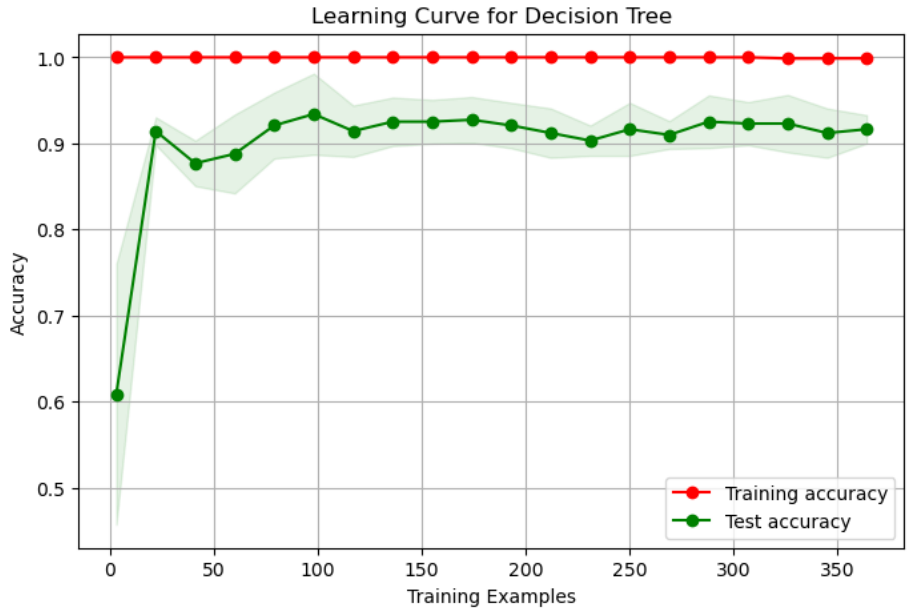
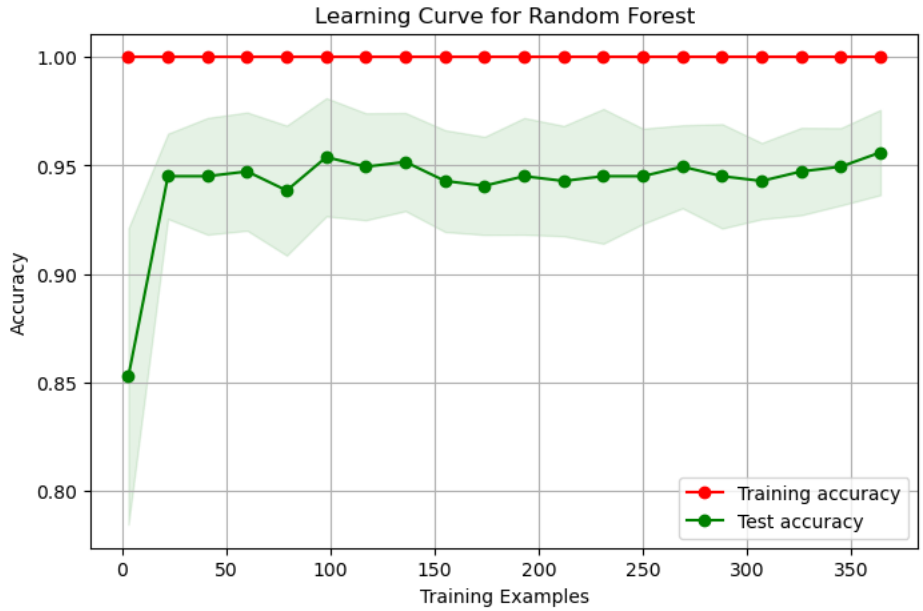
1.1) Decision Tree - Training Accuracy: 1.0000, Test Accuracy: 0.8947

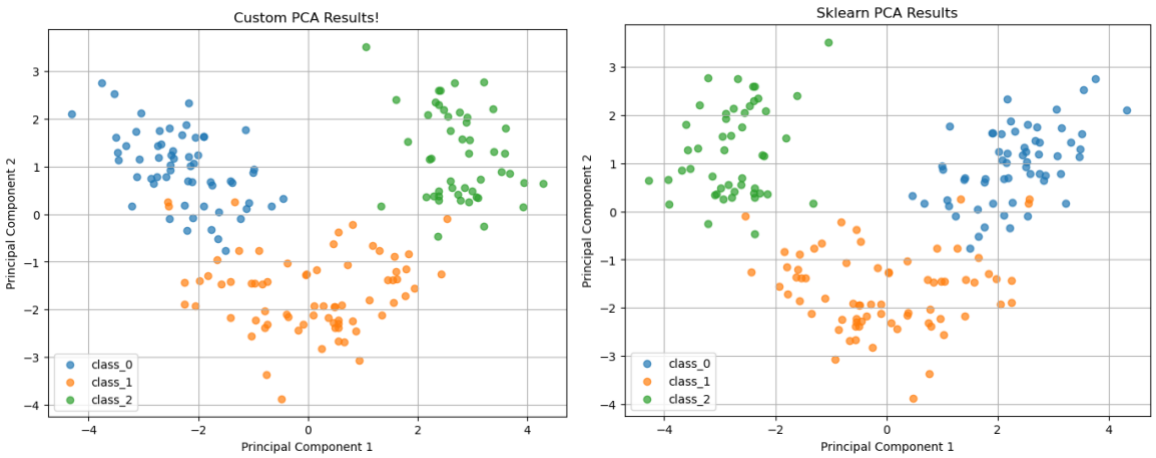


Random Forest - Training Accuracy: 0.9978, Test Accuracy: 0.9561



1.2) 두 모델 모두 학습 초기의 Test accuracy 상승률이 빠르다. 그러나 DT 의 경우 0.9 주변에서 정확도가 움직이고, RF 의 경우가 더 높은 정확도 구간인 0.95 주변에서 움직이기 때문에 RF 가 더 좋은 모델이라고 말할 수 있다.

2.1)



2.2)

PCA 의 장단점

차원 축소를 통해 PCA 이후 적용할 모델에서 과도한 연산량을 줄이고, 더 낮은 차원에서 데이터를 효율적으로 표현할 수 있다.

그러나 PCA 는 데이터가 선형 형태일 때 효과적이고 비선형 구조일 때에는 중요한 정보를 잘 포착하지 못한다. 또한 데이터에 스케일링을 진행하지 않는 경우에는 잘못된 결과를 낳을 가능성이 있다.

이외 차원 축소 방법론으로는 t-SNE 가 있다. 고차원 데이터를 저차원 공간으로 임베딩할 때 확률적 방법론을 활용한다. 클러스터 형성을 잘 시각화할 수 있으며, PCA 와 달리 비선형 구조를 잘 포착할 수 있다는 장점이 있다.

3.1)

하드 마진 SVM 의 가정은 데이터가 완전히 선형으로 분리될 수 있다고 가정한다. 즉 오분류 없이 모든 샘플을 정확히 분류할 수 있는 초평면을 찾는데, 이 때 오차가 0 이 되도록 결정경계를 학습한다. 이에 반해 소프트 마진 SVM 은 현실적인 데이터의 잡음이나 중복성 등을 고려하여 데이터가 완전히 분리가 불가능하다는 점을 고려한다. 이를 위해 약간의 오분류를 허용하면서 마진을 최대화한다.

하드 마진 SVM 의 장점은 데이터에 대해 정확도가 매우 높으며 모델이 단순하다는 점이 있다. 단점으로는 현실적으로 실제 데이터는 완전한 선형 분리가 어려우며, 모델이 잡음에 매우 민감해서 오버피팅이 일어나는 등 모델 불안정성의 위험이 있다.

소프트 마진 SVM 은 일부 outlier 가 있어도 오분류를 어느정도 허용하여 일반화 성능을 높이고, 모델을 안정적으로 학습할 수 있다는 장저을 가진다. 단점은 오분류 허용도를 하이퍼파라미터로 조정해야 하고, 모델이 복잡하며 최적 파라미터를 찾기 위해 search 의 과정이 더 필요하다는 점이 있다.

3.2)