머신러닝 기반
성적 예측 프로젝트

빅데이터 6기

3조 박수아 윤소현 이도현 이신영 이효준

요약

본 프로젝트는 포르투갈 학생들의 수학 성적 데이터를 바탕으로 학생들의 학업 성적 예측 모델을 개발하여 학업 성취도 향상을 위한 방안을 제시하고 학생 개개인의 특성에 맞춘 맞춤형 교육 프로그램 제공하는 데 목적이 있다. 모델 개발 과정에서 '성별, 나이, 과락 횟수, 학교 지원, 진학 의향, 가족 관계, 외출 빈도, 결석 횟수' 및 새롭게 생성한 '보호자의 교육 수준'을 변수로 선정하였다. 이를 바탕으로 'G3의 성적을 유의하게 예측을 할 수 있다.'를 가설로 설정하였다.

8개의 회귀 모델과 4개의 분류 모델을 돌려본 결과, 회귀모델은 LinearRegression 모델이 다른 것에 비해 성능지표가 높았지만 일반적인 성능(0.5-1.0)에 미치지 못 했다. RandomForestClassifier 모델은 정밀도를 높임으로써 모델이 예측하는 클래스에 속해있는 학생이 올바르게 속해있을 확률을 높였다.

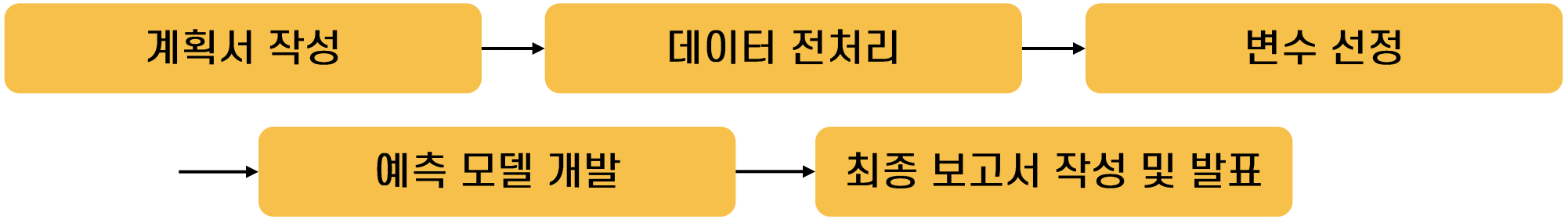
이 결과를 바탕으로 중상위권(B)을 대상으로는 상위권 도약을 위한 고난도 문제를, 하위권(D)을 대상으로는 학습 흥미를 유발할 수 있는 자료를 부여하는 수준별 학습을 제공함으로써 학업 성취도 향상과 교육 효율성 증대를 기대한다.

프로젝트 개요

1. 프로젝트명 : 학생의 학업 성적 예측 모델 분석
2. 프로젝트 진행 기간 : 2024.02.13 ~ 2024.02.14 16:00(총 2일)
3. 역할 분배

이름	책임 파트	상세 업무
박수아	데이터 전처리	데이터 전처리 메인, 변수 선정 서브
윤소현	변수 선정	데이터 전처리 서브, 변수 선정 메인
이도현	예측 모델 개발	데이터 전처리 서브, 예측 모델 개발 메인
이신영	최종 보고서 작성	최종 보고서 작성 메인, 예측 모델 개발 서브
이효준	발표	예측 모델 개발 서브, 발표

4. 프로젝트 절차



비즈니스 문제 정의

학생별 학력과 역량의 개인차로 인해 2022 교육부 교육과정만으로는 학습에 있어 충분한 교육이 제공되지 않을 수 있다. 때문에 각 학생의 특성과 발달에 적절한 교육 프로그램의 계획과 실행이 필요하다. 개별화 교육 계획 관련 연구를 통해 개별화 교육지원팀 운영에 대한 구체적 안내 지침의 부족, 실제 수업과 괴리된 비현실적 운영, 구성원들 간의 의사소통 부족 등 여러 문제점들이 지적되어 왔다. 이러한 문제점을 해결하는 데 있어 학생의 학업 성적을 예측하는 것은 교육과정에 대한 개별 이해도를 예측할 수 있다는 점에서 도움이 될 것이다.

학생들의 학업 성적을 예측하는 것은 성적이 낮은 학생을 예측하여 성적을 높이기 위한 다양한 교육 프로그램을 제공할 수 있다는 점에서 도움이 될 것이다. 본 조는 포르투갈 학생들의 성별, 나이, 가정 환경, 학교 관련 정보 등 다양한 특성을 포함하고 있는 데이터로 학생들의 학업 성적을 예측하는 모델을 개발하고자 했다. 이를 바탕으로 학생들의 성적을 예측하고 개개인의 학생 특성을 고려한 맞춤형 교육 및 지원 프로그램을 제공함으로써 학생들의 학업 성취도를 향상시킬 수 있을 것이다.

데이터 선정 기준

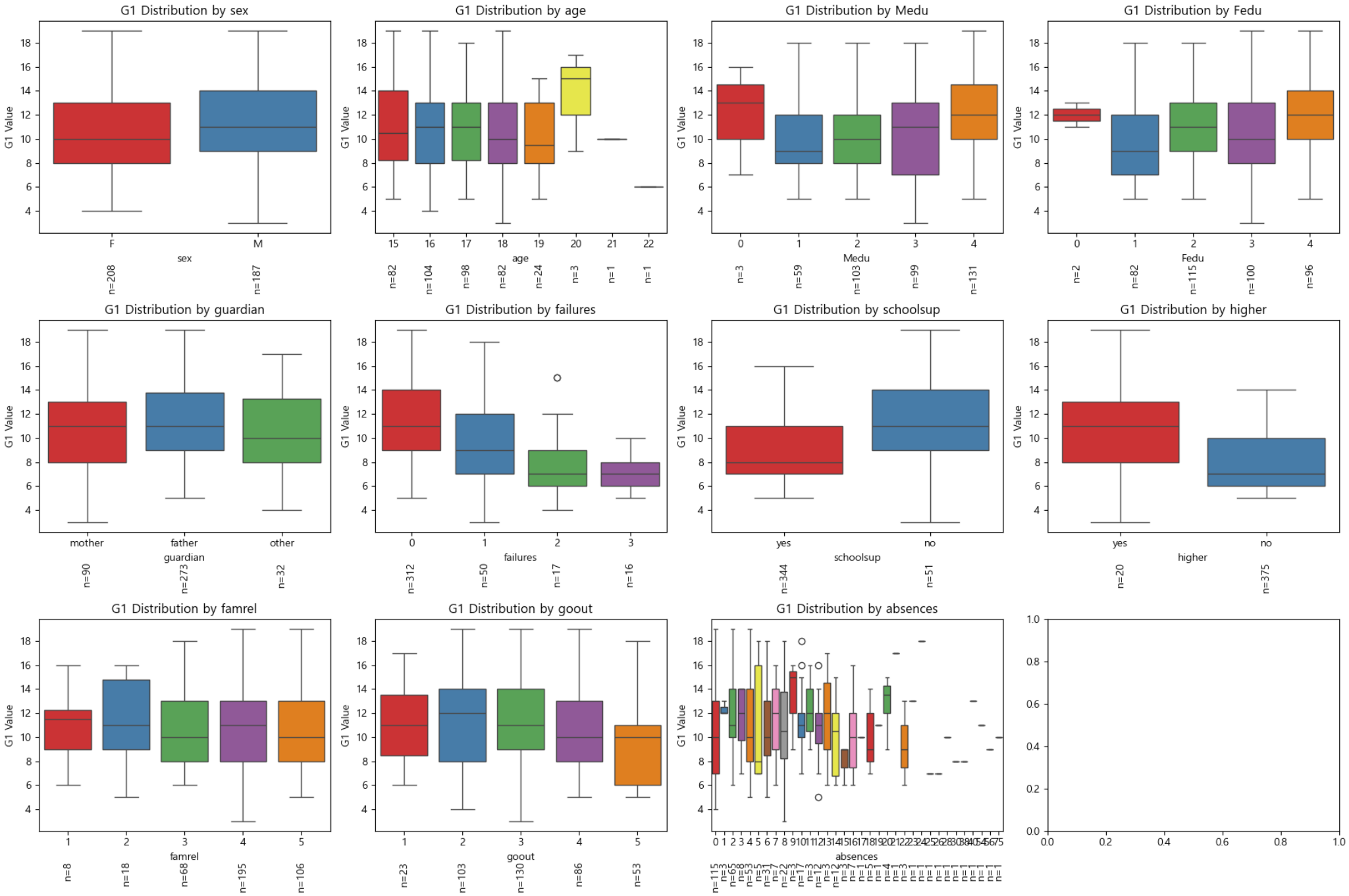
본 프로젝트에서 제공된 데이터는 포르투갈 학생들의 포르투갈어와 수학 과목에 대한 성적(G1)이다. 이 중 포르투갈어는 포르투갈의 모국어이기 때문에 본사에 시사점을 제공하기 어렵다고 판단하여 분석 대상에서 제외하였다. 수학은 국가에 상관 없이 교육이 이루어지는 공통 과목이므로 수학 데이터만을 분석함으로써 본사의 데이터를 사용했을 때 더 정확하게 예측하여 유용한 결과를 얻을 수 있을 것으로 기대한다.

변수 선정

sex	학생의 성별
age	학생의 나이
failures	과거 수업 낙제 횟수
schoolsup	추가 교육 지원 여부
higher	고등 교육 희망 여부
famrel	가족 관계 상황
goout	친구들과 노는 정도
absences	학교 결석 횟수
guardian_power	보호자의 학력

주어진 데이터 내 모든 변수에 대해서
각각의 변수의 값들에 대한 그룹화를 진행한 후
각각의 그룹 별 종속변수(G1) 값의 분포를
boxplot 및 t-test를 이용하여 확인 결과,
유의미한 결과가 도출된
'sex', 'age', 'failures', 'schoolsup', 'higher',
'famrel', 'goout', 'absences'를 변수로 선정함

변수별 데이터 분포


G1 Distribution by sex
G1 Distribution by age
G1 Distribution by Medu
G1 Distribution by Fedu
G1 Distribution by guardian
G1 Distribution by failures
G1 Distribution by schoolsup
G1 Distribution by higher
G1 Distribution by famrel
G1 Distribution by goout
G1 Distribution by absences
G1 Value
sex
age
Medu
Fedu
guardian
failures
schoolsup
higher
famrel
goout
absences

변수 선정

sex	학생의 성별
age	학생의 나이
failures	과거 수업 낙제 횟수
schoolsup	추가 교육 지원 여부
higher	고등 교육 희망 여부
famrel	가족 관계 상황
goout	친구들과 노는 정도
absences	학교 결석 횟수
guardian_power	보호자의 학력

도메인 지식을 활용하여 'Medu', 'Fedu', 'guardian'가 모델 개발에 유의미한 변수가 될 수 있다고 판단했으나, 세 변수를 모두 활용하여 One-hot encoding을 진행하면 많은 변수를 만들어야 함

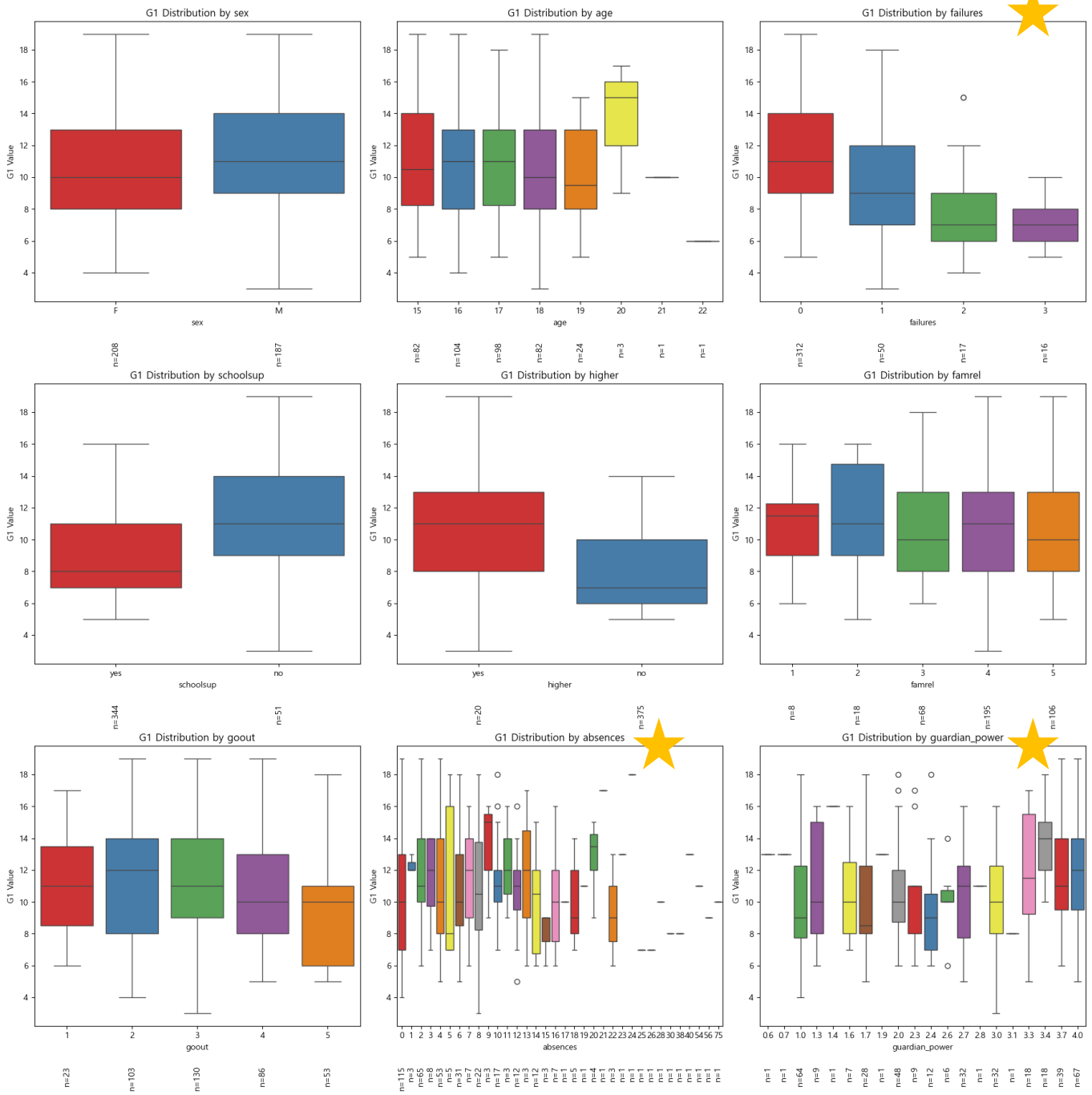
→ Feature engineering을 실시하여 세 변수를 활용한 하나의 변수를 생성

guardian_power	Guardian = mother → Medu X 0.7 + Fedu X 0.3
	Guardian = father → Medu X 0.3 + Fedu X 0.7

변수별 G1과의 관계 분석

	DF	F-value	P-value
sex	1.0	0.39	0.53
age	7.0	0.85	0.54
failures	3.0	2.40	0.10
schoolsup	1.0	0.59	0.45
higher	1.0	0.23	0.63
famrel	4.0	1.04	0.38
goout	4.0	0.55	0.65
absences	2.0	2.73	0.08
guardian_power	2.0	1.59	0.22

ANOVA 분석을 실시하여 각각의 변수가
성적에 유의미한 영향을 끼치는지 추가로 검증



가설 설정

귀무가설(H0)

'sex', 'age', 'failures', 'schoolsup', 'higher', 'famrel', 'goout',
'absences', 'guardian_power'
총 9개의 변수는 G3 성적을 유의하게 예측할 수 없다.

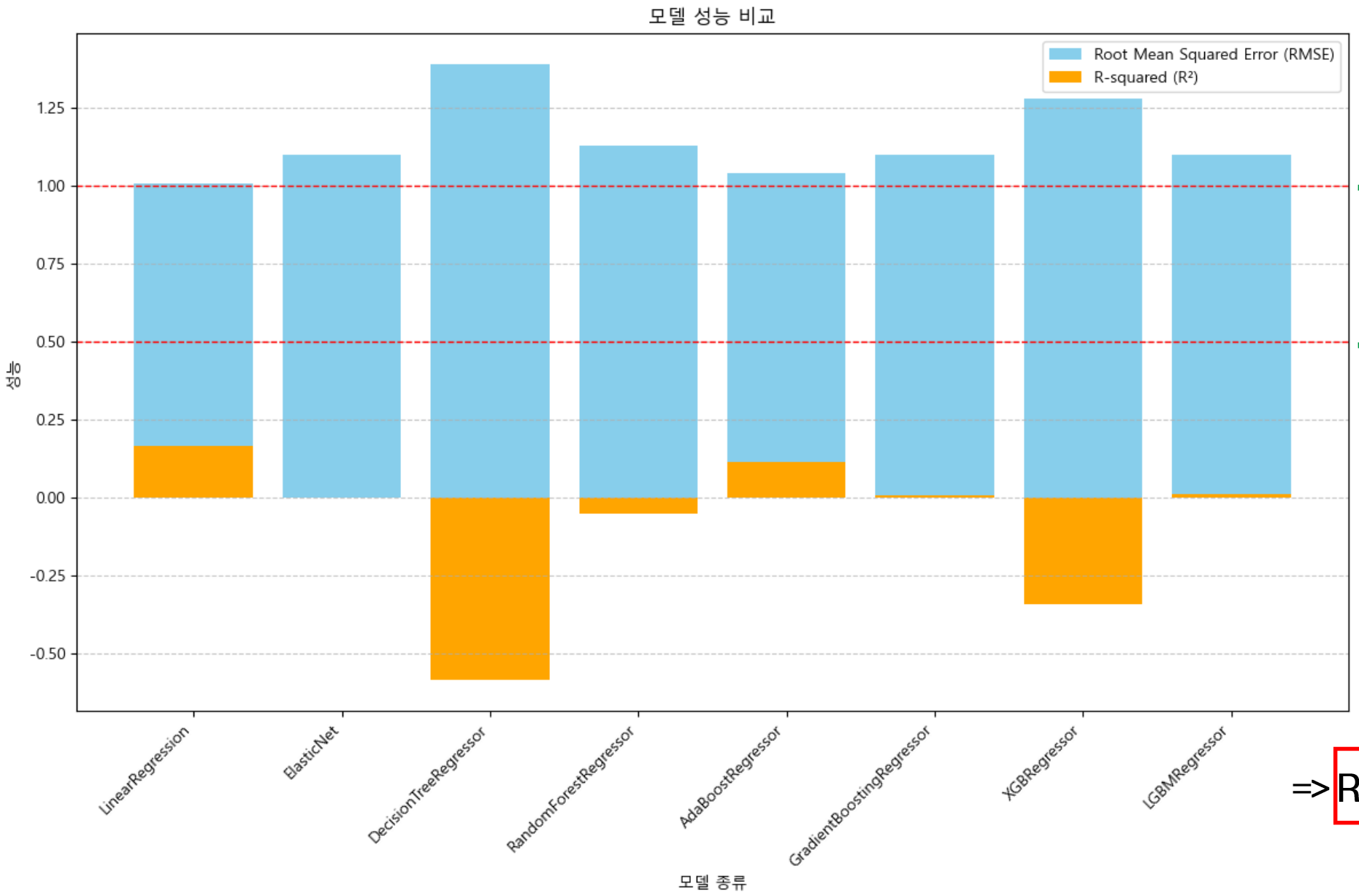
대립가설(H1)

'sex', 'age', 'failures', 'schoolsup', 'higher','famrel', 'goout',
'absences', 'guardian_power'
총 9개의 변수는 G3 성적을 유의하게 예측할 수 있다.

회귀, 분류 모델 비교

	모델	과적합 방지	선형성	변수중요도	계산속도	모델복잡성
회귀	Linear Regression	X	O	O	빠름	낮음
	Elastic Net	O	O	O	빠름	중간
	Decision Tree	O	X	O	느림	낮음
	Random Forest	O	X	O	느림	중간
	Ada Boost	O	X	O	느림	중간
	Gradient Boosting	O	X	O	느림	높음
	XGBoost	O	X	O	빠름	높음
	LGBM	O	X	O	빠름	높음
분류	LGBM	O	X	O	빠름	높음
	Gradient Boosting	O	X	O	느림	높음
	Random Forest	O	X	O	느림	중간
	LGBM + Random Forest	O	X	O	느림	높음

회귀 모델 성능 비교



평가 지표

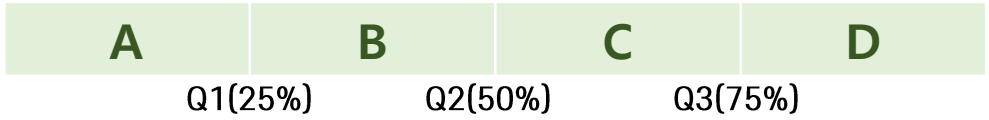
- RMSE
  (Root Mean Squared Error)
- R-squared(R^2)

→ 좋은 모델 성능 구간

=> RandomForestClassifier 모델 사용

모델 선정 기준

- Y(G1)을 Q1, Q2, Q3로 나누어 4개의 클래스로 분류

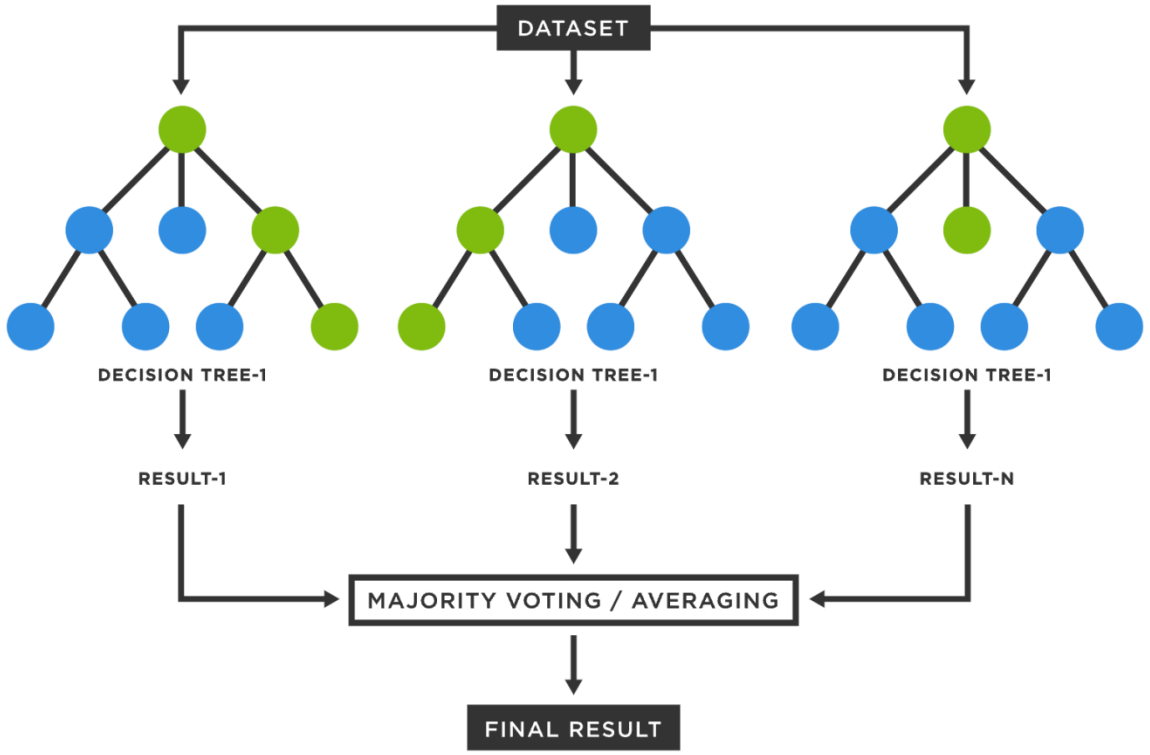


- 다양한 분류 모델을 사용하여 accuracy가 높은 모델을 선택

LGBM : 0.3671
Gradient Boosting : 0.4304
Random Forest: 0.4879
앙상블 모델(LGBM+RF): 0.3544

최종 모델 선정: Random Forest

모델 선정



Random Forest Classifier

- 트리 기반 알고리즘

평균을 사용 or 범주형 기능의 모드

높은 정확성 , 안정성 및 해석 용이성

- 랜덤 포레스트

수백 개의 결정 트리를 결합한 다음 서로 다른

관찰 샘플에 대해 각 결정 트리를 훈련

\+ 과적합 완화

\- 메모리 및 시간 소모

https://www.incodom.kr/Random_Forest

https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/

Grid Search Cross-Validation

Grid Search CV는 가능한 모든 하이퍼파라미터의 조합을 시도하여 최적의 조합을 찾는 방법

1) 가능한 모든 하이퍼파라미터 조합을 그리드 형태로 나열
2) 각각의 조합에 대해 교차 검증을 수행하여 모델의 성능을 측정
3) 가장 좋은 성능을 보이는 하이퍼파라미터 조합을 선택

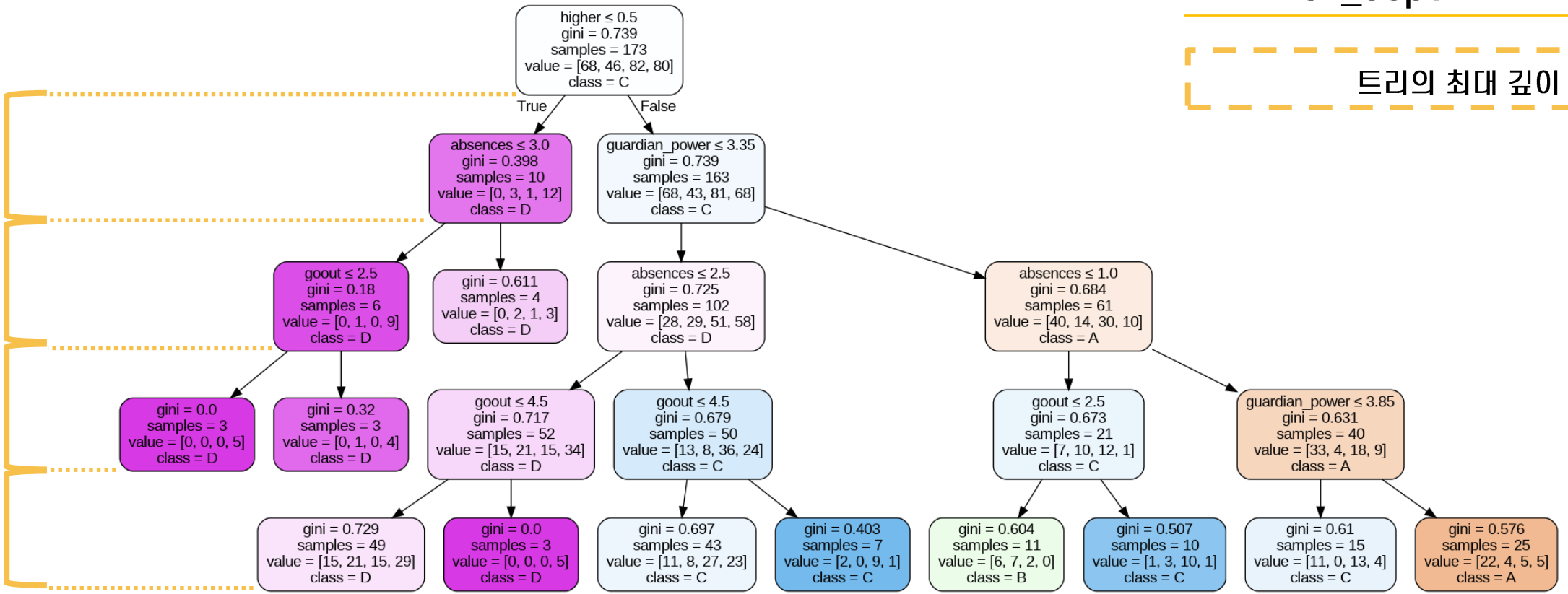
	Best Parameters	
트리의 최대 깊이 ←	max_depth	4
리프 노드가 가져야 할 최소 smaple 수 ←	min_samples_leaf	2
노드를 분할하기 위한 최소 sample 수 ←	min_samples_split	5
결정트리의 개수 ←	n_estimators	200

cross validation	5

Random Forest Classifier

max_depth	4

트리의 최대 깊이



총 395명의 데이터를 7:3의 비율로 Train set과 Test set으로 분할

Grid Search CV를 통해 가장 좋은 성능을 보이는 하이퍼파라미터 조합으로

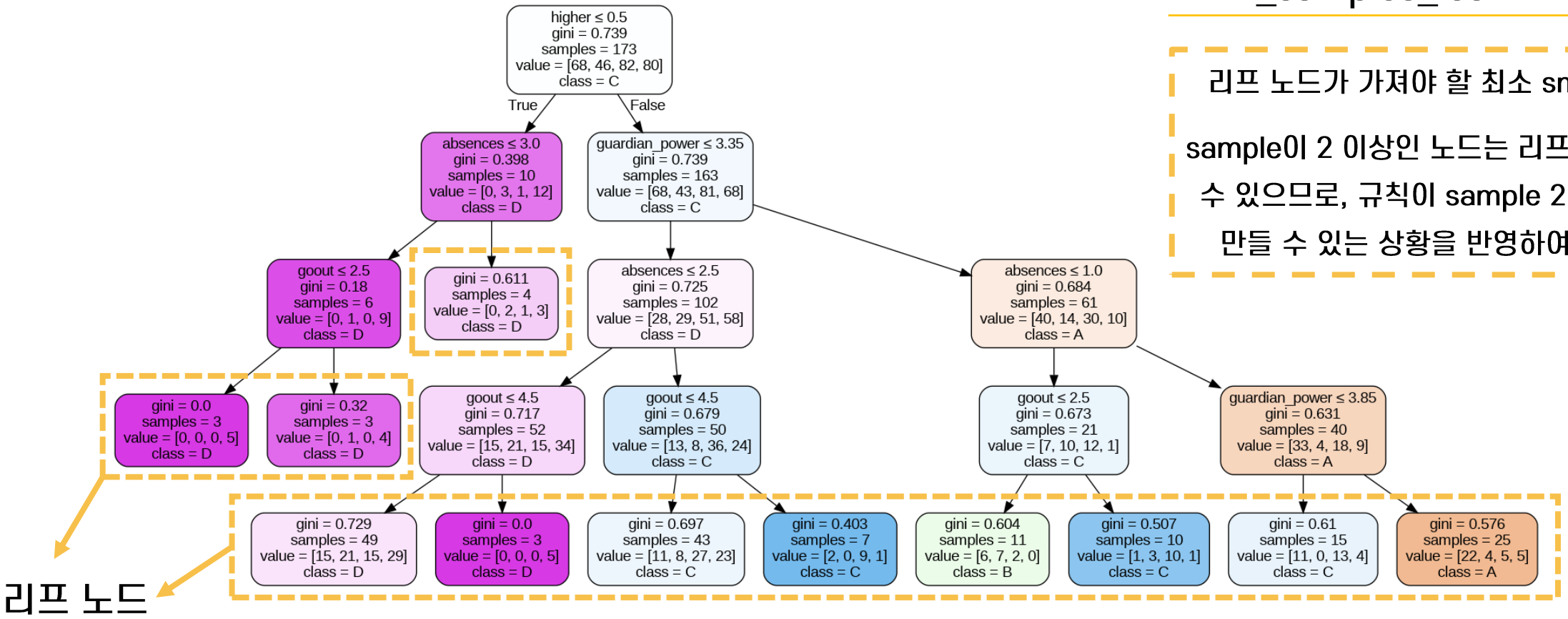
Random Forest Classifier로 학습한 결과

Random Forest Classifier

min_samples_leaf	2

리프 노드가 가져야 할 최소 smaple 수

sample이 2 이상인 노드는 리프 노드가 될 수 있으므로, 규칙이 sample 2인 노드를 만들 수 있는 상황을 반영하여 변경됨



총 395명의 데이터를 7:3의 비율로 Train set과 Test set으로 분할

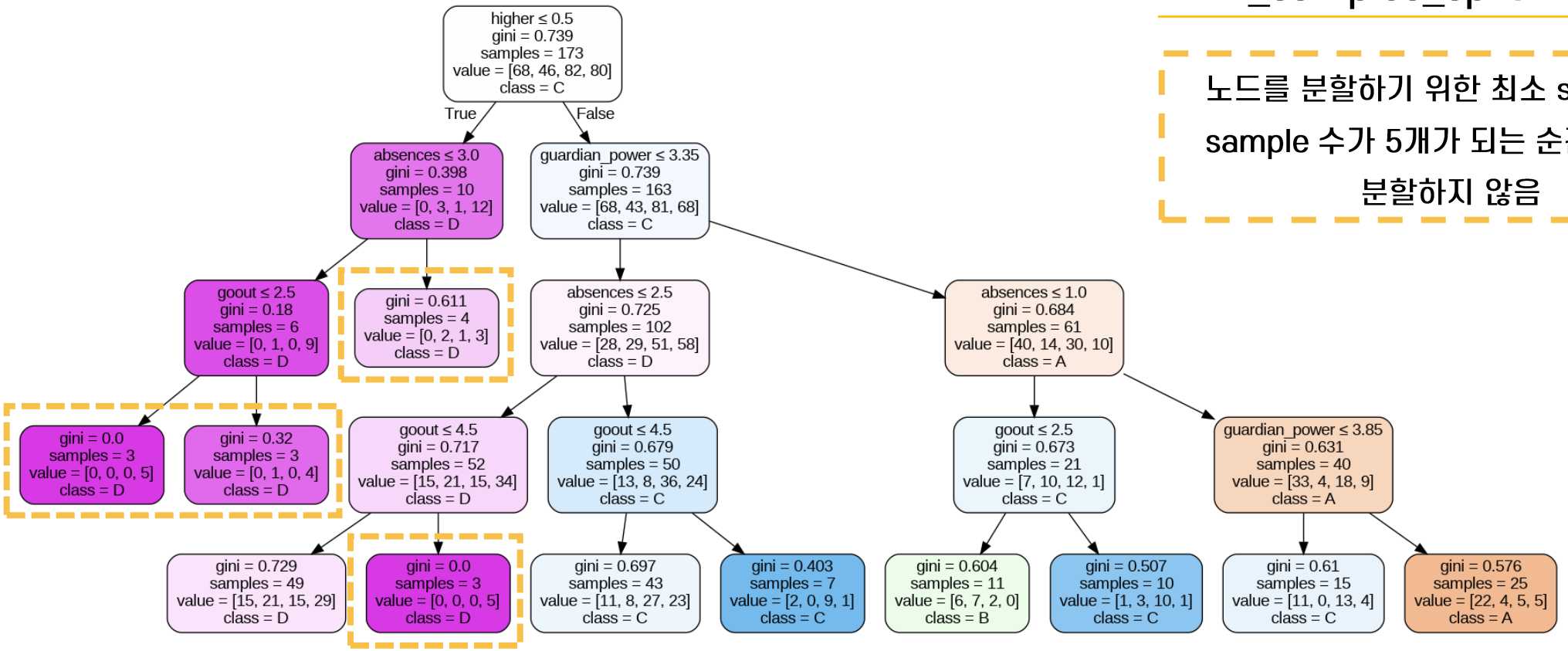
Grid Search CV를 통해 가장 좋은 성능을 보이는 하이퍼파라미터 조합으로

Random Forest Classifier로 학습한 결과

Random Forest Classifier

min_samples_split	5

노드를 분할하기 위한 최소 sample 수
sample 수가 5개가 되는 순간 더 이상 분할하지 않음



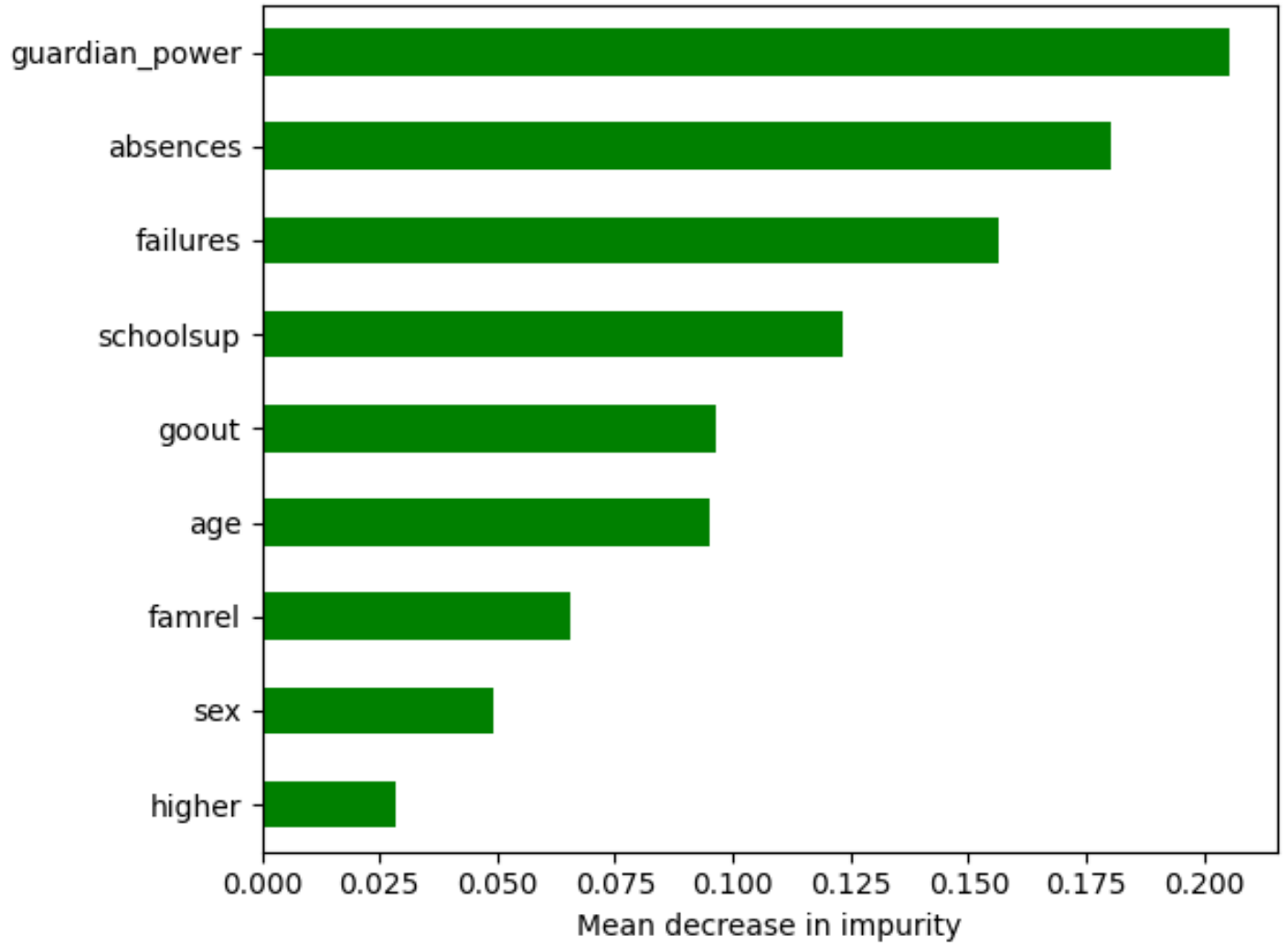
총 395명의 데이터를 7:3의 비율로 Train set과 Test set으로 분할

Grid Search CV를 통해 가장 좋은 성능을 보이는 하이퍼파라미터 조합으로
Random Forest Classifier로 학습한 결과

Random Forest Classifier

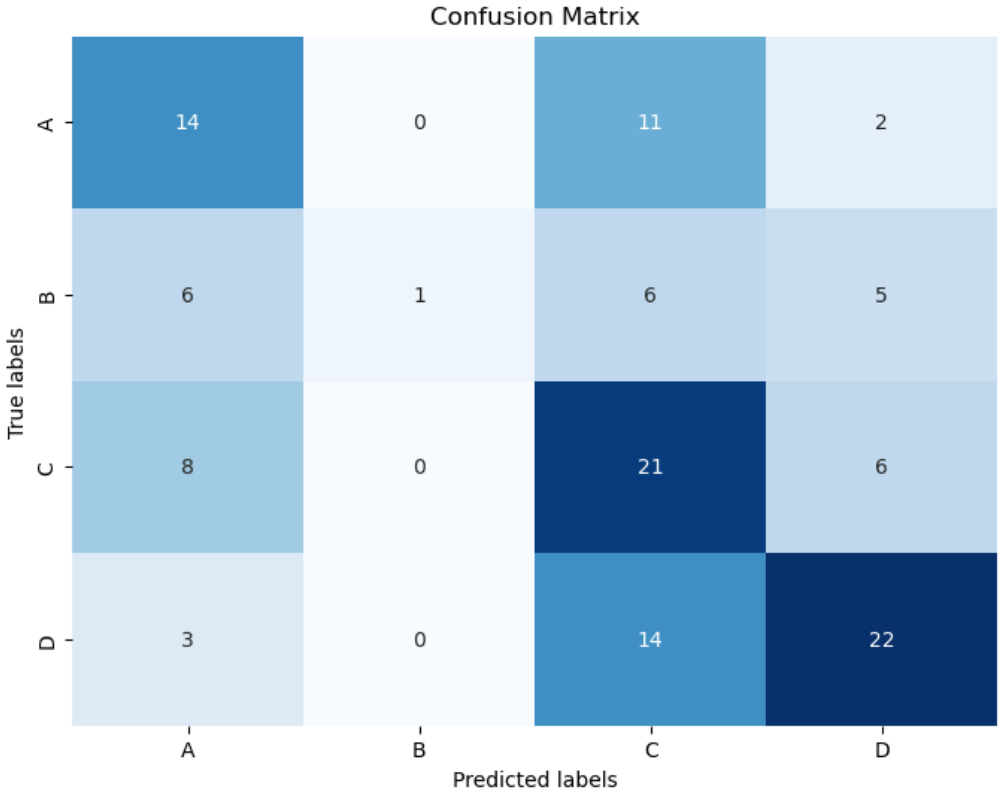
Feature Importance

guardian_power	보호자의 학력
absences	학교 결석 횟수
failures	과거 수업 낙제 횟수
schoolsup	추가 교육 지원 여부
goout	친구들과 노는 정도
age	학생의 나이
famrel	가족 관계 상황
sex	학생의 성별
higher	고등 교육 희망 여부



Random Forest Classifier

Score	
Precision	**0.6210**
Recall	**0.4346**
F1 score	**0.4164**
Accuracy	**0.4874**



모델의 정밀도(Precision)를 높임으로써

모델이 예측하는 클래스에 속해있는 학생이 올바르게 속해있을 확률을 높일 수 있음

모델의 예측 효율을 바탕으로 적절한 개별화 교육이 이루어질 수 있기를 기대

기대효과

1

모델 결과
→ B/D 집단에서
높은 정밀도를
보임

2

수학 성적
B(중상)/D(하위)
집단에 대하여
수준별 학습 제공

3

상위권 도약을
위한 고난도 문제
제공

3

공부에 흥미를
가질 수 있는
흥미유발 자료 제공

감사합니다.