DATE. 2024.07.03

# 머신러닝 개인 프로젝트

빅데이터 8기 이수민

CONTENTS

# 목차

# 데이터

- 데이터: train.csv
- 33개의 컬럼
- 총 317개의 데이터

- school: 학생의 학교
- sex: 성별
- age: 나이
- address: 거주지
- famsize: 가족 규모

...

- Dalc: 평일 알코올 섭취
- Walc: 주말 알코올 섭취
- Health: 현재 건강 상태
- Absence: 학교 결석 일수
- G1: 1학기 점수
- G2: 2학기 점수
- G3: 3학기 점수

- 분석 목표

- Data: 'G3'를 제외한 모든 컬럼
- Target: 'G3'

총 샘플 수는 397개로, 무작위로
섞어 317개의 샘플을 학습 데이터로서 배포하고,
함께 첨부한 80개의
평가 샘플은 암호화되어 있으며, 평가일 당일에
패스워드를 배포합니다.
학습 데이터에서 교차검증을 통해 최적의 모델을
학습한 후, 제출일 오후 5시에 배포한 평가 샘플의
정확도를 출력합니다.

# 데이터 전처리

## 1. 원핫인코딩

분석을 위해 범주형으로 구성되어 있는 컬럼들
의 변수들에 원핫인코딩을 적용하였다.

ex) 'school' 컬럼 - 'GP', 'MS' 로 구성
↓
'school_GP', 'school_MS' 컬럼 - 0, 1로 구성

## 1. 변수값 변경

'yes', 'no'로 구성되어 있는 컬럼들의 변수들을
'yes'는 1, 'no'는 0으로 변경하였다.

# 머신러닝 - 회귀 분석 과정(train)

- 사용한 알고리즘: RandomForestRegressor(랜덤포레스트 회귀 모델)
- 하이퍼파라미터 탐색: RandomizedSearchCV 사용, n_iter = 300으로 설정하여 300번의 랜덤 샘플링 수행
- 사용된 하이퍼파라미터

  : min_impurity_decrease(과적합 방지, 모델의 일반화 성능을 높이기 위해)

  : max_depth(과적합 방지, 계산 효율성을 높이기 위해)

  : min_samples_split(모델의 복잡도를 조절하여 과적합 방지)

  : min_samples_leaf(과적합 방지, 모델의 안정성을 높이기 위해)

# 머신러닝 - 회귀 분석 과정(train)

- 최적 하이퍼파라미터 출력

| min_impurity_decrease | max_depth | min_samples_split | min_samples_leaf |
|---|---|---|---|
| 0.00028223608778806236 | 45 | 3 | 4 |

- 교차검증: cross_validate 함수를 사용하여 모델의 성능 평가, 5-fold 교차 검증
- 예측 및 평가:  교차 검증 결과 & 검증 데이터에 대한 $R^2$ 값 출력

# 머신러닝 - 회귀 분석 결과(train)

- 검증 데이터에 대한 결과

  Validation $R^2$: 0.8980919011234413

- 교차 검증 결과

  Cross-validated score mean

  : (np.float64(0.9337154754805976), np.float64(0.8606601952670925))

  train_score　　test_score

# 머신러닝 - 회귀 분석 결과(train)

- 특성 중요도

| Feature | Importance |
|---|---|
| G2 | 0.803859 |
| absences | 0.126912 |
| age | 0.014961 |
| reasom_course | 0.005171 |
| farmel | 0.004872 |
| failures | 0.004672 |
| G1 | 0.004206 |

| Feature | Importance |
|---|---|
| health | 0.803859 |
| reason_home | 0.126912 |
| freetime | 01014961 |
| Fedu | 0.005171 |
| Walc | 0.004872 |
| guardian_mother | 0.004672 |
| traveltime | 0.004206 |

...

특성 중요도를 출력해보았을 때,
'G2'가 0.803859로 가장 높은 중요도를 보였다.
그 다음으로는 0.126912의 'absences'가 높은
중요도를 가지고 있음을 확인할 수 있다.

# 머신러닝 - 테스트 데이터 적용 결과

- 최종 결과

  Test $R^2$: 0.917305263515573