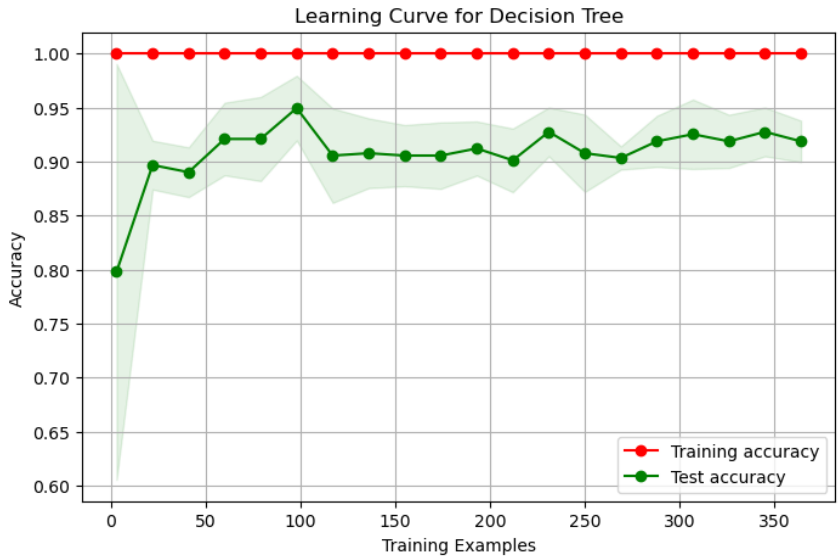
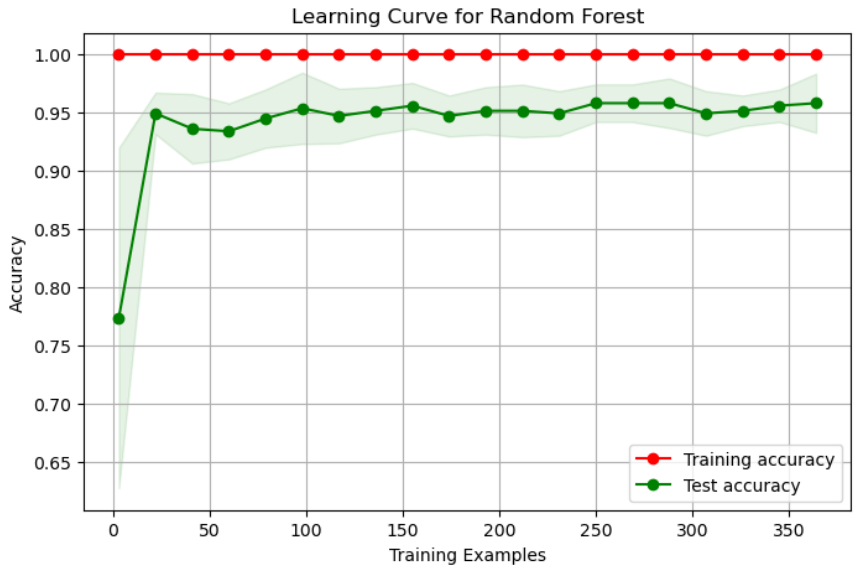
1.1

(1) Decision Tree



Training Accuracy: 1.0000 (100%) / Test Accuracy: 0.8860 (88.6%)

(2) Random Forest



Training Accuracy: 1.0000 (100%) / Test Accuracy: 0.9386 (93.86%)

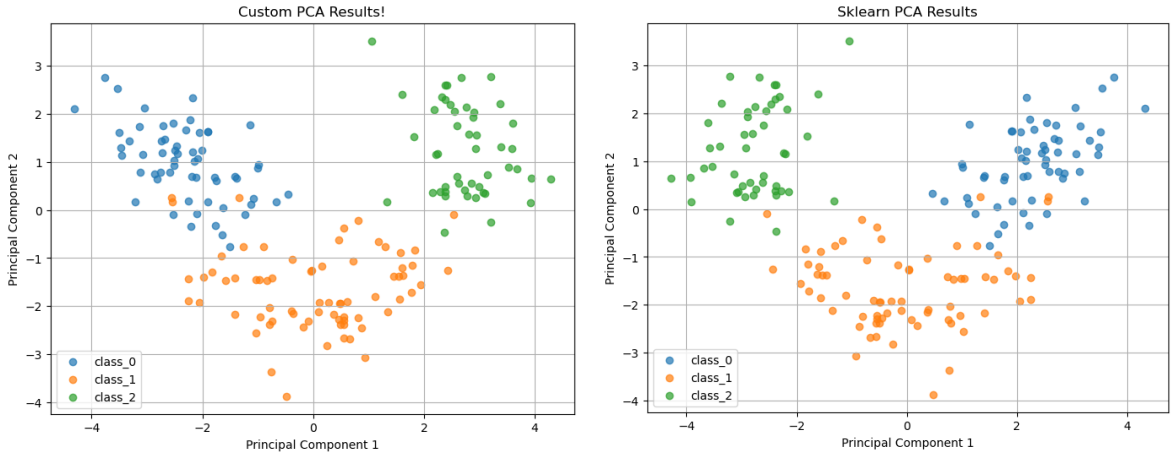
1.2

Random Forest 모델이 더 나은 모델로 판단됩니다.

우선 Random Forest의 테스트 정확도(0.9386)가 Decision Tree(0.8860)보다 더 높습니다. 이는 Random Forest가 더 일반화 성능이 좋고, 새로운 데이터를 더 정확히 예측한다는 것을 의미합니다. 또한, Training Accuracy와 Test Accuracy의 차이를 통해서 Decision Tree는 과적합이 발생하기 쉬운 알고리즘임을 알 수 있습니다. Training Accuracy가 1.0000인 반면, Test Accuracy는 0.8860으로 감소한 것은 학습 데이터에 너무 집중되었음을 보여주기 때문입니다. 반면, Random Forest는 여러 Decision Tree를 조합하여 과적합을 방지하며, 테스트 데이터에서도 더 좋은 성능을 보였습니다.

그래프적으로도, Random Forest의 테스트 정확도 그래프는 Decision Tree보다 안정적이고 높은 정확도를 유지합니다. Decision Tree의 테스트 정확도는 불안정하고 편차가 크지만, Random Forest는 분산을 줄이는 앙상블 기법의 효과로 인해 일관된 결과를 제공한 것으로 보여집니다.

2.1



2.2

우선 PCA의 가장 큰 장점은 차원 축소입니다. 이를 통해 고차원 데이터를 저차원으로 변환하여 분석과 시각화를 더 쉽게 만들 수 있고. 데이터가 차원이 높아질수록 발생하는 "차원의 저주(Curse of Dimensionality)"를 완화할 수 있습니다. 또한, 데이터의 분산이 작은 축(주로 노이즈로 간주)을 제거하여 데이터 품질을 개선하거나 데이터의 차원을 줄임으로써 머신러닝 모델의 학습 및 예측 시간을 단축할 수 있습니다.

반면 PCA는 주성분을 선형 결합으로 변환하기 때문에, 결과 변수가 원래 변수와의 명확한 해석이 어렵다는 단점을 가지고 있습니다. 또한 PCA는 데이터 간의 선형적인 관계만 반영하므로, 비선형적인 패턴을 설명하기에는 한계가 있습니다. 그 외에도 일부 주성분만 선택하면 원래 데이터의 정보가 일부 손실될 수 있다는 점, 데이터가 정규화되지 않으면 결과가 왜곡될 수 있다는 단점을 가지고 있습니다.

PCA를 대체할 수 있는 대표적인 차원 축소 기법은 t-SNE입니다. t-SNE는 고차원 데이터를 저차원(주로 2D 또는 3D)으로 변환하여 시각화에 뛰어난 성능을 발휘하며 데이터 포인트 간의 지역적 구조를 잘 유지하여, 데이터 클러스터(군집) 간의 관계를 효과적으로 보여줍니다. 그러나 계산 비용이 높아 대규모 데이터에는 비효율적이며 결과가 랜덤이어서, 매번 다른 결과를 생성할 수 있다는 단점을 갖고 있습니다.

3.1, 3.2

Soft Margin SVM과 Hard Margin SVM은 SVM의 두 가지 방법으로, 데이터 특성과 상황에 따라 다르게 사용됩니다. Hard Margin SVM은 데이터가 선형적으로 완벽히 구분 가능하다는 가정 하에 작동하며, 모든 데이터 포인트가 결정 경계 밖에 위치하도록 학습합니다. 이 방식은 노이즈가 없고 선형적으로 구분 가능한 경우에 적합하지만, 현실적인 데이터에서는 과적합(overfitting)이 발생하거나 유연성이 부족하다는 단점이 있습니다.

반면, Soft Margin SVM은 현실적인 데이터셋에서 노이즈나 일부 예외를 허용하도록 설계되었습니다. 데이터를 완벽히 구분하지 않아도 되며, 일부 오차를 허용하면서도 결정 경계를 학습합니다. 이로 인해 노이즈가 포함된 데이터에서도 잘 작동하며, 일반화 성능이 더 우수합니다. 하지만 하이퍼파라미터를 적절히 설정해야 하고, 계산량이 더 많아질 수 있다는 단점이 있습니다.

결론적으로, Hard Margin SVM은 단순하고 계산이 효율적이지만, 현실적인 데이터에서는 Soft Margin SVM이 유연성과 안정적인 성능 면에서 더 적합한 방식으로 널리 사용됩니다.