Machine Learning 과제 보고서

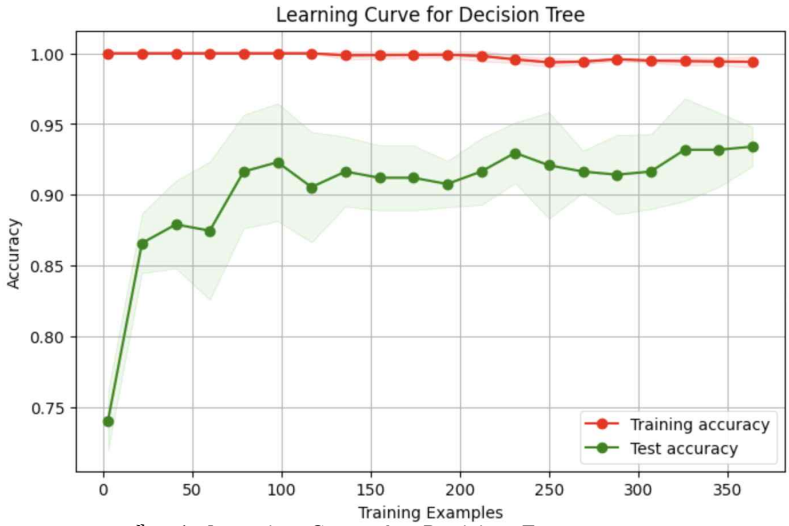
2022148014 엄윤희

1. Ensemble Method (앙상블 방법)

1.1 Decision Tree vs. Random Forest

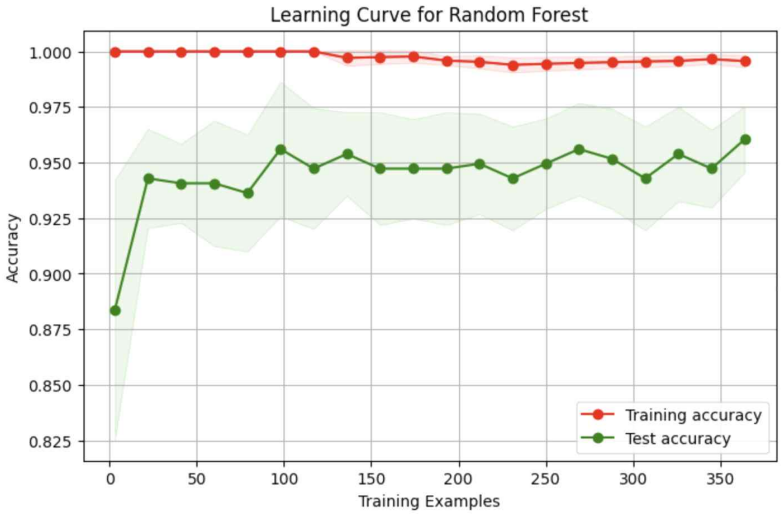
앙상블 기법 중에서 결정 트리(Decision Tree) 모델과 랜덤 포레스트(Random Forest) 모델을 사용하여 유방암 데이터셋에 대해 이진 분류를 수행했습니다. 학습곡선 및 정확도 평가를 통해 두 모델의 성능을 비교했습니다.

첫 번째 그래프는 결정 트리 모델에 대한 학습곡선입니다. 학습 초기에는 훈련 정확도와 테스트 정확도 모두 낮지만, 데이터가 증가함에 따라 정확도가 점차 상승합니다. 훈련 정확도는 1에 가까운 값을 유지하며, 테스트 정확도는 안정적으로 증가하는 모습을 보입니다.



그래프 1: Learning Curve for Decision Tree

두 번째 그래프는 랜덤 포레스트 모델에 대한 학습곡선입니다. 랜덤 포레스트는 여러 결정 트리를 결합하여 예측하는 모델로, 결정 트리보다 더 안정적인 성능을 보여줍니다. 훈련 정확도와 테스트 정확도 모두 높은 값을 유지하고 있으며, 훈련 정확도의 과적합을 방지하는 경향을 보입니다.



그래프 2: Learning Curve for Random Forest

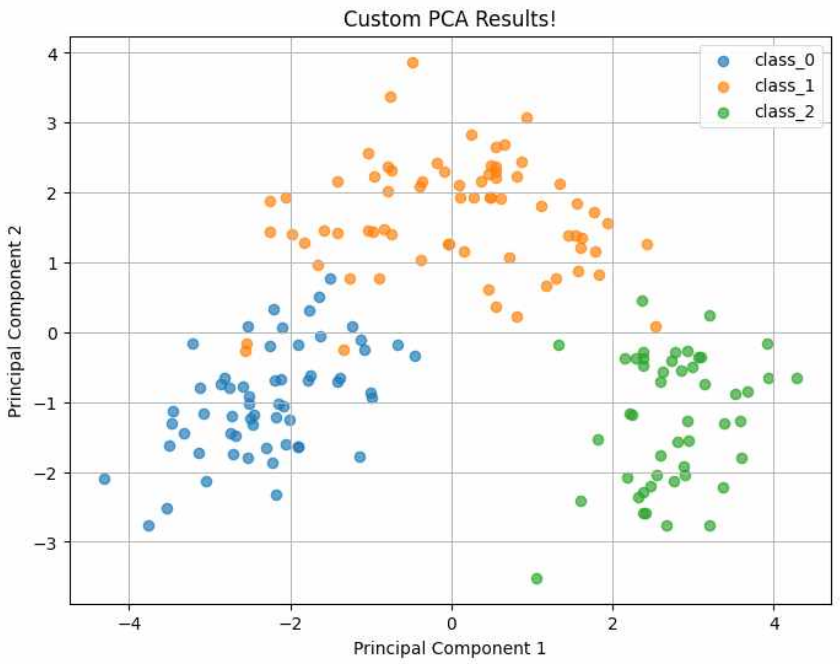
1.2 모델 비교

결정 트리는 학습 데이터에 대해 과적합이 발생하는 경향이 있으며, 테스트 데이터에 대한 정확도가 약간 낮은 모습을 보였습니다. 랜덤 포레스트는 여러 트리를 결합하여 상대적으로 높은 정확도와 안정적인 성능을 보여주었습니다. 따라서, 랜덤 포레스트(Random Forest) 모델이 결정 트리보다 더 나은 성능을 보였으며, 더 적합한 모델이라고 판단됩니다.

2. PCA (주성분 분석)

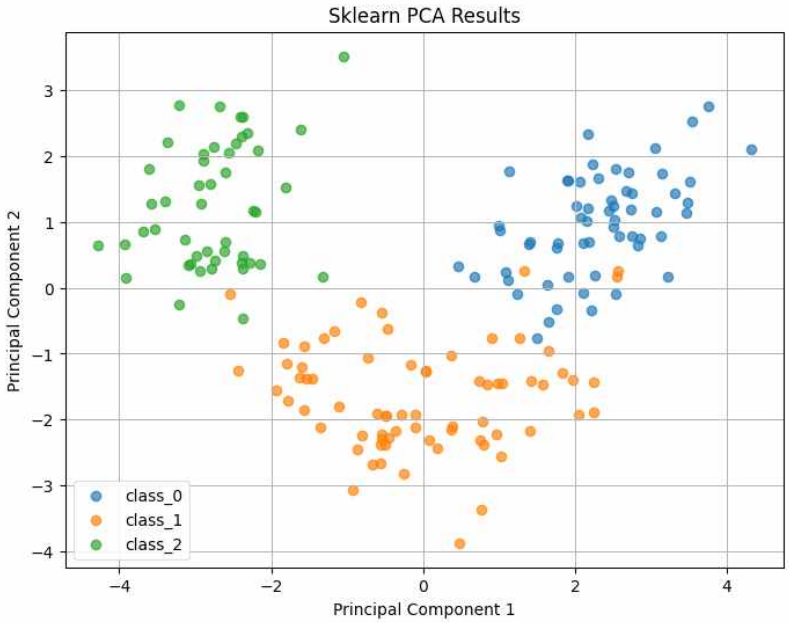
PCA는 고차원 데이터를 저차원으로 축소하여 데이터를 시각화하고 분석하기 위한 방법입니다. 여기서는 와인 데이터셋을 사용하여 주성분 분석을 수행하고, 차원 축소된 데이터에 대한 결과를 시각화했습니다.

직접 구현한 PCA를 통해 데이터를 2차원으로 축소하여 클래스별로 분포를 시각화했습니다. 각 클래스는 잘 구분되며, PCA가 차원 축소 후에도 데이터의 구조를 잘 보존한 것을 확인할 수 있었습니다.



그래프 3: Custom PCA Results

Sklearn의 PCA 라이브러리를 사용하여 동일한 데이터를 차원 축소한 결과입니다. sklearn을 사용한 결과 역시 세 클래스가 명확히 구분되며, 두 PCA 방법 간의 차이는 거의 없는 것을 확인할 수 있었습니다.

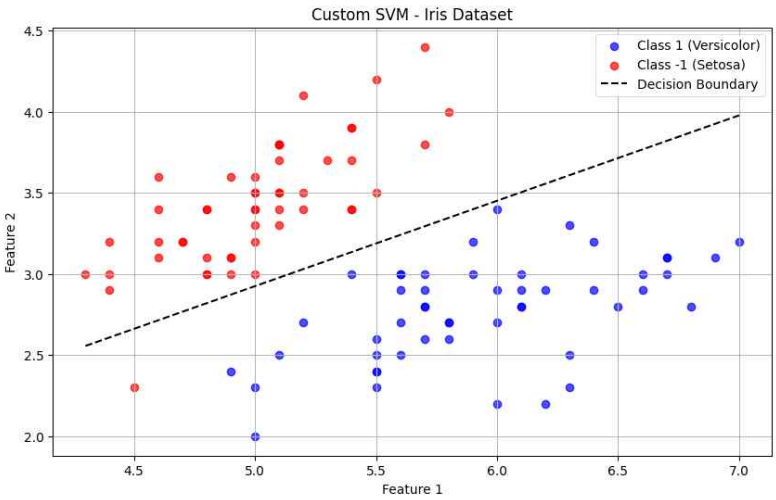


그래프 4: Sklearn PCA Results

3. SVM (서포트 벡터 머신)

서포트 벡터 머신(SVM)은 주어진 데이터에 대해 최적의 결정 경계를 찾는 모델입니다. 이번 과제에서는 이진 분류를 위해 Setosa와 Versicolor 두 클래스를 선택하여 하드 마진 SVM을 구현했습니다.

SVM 모델을 적용한 후, 결정 경계를 시각화한 결과, 두 클래스는 명확히 구분되었으며, 모델은 두 클래스 간의 최적의 경계를 찾아냄을 확인할 수 있었습니다.



그래프 5: Custom SVM - Iris Dataset

각 모델에 대한 학습곡선과 성능을 분석한 결과 랜덤 포레스트 모델이 결정 트리보다 더 높은 정확도와 안정성을 보여주었습니다. PCA는 차원 축소 후에도 데이터의 구조를 잘 보존하여, 고차원 데이터를 시각화하는 데 효과적이었습니다. SVM은 두 클래스의 구분을 명확하게 잘 수행하여 좋은 성능을 보였습니다. 이 결과들을 바탕으로 각 모델을 적용하는 데 있어서, 랜덤 포레스트와 SVM이 가장 효율적인 선택임을 알 수 있었습니다.