정보통신대학원

# - Kaggle 데이터를 이용한 머신러닝 분석과 예측 -

(AI 개인화 학습 데이터를 활용한 학생 성과 예측 및 최적화 전략 수립)

[머신러닝기반의연구방법론]
지도교수: 양희정 교수님


전공: 인공지능
학번: 202576626
이름: 오경주

# 목차

# 1. 개요

1.1 연구 배경

현대 교육 환경에서 AI 기술을 활용한 개인화 학습은 학생들의 학습 효과를 극대화하는 핵심 전략으로 부상하고 있다. 본 연구는 학생들의 다양한 학습 특성과 행동 패턴을 분석하여 개인별 맞춤형 학습 전략을 수립하고자 한다.

(실습 데이터셋: https://www.kaggle.com/datasets/ziya07/ai-powered-personalized-learning-dataset)

1.2 연구 목표

학생 성과를 정확히 예측할 수 있는 머신러닝 모델 개발

군집 분석을 통한 학습자 유형별 최적 학습 경로 도출

실시간 적응형 콘텐츠 추천 시스템 구축

다양한 머신러닝 모델의 성능 비교 및 최적화

1.3 분석 방법론

예측 모델: 선형회귀, 의사결정 트리, 랜덤 포레스트, 앙상블

성능 지표: 결정계수($R^2$), 평균 제곱오차(MSE), 평균 절대오차(MAE)

최적화 기법: GridSearchCV를 활용한 하이퍼파라미터 튜닝

군집 분석: K-means 클러스터링

추천 시스템: 협업 필터링 기반 KNN 알고리즘

# 2. 데이터 분석 개요

2.1 데이터셋 구성

분석에 사용된 데이터셋은 다음과 같은 18개의 변수로 구성되어 있다.

| 변수명 | 설명 | 데이터 타입 |
|---|---|---|
| student_id | 학생 고유 식별자 | 범주형 |
| age | 학생 나이 | 수치형 |
| gender | 성별 | 범주형 |
| education_level | 교육 수준 | 범주형 |
| learning_style | 학습 스타일 (시각적/청각적/운동감각적) | 범주형 |
| previous_gpa | 이전 평점 | 수치형 |
| completed_modules | 완료한 모듈 수 | 수치형 |
| avg_time_per_module | 모듈당 평균 학습 시간 | 수치형 |
| engagement_score | 참여도 점수 | 수치형 |
| distraction_events | 방해 이벤트 횟수 | 수치형 |
| quiz_accuracy | 퀴즈 정확도 | 수치형 |
| feedback_score | 피드백 점수 | 수치형 |
| contextual_difficulty_level | 난이도 수준 | 범주형 |
| recommended_path | 추천 학습 경로 | 범주형 |

| | | |
|---|---|---|
| actual_path_followed | 실제 학습 경로 | 범주형 |
| path_efficiency_score | 경로 효율성 점수 | 수치형 |
| final_assessment_score | 최종 평가 점수 (목표 변수) | 수치형 |
| learning_outcome | 학습 결과 | 범주형 |

2.2 데이터 전처리

범주형 변수에 대한 레이블 인코딩 수행

결측값 처리 (데이터셋에 결측값 없음 확인)

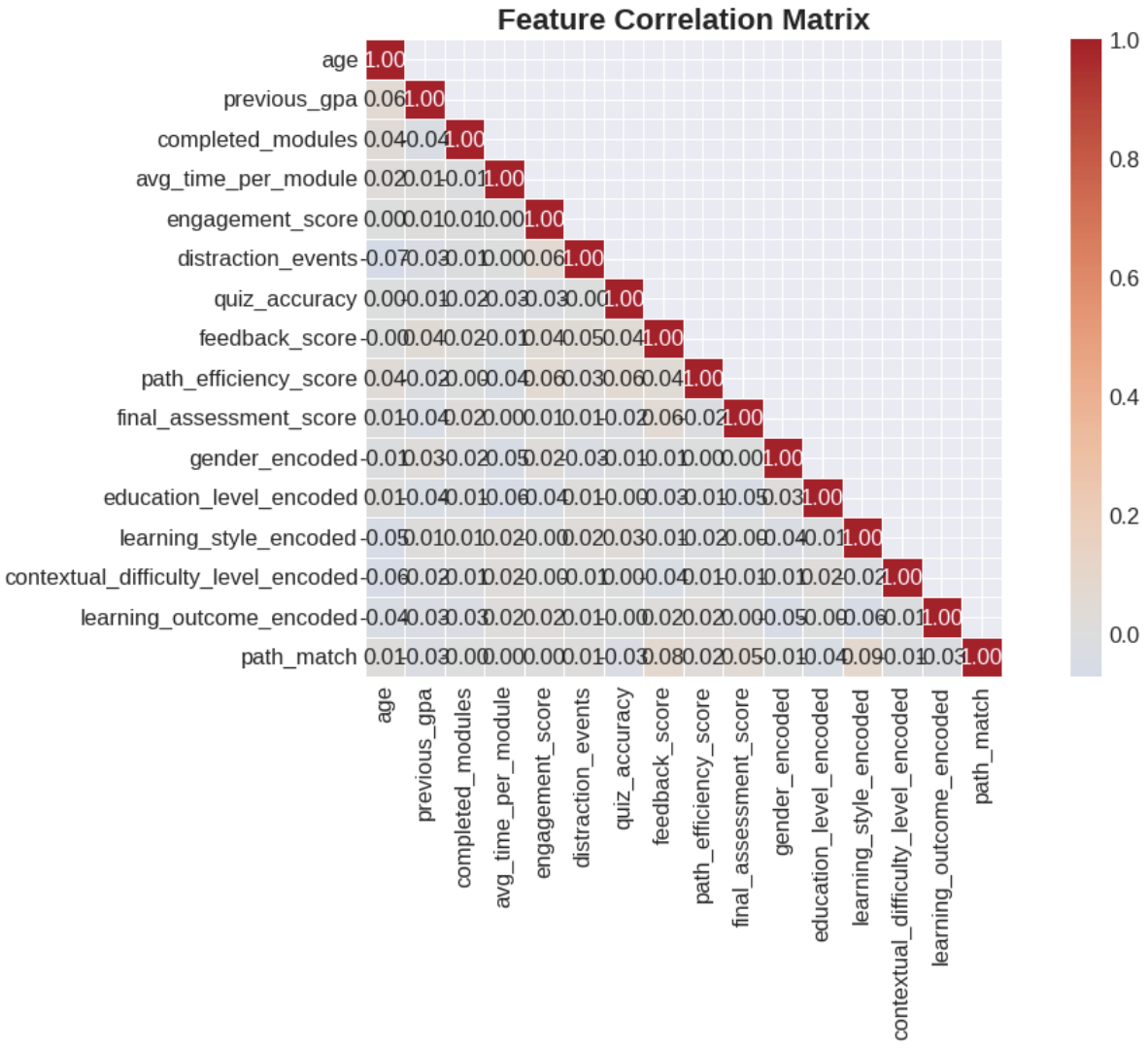
특성 스케일링 (StandardScaler 사용)

파생 변수 생성: path_match (추천 경로와 실제 경로 일치 여부)

2.3 탐색적 데이터 분석 (EDA)

상관관계 분석

주요 발견사항:



## 3. 학생 성과 예측 모델링

3.1 특성 선택

모델 학습에 사용된 14개 특성:

인구통계학적 특성: age, gender_encoded, education_level_encoded
학습 특성: learning_style_encoded, previous_gpa
행동 특성: completed_modules, avg_time_per_module, engagement_score
성과 지표: distraction_events, quiz_accuracy, feedback_score
환경 요인: contextual_difficulty_level_encoded, path_efficiency_score, path_match

3.2 데이터 분할

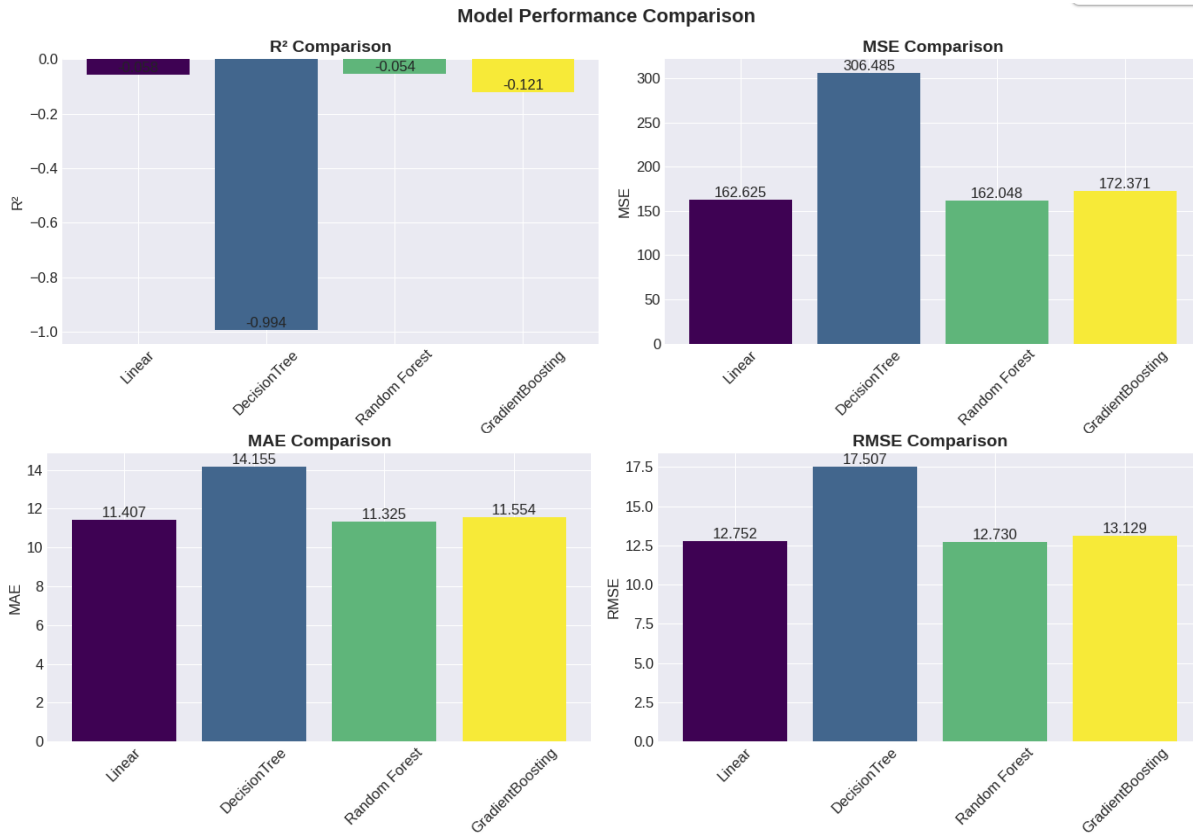
훈련 세트: 80%
테스트 세트: 20%
무작위 시드: 42 (재현성 보장)

## 4. 모델 성능 비교 및 평가

4.1 베이스라인 모델 성능

| 모델 | $R^2$ | MSE | MAE | RMSE |
|---|---|---|---|---|
| 선형회귀 (Linear) | -0.0579 | 162.6254 | 11.4071 | 12.7525 |
| 의사결정트리 (DecisionTree) | -0.9938 | 306.4850 | 14.1550 | 17.5067 |
| 랜덤포레스트 (Random Forest) | -0.0542 | 162.0483 | 11.3245 | 12.7298 |
| 그래디언트부스팅 (GradientBoosting) | -0.1213 | 72.3715 | 11.5539 | 13.1290 |

### 4.2 성능 분석

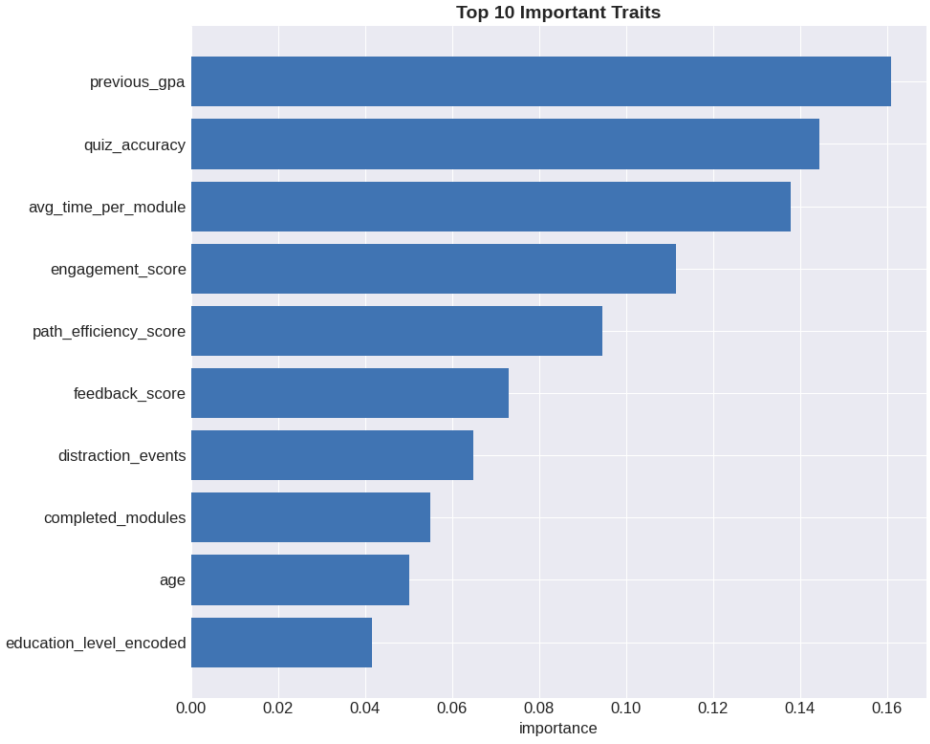



## 5. 하이퍼파라미터 튜닝

### 5.1 튜닝 전략

각 모델에 대해 GridSearchCV를 활용하여 최적의 하이퍼파라미터 조합을 탐색했다.

특성 중요도 분석 (랜덤포레스트 특성 중요도)



상위 5개 중요 특성

| index | feature | importance |
|---|---|---|
| 4 | previous_gpa | 0.160967 |
| 9 | quiz_accuracy | 0.144317 |
| 6 | avg_time_per_module | 0.137920 |
| 7 | engagement_score | 0.111506 |
| 12 | path_efficiency_score | 0.094527 |

## 6. 군집 지능 기반 학습 경로 최적화

6.1 클러스터링 분석

K-means 알고리즘을 사용하여 학생들을 4개의 군집으로 분류했다.

| | previous_gpa | engagement_score | quiz_accuracy | final_assessment_score |
|---|---|---|---|---|
| 0 | 3.01 | 69.81 | 62.47 | 80.10 |
| 1 | 2.94 | 77.74 | 87.45 | 80.54 |
| 2 | 3.00 | 91.24 | 62.30 | 80.85 |
| 3 | 3.00 | 81.65 | 87.18 | 80.32 |

시각화