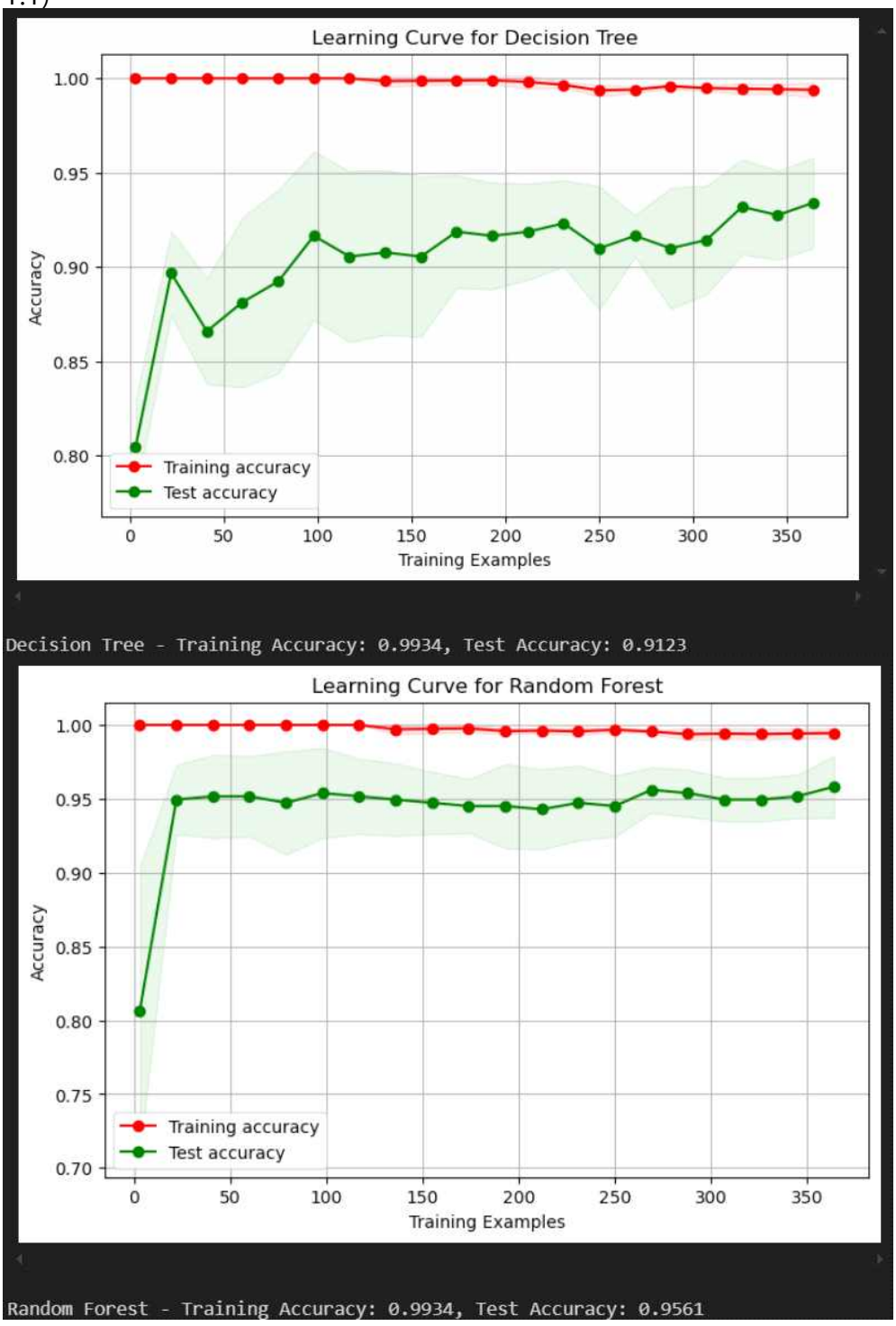
1.1)

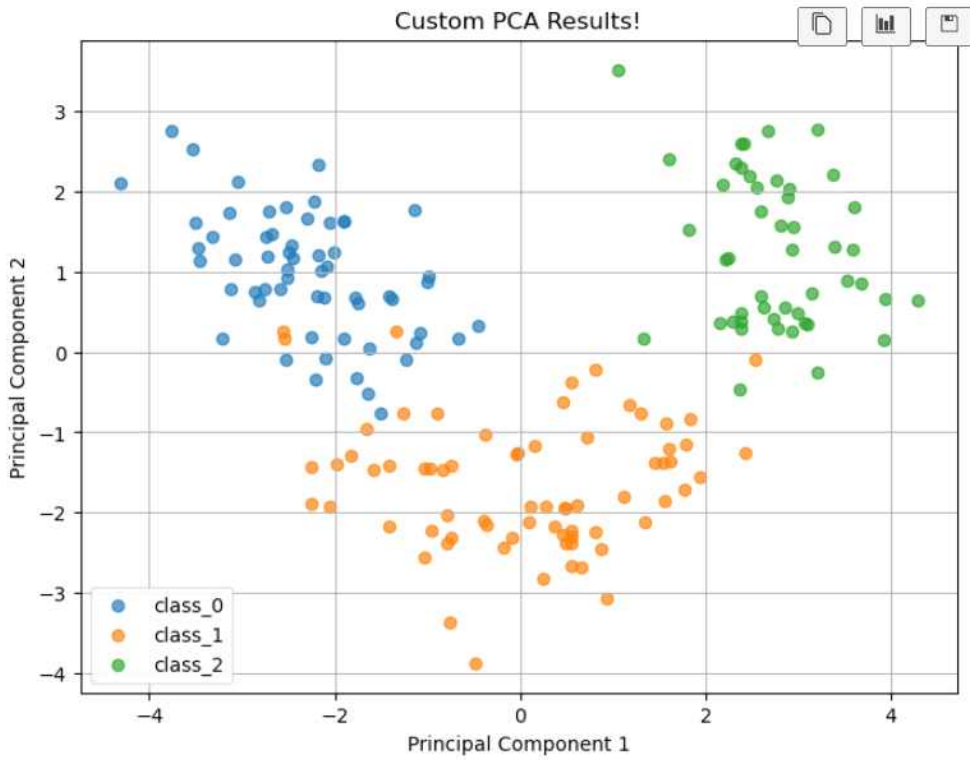


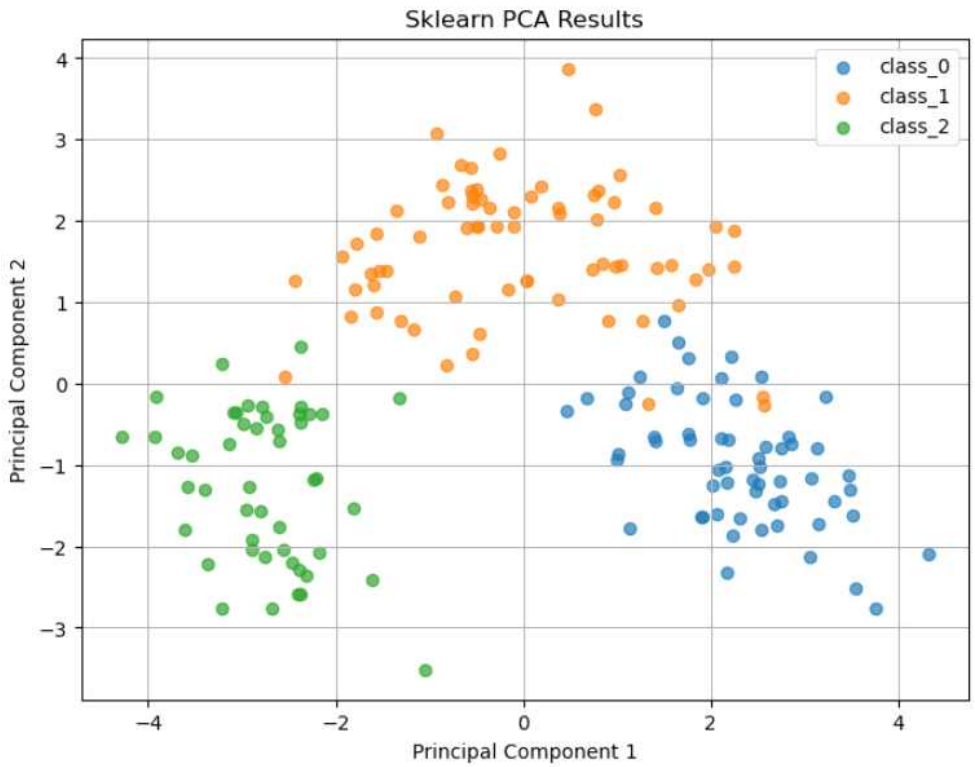

1.2)

아래의 Random Forest가 더 나은 모델이라고 생각합니다. train accuracy는 두 모델 모두 99.3%로 동일하지만, test accuracy에서 Random Forest가 95.6%로 약 4.4% 더 높은 정확도를 보여주었습니다. 또한, learning curve를 살펴보면, Decision Tree의 테스트 정확도 곡선은 진동하며 불안정한 모습을 보이는 반면, Random Forest는 테스트 정확도 곡선이 안정적으로 수렴하는 모습을 확인할 수 있습니다. 이는 여러 개의 Decision Tree를 앙상블하여

분산을 줄이고, 단일 Decision Tree가 가질 수 있는 overfitting 문제를 효과적으로 완화했습니다.

2.1)



2.2)
PCA는 데이터의 차원을 줄여주고, 최대한의 분산 즉, 정보를 유지하면서 차원을 줄여주는데, 이는 모델의 학습 속도 향상에 기여할 수 있습니다. 그리고 분산이 적은 즉, 중요도가 낮은 차원을 제거할 수 있기 때문에, 다중공선성과 중복된 특성을 제거하여 머신러닝 모델의 성능을 향상시킬 수 있습니다.
하지만, PCA는 데이터가 선형적이라는 가정을 기반으로 하기 때문에, 비선형적인 구조를 가진 데이터에서는 성능이 저하될 수 있으며, 차원을 축소하는 과정에서 일부 정보의 손실을 감안해야하는 단점이 있습니다.
PCA를 대체하여 쓸 수 있는 차원 축소 기법에는 t-SNE 방식이 있는데, 이는 PCA가 놓칠 수 있는 비선형 관계를 잡아낼 수 있다는 장점이 있습니다. 물론 t-SNE 방식은 PCA에 비해 더 많은 계산량이 필요해 느리다는 단점이 있습니다.

3.1)
Hard Margin SVM는 데이터가 완벽히 선형적으로 분리 가능한 경우를 가정하고, 모든 데이터를 올바르게 분류하면서 2개의 클래스 사이의 마진을 최대화하는 hyperplane을 찾습니다. 그렇기 때문에, 데이터가 선형적으로 완벽히 분리가능하다면 완벽한 분류를 보장하고 클래스 사이에 마진을 최대화하여 가장 안정적인 결정 경계를 제공한다는 장점이 있지만, 노이즈와 이상치에 민감하다는 단점이 있고, 데이터가 선형적으로 분리되지 않는 경우에는 작동하지 않는다는 단점이 있다.
Soft Margin SVM는 일부 데이터가 잘못 분류되거나 마진 내에 위치하는 것을 허용하고, 선형적으로 분리되지 않는 데이터에도 유연하게 대처할 수 있습니다. 선형적으로 분리되지 않는 데이터에도 적용이 가능하며 Hard Margin SVM보다 노이즈와 이상치에 대해 robust하며 일반화 성능이 우수하는 장점이 있습니다. 하지만 마진과 분류 오류 간의 밸런스를 조정하는 람다라는 하이퍼파라미터를 튜닝해야한다는 단점이 있습니다.

3.2)