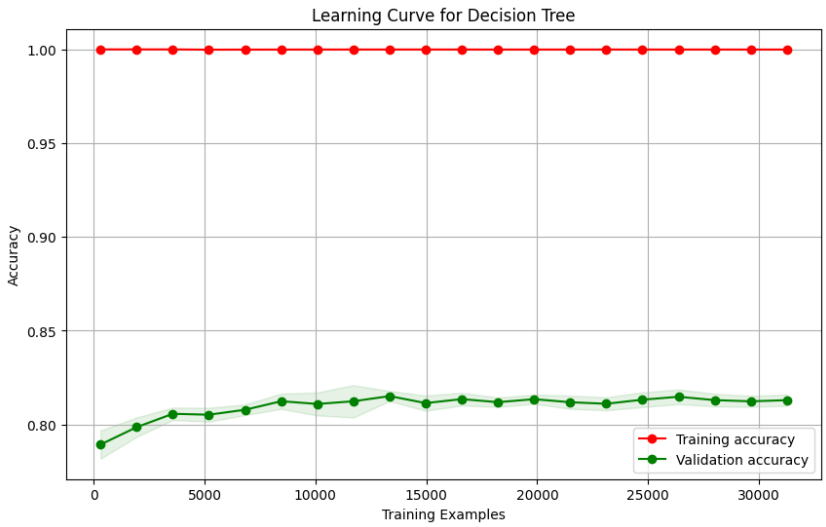
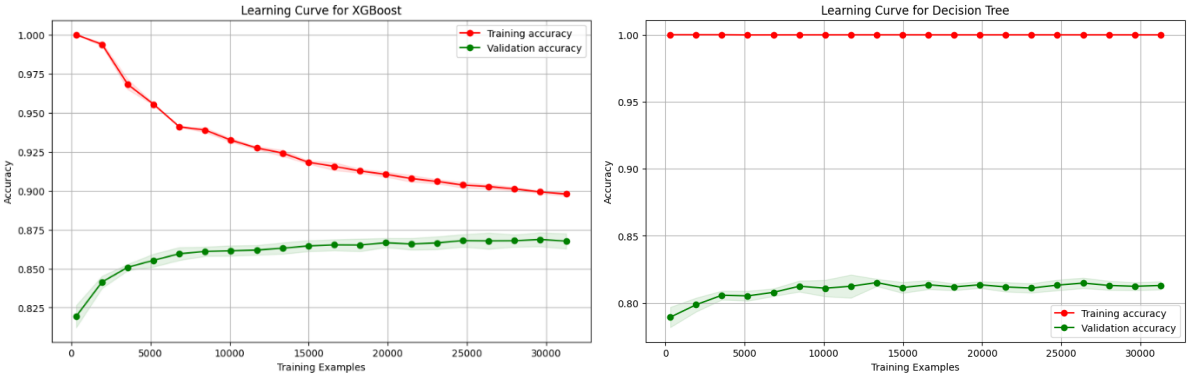
TODO 2-1 learning curve 및 성능 평가 결과를 참고하여 Decision Tree 모델이 오버피팅 되었는지 판단해주세요. 판단의 근거를 제시하고, ML 모델에서 오버피팅을 완화할 수 있는 방안을 찾아 함께 작성해주세요.



의사결정 나무 모델의 학습 곡선을 보면, 학습 데이터에서의 예측 정확도는 1.00을 기록하며 이는 완벽하게 에측해내고 있음을 알 수 있다. 반면에, 검증 데이터에서의 예측 정확도는 학습 데이터에서의 예측률 대비 저조한 성적을 보이고 있으며, 예측을 계속해서 진행해도 초반 부분을 제외한 이후 검증 세션에서는 유의미한 예측률 상승을 보이지 않고 있음을 알 수 있다. 따라서, 현재 의사결정 나무 모델은 과적합 상태라고 판단할 수 있다.

이를 완화할 수 있는 방안으로는 bagging 방식과 같은 앙상블 기법을 활용하여 여러 의사결정 나무를 생성한뒤 이를 평균화하여 과적합을 줄이는 방안, 인스턴스 분류에 큰 의미를 가지지 않는 섹션을 제거하는 차원 축소 방안을 생각해볼 수 있다.

TODO 2-2 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.



(좌) XGBoost (우) Decision Tree

XGBoost와 같은 앙상블 방법은 여러 분석 모델들의 예측을 평균화하여 분산을 줄여서 더 안정적인 성능을 보이고, Variance-Bias 트레이드오프의 균형점을 잘 잡아서 전체 오류를 줄일 수 있기 때문에 일반화 성능이 좋다고 할 수 있다.

이번 XGBoost 모델의 경우, Decision Tree보다 검증 데이터에서의 예측률 자체만으로도 더 나은 성능을 보이고 있음을 알 수 있다. 또한 훈련-검증 데이터 간 예측률 격차가 Decision tree 모델에서는 전혀 좁혀지지 않는 반면, XGBoost 모델에서는 점차 좁혀지는 모습을 볼 때, 일반화가 잘 이뤄졌음을 알 수 있다.