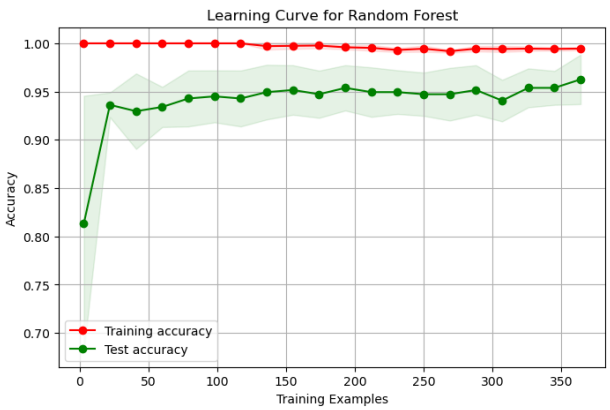
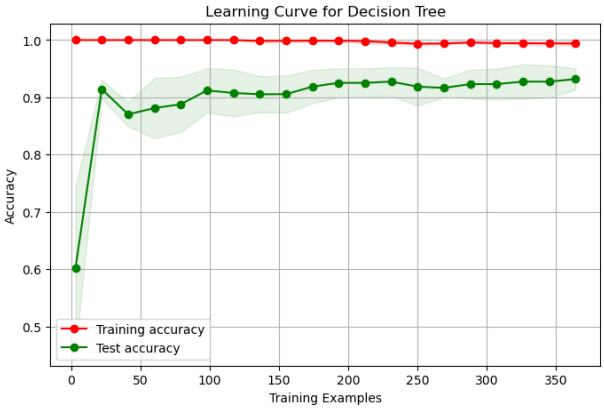
6 회차 ML Report

1

1.1)





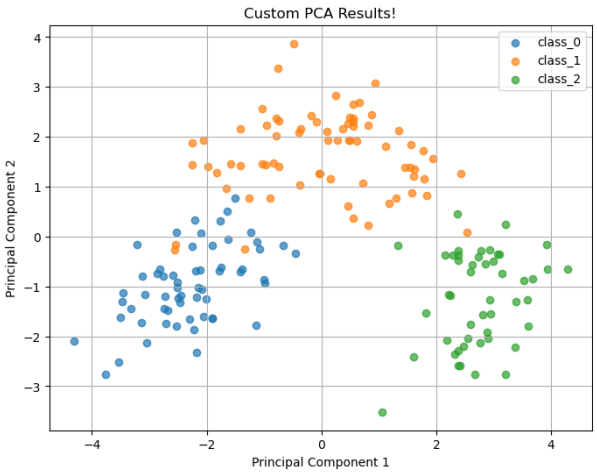
1.2)

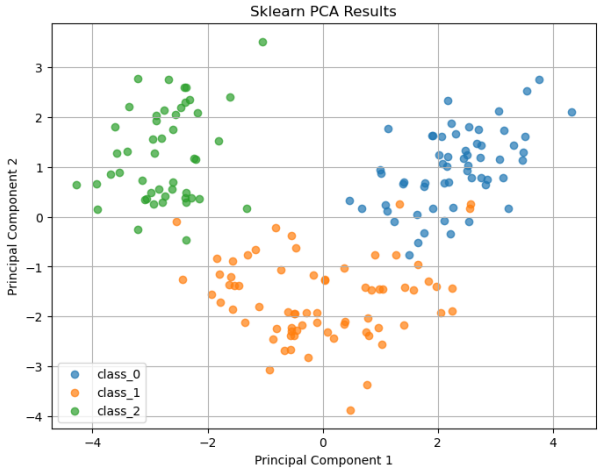
두 모델 중 더 나은 모델 : Random Forest

이유 : Decision tree 은 test dataset 에서 0.9 부근에서 test accuracy 가 형성되지만 random forest 는 더욱 정확도가 높으며, training accuracy 는 Decision tree 가 더 높으나 과적합으로 인하여 test dataset 에서는 정확도가 더 낮음을 볼 수 있다. 이는 random forest 에서 여러 개의 트리를 앙상블하여 과적합 문제를 비교적 해결했기 때문으로 생각된다.

2

2.1)





2.2)

PCA 의 장점: 데이터의 차원 감소 / 최대한의 분산을 유지하며 차원 감소 (모델 계산 속도 향상)

다중공산성 제거 / 모델 성능 향상

PCA 의 단점: 정보가 손실되어 모델 성능이 오히려 하락할 수 있음

비선형적인 구조에서 성능 저하 가능

PCA 이외의 차원 축소 방법 : t-SNE LDA 등

3

3.1) 두 모델의 차이점: 데이터가 선형적으로 완전히 분리 가능한지 여부

-> 하드 마진의 경우 오차가 없는 마진을 설정하여 데이터와 마진 경계 사이의 거리를 최대화하는 것이 목표 / 마진 내 데이터 없음

-> 소프트 마진의 경우 말그대로 부드럽게 선형적으로 분리되지 않는 경우를 허용 이 때 규제 파라미터 C 를 통해 마진과 오차의 균형을 조절 가능 / 마진 내 일부 데이터 허용 / 이상치에 강건

3.2)