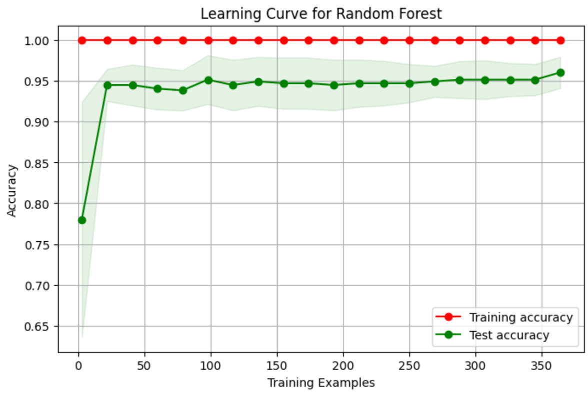
ML Report

26기 조석희

1.1)



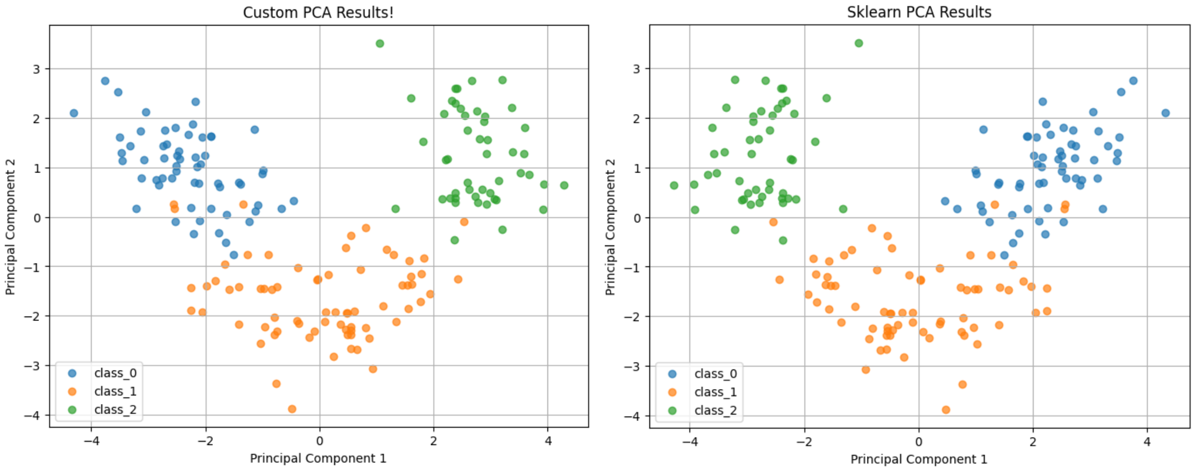
Decision Tree - Training Accuracy: 0.9934, Test Accuracy: 0.9298

Random Forest - Training Accuracy: 1.0000, Test Accuracy: 0.9561

1.2)

The better model is Random Forest. First, the test accuracy of RF is higher than that of the DT, indicating that RF generalizes better to unseen data. Second, RF's learning curve is smoother, with higher and more stable test accuracy, indicating the model's robustness and consistency.

2.1)



2.2)

PCA reduces the number of features while retaining as much variability in the data as possible, making computations faster and less resource-intensive. It also removes multicollinearity by transforming correlated features into uncorrelated principal components. But the transformed features are linear combinations of the original features, making it difficult to interpret their meaning in the context of the original data.

One alternative to PCA is t-SNE (t-Distributed Stochastic Neighbor Embedding). t-SNE is a non-linear dimensionality reduction technique that is particularly effective for data visualization by embedding high-dimensional data into a lower-dimensional space.

3.1)

A Hard Margin SVM is designed to find a hyperplane that perfectly separates the data into two classes, with no misclassifications allowed. This method requires the data to be linearly separable, as it does not tolerate any points falling within the margin or on the wrong side of the hyperplane. The primary advantage of Hard Margin SVM is its simplicity and high accuracy when the data is perfectly separable. However, it is not robust to outliers; a single outlier can drastically affect the decision boundary. Additionally, its applicability is limited to datasets that are linearly separable, which is often not the case in real-world scenarios.

On the other hand, a Soft Margin SVM introduces flexibility by allowing some misclassification through the use of slack variables. This method penalizes data points that violate the margin, enabling it to handle datasets that are not linearly separable. The main advantage of Soft Margin SVM is its ability to handle noisy data and outliers better than Hard Margin SVM. It is also applicable to a wider range of datasets, as it does not require perfect separability. However, Soft Margin SVM is computationally more expensive due to the added complexity of balancing the margin maximization with the penalty for violations. Moreover, it requires careful tuning of the regularization parameter C which determines the trade-off between maximizing the margin and minimizing the classification error.

3.2)