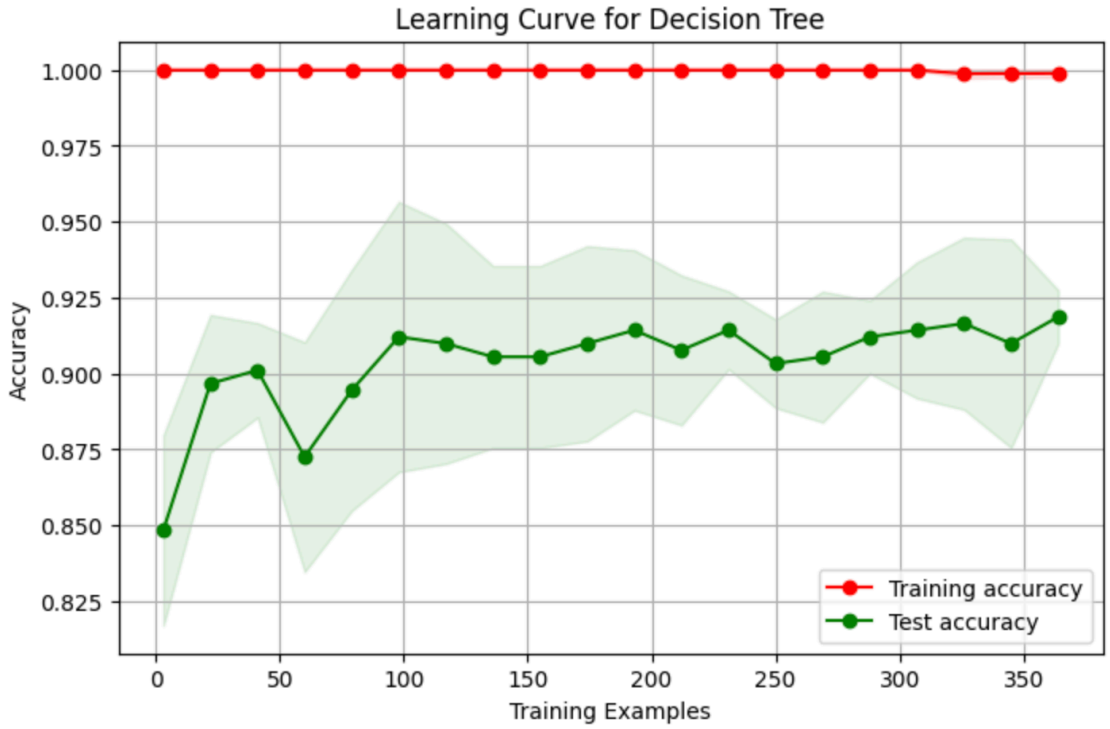
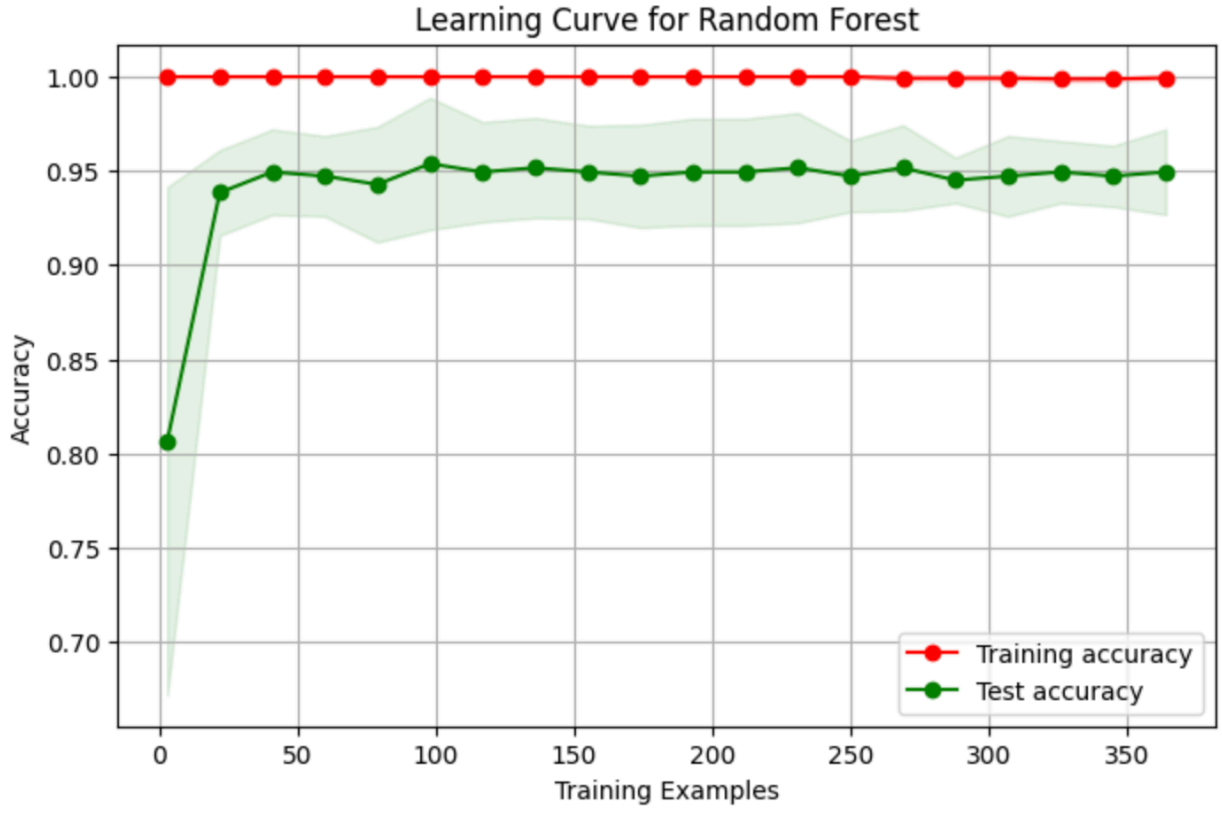
1) Ensemble Method




```
Decision Tree - Training Accuracy: 1.0000, Test Accuracy: 0.9035
```




```
Random Forest - Training Accuracy: 0.9978, Test Accuracy: 0.9474
```

What is the better model and please provide evidence and supporting arguments that back your decision? State your answer in the report.pdf.

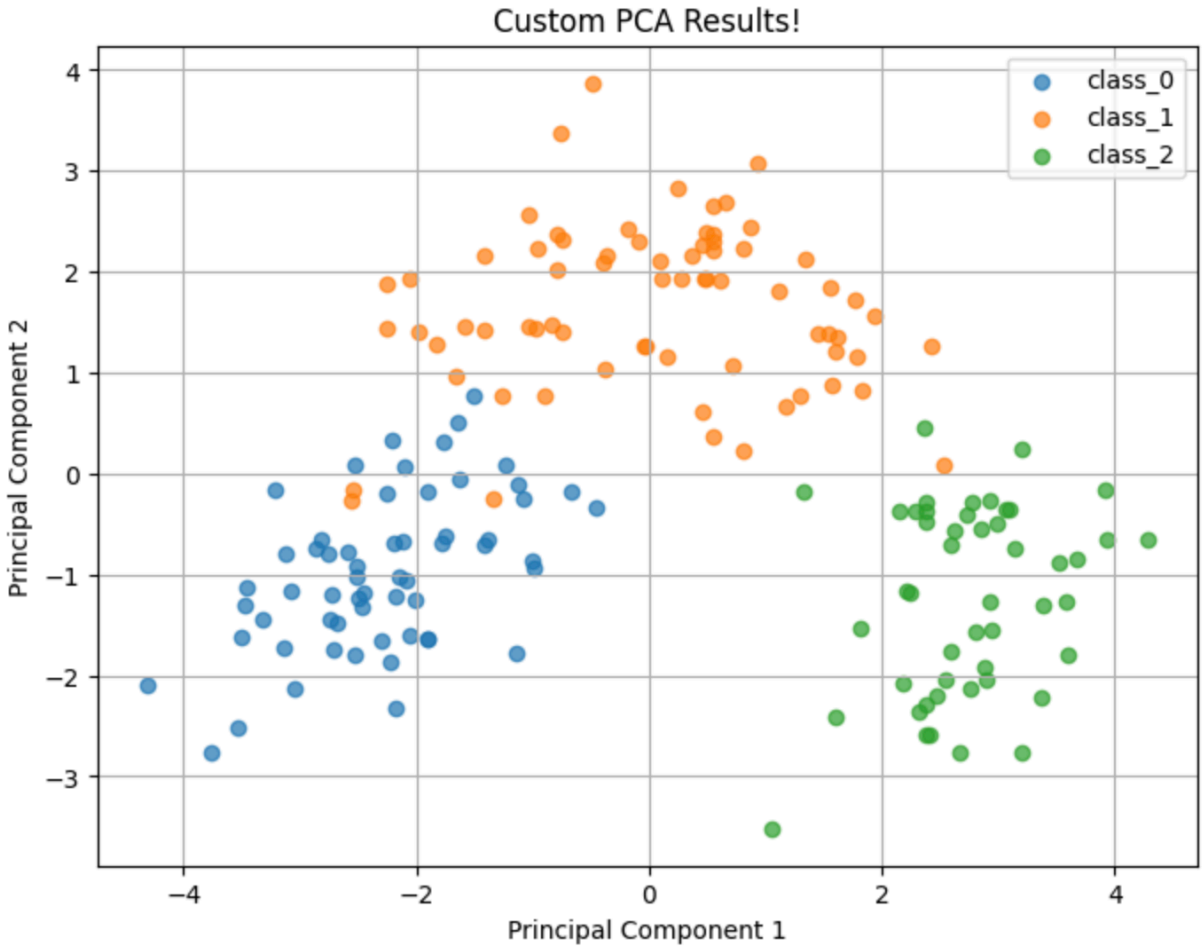
Depth 7기준으로 random forest모델이 더 낫다

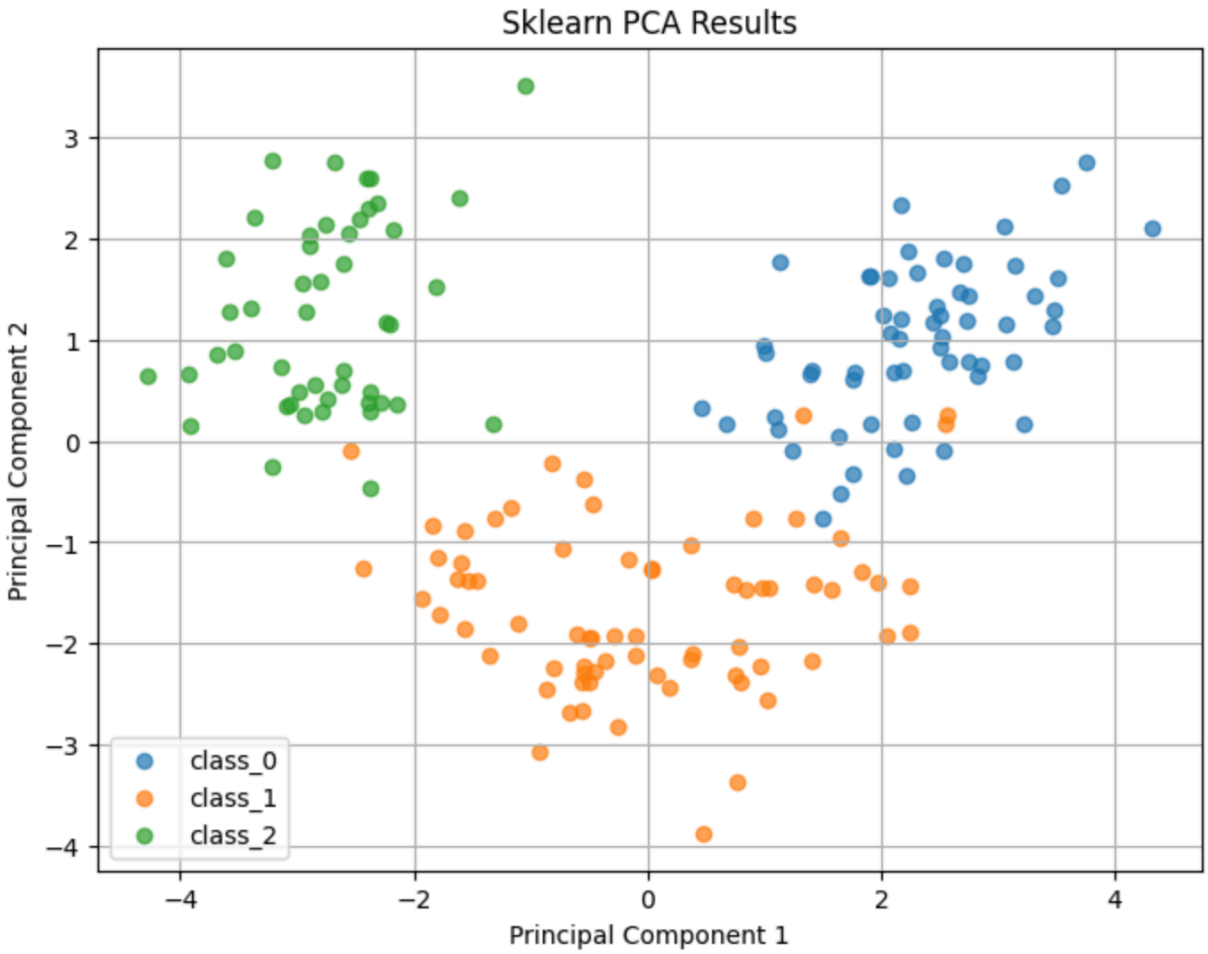
Accuracy:

- decision tree: 1.00(training) 0.9035(test)
- Random forest: 0.9978(training), 0.9474(test)

random forest가 test accuracy가 더 높기도 하고, 둘 다 overfitting된 경향이 있지만 decision tree에 비하면 살짝 덜하다. 추가호 그래프 상으로도 random forest의 learning curve가 더 안정적이다.

2) PCA





What are the benefits and the disadvantages of PCA? Could you provide another dimensionality reduction method that can be used apart from PCA? State your answer in the report.pdf.

Benefit: 연산 효율성 증가, sparsity 제거 가능

Disadvantages: 정보 손실 가능성 존재. 선형성을 가정하기에 비선형 구조에 적용하기 어려움.

Autoencoder

- 비선형 데이터에도 효과적으로 작용하며, 구조를 학습한 뒤 데이터를 축소시킴

3) SVM

What is the difference between a soft margin SVM and hard margin SVM? Furthermore, can you provide advantages and disadvantages of both methods? State your answer in the report.pdf.

Soft margin

- Advantages: 노이즈와 이상치에 강함
- Disadvantages: support vertor의 수 증가 -> decision boundary 결정하는 데 많은 sample이 반영됨 -> underfitting 가능성 존재

Hard margin

- Advantages: 선형 분리가 가능한 데이터면 거의 완벽하게 분리 가능
- Disadvantages: support vector 수 감소 -> overfitting 우려됨