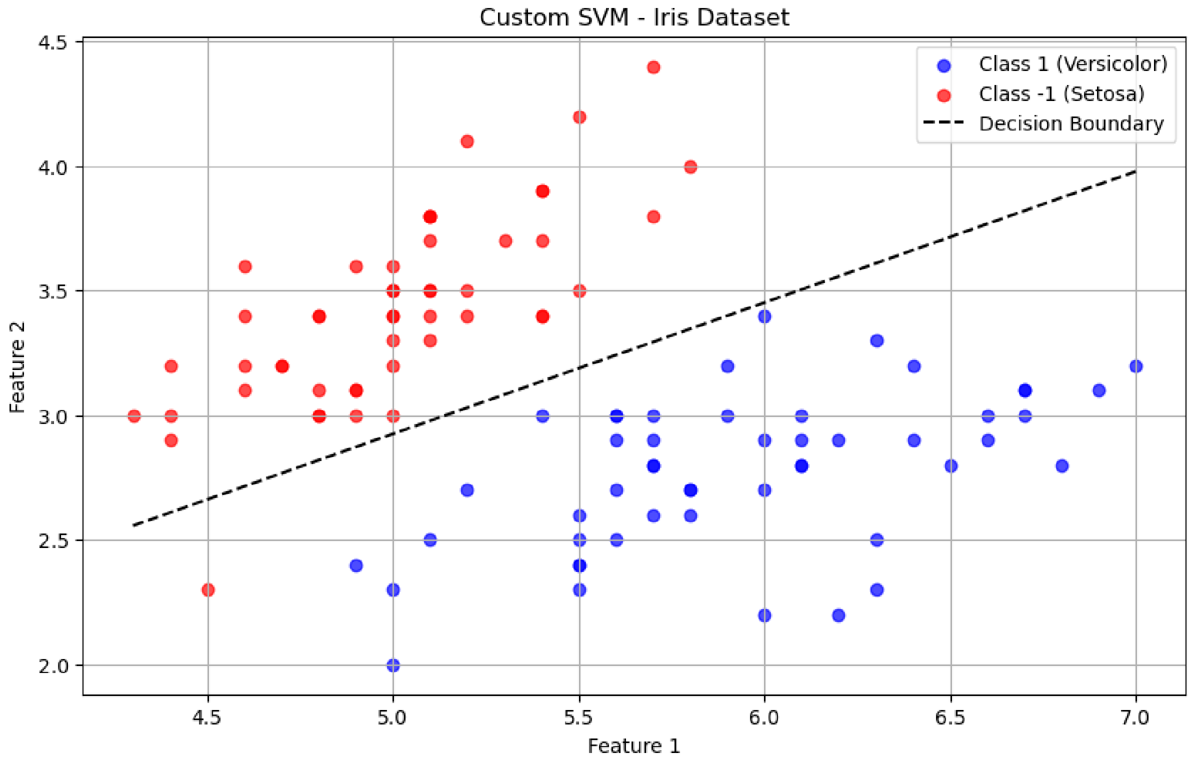
Provides flexibility through the regularization parameter CC.

Disadvantages:

More complex to tune (requires selecting CC).

Risk of overfitting or underfitting depending on the choice of CC.

3.2



Custom SVM - Iris Dataset
Class 1 (Versicolor)
Class -1 (Setosa)
Decision Boundary
Feature 2
Feature 1
4.5
4.0
3.5
3.0
2.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0