

# 효율적 성적 상승을 위한 전략 모델링

통계분석 및 머신러닝을 통한  
성적예측 모델링

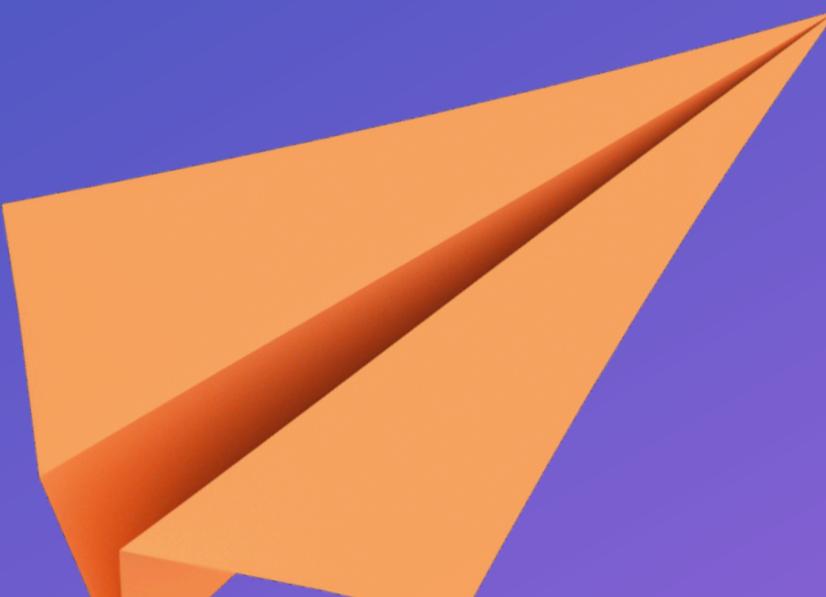
- 김새한
- 2026.01.15
- Kaggle Predicting Student Test Scores



# 목차



- 01. 프로젝트 개요
  - 02. 데이터 소개
  - 03. EDA
  - 04. 통계 분석
  - 05. 머신러닝
  - 06. 요인분석
  - 07. 결론
- Q&A

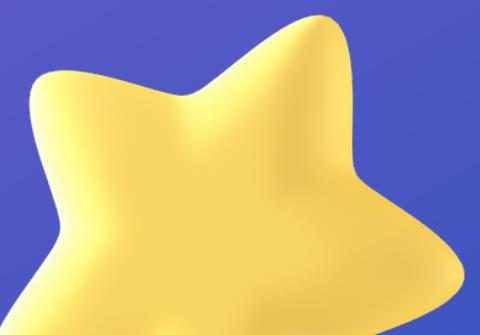




01

# 프로젝트 개요

배경, 목적 및 기대효과





## 01. 프로젝트 개요



### 문제 정의

“ 학생 성적에 결정적 영향을 미치는 핵심 동인은 무엇이며, 복합요인 간의 비선형 관계를 학습하여 성적을 정밀하게 예측할 수 있는 최적의 모델은 무엇인가? ”

높은 교육열로 인해 학생들의 성적은 큰 관심의 대상이며 학생들은 효과적 성적 향상을 위해 다양한 노력을 한다.  
하지만 현대 교육환경에서 성적은 다양한 요인들과 복잡한 상호작용과 비선형적 관계를 가지고 있다.  
특정 요인이 성적에 미치는 진정한 기여도를 산출하고 정확한 예측을 하기 위해서는 이러한 관계를 고려해야 한다.





## 01. 프로젝트 개요

### 목표

성적에 영향을 주는 핵심 요인을 도출하고, 성적 예측 모델을 만들어 효율적인 학습처방을 제공해 학업성취도를 전략적으로 극대화한다.

### 분석 범위

EDA(탐색적 분석) → 전처리/특성공학 → 모델링 → 성능 평가 → 요인 분석  
→ 인사이트 도출

### 기대효과

데이터 기반 의사결정 지원:

- 성적 하락 위험 학생 조기 식별 및 선제적 개입
- 한정된 학습 지원 자원의 최적 배분
- 맞춤형 학습 지도 가이드라인의 근거 마련





02

## 데이터셋 소개

변수 정의 및 데이터 구조





## 02. 데이터셋 소개

**900000**

SAMPLES

**12**

FEATURES

### TARGET FEATURE

**exam\_score**

Range: 19.6 - 100

### DATA SPLIT

train

validation

test

56%

14%

30%



개인 특성

age, gender



생활습관

sleep\_hours, sleep\_quality



학습 습관

study\_hours, study\_method, class\_attendance



학습 환경

internet\_access, facility\_rating,  
exam\_difficulty, course

## FEATURE TABLE

<b>id</b>	NUMERIC 학생 고유 학번	<b>class_attendance</b>	NUMERIC 수업참여도
<b>age</b>	NUMERIC 나이	<b>sleep_hours</b>	NUMERIC 하루수면시간
<b>study_hours</b>	NUMERIC 학습시간	<b>facility_rating</b>	ORDINAL 환경점수
<b>internet_access</b>	CATEGORICAL 인터넷 가능 여부	<b>sleep_quality</b>	ORDINAL Poor/Average/Good
<b>cours</b>	CATEGORICAL 전공과목	<b>exam_difficulty</b>	ORDINAL 시험 난이도
<b>gender</b>	CATEGORICAL 성별	<b>study_method</b>	CATEGORICAL self-study, online-videos, Other



03

## EDA

데이터 분포 확인 및 이상치 탐지





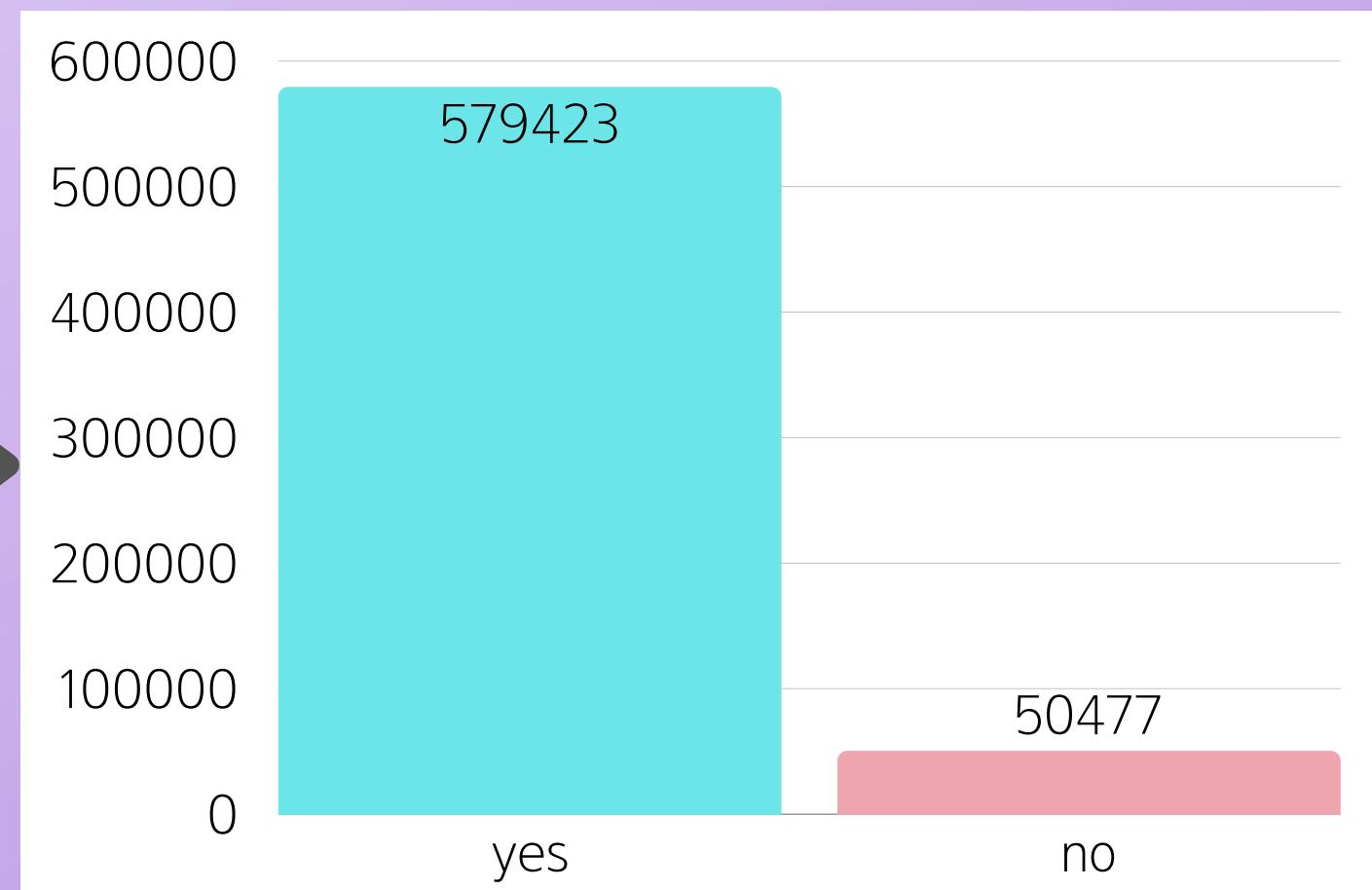
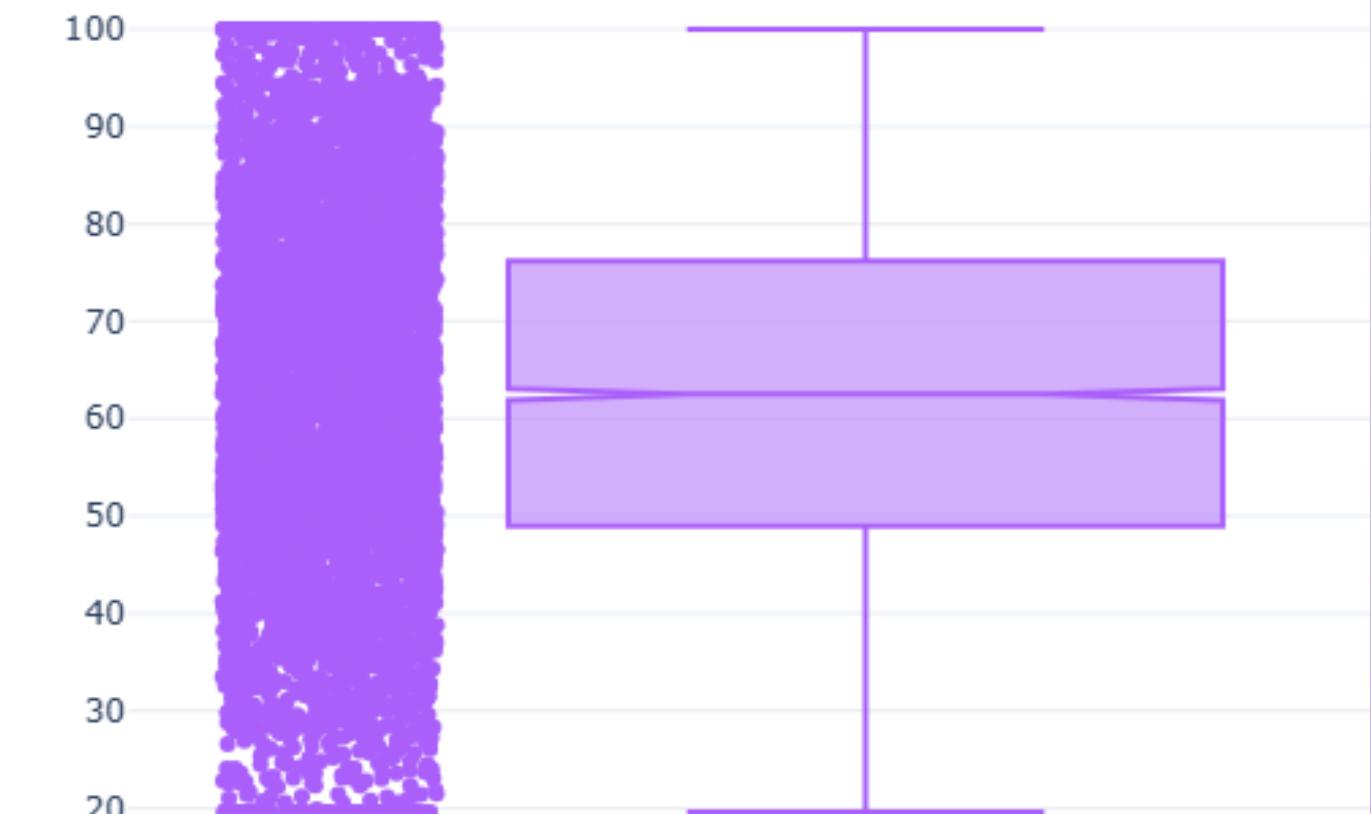
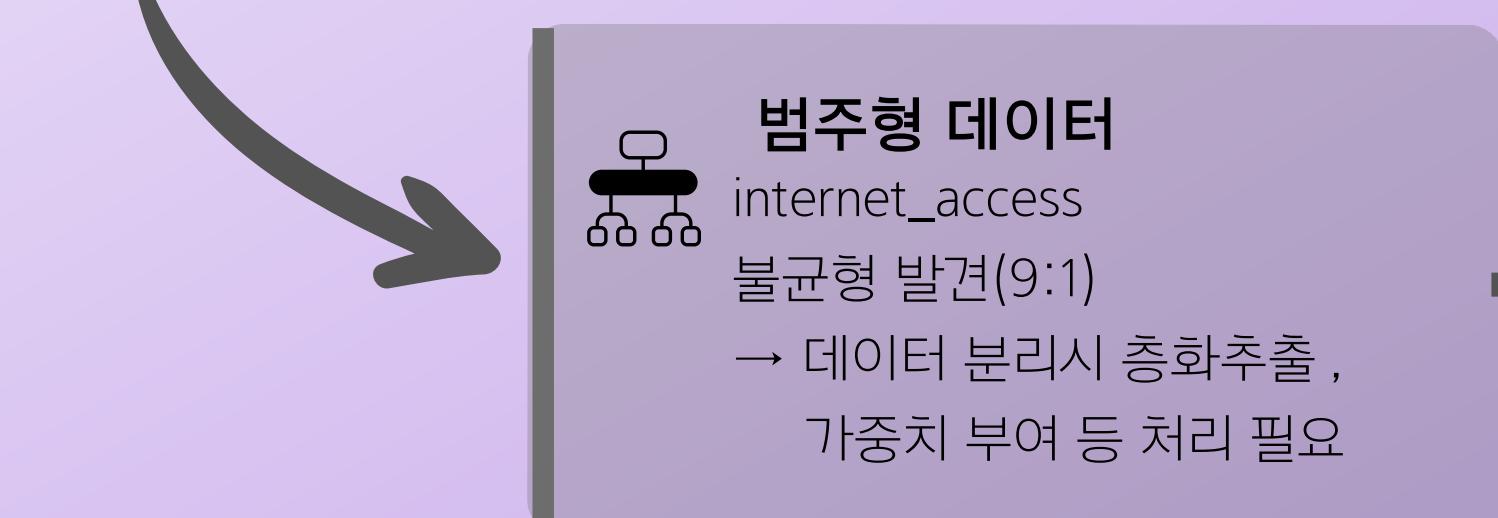
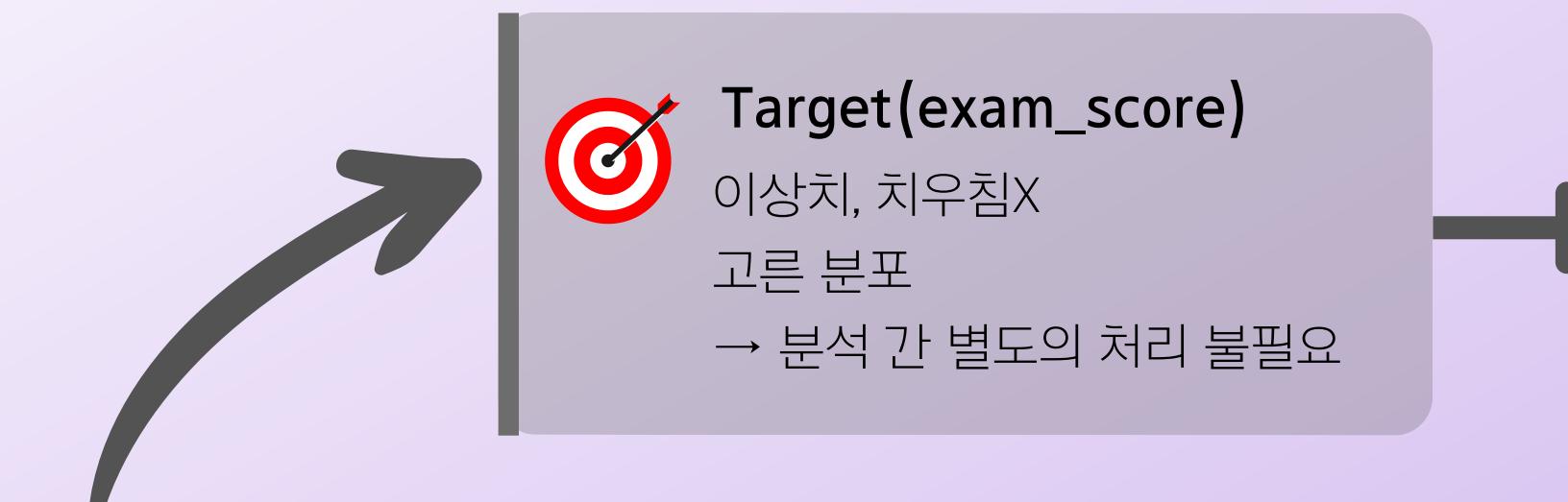
## 03. EDA

train데이터 구조 분석



### 전체 데이터

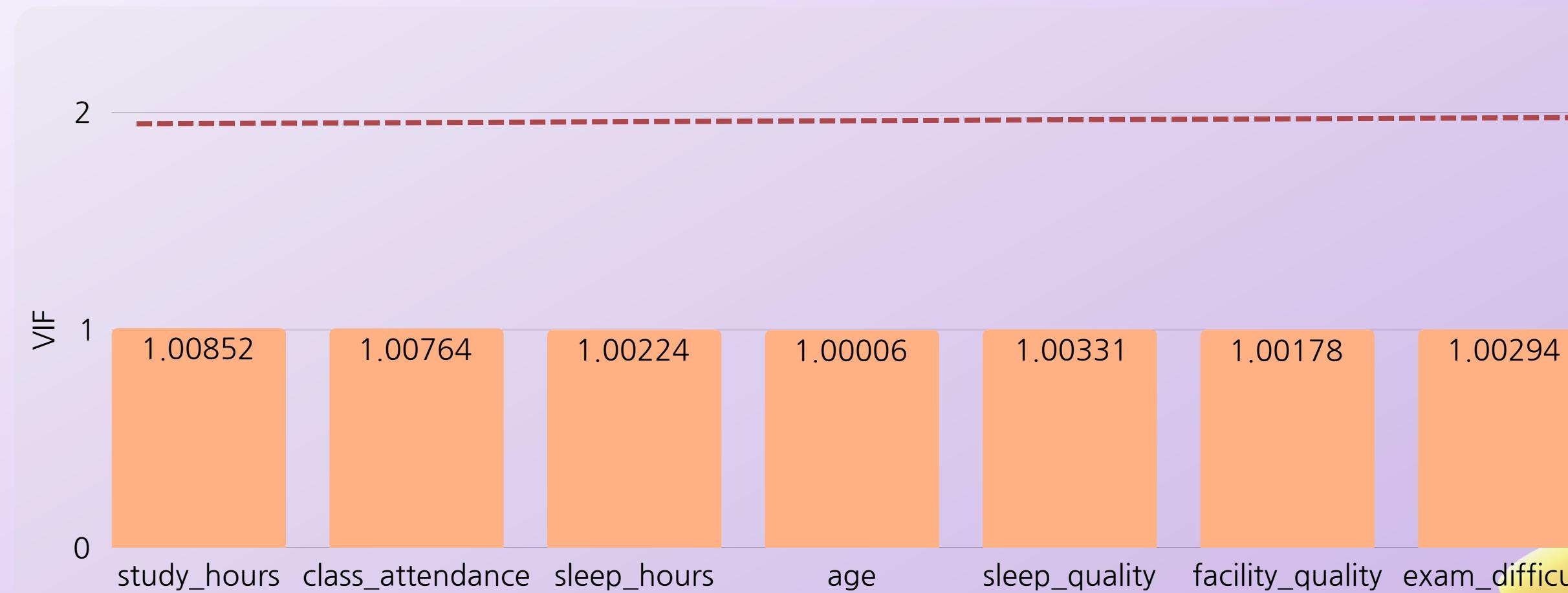
630000 row / 12 features  
결측치, 중복값 X





## 03. EDA

train데이터 구조 분석



### 변수 간 상관관계(VIF)

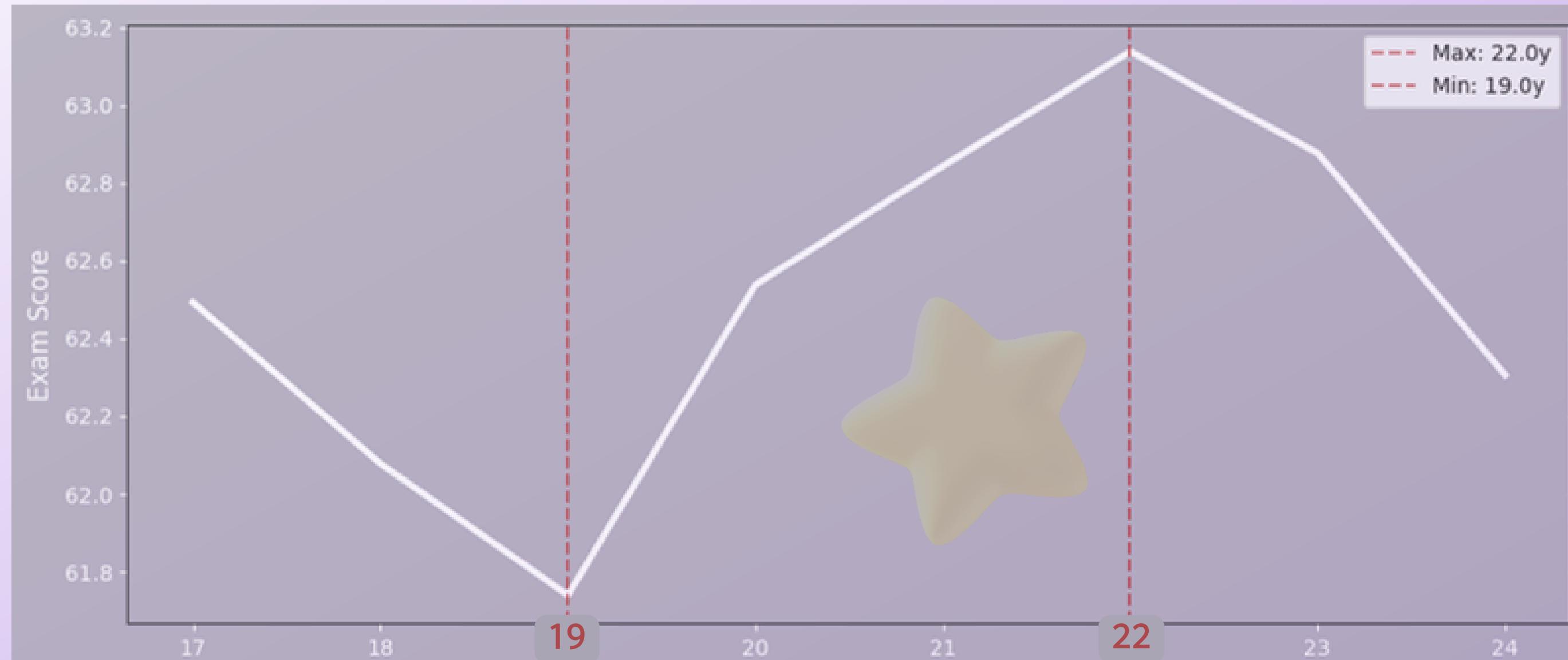
- 최대 VIF : 1.009(study\_hours)  
변수 간 상관관계 무시 가능 (5 이상 제거 고려)



## 03. EDA

### 나이에 따른 점수 변화 분석

#### ⟳ 연령별 점수 중앙값 변화 패턴 분석



- 19 ~ 22세 성적 증가 경향
  - 학업에 대한 관심 증대 시기
  - 입시 시기와의 연관성 존재 가능

- 파생변수 생성 'age\_ord'
  - 19, 22 기준 구간별 중앙값(범주형)
  - 비선형성 반영



## 03. EDA

# 전처리 전략

nominal : OneHot  
ordinal : Ordinal

Encoding

각 변수별 단위가 다르므로  
standard scaling

Scaling

파생변수 생성/노이즈 제거

변수 검정/ 선정

catboost 별도의 전처리

모델별 전처리

안건으로 돌아가기



04

## 통계분석

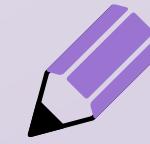
다양한 통계기법을 활용한 변수 검증



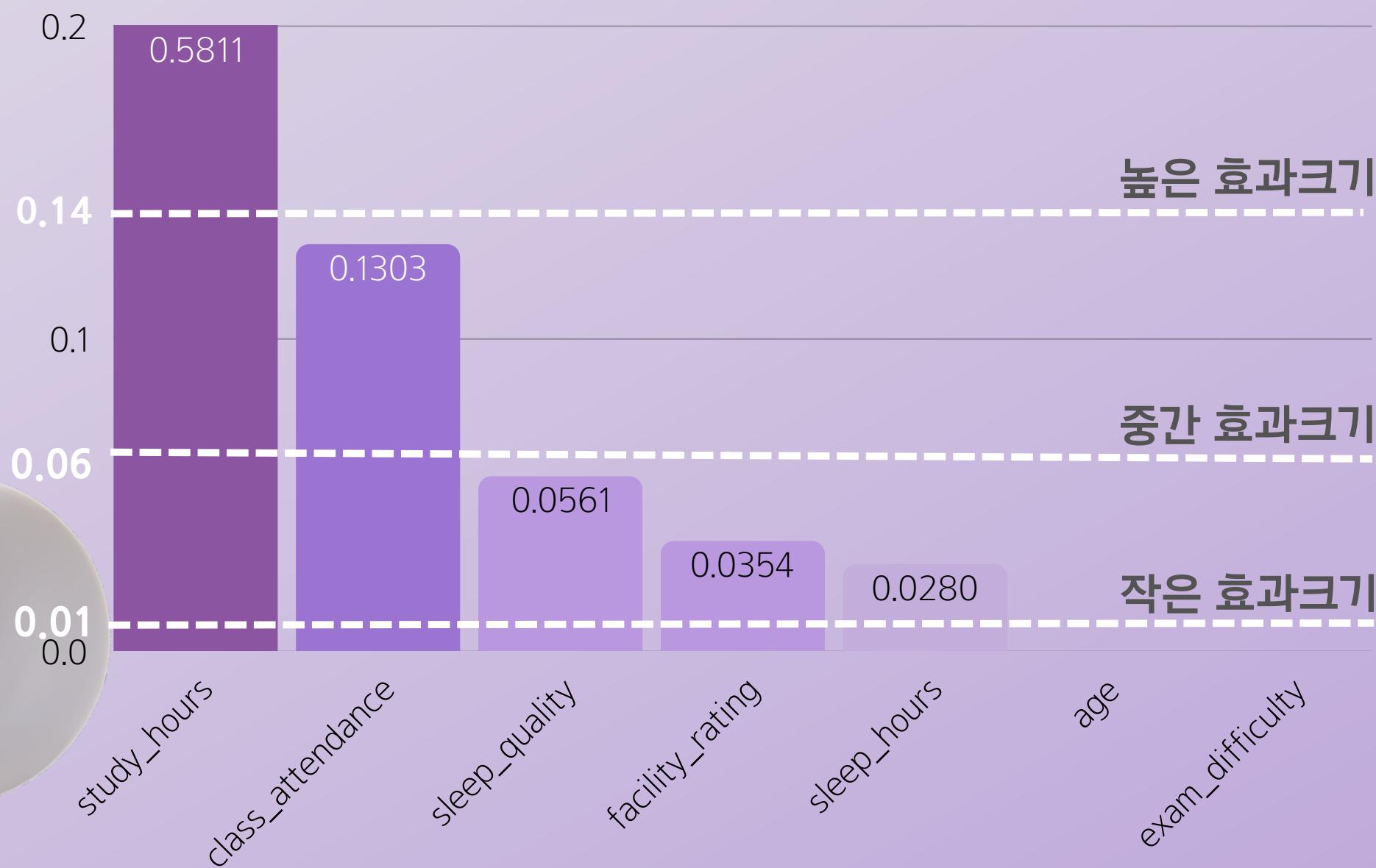


## 04. 통계분석

수치형, 순위형 변수(효과크기, 상관분석)



### 수치형 - 효과크기(에타제곱( $\eta^2$ ))



### 결과

**높은 효과크기(0.14 이상)**

학습시간(0.5811)

**중간 효과크기(0.06 이상)**

수업 참석률

**작은 효과크기(0.01 이상)**

수면의 질, 시설물 질, 수면시간

**유의하지 않은 효과크기**

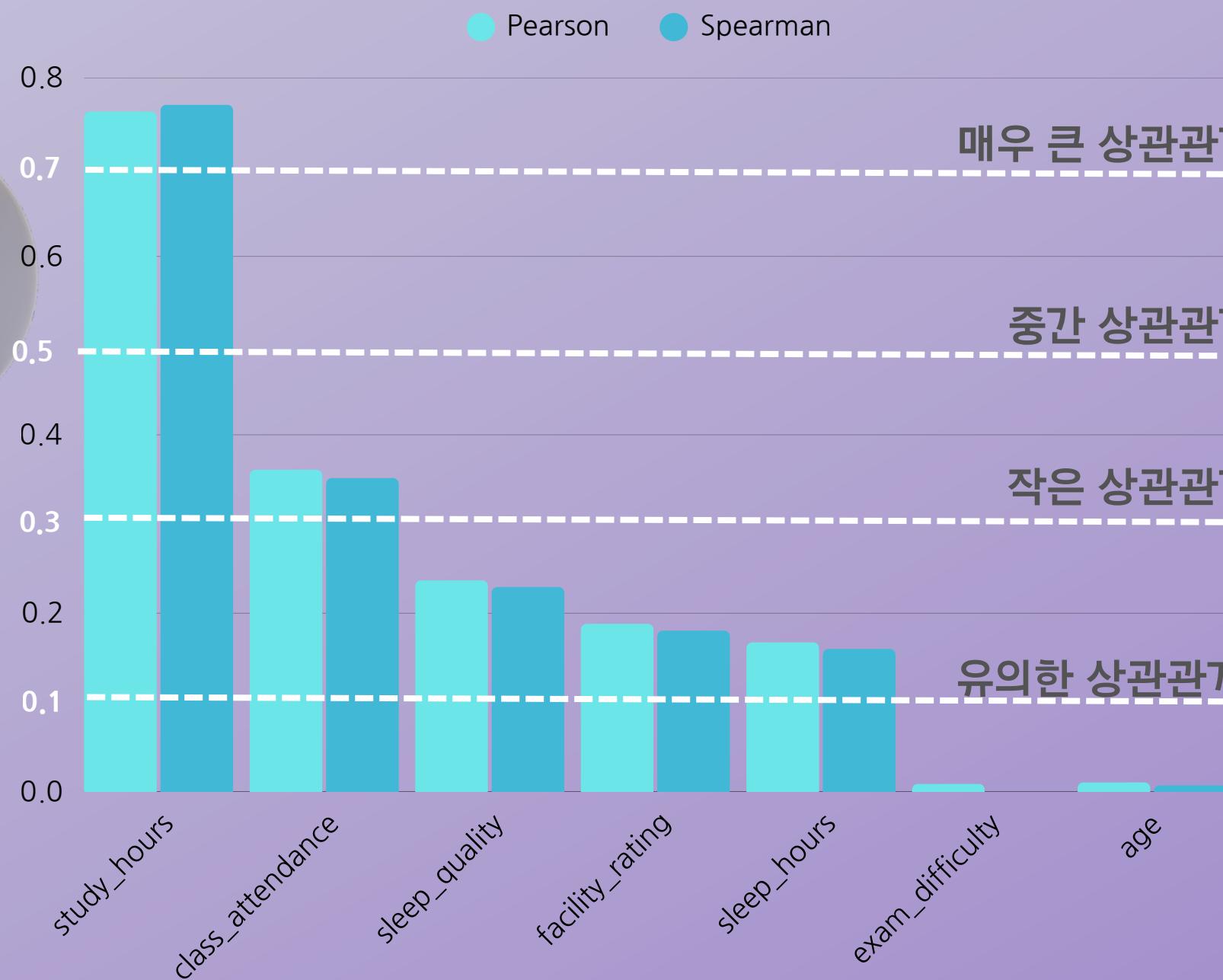
나이, 시험 난이도



## 04. 통계분석

수치형, 순위형 변수(효과크기, 상관분석)

### 상관분석



Pearson : 선형적 관계

Spearman: 순위기반 / 비선형성도 포함



### 결과

매우 큰 상관관계(0.7 이상)

학습시간

작은 상관관계(0.3이상)

수업 참석률

유의하지만 작은 상관관계(0.1 이상)

수면의 질, 시설물 질, 수면시간

유의하지 않은 상관관계

시험난이도, 나이



## 04. 통계분석

범주형 변수(효과크기, welch's ANOVA)



### 범주형 - 효과크기(에타제곱( $\eta^2$ ))

유의한 효과 : study\_method(0.04690)

미미한 효과 : course, age\_ord, gender (0.001 이하)

유의하지 않은 효과: internet\_access(0)

variable	Effect Size
study_method	0.0469
course	0.00031
age_ord	0.00017
gender	0.00013
internet_access	0.00001



### Welch's ANOVA

유의한 효과 : study\_method(설명력 0.05)

미미한 효과 : course, age\_ord, gender (0.001 이하)

유의하지 않은 효과: internet\_access(pvalue > 0.5)

variable	P-value	F-score
study_method	0	8271.34401
course	0	32.12757
age_ord	0	91.73659
gender	0	55.5387
internet_access	0.72717	0.12172

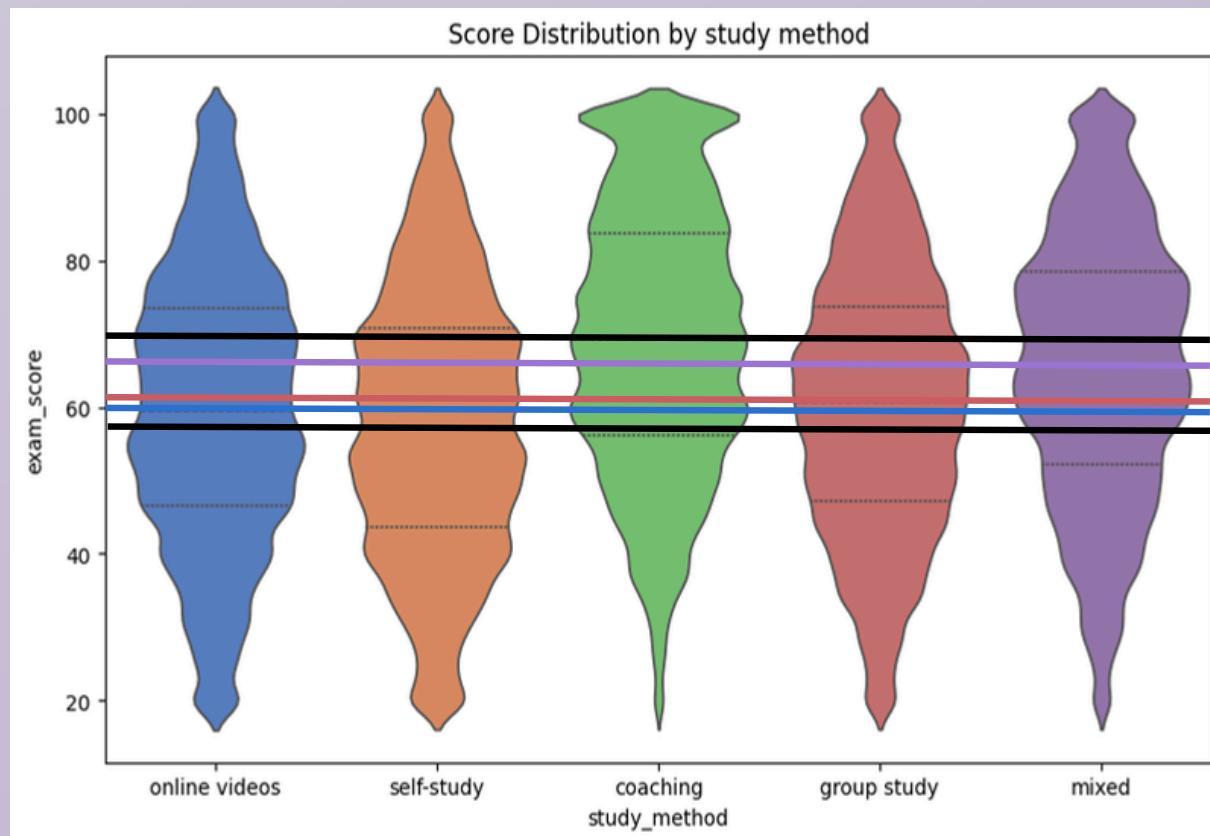


## 04. 통계분석

범주형 변수- 학습방법 별 성적분포



### 학습 방법별 점수 분포



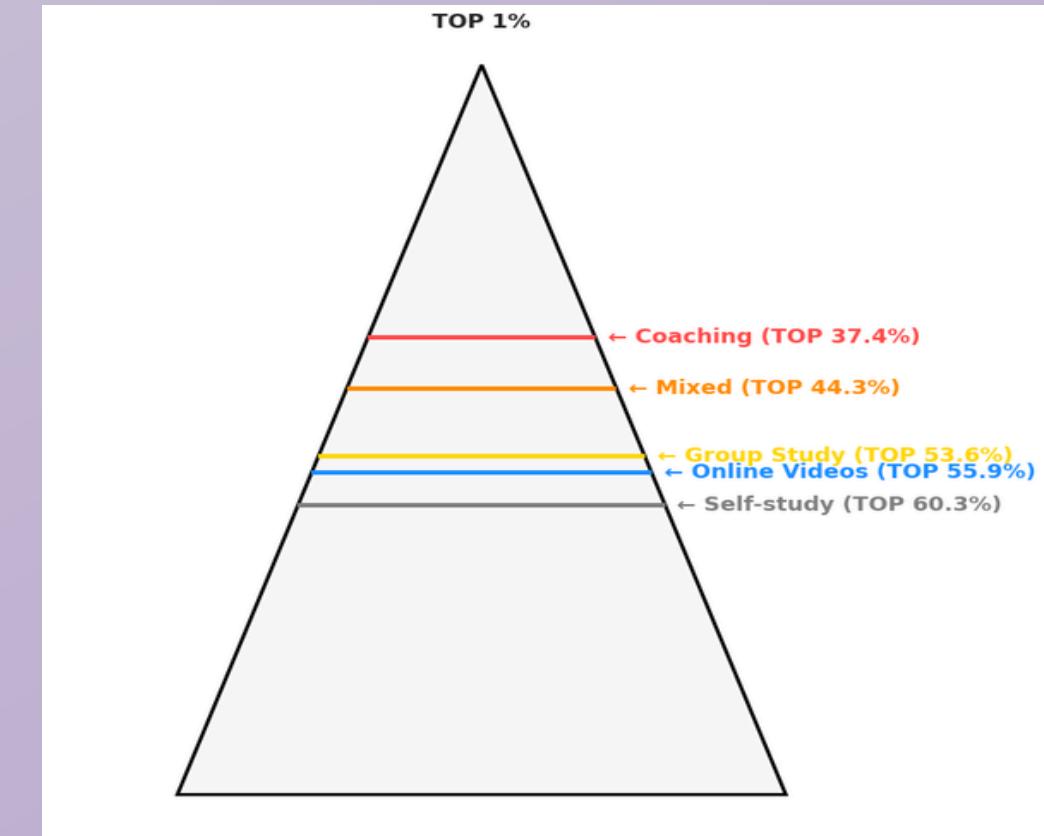
상위권(coaching)의 성적 중앙값 : 85

하위권(self)성적 중앙값 : 60

→ 약 25점 차이



### 학습 방법별 등수 분포(상위 %)



상위권(coaching)의 등수 중앙값 : 37.4%

하위권(self)의 등수 중앙값 : 60.3%(하위39.7 %)

→ 약 23% 차이



결과

학습방법의 효율은  
coaching  
mixed  
grouped study  
online-videos  
self study

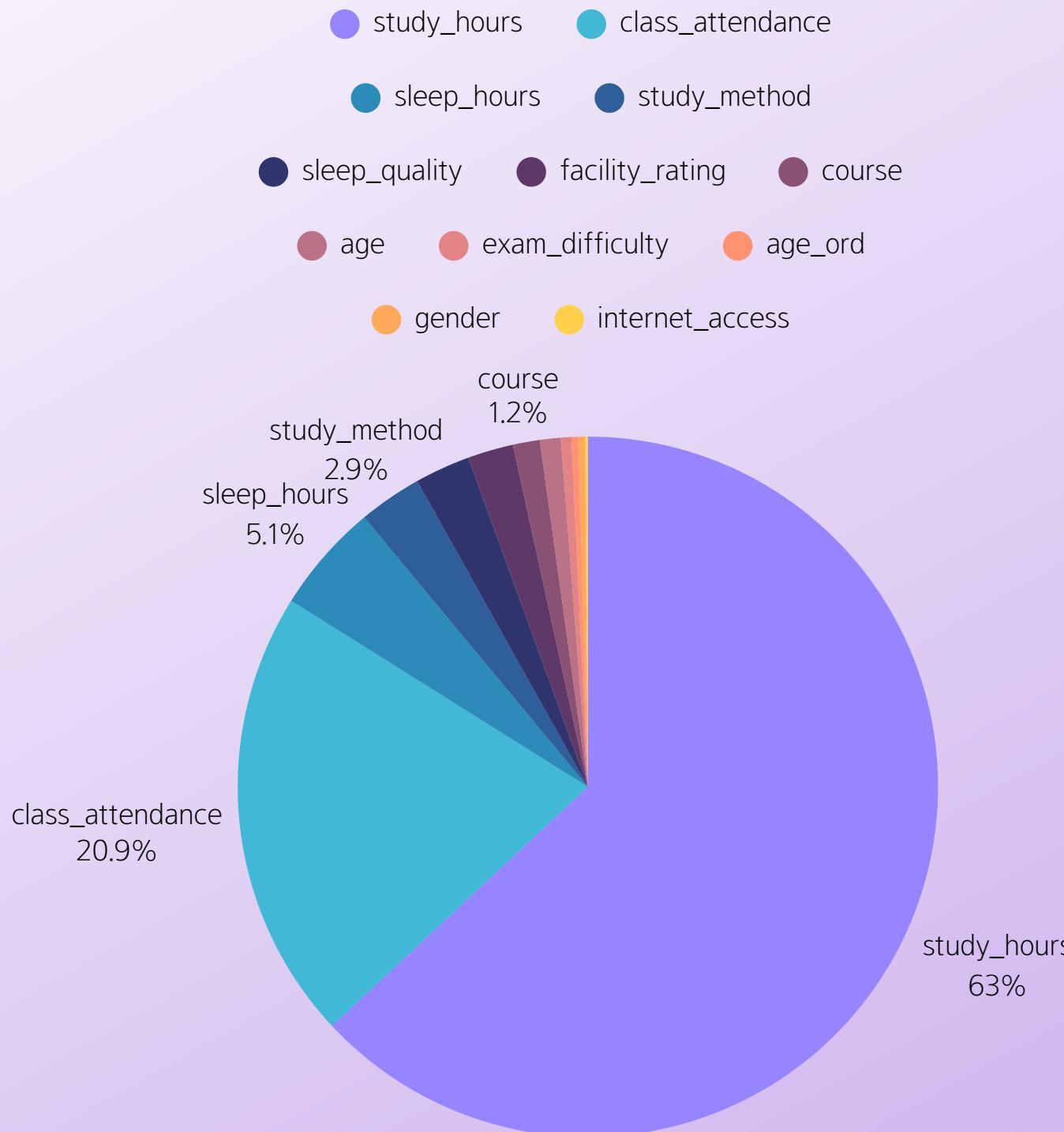
coaching과 self study는  
약 25점 차이가 난다.



## 04. 통계분석

전체 변수(Mutual Information Score)

“변수가 타겟에 대해 얼마나 많은 정보를 담고 있는가?” (선형 + 비선형)



### 해석



#### 양적 요인

study\_hours(0.91): 성적의 변동을 설명하는 최대 요인. 학습시간이 성적과 직결됨  
class\_attendance(0.30): 높은 설명력을 가짐. 성실성이 성적에 영향을 줌



#### 질적 요인

study\_method(0.04) : 유의미한 정보량을 보유. 학습 방식의 중요성 설명



#### 환경적 요인

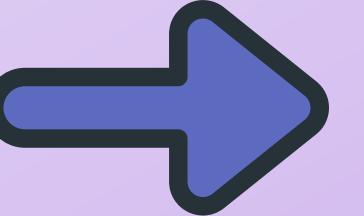
인터넷 접근성이나 성별 등 환경적/인구통계적 요인은 성적 예측에 영향력이 적음



## 04. 통계분석

결론 요약 및 활용 방안

핵심요인	주요요인	보조요인	기타요인
학습시간	수업참석률 학습방법	수면시간 수면의 질 시설물의 질	나이 성별 시험난이도



1. 학습시간이 낮고 자습위주의 학습, 수업참석률이 낮은 학생  
을 고위험군으로 판단, 집중관리
2. 19- 22세 입시 시기의 학생들 주요 관리대상에 포함
1. 공부하는 습관 키우기
  - 공부하는 습관을 키워 학습시간을 점차적으로 증가 시키기
  - 수업에 참석하도록 독려해 학습시간과 수업참석률 향상
  - 타이머 등을 활용해 학습시간 관리 및 보상
2. 공부방법
  - 자습보다는 과외, 학원 등의 방법 선택, 수업시간에 집중하기
  - 가계 곤란 시 방과후 수업, 무료 인터넷 강의, 국가에서 시행하는 무료 과외서비스 등 활용
3. 생활습관 개선
  - 적절한 수면과 휴식을 취하되, 수면시간에는 수면에만 집중
4. 황금시기 집중관리
  - 19 -22세에는 의지가 강해 성적이 오를 가능성이 높으므로  
보다 체계적이고 적극적인 관리



05

# 머신러닝

머신러닝을 통한 성적 예측모델





## 05. 머신러닝

모델링 전략 및 실험 설계

사용 모델

xgboost, light gbm,  
catboost

앙상블

Gradient Boosting Decision Tree 계열

비선형 및 복잡한 상호작용 포착  
변수 중요도제공



K-Fold CV

5-Fold 교차 검증 수행 안정적인 평가



RMSE

kaggle 평가지표 활용



RandomSearch

광범위한 파라미터 공간을 효율적으로 탐색.

Iter: 10회 무작위 샘플링

Params: 학습률, 트리 깊이, 정규화 계수 등

## 예측 모델 및 전처리 (Models & Preprocessing)

### Category

### Details



Ensemble Models  
(고성능 부스팅 알고리즘)



XGBoost



LightGBM



CatBoost



Preprocessing  
(Column Transforme)



StandardScaler (표준화)



OneHotEncoder /  
OrdinalEncoder

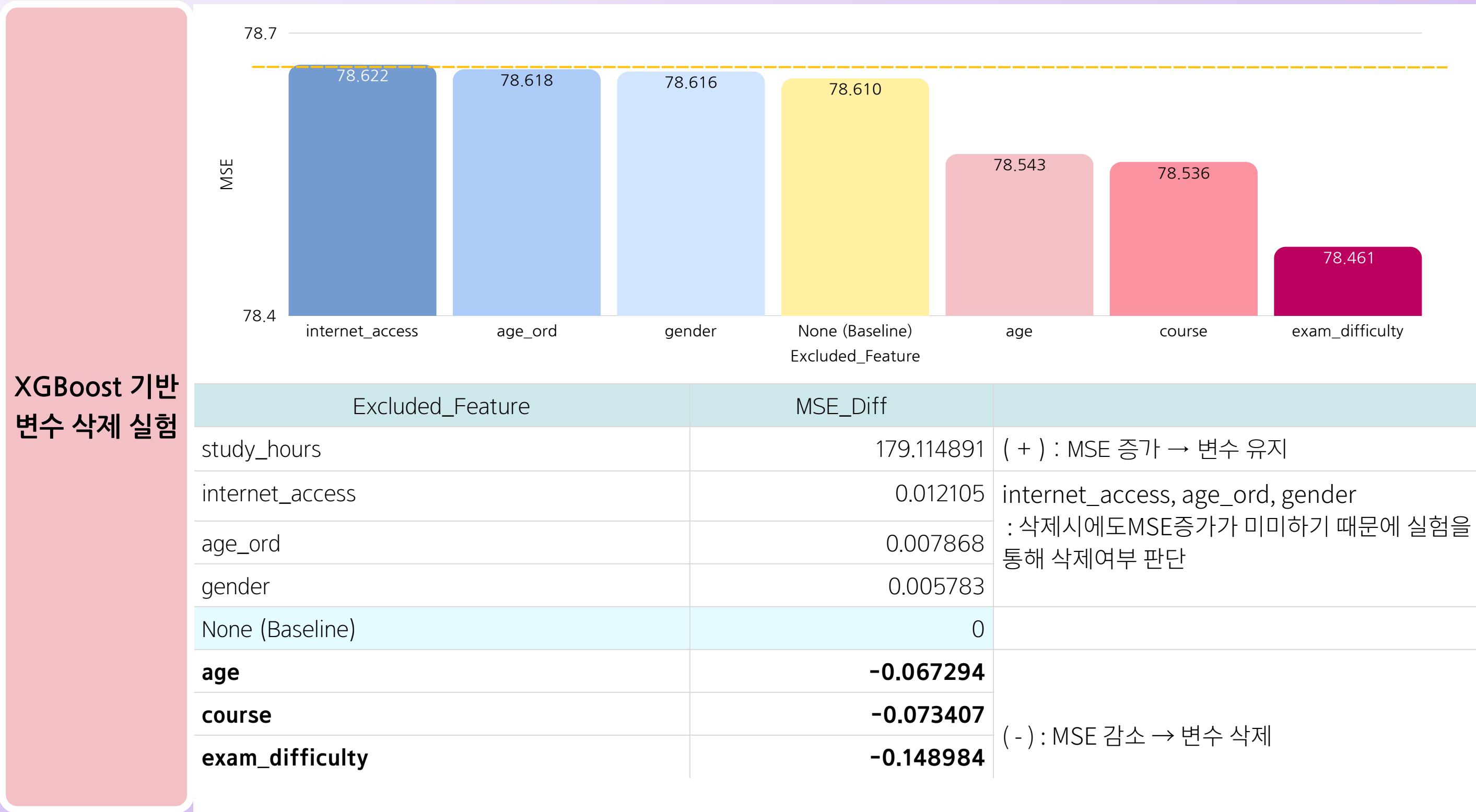


catboost 내장 전처리기 활용



## 05. 머신러닝

변수선택( 삭제변수 선정 )





## 05. 머신러닝

변수 Set별 점수차이/모델성능 평가 및 가중치 부여

### 변수선택

Model	Drop	RMSE
XGBoost	base (age, exam_difficulty, course)	8.830941
	base + gender	8.829563
	base + gender,internet_access	8.827545
	base + gender,internet_access , age_ord	8.82975

### 최종 선택 변수 Set

study\_hours  
class\_attendance  
study\_method  
sleep\_hours  
sleep\_quality  
facility\_rating  
age\_ord

총 7개 변수



## 05. 머신러닝

모델성능 평가 및 가중치 부여

Model	RMSE	Time	★Weight
Lightgbm	8.849688	0:03:13	0.333
XGBoost	8.827545	0:04:11	0.3336
CatBoost	8.828255	0:33:42	0.3334



### 최종 양상을 점수

8.6553

최상의 단일 모델 대비

RMSE 0.17 이상 추가 개선



### 성능향상 분석

예측의 견고함(Robustness) 향상

일반화 성능(Generalization) 극대화

변동성 감소



06

## 요인분석

머신러닝 결과를 통한 요인별 중요도 분석





## 06. 요인분석

SHARP를 통한 요인분석



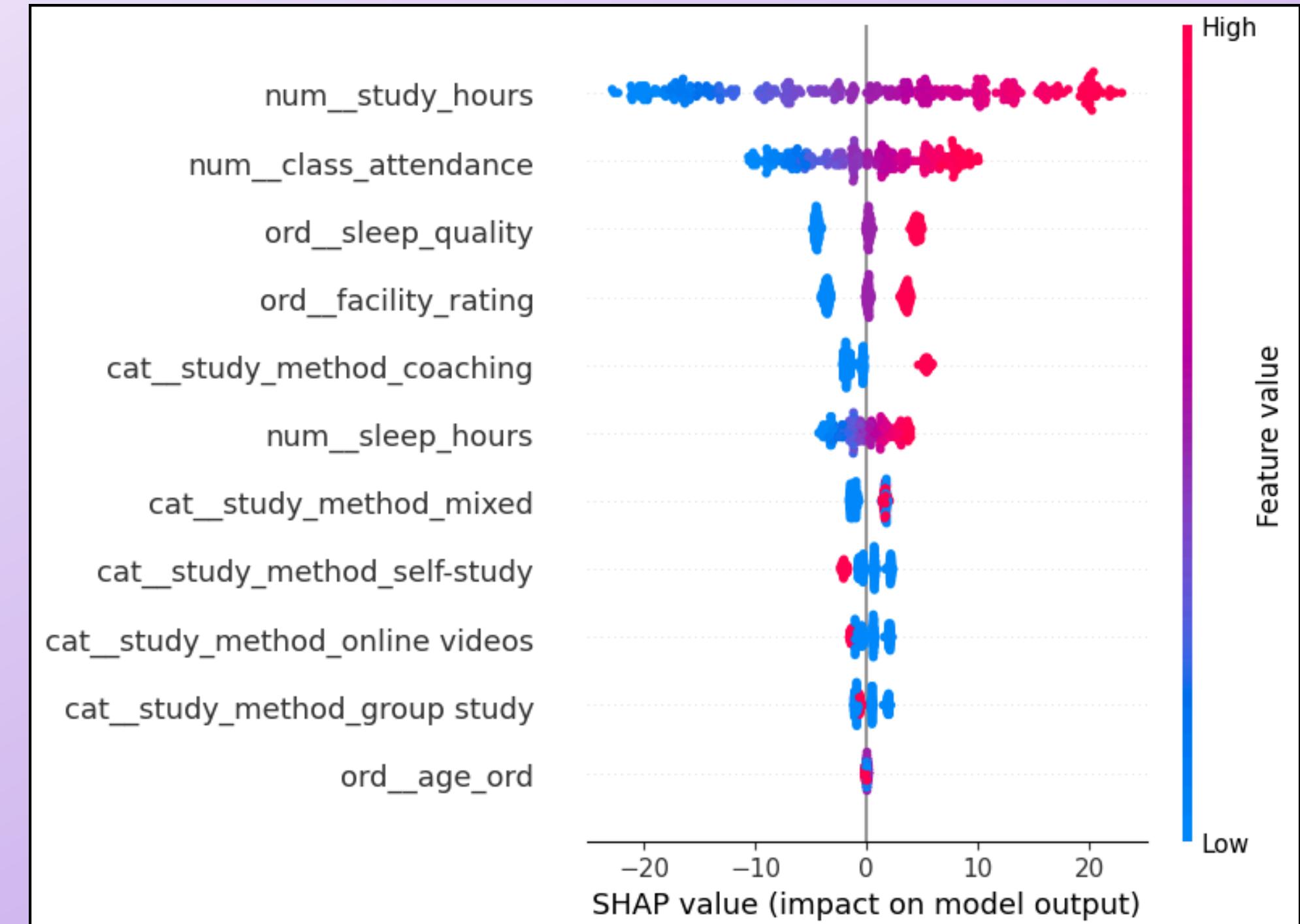
### 모델 내부 의사결정 메커니즘 분석

#### 지배적 요인:

학습시간, 출석율이 전체 예측 기여도의 70% 이상을 차지함.

#### 비선형적 특징

학습방법중 coaching은 특정 구간에서 점수를 도약시키는 '트리거' 역할을 수행함.  
반대로 자습은 특정구간 점수를 급하시킴





 07

# 결 론

Kaggle 점수 및 인사이트 도출





## 07. 결론

### 제출 프로세스 (Submission Pipeline)

```
graph LR; A[LightGBM Model<br/>학습된 모델 로드] --> B[Predict Probability<br/>예측 확률 산출 (Not Label)]; B --> C[Submission.csv<br/>ID, diagnosed_diabetes]
```

### 리더보드 점수 (Scores)

Final score  
**8.72952**

Public LB

Private LB

### 평가 지표 (Evaluation Metric)

RMSE(Root Mean Squared Error)

이 대회는 회귀(Regression)문제로, 모델이 시험점수를 얼마나 잘 예측하는지 평가합니다.



## 07. 결론

### 기대효과

- 성적 하락 위험 학생 조기 식별 및 선제적 개입
- 한정된 학습 지원 자원의 최적 배분
- 맞춤형 학습 지도 가이드라인의 근거 마련

### 타겟 학생

- 학습시간이 낮고 자습위주의 학습, 수업참석률이 낮은 학생을 고위험군으로 판단, 집중관리
- 19- 22세 입시 시기의 학생들 주요 관리대상에 포함

### 맞춤학습지도 가이드라인

#### 공부하는 습관 키우기

- 공부하는 습관을 키워 **학습시간을 점차적으로 증진**
- 수업에 참석하도록 독려해 **수업참석률** 향상
- 타이머 등을 활용해 학습시간 관리 및 보상



## 07. 결론

### 기대효과

- 성적 하락 위험 학생 조기 식별 및 선제적 개입
- 한정된 학습 지원 자원의 최적 배분
- 맞춤형 학습 지도 가이드라인의 근거 마련

### 타겟 학생

- 학습시간이 낮고 자습위주의 학습, 수업참석률이 낮은 학생을 고위험군으로 판단, 집중관리
- 19- 22세 입시 시기의 학생들 주요 관리대상에 포함

### 맞춤학습지도 가이드라인

#### 공부방법

- 자습보다는 **과외, 학원** 등의 방법 선택
- 가계 곤란 시 **방과후 수업, 무료 인터넷 강의**, 국가에서 시행하는 **무료 과외** 등 활용



## 07. 결론

### 기대효과

- 성적 하락 위험 학생 조기 식별 및 선제적 개입
- 한정된 학습 지원 자원의 최적 배분
- 맞춤형 학습 지도 가이드라인의 근거 마련

### 타겟 학생

- 학습시간이 낮고 자습위주의 학습, 수업참석률이 낮은 학생을 고위험군으로 판단, 집중관리
- 19- 22세 입시 시기의 학생들 주요 관리대상에 포함

## 맞춤학습지도 가이드라인

### 생활습관 개선

적절한 수면과 휴식을 취하되,  
수면시간에는 수면에만 집중

### 황금시기 집중관리

19 -22세에는 성적이 오를 가능성이 높으므로  
학습을 독려하고, 보다 체계적이고 공격적 관리



## 참고 자료



## 참고 문헌

- Crede, M., & Kuncel, N. R. (2008). Study habit in relation to academic performance: A meta-analysis. *Perspectives on Psychological Science*.
- Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, 11(2), 114-126.
- Selwyn, N. (2004). Reconsidering political and popular understandings of the digital divide. *New Media & Society*, 6(3), 341-362.
- Haslwanter, T. (2016). An Introduction to Statistics with Python: With Applications in the Life Sciences. Springer International Publishing.
- Unpingco, J. (2016). Python for Probability, Statistics, and Machine Learning. Springer International Publishing.