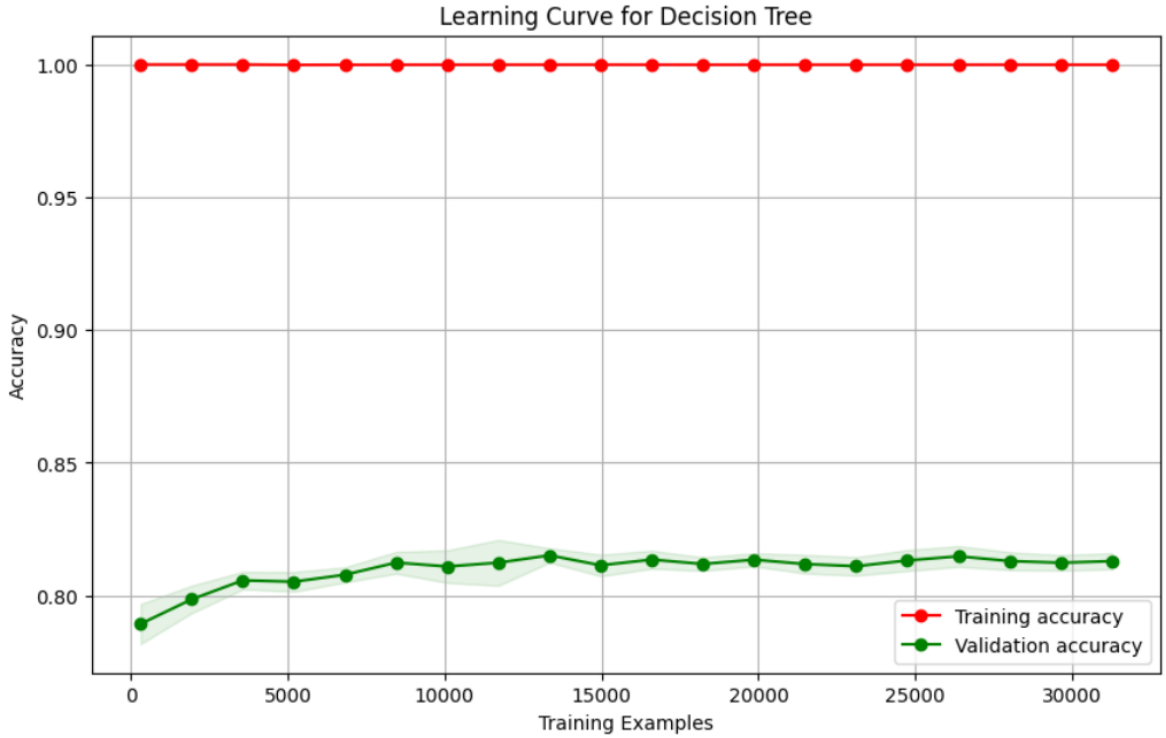
Report

김준호

과제1 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_curve의 코드를 바탕으로 learning curve가 의미하는 바가 무엇인지 생각해보세요.
- 오버피팅인지 아닌지의 판단은 성능 평가 결과를 바탕으로 이루어져야 합니다.




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

답: Decision Tree 모델의 오버피팅 여부 판단

(1) Learning Curve 분석:

Training accuracy: Decision Tree 모델의 학습 정확도는 거의 1.00으로, 훈련 데이터에 대해 거의 완벽한 성능을 보인다.

Validation accuracy: 검증 정확도는 약 0.81로, 훈련 정확도와 큰 차이를 보인다.

(2) 판단 근거:

학습 정확도가 매우 높고, 검증 정확도가 상대적으로 낮은 경우, 모델이 훈련 데이터에 지나치게

맞춰져 일반화 성능이 떨어지는 오버피팅 현상이 발생했다고 볼 수 있다.
functions.py 파일에서 구현된 plot_learning_curve 함수는 모델의 학습 및 검증 정확도를 시각화하여 모델의 성능을 평가하는 데 도움을 준다. 학습 데이터 크기에 따른 학습 정확도와 검증 정확도의 변화를 보여줌으로써 모델이 오버피팅되었는지, 언더피팅되었는지, 적절하게 학습되었는지 판단할 수 있다.

(3) 오버피팅 완화 방안:

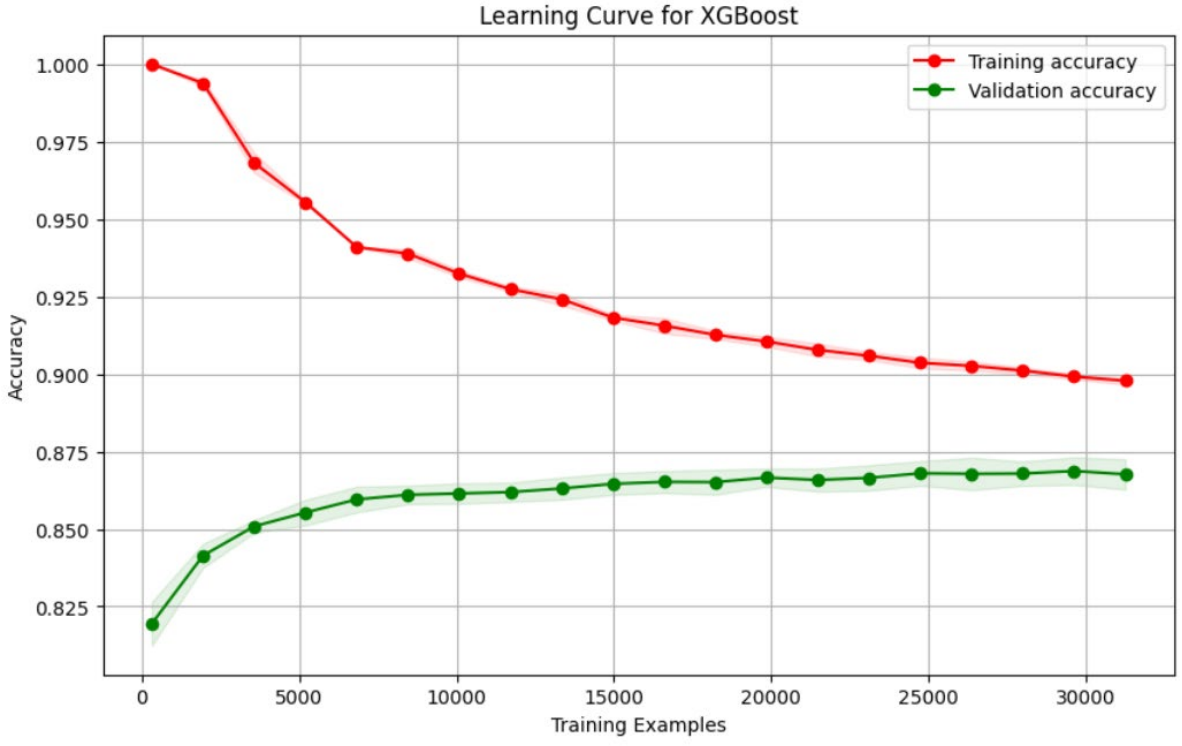
프루닝 (Pruning): Decision Tree의 복잡도를 줄이기 위해 트리 가지치기(pruning)를 적용한다. 이는 최대 깊이(max_depth)를 제한하거나, 최소 샘플 수(min_samples_split)를 조정하는 방식으로 수행할 수 있다.

앙상블 기법: 여러 개의 Decision Tree를 사용하는 랜덤 포레스트(Random Forest)나 부스팅(Boosting) 방법을 적용한다.

데이터 증강: 더 많은 데이터를 수집하거나 데이터 증강 기법을 사용하여 모델의 일반화 성능을 향상시킨다.

K-Fold Cross Validation: 모델의 성능을 보다 안정적으로 평가하고, 과적합을 방지한다.

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
```

답: XGBoost 모델의 일반화 성능 평가 및 이유

(1) 일반화 성능이 좋은 이유:

앙상블 기법: XGBoost는 여러 개의 약한 학습기(weak learner)를 결합하여 강력한 학습기(strong learner)를 만드는 부스팅(Boosting) 기법을 사용한다. 이는 모델의 분산을 줄이고, 과적합을 방지하며, 일반화 성능을 향상시킨다.

정규화: XGBoost는 정규화(term을 추가하여 모델의 복잡도를 제어하고, 과적합을 방지한다.

트리 기반 모델: 여러 개의 결정 트리를 사용하여 학습하므로, 다양한 데이터 패턴을 잘 포착할 수 있다.

(2) 성능 평가 결과:

Training accuracy: XGBoost 모델의 학습 정확도는 약 0.89로 Decision Tree 모델에 비해 다소 낮지만, 여전히 높은 편이다.

Validation accuracy: 검증 정확도는 약 0.87로 Decision Tree 모델에 비해 높다. 이는 XGBoost가 더 나은 일반화 성능을 보인다는 것을 의미한다.

(3) 결론:

성능 평가 결과, XGBoost가 Decision Tree보다 나은 일반화 성능을 보이고 있다. 이는 학습 및 검증 정확도의 차이가 적고, 검증 정확도가 더 높기 때문이다.