데 이 터 셋 분 석

STUDENTS PERFORMANCE IN EXAM

양상우

컴퓨터공학과 3학년 - 2016108271

STUDENTS PERFORMANCE IN EXAM

데이터셋 분석

목차

01

데 이 터 셋 분 석

데이터 소개

데이터 소개

STUDENTS PERFORMANCE IN EXAM

다양한 과목에서 학생들이 얻은 점수로 구성 되어 있는 데이터 셋 입니다.

이를 통해

부모 배경, 시험 준비 등이 학생 수행에 미치는 영향을 이해 할수 있습니다.

데이터 확인

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

- gender : 성별(남/녀)
- race/ethnicity : 인종/민족
- parent level of education : 부모 학력 수준
- lunch : 점심식사 여부
- test preparation course : 시험 준비학습 여부
- math score : 수학 성적
- reading score : 읽기 성적
- writing score : 쓰기 성적

인공 지능

데이터 확인

```
# 데이터 내용 간략히 정리

print("전체 데이터 수:",student.shape[0] * student.shape[1])
print("결측치 수:",student.isnull().sum().sum())
print("전체 학생 수:", student["gender"].count())
```

- 전체 데이터 수 : 8000
- 결측치 수 : 0
- 전체 학생 수 : 1000

```
<bound method NDFrame.head of      gender race/ethnicity parental level of education         lunch  \
0    female        group B           bachelor's degree      standard
1    female        group C                some college      standard
2    female        group B             master's degree      standard
3      male        group A          associate's degree  free/reduced
4      male        group C                some college      standard
..      ...            ...                         ...           ...
995  female        group E             master's degree      standard
996    male        group C                 high school  free/reduced
997  female        group C                 high school  free/reduced
998  female        group D                some college      standard
999  female        group D                some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72             72             74
1                 completed          69             90             88
2                      none          90             95             93
3                      none          47             57             44
4                      none          76             78             75
..                      ...         ...            ...            ...
995               completed          88             99             95
996                    none          62             55             55
997               completed          59             71             65
998               completed          68             78             77
999                    none          77             86             86

[1000 rows x 8 columns]>
```

STUDENTS PERFORMANCE IN EXAM

데이터셋 분석

2016108271

인공지능

분석에 앞서..

01 **부모 학벌이 자녀에 학생 성적에 가장 큰 영향을 끼치지 않을까?**

아무래도 "스카이캐슬"에서 보았듯이 부모들이 자기 자식 또한 더 좋은 학벌을 만들기 위해 노력을하고, 더 많은 사교육 하려고 하지 않을까?

02 **인종/민족과 성적에서 성적의 차이가 없지 않을까?**

같은 교육을 받았다면, 인종/민족간의 차이가 없을것 같다. 같은 '인간'이라는 특성상 큰 차이가 없을것 같다

03 **성적에 가장 중요한 특성은 무엇일까?**

항상 궁금해 왔다, 여러 특성중에 어떤게 가장 중요한 특성일까? 이를 통해 내 기말도 잘 보았으면 좋겠다...

04 **아무래도 여자가 좀 더 성적이 좋지 않을까?**

살아오면서 보았던 사람들중 여성분들이 좀 더 학업을 열심히 한 경우가 많았던것 같다(경험) 이 데이터 역시 여성분들이 좀 더 좋은 결과가 나올 것 같다

STUDENTS PERFORMANCE IN EXAM

데이터셋 분석

2016108271

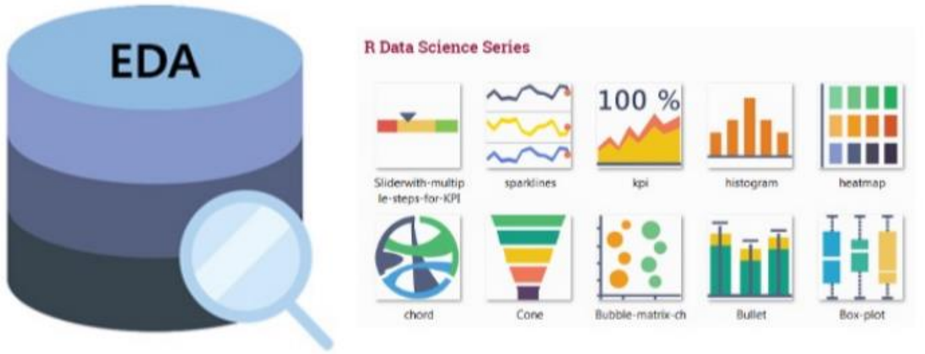
03

데 이 터 셋 분 석

EDA & Visualization

PRESENTATION

EDA?



EDA의 정의

EDA(Exploratory Data Analysis)란, 탐색적 데이터 분석을 의미한다.

데이터 분석에 있어서 매우 중요한, 초기 분석의 단계이자 해야하는 일이다.

데이터에 대한 일종의 견적을 내는 일이라고 비유할 수 있겠다.

주어진 데이터의 특성을 알아야 내가 이 데이터로 해결하고자 하는 문제를 해결할 수 있는 방법을 찾아볼 수 있기 때문이다..

EDA의 목적

시각화 및 통계 도구를 사용하여 데이터를 이해할 수 있다.

도출하고자 하는 결과의 기본이 되는 가정에 접근하고 가정을 검증할 수 있다.

모델을 만들기 전에 데이터를 이해합니다.

03 : EDA

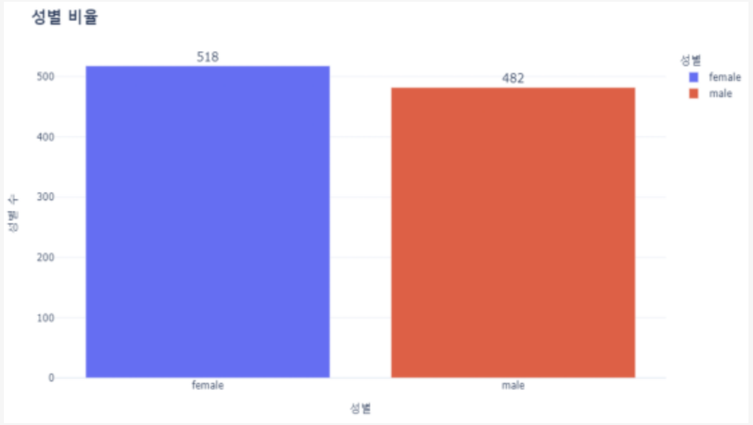
a. 성별과 성적

b. 인종/민족과 성적

c. 부모 교육 수준과 성적

d. 점심 식사 여부와 성적

e. 시험 준비 여부와 성적



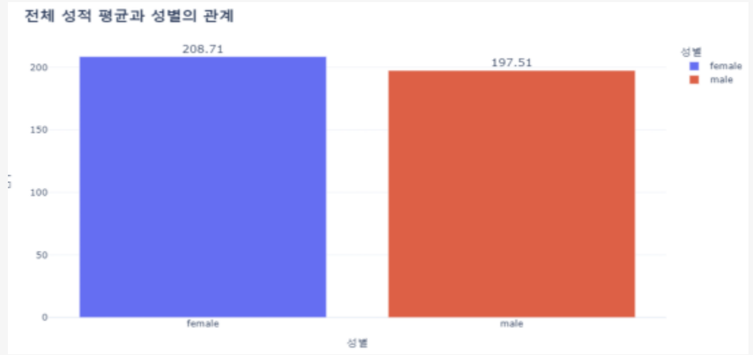
성별의 비율

간단하게 성별에 대한 성별수로 막대 그래프로 나타냈습니다

사용한 코드는 px.bar입니다.

현재 여학생이 518명, 남학생이 482명으로 여학생이 조금더 많은 데이터

라는 것을 알수 있습니다.



전체 성적 평균과 성별의 관계

전체 성적 합계를 total score로 컬럼으로 등록한 뒤에

막대 그래프로 나타낸것입니다

전체 성적인 경우

현재 여학생이 208.71점, 남학생이 197.51점으로

여학생이 남학생보다 조금 더 높습니다.

개별 과목 평균과 성별의 관계

전체 데이터 컬럼에 있는 수학, 읽기, 쓰기 성적을 성별로 구분했습니다

수학의 경우

여학생 63.63 남학생 68.73으로 남학생이 높고

읽기, 쓰기의 경우

대체적으로 여학생 성적이 더 높다는 것을 알 수 있습니다.

02 : EDA

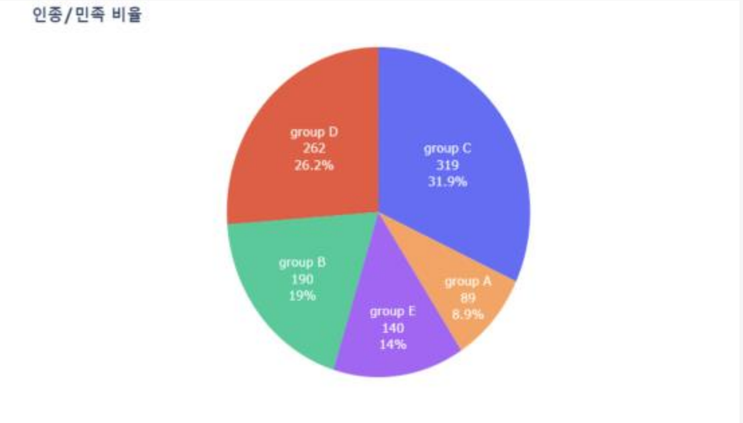
a. 성별과 성적

b. 인종/민족과 성적

c. 부모 교육 수준과 성적

d. 점심 식사 여부와 성적

e. 시험 준비 여부와 성적

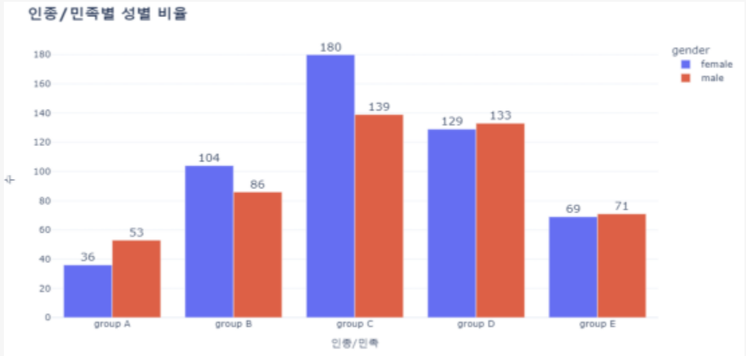


인종/민족의 분포도

기존 데이터가

특정 인종/민족을 지정하진 않고 같은 인종/민족끼리 그룹화 돼있습니다.

그룹 C가 가장 많고, 그룹 A가 가장 적다는 것을 알수 있습니다



인종/민족의 성별 비율

- 그룹 A : 남학생이 더 많습니다
- 그룹 B : 여학생이 더 많습니다
- 그룹 C : 여학생이 더 많습니다
- 그룹 D : 남학생이 더 많습니다
- 그룹 E : 남학생이 더 많습니다

02 : EDA

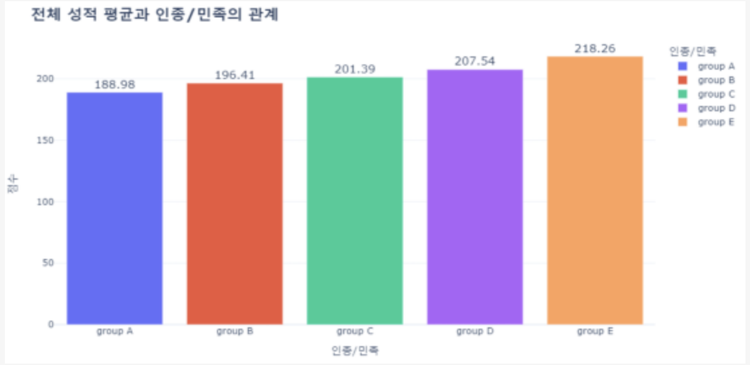
a. 성별과 성적

b. 인종/민족과 성적

c. 부모 교육 수준과 성적

d. 점심 식사 여부와 성적

e. 시험 준비 여부와 성적



전체 성적 평균과 인종/민족의 관계

- 그룹 A : 189.98
- 그룹 B : 196.41
- 그룹 C : 201.39
- 그룹 D : 207.54
- 그룹 E : 218.26

으로 그룹 E가 가장 높은 성적, 그룹 A가 가장 낮은 성적을 가지고 있습니다.
하지만 전체적으로 큰차이가 나지 않습니다.

개별 과목별 성적과 인종/민족의 관계

- 그룹 A : 189.98
- 그룹 B : 196.41
- 그룹 C : 201.39
- 그룹 D : 207.54
- 그룹 E : 218.26

모든 과목에서 그룹 E가 가장 높으며, 그룹A가 가장 낮으나
전체 그룹간 차이가 크게 나지않습니다.

02 : EDA

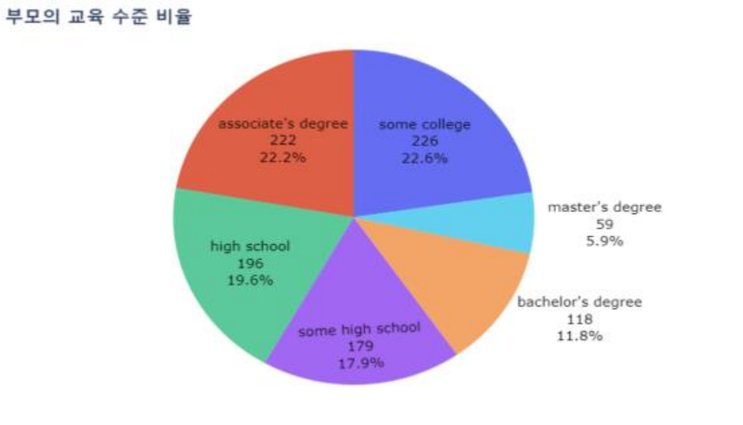
a. 성별과 성적

b. 인종/민족과 성적

c. 부모 교육 수준과 성적

d. 점심 식사 여부와 성적

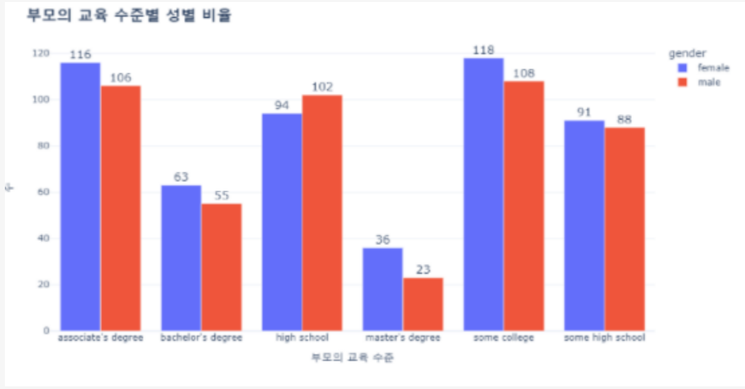
e. 시험 준비 여부와 성적



부모의 교육 수준 비율

-학사 수준 이상의 교유글 받은 비율이 약 62.5% 이며

나머지는 고등 교육까지 이수 했습니다.



부모의 교육 수준 성별 비율

부모의 교육 수준이 some college 인 여학생/ 남학생의 수가 가장 많고

부모의 교육 수준이 high school 인 경우를 제외하고,

여학생 수가 남학생 수 보다 많습니다.

02 : EDA

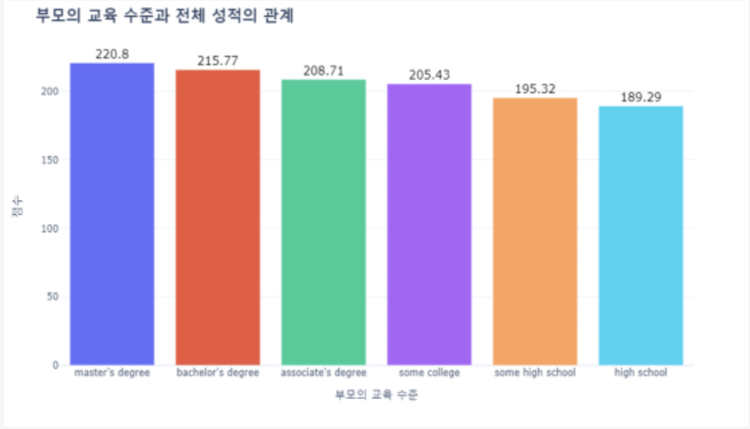
a. 성별과 성적

b. 인종/민족과 성적

c. 부모 교육 수준과 성적

d. 점심 식사 여부와 성적

e. 시험 준비 여부와 성적

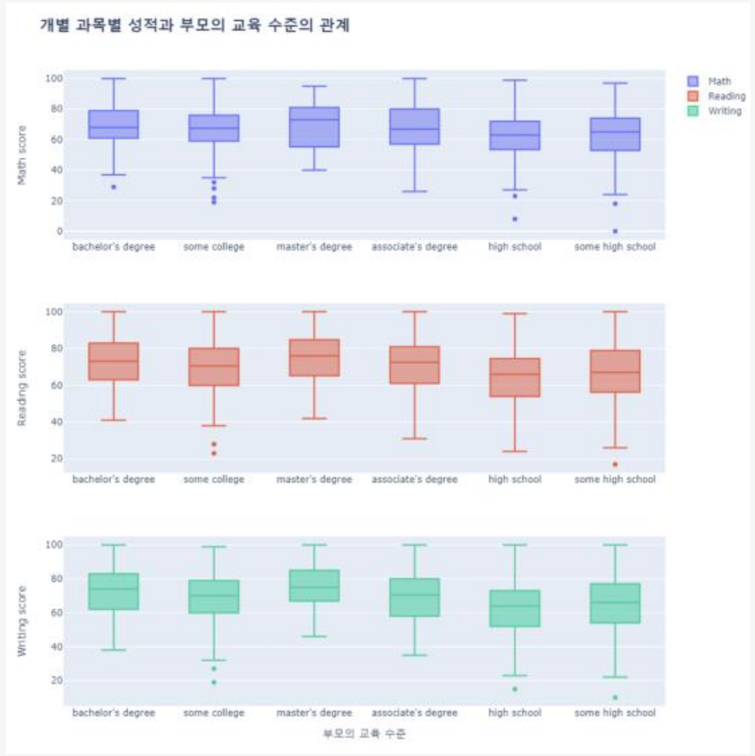


부모의 교육 수준과 전체 성적의 관계

부모의 교육 수준이 master's degree인 경우가 성적이 높고

high school인 경우 성적이 가장 낮지만

전반적으로 큰 차이가 나지 않습니다.



개별 과목별 성적과 부모의 교육 수준의 관계

- math의 경우

high school이 조금 낮지만

전반적으로 크게 차이나지 않습니다

02 : EDA

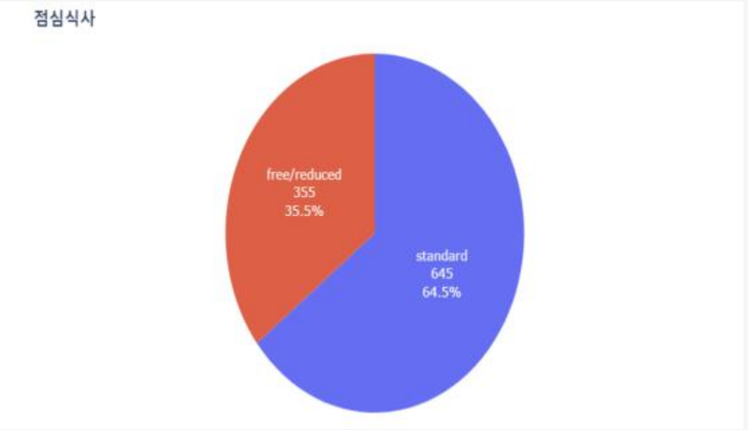
a. 성별과 성적

b. 인종/민족과 성적

c. 부모 교육 수준과 성적

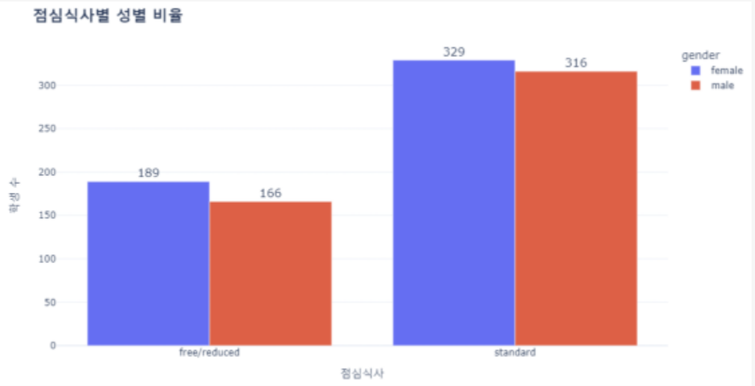
d. 점심 식사 여부와 성적

e. 시험 준비 여부와 성적



점심식사를 하는 학생

급식을 하는 학생이 64.5%로
더 많습니다.



점심식사별 성별 비율

점심식사 종류와 상관없이 여학생이 더 많지만
남학생과 차이가 크지 않습니다.

02 : EDA

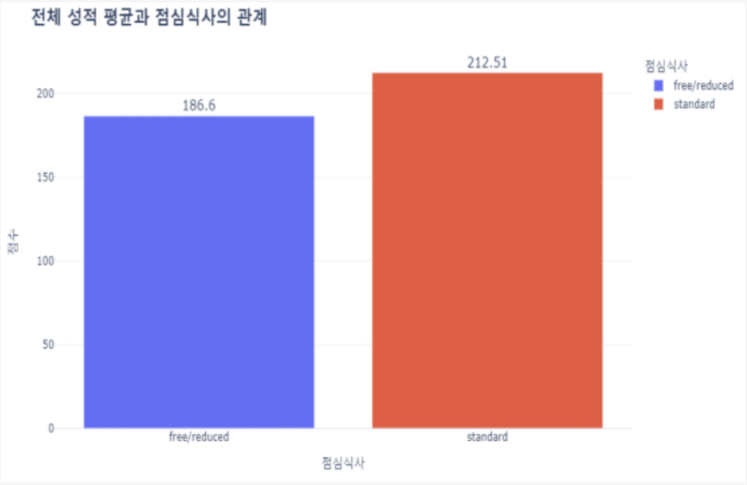
a. 성별과 성적

b. 인종/민족과 성적

c. 부모 교육 수준과 성적

d. 점심 식사 여부와 성적

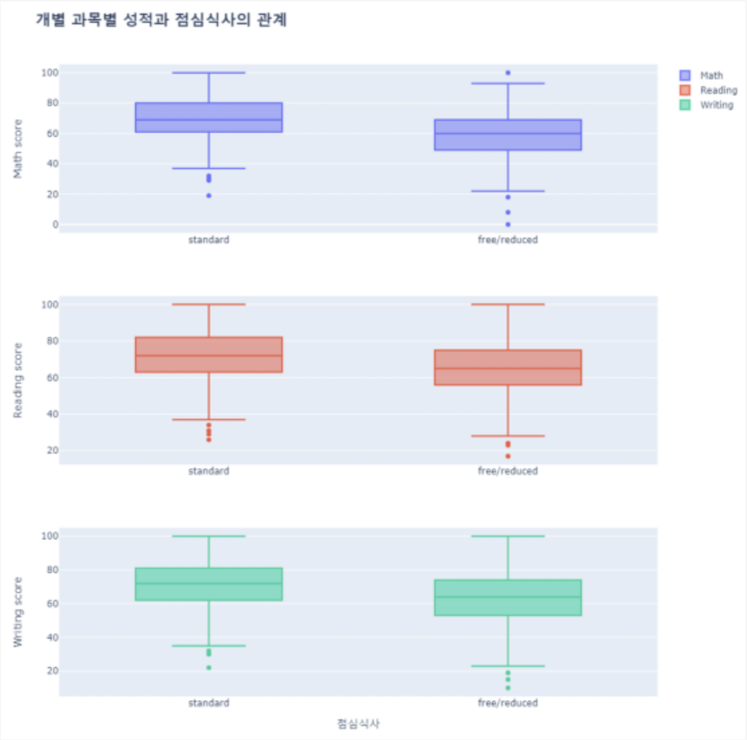
e. 시험 준비 여부와 성적



전체 성적 평균과 점심식사의 관계

- 점심 식사를 하지 않는 그룹 : 186.6
- 점심 식사를 하는 그룹 : 212.51

급식을 하는 경우가 하지 않는 경우보다 성적이 더 높단는걸 확인할 수 있습니다



개별 과목별 성적과 점심식사의 관계

모든 과목에서 급식을 하는 경우가 하지 않는 경우보다 성적이 더 높습니다.

02 : EDA

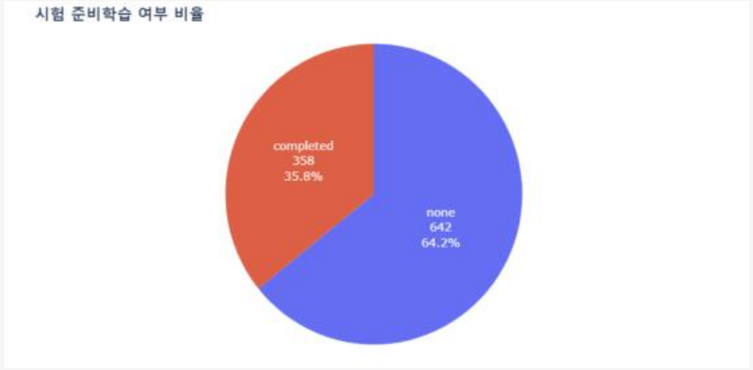
a. 성별과 성적

b. 인종/민족과 성적

c. 부모 교육 수준과 성적

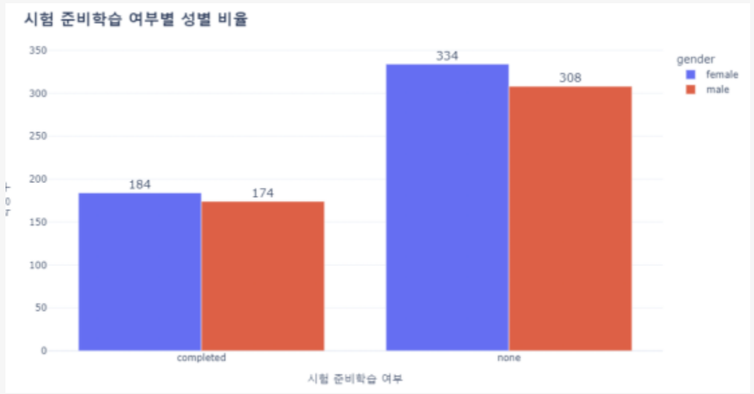
d. 점심 식사 여부와 성적

e. 시험 준비 여부와 성적



시험 준비학습 여부 비율

- 준비하지 않는 그룹



시험 준비학습 여부 별 성별 비율

- 시험 준비학습 여부에 상관없이 여학생의 수가 남학생보다 더 많습니다.

02 : EDA

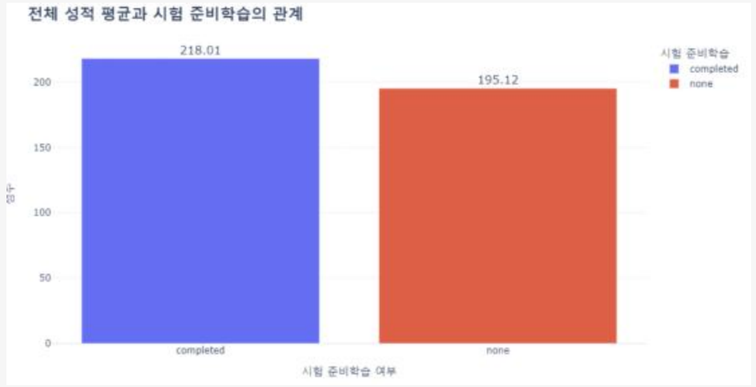
a. 성별과 성적

b. 인종/민족과 성적

c. 부모 교육 수준과 성적

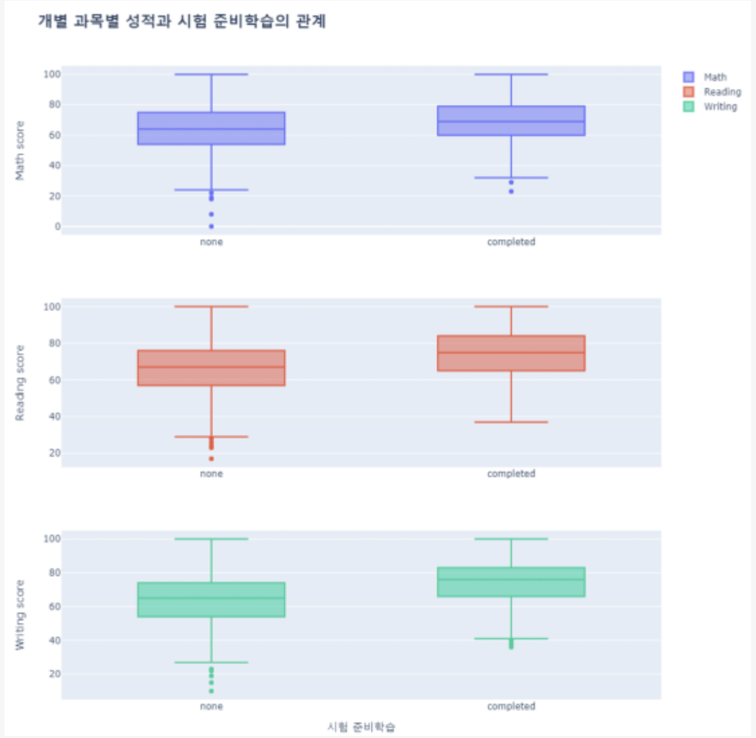
d. 점심 식사 여부와 성적

e. 시험 준비 여부와 성적



전체 성적 평균과 시험 준비학습의 관계

- 시험 준비학습을 한 경우가 하지 않은 경우 보다 점수가 더 높습니다.

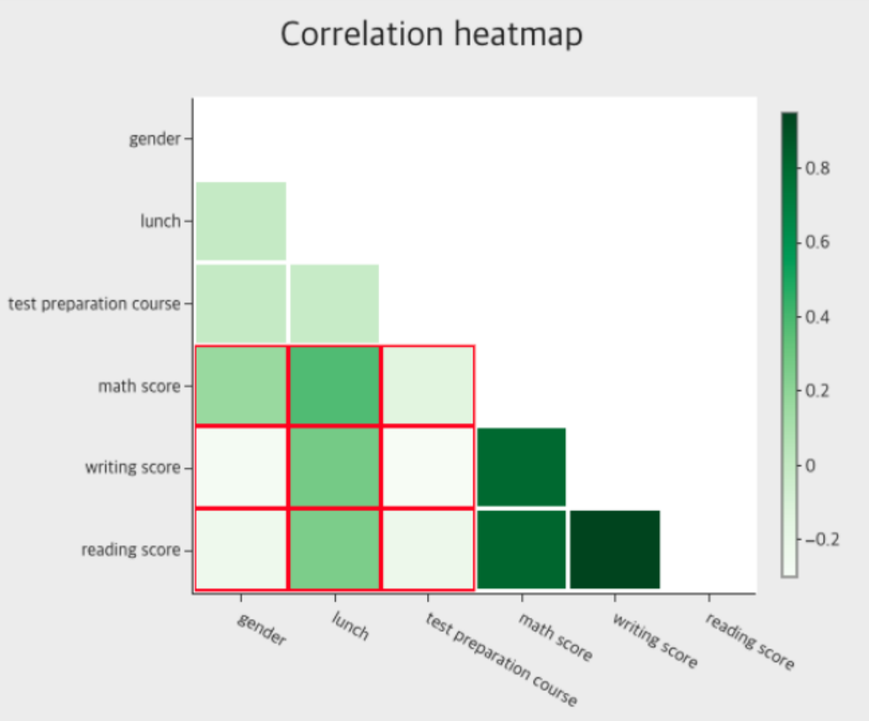


개별 과목별 성적과 시험 준비학습의 관계

모든 과목에서 시험 준비학습을 하는 경우가 하지 않는 경우보다 성적이 더 높습니다.

PRESENTATION

EDA : 상관 관계 히트맵



상관관계 히트맵

- 여기서는 점-양분 상관 계수를 사용하여
독립 변수(범주) vs 연속 타겟 사이의 상관 관계를 확인합니다.

- 우리는 상관관계가 약하다는 것을 알 수 있습니다.
가장 높은 상관관계는 0.38로 점심과 수학 시험 점수입니다.

03

데 이 터 셋 분 석

REVIEW

PRESENTATION

REVIEW : 성별과 성적의 관계는?

여학생 518명 남학생 482명

전체 성적의 경우

- 여학생이 남학생보다 조금 더 높습니다

개별 과목

- 수학 : 남학생이 여학생보다 더 높음
- 읽기 : 여학생이 남학생보다 더 높음
- 쓰기 : 여학생이 남학생보다 더 높음

PRESENTATION

REVIEW : 인종/ 민족과 성적의 관계는?

비율 : 그룹 C가 가장많고 A가 가장 적음
인원 : 그룹 C가 가장 많음

전체 성적의 경우

- 그룹 E가 가장 성적 평균이 높고, A가 가장 낮으면 큰 차이가 없습니다.

개별 과목

- 수학 : 그룹 E가 가장 성적 평균이 높고, A가 가장 낮으면 큰 차이가 없습니다.
- 읽기 : 그룹 E가 가장 성적 평균이 높고, A가 가장 낮으면 큰 차이가 없습니다.
- 쓰기 : 그룹 E가 가장 성적 평균이 높고, A가 가장 낮으면 큰 차이가 없습니다.

PRESENTATION

REVIEW : 부모의 교육 수준과 성적의 관계는?

학사 수준 이상의 교육을 받은 비율이 약 62.5% 이며
나머지는 고등 교육까지 이수

전체 성적의 경우

- 부모의 교육 수준이 master's degree 인 경우가 성적이 가장 높고
- high school 인 경우가 성적이 가장 낮다
= 하지만 전반적으로 큰 차이가 나지 않는다

개별 과목

- 수학 : high school 인 경우가 조금 낮다
- 읽기 : master's degree 인 경우가 성적이 조금 높다
- 쓰기 : master's degree 인 경우가 성적이 조금 높다
= 하지만 전반적으로 큰 차이가 나지 않는다

PRESENTATION

REVIEW : 점심식사 여부와 성적의 관계는?

급식을 하는 학생이 64.5%

전체 성적의 경우

- 급식을 하는 경우 > 급식을 하지 않는 경우

개별 과목

- 수학 : 급식을 하는 경우 > 급식을 하지 않는 경우
- 읽기 : 급식을 하는 경우 > 급식을 하지 않는 경우
- 쓰기 : 급식을 하는 경우 > 급식을 하지 않는 경우

= 전체적으로 급식을 하는 경우가 급식을 하지 않는 경우보다 높습니다.

PRESENTATION

REVIEW : 시험 준비 학습 여부와 성적의 관계는?

시험을 준비하지 않는 학생은 64.2%이다

전체 성적의 경우

- 시험 준비 학습을 하는 경우 > 시험 준비 학습을 하지 않는 경우

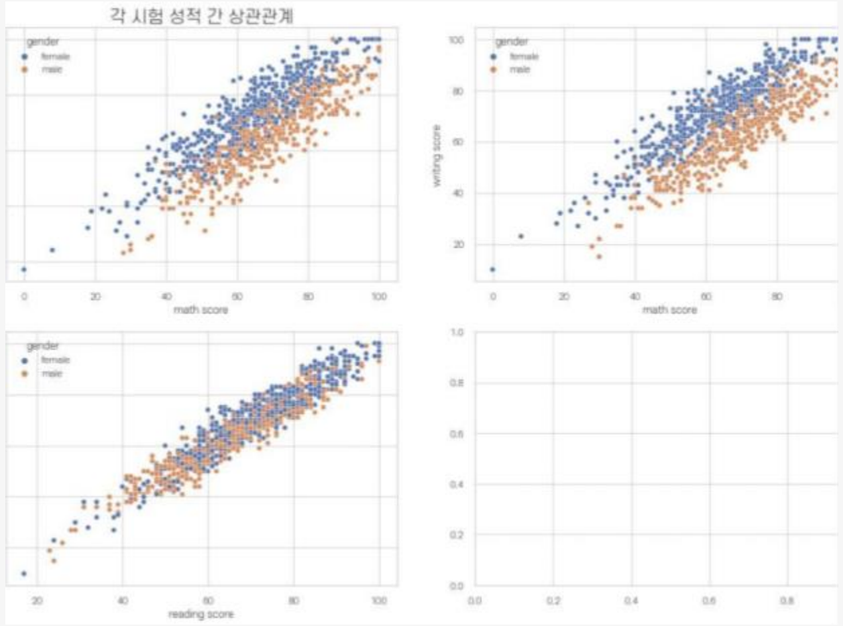
개별 과목

- 수학 : 시험 준비 학습을 하는 경우 > 시험 준비 학습을 하지 않는 경우
- 읽기 : 시험 준비 학습을 하는 경우 > 시험 준비 학습을 하지 않는 경우
- 쓰기 : 시험 준비 학습을 하는 경우 > 시험 준비 학습을 하지 않는 경우

= 전반적으로 시험 준비 학습을 하는 경우 > 시험 준비 학습을 하지 않는 경우

PRESENTATION

REVIEW : 각 시험 성적간의 상관관계는?



성적간의 상관 관계

- 성별에 상관 없이 양의 상관관계를 나타내고 있다

04

데 이 터 셋 분 석

모델링

04 : 모델링

a. 카이제곱 검정

b. LINEAR REGRESSION

c. RIDGE REGRESSION

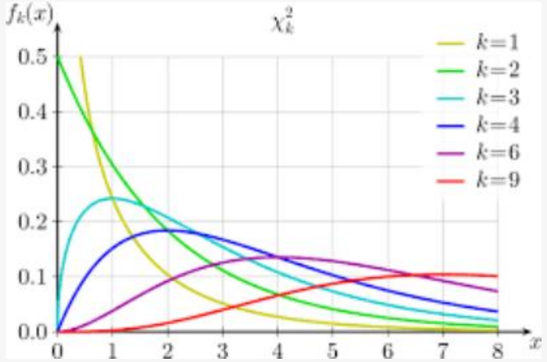
d. LASSO REGRESSION

e. RANDOM FOREST REGRESSOR

f. 부트스트랩(Bootstrap)

g. 결론

카이제곱 검정?



카이제곱 검정의 정의

범주형 자료의 집단간의 동질성 여부를 통계적으로 검증하거나 두 변인간의 상관성을 통계적으로 검증 하고자 할때 사용.

카이제곱 검증을 이용하여 p-value가 0.05 이상일 경우에 통계적으로 두 변수가 독립적임을 알 수 있습니다.

04 : 모델링

카이제곱 검정

```python
#creating a function to execute chisq test for independence
def chisq(col1,col2):
    #create a contingency table
    table=pd.crosstab(new_data[col1],new_data[col2])
    #get chi_Sq statistics,p-value,degrees of freedom and expected frequencies.
    stat, p, dof, expected = chi2_contingency(table)
    #set significance level
    alpha=0.05
    if p<=0.05:
        print('Features are associated')
    else:
        print('Features are not associated')


chisq('gender','lunch_type')
chisq('gender','parent_ed_level')
chisq('gender','race')
chisq('gender','test_prep')
chisq('lunch_type','test_prep')
chisq('lunch_type','parent_ed_level')
chisq('lunch_type','race')
chisq('parent_ed_level','race')
chisq('parent_ed_level','test_prep')
chisq('race','test_prep')
```

```
Features are not associated
Features are not associated
Features are not associated
Features are not associated
Features are not associated
Features are not associated
Features are not associated
Features are not associated
Features are not associated
Features are not associated
```

카이제곱 검정 결론

모두 카이제곱 검정을 통과하였기에 모든 feature을 이용하여 모델을 학습하겠습니다.

04 : 모델링

a. 카이제곱 검정

b. LINEAR REGRESSION

c. RIDGE REGRESSION

d. LASSO REGRESSION

e. RANDOM FOREST REGRESSOR

f. 부트스트랩(Bootstrap)

g. 결론

모델링

모델 학습 시키기전에..

```
def reg_metrics(actual,predicted):
    mae=metrics.mean_absolute_error(actual,predicted)
    mse=metrics.mean_squared_error(actual,predicted)
    rmse=np.sqrt(metrics.mean_squared_error(actual,predicted))
    r2=r2_score(actual,predicted)
    print("MAE:",mae)
    print("MSE:",mse)
    print("RMSE:",rmse)
    print("R2:",r2)
```

어떤 값을 확인할까?

여러가지 모델을 학습시켜 볼것인데요.

저는 reg_metrics 함수로 코드 추상화를 통해 모델을 학습시킬 경우 해당 함수를 이용하여서

MAE와 MSE, RMSE,R2 값을 확인해 보겠습니다..

04 : 모델링

a. 카이제곱 검정

b. LINEAR REGRESSION

c. RIDGE REGRESSION

d. LASSO REGRESSION

e. RANDOM FOREST REGRESSOR

f. 부트스트랩(Bootstrap)

g. 결론

모델링

LINEAR REGRESSION

```
X_train, X_holdout, y_train, y_holdout =
train_test_split(new_data.values,y,test_size=0.3,random_state=17)
reg=LinearRegression(normalize=True)
reg.fit(X_train,y_train)
pred=reg.predict(X_holdout)

reg_metrics(y_holdout,pred)
```

```
MAE: 0.051054370806356546
MSE: 0.004043699097831908
RMSE: 0.0635900864744805
R2: 0.8336765595537288
```

04 : 모델링

a. 카이제곱 검정

b. LINEAR REGRESSION

c. RIDGE REGRESSION

d. LASSO REGRESSION

e. RANDOM FOREST REGRESSOR

f. 부트스트랩(Bootstrap)

g. 결론

모델링

RIDGE REGRESSION

```
ridge=Ridge(alpha=0.04)
ridge.fit(X_train,y_train)
pred1=ridge.predict(X_holdout)
reg_metrics(y_holdout,pred1)
```

```
MAE: 0.05109392129629646
MSE: 0.004049491584354413
RMSE: 0.06363561569085674
R2: 0.8334383058499163
```

04 : 모델링

a. 카이제곱 검정

b. LINEAR REGRESSION

c. RIDGE REGRESSION

d. LASSO REGRESSION

e. RANDOM FOREST REGRESSOR

f. 부트스트랩(Bootstrap)

g. 결론

모델링

LASSO REGRESSION

```
lasso=Lasso(normalize=True,alpha=0)
lasso.fit(X_train,y_train)
pred2=lasso.predict(X_holdout)
reg_metrics(y_holdout,pred2)
```

```
MAE: 0.05105437080635654 6
MSE: 0.004043699097831909
RMSE: 0.06359008647448051
R2: 0.8336765595537288
```

04 : 모델링

a. 카이제곱 검정

b. LINEAR REGRESSION

c. RIDGE REGRESSION

d. LASSO REGRESSION

e. RANDOM FOREST REGRESSOR

f. 부트스트랩(Bootstrap)

g. 결론

모델링

RANDOM FOREST REGRESSOR

```
from sklearn.ensemble import RandomForestRegressor

forest=RandomForestRegressor(n_estimators=100)
forest.fit(X_train,y_train)
pred3=forest.predict(X_holdout)
reg_metrics(y_holdout,pred3)
```

```
MAE: 0.05758223849206347 4
MSE: 0.005187352899855296
RMSE: 0.07202328026308782
R2: 0.7866363544271959
```

성능 향상을 해보자

저는 랜덤포레스트의 모델을 성능을 향상시키기위해 하이퍼 파라미터 튜닝을 통해 최적의 파라미터를 찾아보겠습니다.

제가 사용한 GridSearchCV 메서드는 Brute Force 알고리즘처럼

제가 지정한 파라미터값의 모든 조합을 순회하며 최적의 하이퍼 파라미터 조합을 찾아냅니다.

04 : 모델링

RANDOM FOREST REGRESSOR

```python
param_grid = {
    'bootstrap': [True],
    'max_depth': [3,4,5],
    'max_features': [3,4,5],
    'min_samples_leaf': [3,4,5],
    'min_samples_split': [8,10],
    'n_estimators': [100, 200]
}
forest_cv=RandomForestRegressor(criterion='mae')
# Instantiate the grid search model
grid=GridSearchCV(estimator=forest_cv, param_grid=param_grid, cv=6, n_jobs=-1, verbose=2)

grid.best_params_
```

```
{'bootstrap': True,
 'max_depth': 5,
 'max_features': 4,
 'min_samples_leaf': 3,
 'min_samples_split': 8,
 'n_estimators': 100}
```

04 : 모델링

a. 카이제곱 검정

b. LINEAR REGRESSION

c. RIDGE REGRESSION

d. LASSO REGRESSION

e. RANDOM FOREST REGRESSOR

f. 부트스트랩(Bootstrap)

g. 결론

모델링

RANDOM FOREST REGRESSOR

n_estimators	- 결정트리의 갯수를 지정 - Default = 10 - 무작정 트리 갯수를 늘리면 성능 좋아지는 것 대비 시간이 걸릴 수 있음
min_samples_split	- 노드를 분할하기 위한 최소한의 샘플 데이터수 → 과적합을 제어하는데 사용 - Default = 2 → 작게 설정할 수록 분할 노드가 많아져 과적합 가능성 증가
min_samples_leaf	- 리프노드가 되기 위해 필요한 최소한의 샘플 데이터수 - min_samples_split과 함께 과적합 제어 용도 - 불균형 데이터의 경우 특정 클래스의 데이터가 극도로 작을 수 있으므로 작게 설정 필요
max_features	- 최적의 분할을 위해 고려할 최대 feature 개수 - Default = 'auto' (결정트리에서는 default가 none이었음) - int형으로 지정 →피처 갯수 / float형으로 지정 →비중 - sqrt 또는 auto : 전체 피처 중 √(피처개수) 만큼 선정 - log : 전체 피처 중 log2(전체 피처 개수) 만큼 선정
max_depth	- 트리의 최대 깊이 - default = None → 완벽하게 클래스 값이 결정될 때 까지 분할 또는 데이터 개수가 min_samples_split보다 작아질 때까지 분할 - 깊이가 깊어지면 과적합될 수 있으므로 적절히 제어 필요
max_leaf_nodes	리프노드의 최대 개수

04 : 모델링

a. 카이제곱 검정

b. LINEAR REGRESSION

c. RIDGE REGRESSION

d. LASSO REGRESSION

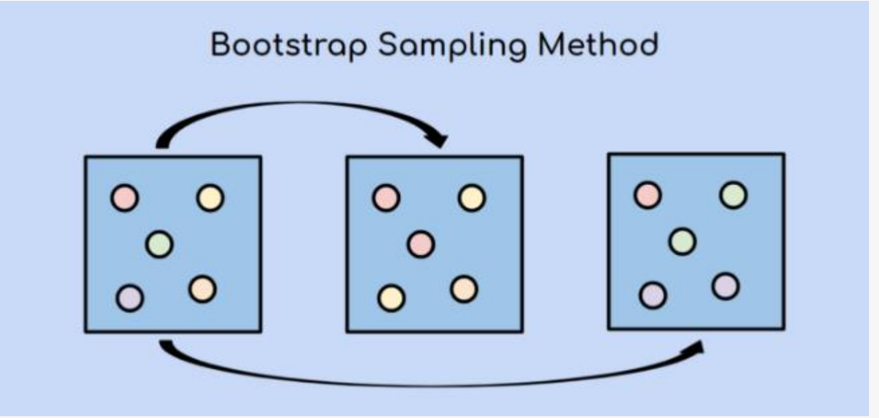
e. RANDOM FOREST REGRESSOR

f. 부트스트랩(Bootstrap)

g. 결론

모델링

부트스트랩(BOOTSTRAP)



부트 스트랩이란?

부트스트랩이란 데이터를 조금은 편향되도록 샘플링하는 기법입니다.

보통 의사결정 트리처럼 과대적합되기 쉬운 모델을 앙상블할 때 많이 사용됩니다.

저는 Decision Tree를 이용한 RandomForest 모델을 만들기 때문에 Bootstrap = 'true'를 지정해주어 튜닝하였습니다.

특징

부트스트랩은 주어진 자료에서 단순랜덤 복원추출 방법을 활용하여 동일한 크기의 표본을 여러개 생성하는 샘플링 방법이다.

부트스트랩을 통해 100개의 샘플을 추출하더라도 샘플에 한번도 선택되지 않는 원 데이터가 발생할 수 있는데

전체 샘플의 약 36.8%가 이에 해당한다.

04 : 모델링

a. 카이제곱 검정

b. LINEAR REGRESSION

c. RIDGE REGRESSION

d. LASSO REGRESSION

e. RANDOM FOREST REGRESSOR

f. 부트스트랩(Bootstrap)

g. 결론

모델링

결론

```
forest_cv=RandomForestRegressor(criterion='mae',bootstrap=True,max_depth=5,max_features=4,min_samples_l
eaf=3,min_samples_split=10,n_estimators=100)
forest_cv.fit(X_train,y_train)
pred4=forest_cv.predict(X_holdout)

reg_metrics(y_holdout,pred4)
```

```
MAE: 0.05372166666666675
MSE: 0.004480081933333354
RMSE: 0.06693414176578O3
R2: 0.8157274756094729
```

결론?

r2 접구사 78에서 81으로 더 높은 모델로 튜닝이 가능하였습니다. 우리는 위의 모델을 이용하여서 학생들의 여러 특성들을 가지고 그 학생의 평균 절대 오차(MAE) 5%의 시험 점수를 회귀모델로 계산할 수 있습니다

05

데 이 터 셋 분 석

정리

데이터 확인

Students Performance in Exam 에서
받은 데이터를 확인했습니다
1. 데이터의 컬럼
2. 전체 데이터 정리
3. 데이터 전처리

데이터 분석

질문에 따른 데이터 분석을 시행했습니다
1. 성별과 성적의 관계는?
2. 인종/민족과 성적의 관계는?
3. 부모 교육 수준과 성적의 관계는?
4. 점심 식사 여부와 성적의 관계는?
5. 시험 준비 학습 여부와 성적의 관계는?
6. 각 시험 성적의 상관 관계는?

모델링

데이터를 분류 하기 위해
모델링을 진행하였습니다
1. Lunear Regression
2. RIDGE REGRESSION
3. LASSO REGRESSION
4. RANDOM FOREST REGRESSOR
모델로 데이터를 학습 시켰습니다.

모델 평가 & 향상

모델을 평가 했습니다
평가 기준은
1. MSE
2. RMSE
3. MAE
4. R2
입니다.
랜덤포레스트의
모델을 성능을 향상시키기위해
하이퍼 파라미터 튜닝을 통해
최적의 파라미터를 찾았다

STUDENTS PERFORMANCE IN EXAM

데이터셋 분석

2016108271

인공지능

다시 한번..

01 **부모 학벌이 자녀에 학생 성적에 가장 큰 영향을 끼치지 않을까?**

아무래도 "스카이캐슬"에서 보았듯이 부모들이 자기 자식 또한 더 좋은 학벌을 만들기 위해 노력을하고, 더 많은 사교육 하려고 하지 않을까?

02 **인종/민족과 성적에서 성적의 차이가 없지 않을까?**

같은 교육을 받았다면, 인종/민족간의 차이가 없을것 같다. 같은 '인간'이라는 특성상 큰 차이가 없을것 같다

03 **성적에 가장 중요한 특성은 무엇일까?**

항상 궁금해 왔다, 여러 특성중에 어떤게 가장 중요한 특성일까? 이를 통해 내 기말도 잘 보았으면 좋겠다...

04 **아무래도 여자가 좀 더 성적이 좋지 않을까?**

살아오면서 보았던 사람들중 여성분들이 좀 더 학업을 열심히 한 경우가 많았던것 같다(경험) 이 데이터 역시 여성분들이 좀 더 좋은 결과가 나올 것 같다

STUDENTS PERFORMANCE IN EXAM

데이터셋 분석

2016108271

인공지능

가설에 대한 결론

01 **생각보다 부모 학벌이 자녀에 학생 성적에 영향이 없었다.**

부모의 교육 수준이 master's degree 인 경우가 성적이 가장 높고 high school 인 경우가 성적이 가장 낮게 나타났지만

큰 차이점이 없었다.

02 **인종/민족과 성적에서 성적의 차이가 없었다.**

E그룹이 가장 높아고 A가 가장 낮았지만

큰 차이점이 없었다

03 **성적에 가장 중요한 특성은 점심식사 여부 였다**

점심식사와 수학성적이 상관관계가 0.38로 가장 높았으며, 이는 양의상관계를 이루는 다른 성적에도 영향이 있다

04 **평균적으로 여자가 성적이 좋았다.**

전체 성적이나, 개별 성적이나 모두 여자가 높았다

가설에 대한 결론

05

이러한 예측, 분석을 위해서 부트스트랩을 활용한 랜덤 포레스트 모델을 사용하는게 가장 적합하는것을 알수 있었습니다

r2 접구사 78에서 81으로 더 높은 모델로 튜닝이 가능하였습니다. 우리는 위의 모델을 이용하여서 학생들의 여러 특성들을 가지고 그 학생의 평균 절대 오차(MAE) 5%의 시험 점수를 회귀모델로 계산할 수 있습니다

```
{'bootstrap': True,
 'max_depth': 5,
 'max_features': 4,
 'min_samples_leaf': 3,
 'min_samples_split': 8,
 'n_estimators': 100}
```

06

데 이 터 셋 분 석

느낀점

느낀점

데이터 분석은 시나리오를 잘 짜야한다.

데이터는 누구나 같은 데이터를 사용한다. 하지만 kkagle을 보면 모두 다른 관점으로 접근하는것을 알 수 있다

나또한 그렇다 해당 데이터를 가지고 어떻게 데이터를 분석하고, 어떤 모델을 사용할것인지 고민하고 찾아서 좋은 시나리오로 데이터 분석을 한다

모델링도 추상화!

결국 데이터 분석도 코드를 이용한다. 중복되는 코드를 최대한 정리하고,

추상화시켜서 언제든지 코드를 재사용 할 수 있게 해야한다.(모델 평가할때 유용하게 사용했다)

그리고 전역에서 잘못사용되는 일 없게, 잘 포장해야한다.

내가 생각한건 정답이 아니다!!

사람이 그저 직관적으로 생각할때와 실제 데이터 분석을 할때 차이점이 많이 난다는것을 알수있었다.

조심해야하는 변인도 찾아봤었고(카이제곱) 어떤 변인이 상관 관계가 높았는지도 데이터 분석을 통해서 알 수 있었다.


Q & A
STUDENTS PERFORMANCE IN EXAM
데이터셋 분석