ML 과제

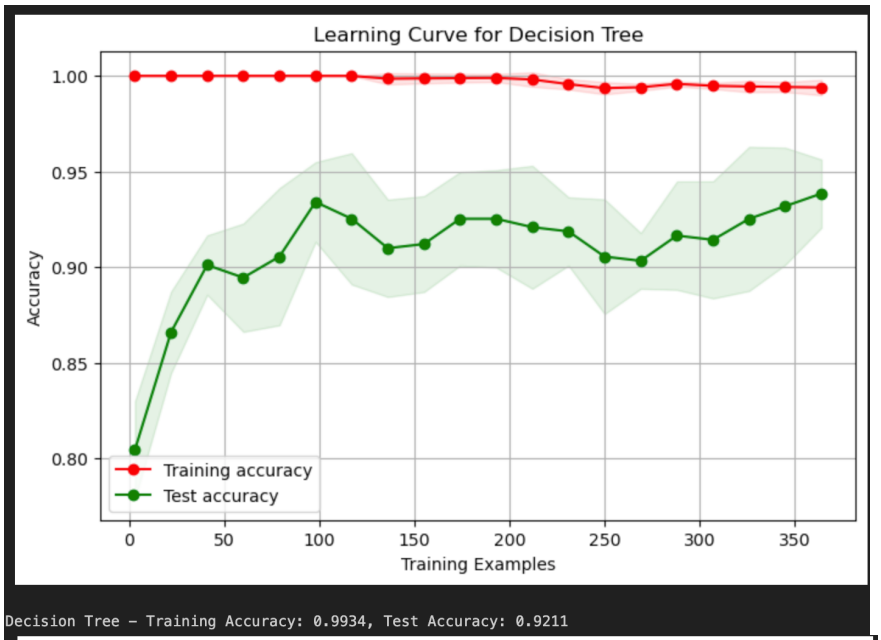
2021122063 이재영

1.1

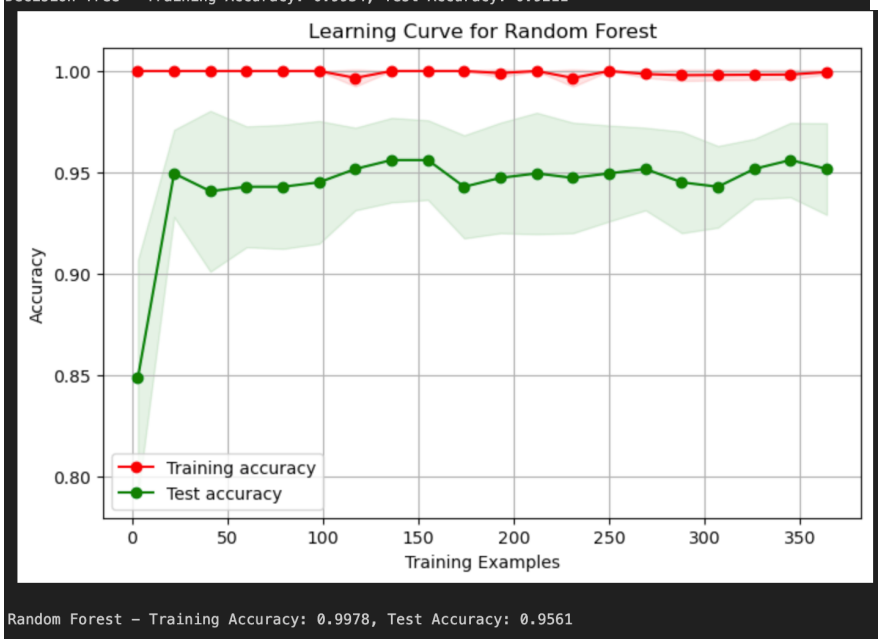
Decision Tree : max_depth 5 , training accuracy 0.9934, test accuracy 0.9211

Random Forest : max_depth 100, n_estimator 20,

train accuracy 0.9978, test accuracy 0.9561




```
Decision Tree – Training Accuracy: 0.9934, Test Accuracy: 0.9211
```

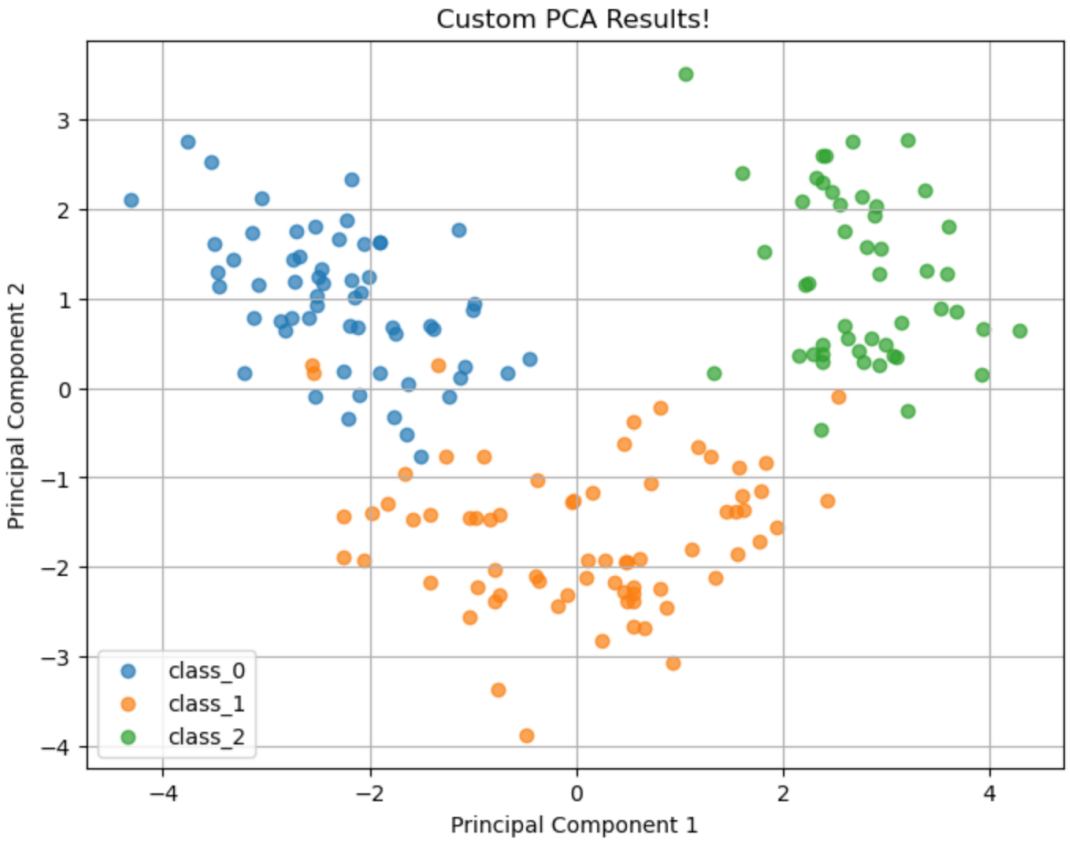


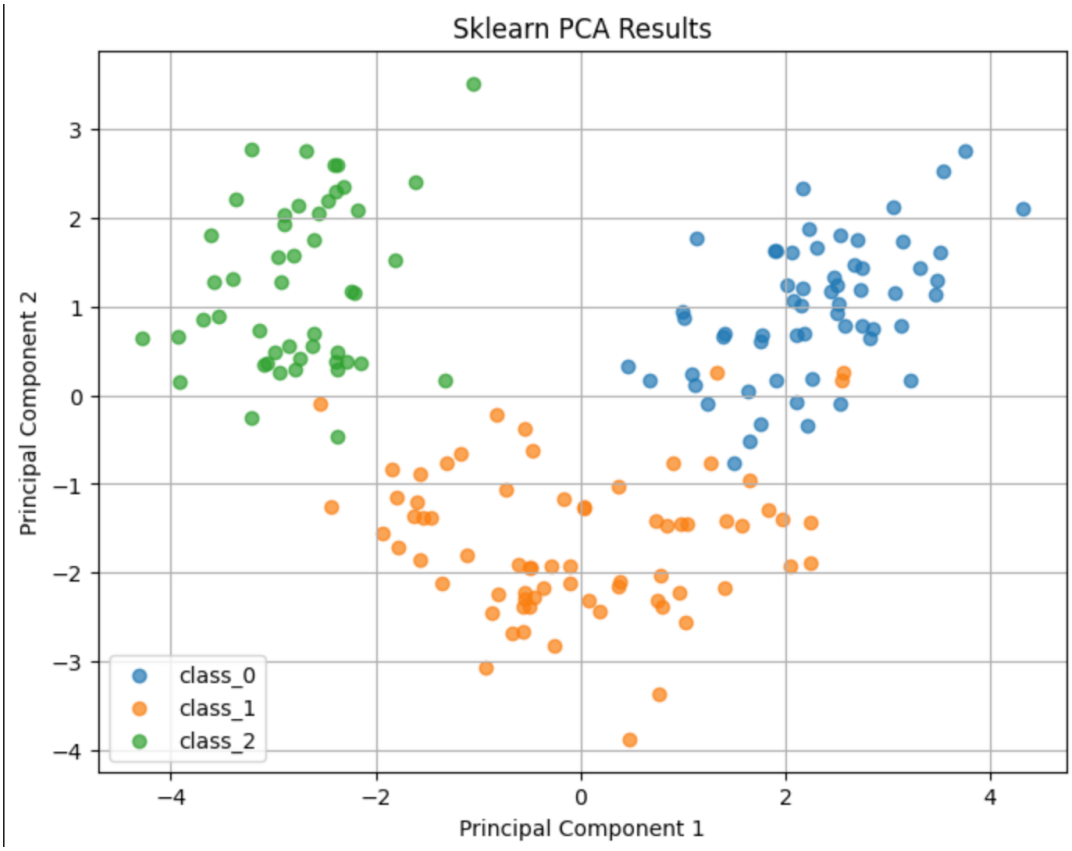

```
Random Forest – Training Accuracy: 0.9978, Test Accuracy: 0.9561
```

1.2

From our results, it is evident that the random forest model outperformed the decision tree. It is easy to know this since ensemble approach utilized by random forest, where multiple decision trees are trained on subsets of the data, and their predictions are aggregated to achieve better generalization and robustness. By examining the test accuracy, we observe that the random forest consistently achieved higher performance, indicating its superior ability to handle overfitting and variability in the data compared to a single decision tree.

2.1





2.2

The benefit of dimension reduction is that we can simplify the data into lower dimension making it easier to analyze while also reducing computational requirements. However it will always have a jump from the original data resulting in loss of information. There is also t-SNE to reduce dimension. Simple difference from PCA is that it does non-linear dimension reduction. Which preserves local relationship inside the data.

3.1

hard margin SVM tries to perfectly separate the data where else soft margin SVM tries to generalize the classification even though having possibility to not perfectly separating the data. So, if we have noise free data that can be perfectly separated hard margin SVM would be a good way to maximize the margin. However, if the data can't be linearly separated, using soft margin SVM will be flexible to generalize the separation for the data. The trade-off is that the soft margin SVMs may sometimes overfit to noisy data if the penalty parameter is not appropriately tuned.

3.2