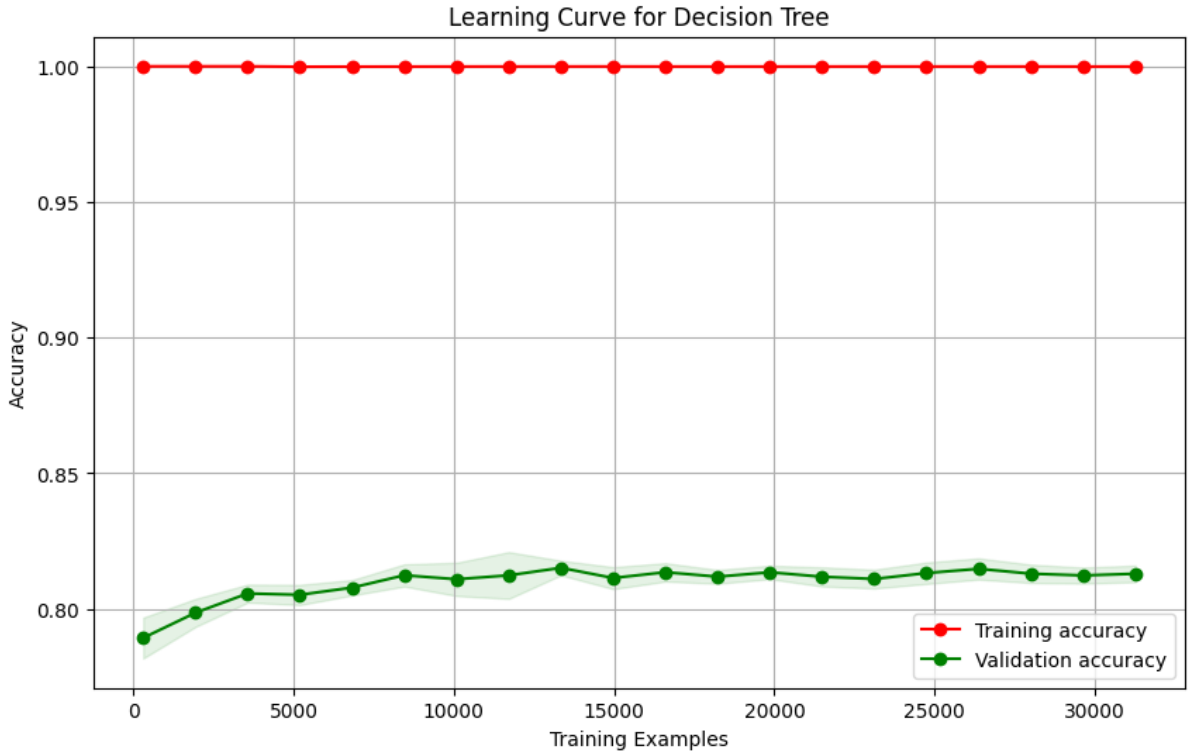
ML 과제

백준호

과제1: 고수입자 분류 문제

[2] 성능 평가 결과 해석

1. *learning curve* 및 성능 평가 결과를 참고하여 *Decision Tree* 모델이 오버피팅 되었는지 판단해주세요. 판단의 근거를 제시하고, *ML* 모델에서 오버피팅을 완화할 수 있는 방안을 찾아 함께 작성해주세요.
   - *- functions.py* 파일에 구현된 *plot_learning_curve*의 코드를 바탕으로 *learning curve*가 의미하는 바가 무엇인지 생각해보세요.
   - - 오버피팅인지 아닌지의 판단은 성능 평가 결과를 바탕으로 이루어져야 합니다.



Decision Tree - Training Accuracy: 0.9999, Test Accuracy: 0.8147

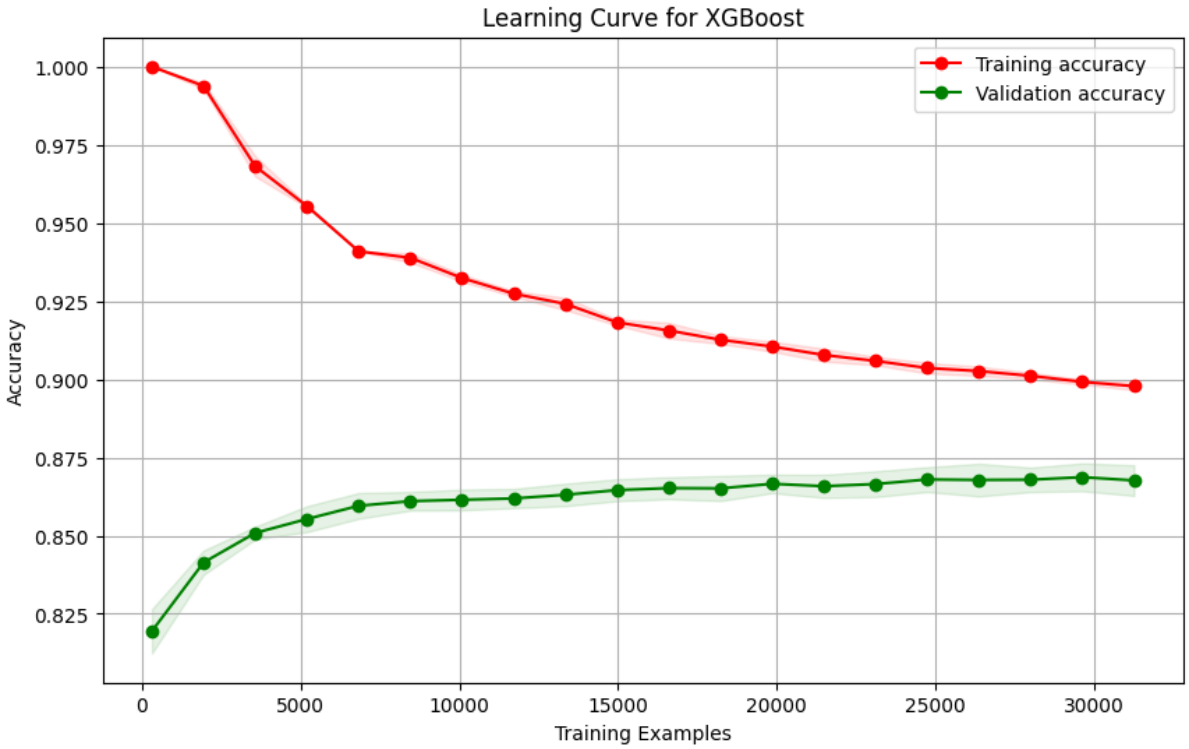
답:

학습정확도는 0.9999이고 검증 정확도는 0.8147로 도출되었다. 학습 정확도가 매우 높고 검증 정확도가 상대적으로 낮으므로, 오버피팅 현상이 발생했다고 추론할 수 있다.

functions.py 파일에 구현된 plot_learning_curve 함수는 학습 과정에서 모델의 성능 변화를 시각적으로 나타내는 도구이다. 이 함수는 학습 데이터의 크기 변화에 따른 모델의 학습 정확도와 검증 정확도를 그래프로 보여줌으로써, 모델이 과적합(overfitting) 또는 과소적합(underfitting) 상태인지, 혹은 적절하게 학습되고 있는지를 판단하는 데 유용하다. 이러한 시각화를 통해 모델의 학습 패턴을 파악하고, 필요한 조치를 취할 수 있다. 위의 그래프를 보면 학습정확도와 검증 정확도의 차이를 시각적으로 확인할 수 있어서 과적합 상태임을 알 수 있다.

오버피팅을 완화하기 위해서는 몇 가지 방법을 사용할 수 있다. 첫째, Decision Tree의 복잡도를 줄이기 위해 트리 가지치기(pruning)를 적용하여 최대 깊이(max_depth)를 제한하거나 최소 샘플 수(min_samples_split)를 조정할 수 있다. 둘째, 여러 개의 Decision Tree를 사용하는 랜덤 포레스트(Random Forest)나 부스팅(Boosting) 방법을 적용하면 모델의 분산을 줄이고, 과적합을 방지하며, 일반화 성능을 향상시킬 수 있다. 셋째, 더 많은 데이터를 수집하거나 데이터 증강 기법을 사용하여 모델의 일반화 성능을 향상시킬 수 있다. 마지막으로, K-Fold 교차 검증(K-Fold Cross Validation)을 사용하여 모델의 성능을 보다 안정적으로 평가하고, 과적합을 방지할 수 있다. 이러한 방법들을 통해 모델의 오버피팅 문제를 효과적으로 완화할 수 있다.

2. 일반적으로 앙상블 모델은 다른 모델에 비해 일반화 성능이 좋습니다. 그 이유가 무엇인지 설명하고, 우리의 성능 평가 결과에서도 XGBoost가 Decision Tree보다 나은 일반화 성능을 보이는지 판단해주세요.



XGBoost - Training Accuracy: 0.8933, Test Accuracy: 0.8689

앙상블 모델은 여러 개의 약한 학습기(weak learner)를 결합하여 강력한 학습기(strong learner)를 만드는 기법으로, 일반적으로 다른 모델에 비해 뛰어난 일반화 성능을 보여준다. 그 이유는 앙상블 기법이 모델의 분산을 줄이고, 과적합을 방지하는 데 효과적이기 때문이다. 예를 들어, 부스팅(Boosting) 방법은 순차적으로 약한 학습기를 학습시키고, 각 학습기는 이전 학습기의 오류를 보완하는 방식으로 작동한다. 이 과정에서 모델의 복잡도를 조절하고, 다양한 데이터 패턴을 학습함으로써 모델의 일반화 성능을 향상시킨다. 또한, 앙상블 모델은 정규화(term을 추가하여 모델의 복잡도를 제어하고, 과적합을 방지하는 특성을 가지고 있다.

우리의 성능 평가 결과에서도 XGBoost가 Decision Tree보다 나은 일반화 성능을 보이고 있다. XGBoost 모델의 학습 정확도는 0.8933, 검증 정확도는 0.8689로, Decision Tree 모델의 학습 정확도 0.9999와 검증 정확도 0.8147에 비해 검증 정확도가 더 높다. 이는 XGBoost 모델이 Decision Tree 모델보다 과적합이 덜 발생하고, 더 나은 일반화 성능을 보인다는 것을 의미한다. 또한, XGBoost의 학습 곡선을 보면, 학습 데이터의 크기가 증가함에 따라 학습 정확도와 검증 정확도의 차이가 줄어드는 것을 확인할 수 있다. 이는 XGBoost가 더 많은 데이터를 효과적으로 학습하고, 다양한 데이터 패턴을 포착할 수 있는 능력이 뛰어나다는 것을 보여준다.