**데이터 로드 및 확인**: 네 개의 CSV 파일을 pandas 라이브러리를 사용하여 불러오고, 각 파일의 첫 몇 행을 표시하여 데이터의 구조를 이해한다. 이 파일들은 data\_info, train\_data, test\_data, sample\_submission으로 구성되어 있다. data\_info는 각 컬럼의 설명을 포함하고 있다.

**기초 통계량 확인:** train\_data의 기초 통계량을 확인하여 데이터의 분포, 평균, 표준편차 등을 파악한다.

**데이터 전처리**: 먼저 누락된 값이 있는지 확인하고, 범주형 변수를 수치형으로 인코딩하기 위해 LabelEncoder를 사용한다. 그리고 데이터를 학습용과 검증용으로 분할한다.

**새로운 특성 생성**: 데이터에 대한 인사이트를 높이기 위해 새로운 특성을 생성한다. 여기에는 total\_learning\_time, login\_time\_difference, course\_completion\_rate 등이 포함된다.

**모델 훈련 및 평가**: Logistic Regression, Random Forest, Gradient Boosting 모델을 사용하여 학습을 진행하고, 검증 데이터셋을 사용하여 각 모델의 성능을 평가한다. 성능 평가는 정확도, 정밀도, 재현율, F1 점수를 사용한다.

**하이퍼파라미터 튜닝**: Gradient Boosting 모델의 성능을 향상시키기 위해 RandomizedSearchCV를 사용하여 최적의 하이퍼파라미터를 찾는다.

**최적화된 모델의 평가**: 찾아낸 최적의 하이퍼파라미터로 그래디언트 부스팅 모델을 다시 훈련시키고, 검증 데이터셋에서의 성능을 평가한다.