인공지능

환경과 유전에 따른 키 예측

전공 : 컴퓨터공학과

학번 : 2017108271

이름 : 좌상헌

우리는 살아가면서 어릴 때부터 '키'라는 것에 많은 관심을 가진다.
초등학생때는 키 순서대로 자리를 배치하기도 하고 줄을 설때도 키를 이용해서 나누게 된다.
또 성인이 되서도 '키'는 중요하게 작용하는 것 같다. 이 때문에 키 크는 방법을 찾기도하고, 키 커지는 음식을 먹기도 한다.
대부분의 사람들이 '키'가 유전적 영향과 환경적 영향을 많이 받는다고 생각하는데, 과연 얼마나 그 영향이 클지 궁금해서 코드를 찾아보게 되었고, 주제로 선정하게 되었다.

데이터 수집

총 359 명에게 8가지 항목에 대해 설문조사를 하여 데이터를 수집하였다. 유전과 관련하여 성별과 부모님의 신장에 대해 조사하였다. 환경과 관련한 항목으로는 학창 시절 취침시간, 운동시간, 섭취 음식, 키 성장과 관련한 약 복용 혹은 주사 여부를 조사하였다.

데이터 항목

성별

남성을 0 여성을 1로 설정하였다.

신장

유전적인 요소로 본인과 부모님의 키를 조사하였다.

학창시절 운동시간(exercise)

거의 하지 않았다, 30분 이내, 1시간 이내, 1시간 30분 이내, 1시간 30분 이상으로 구분하여 5가지의 선택지를 제공하였다. 각 선택지는 다시 0, 30, 60, 90, 120 (분)으로 변환하여 사용하였다.

학창시절 취침시간(sleep_time)

오후 11시 이전, 오전 12시 이전, 오전 1시 이전, 오후 2시 이전, 오후 2시 이후로 구분하여 5가지의 선택지를 제공하였다. 중/고등학생의 일반적인 기상시간인 오전 7시를 기준으로 하여 각 선택지를 다시 8, 7, 6, 5, 4 (시간)으로 변환하였다.

학창시절 키 성장에 도움이 되는 음식 섭취 정도(good_food)

우유,콩,고등어 등 일반적으로 키 성장에 도움이 된다고 알려진 음식을 즐겨 먹었는가에 대해 질문하는 항목으로 전혀 그렇지 않다를 1, 매우 그렇다를 5로 설정하여 1 ~ 5 의 선택지를 제공하였다.

학창시절 키 성장에 방해가 되는 음식 섭취 정도(bad_food)

인스턴스, 탄산음료 등 일반적으로 키 성장에 방해가 된다고 알려진 음식을 즐겨먹었는가에 대해 질문하는 항목이다. 마찬가지로 전혀 그렇지 않다를 1, 매우 그렇다를 5로 설정하여 1 ~ 5 의 선택지를 제공하였다.

학창시절 키 성장에 도움이 되는 약/주사 여부(drug)

학창시절에 키 성장을 위해 약을 복용하였거나, 주사를 맞아본 경험이 있는지, 있다면 어느 정도 였는지에 대해 질문하는 항목이다. 위와 같이 전혀 그렇지 않다를 1, 매우 그렇다를 5로 설정하여 1 ~ 5 의 선택지를 제공하였다.

```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression, Ridge
from sklearn.metrics import r2_score

import tensorflow as tf
```

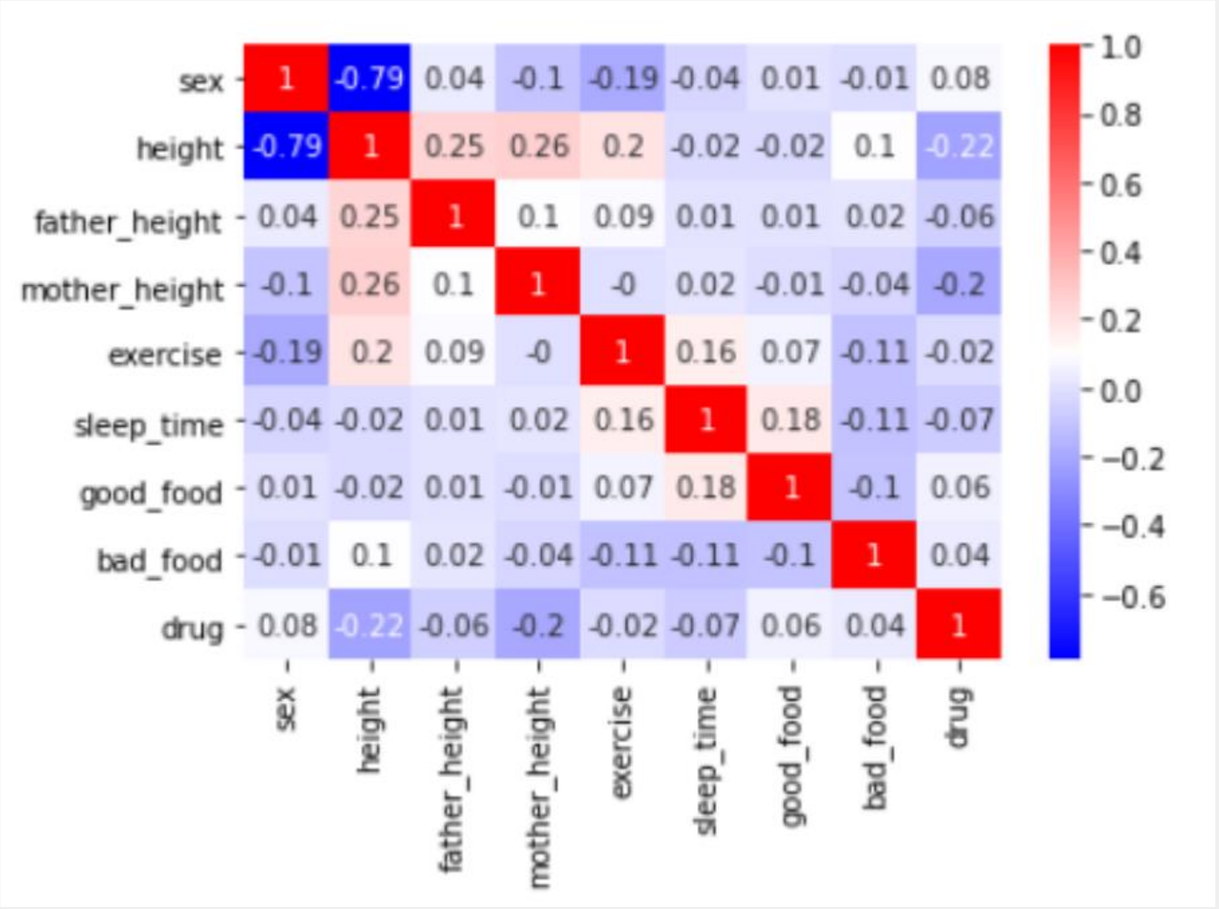
```
file_addr='/kaggle/input/height/height_purged.xlsx'
height=pd.read_excel(file_addr)
height
```

**필요한 모듈들을 import한 후,
데이터를 불러온다.**

데이터의 모습

	timestamp	sex	height	father_height	mother_height	exercise	sleep_time	good_food	bad_food	drug
0	2020-08-12 00:28:04.794	0	179.0	168	162	120	6	4	2	5
1	2020-08-12 00:36:50.504	1	166.0	170	162	60	6	2	5	1
2	2020-08-12 00:39:37.413	0	175.0	174	157	90	7	3	4	1
3	2020-08-12 00:40:47.190	1	168.0	172	163	120	7	5	2	1
4	2020-08-12 00:42:43.345	1	155.0	173	160	30	6	4	5	5
...	...	...	...	...	...	...	...	...	...	...
354	2020-11-15 12:36:53.228	0	177.0	174	167	30	6	3	5	1
355	2020-11-15 12:44:59.310	0	183.0	173	158	90	4	4	3	1
356	2020-11-15 12:48:31.110	0	181.0	176	165	30	6	2	5	1
357	2020-11-15 13:00:25.781	0	173.0	167	159	30	6	3	3	1
358	2020-11-15 13:06:26.016	0	180.0	175	160	60	5	3	4	1

```
correlation_matrix=height.corr().round(2)
sns.heatmap(data=correlation_matrix,annot=True,cmap='bwr')
```



Heatmap을 통해 상관 관계를 나타낸다.

그 결과 부모의 키(유전)이 가장 높은 연관성을 보여준다.

```
height = pd.get_dummies(height, columns=['sex'])
```

```
x=height[['sex_0', 'sex_1', 'father_height','mother_height']]
y=height['height']

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=1)
```

```
lin_reg=LinearRegression()
lin_reg.fit(x_train,y_train)

pred=lin_reg.predict(x_train)
print('train 정확도',r2_score(y_train,pred))
pred=lin_reg.predict(x_test)
print('test 정확도',r2_score(y_test,pred))
```

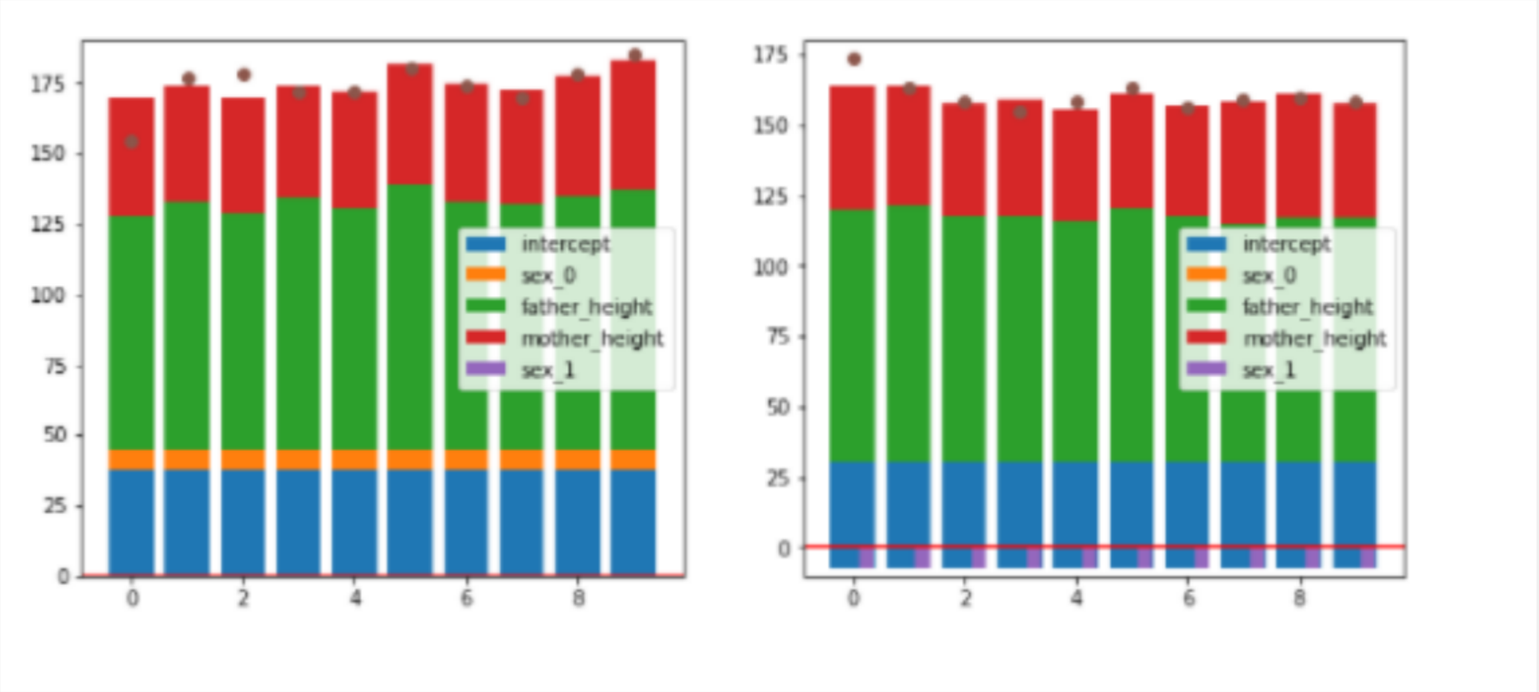
```
train 정확도 0.7322865230192885
test 정확도 0.6864529556028103
```

가장 높은 연관성을 가졌던 유전적 요인을 이용한 키 예측
Linear Regression을 이용
x 값은 성별과 부모 키
y 값은 본인 키
학습시킨 후, 정확도 비교

```python
def axe_make(axe, data, columns, regressor):
    clm = columns

    stack_list = []
    minus_list = []
    label_list = []
    minus_label = []
    bottom = [0 for _ in range(len(data))]

    if lin_reg.intercept_ < 0:
        minus_list.append([regressor.intercept_ for _ in range(len(data))])
        minus_label.append("intercept")
    else:
        stack_list.append([regressor.intercept_ for _ in range(len(data))])
        label_list.append("intercept")

    for i, next_clm in enumerate(clm):
        if regressor.coef_[i] > 0:
            stack_list.append(data[next_clm] * regressor.coef_[i])
            label_list.append(next_clm)
        else:
            minus_list.append(data[next_clm] * regressor.coef_[i])
            minus_label.append(next_clm)
            bottom = [bottom[j] + x for j, x in enumerate(data[next_clm] * regressor.coef_[i])]

    for i, next_stack in enumerate(stack_list):
        axe.bar(range(len(data)), next_stack, bottom=bottom, label=label_list[i])
        for i, val in enumerate(next_stack):
            bottom[i] += val

    bottom = [0 for _ in range(len(data))]

    for i, next_stack in enumerate(minus_list):
        axe.bar([x + 0.25 for x in range(len(data))], next_stack, bottom=bottom, label=minus_la
bel[i], width=0.25)
        for i, val in enumerate(next_stack):
            bottom[i] += val

    axe.scatter(range(len(data)), data['height'], zorder=3)
    axe.axhline(0, color='red')
    axe.legend()

    return axe
```

```python
fig, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 5))
ax1 = ax[0]
ax2 = ax[1]

plot_men = height.loc[height["sex_0"] == 1].sample(10)
plot_women = height.loc[height["sex_1"] == 1].sample(10)


ax1 = axe_make(ax1, plot_men, ['sex_0', 'sex_1', 'father_height','mother_height'], lin_reg)
ax2 = axe_make(ax2, plot_women, ['sex_0', 'sex_1', 'father_height','mother_height'], lin_reg)


ax1.set_ylim(0, 190)
ax2.set_ylim(-10, 180)

plt.show()
```

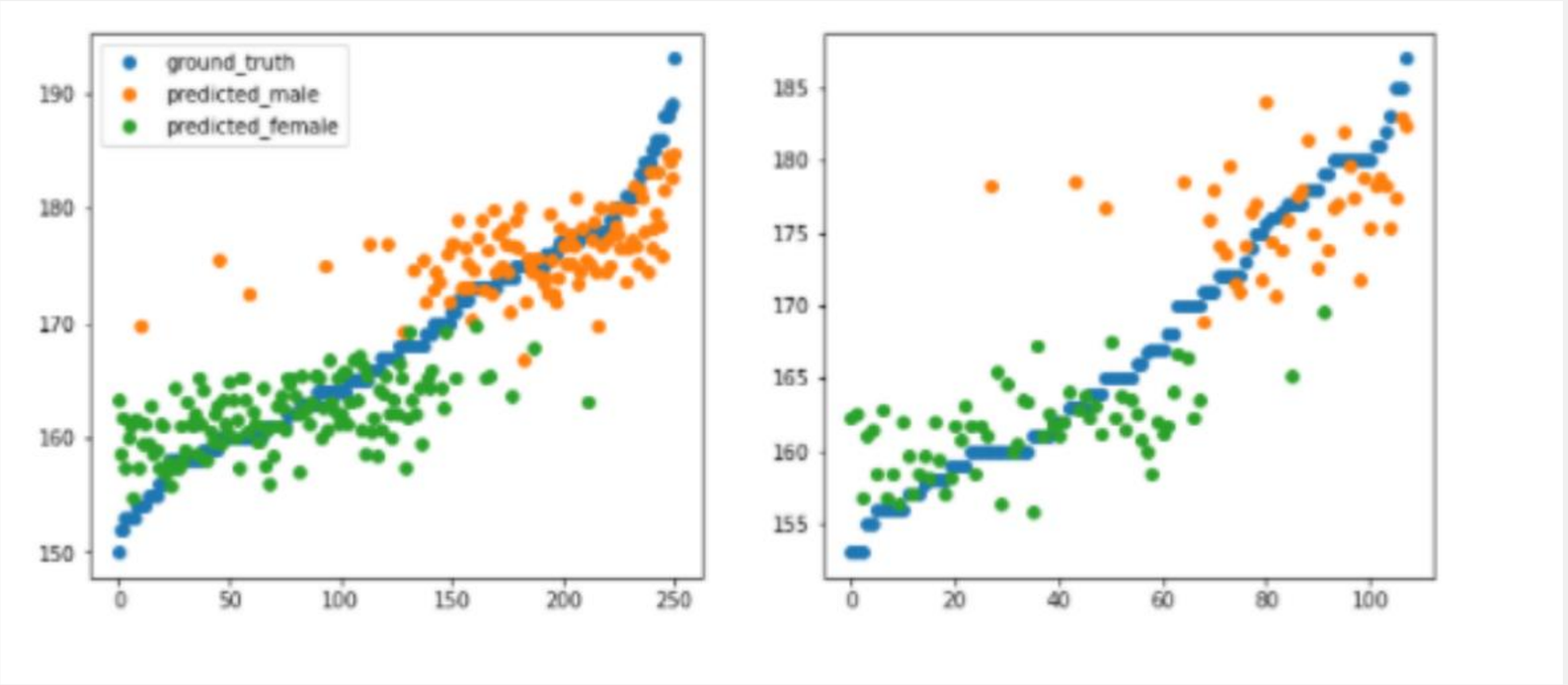
어떤 column(요소)가 영향을 많이 미치는지 확인하기 위해 함수를 만들고 그래프로 나타냈다.



결과 값으로 나온 그래프이다. intercept값이 작을수록 column들의 영향력이 큰것을 의미한다.
보시다시피 부모의 키가 높은 연관성을 보인다.

```
y_train_sorted = y_train.sort_values()
x_train_sorted = x_train.loc[y_train_sorted.index]
x_train_sorted.reset_index(drop=True, inplace=True)
x_train_male = x_train_sorted.loc[x_train_sorted["sex_0"] == 1]
x_train_female = x_train_sorted.loc[x_train_sorted["sex_1"] == 1]
y_train_male_predicted = lin_reg.predict(x_train_male)
y_train_female_predicted = lin_reg.predict(x_train_female)

y_test_sorted = y_test.sort_values()
x_test_sorted = x_test.loc[y_test_sorted.index]
x_test_sorted.reset_index(drop=True, inplace=True)
x_test_male = x_test_sorted.loc[x_test_sorted["sex_0"] == 1]
x_test_female = x_test_sorted.loc[x_test_sorted["sex_1"] == 1]
y_test_male_predicted = lin_reg.predict(x_test_male)
y_test_female_predicted = lin_reg.predict(x_test_female)

fig, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 5))
ax[0].scatter(x_train_sorted.index, y_train_sorted, label="ground_truth")
ax[0].scatter(x_train_male.index, y_train_male_predicted, label="predicted_male")
ax[0].scatter(x_train_female.index, y_train_female_predicted, label="predicted_female")

ax[1].scatter(x_test_sorted.index, y_test_sorted, label="ground_truth")
ax[1].scatter(x_test_male.index, y_test_male_predicted, label="predicted_male")
ax[1].scatter(x_test_female.index, y_test_female_predicted, label="predicted_female")

ax[0].legend()
plt.show()
```

Train 데이터와 Test 데이터를 각각 나누고 정렬하여 예측값과 실제값(Ground Truth)를 비교하는 그래프를 보여주기 위한 코드이다.



결과 값으로 나온 그래프이다.
파란색인 실제 값과 유사한 분포를 보여준다.

```
x=height[['sex_0', 'sex_1', 'exercise', 'sleep_time', 'good_food', 'bad_food', 'drug']]
y=height['height']

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=1)
```

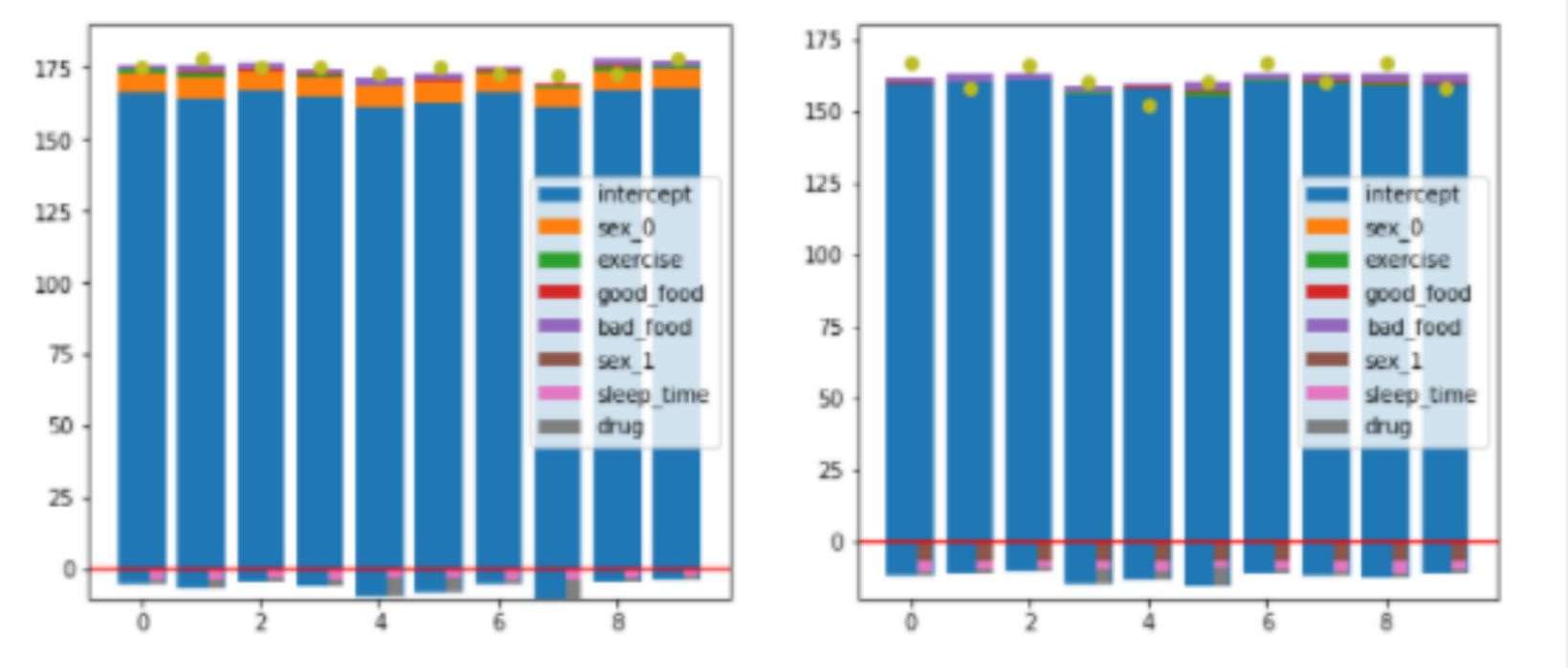
```
lin_reg=LinearRegression()
lin_reg.fit(x_train,y_train)

pred=lin_reg.predict(x_train)
print('train 정확도',r2_score(y_train,pred))
pred=lin_reg.predict(x_test)
print('test 정확도',r2_score(y_test,pred))
```

```
train 정확도 0.6494767605043O5
test 정확도 0.6744506799695542
```

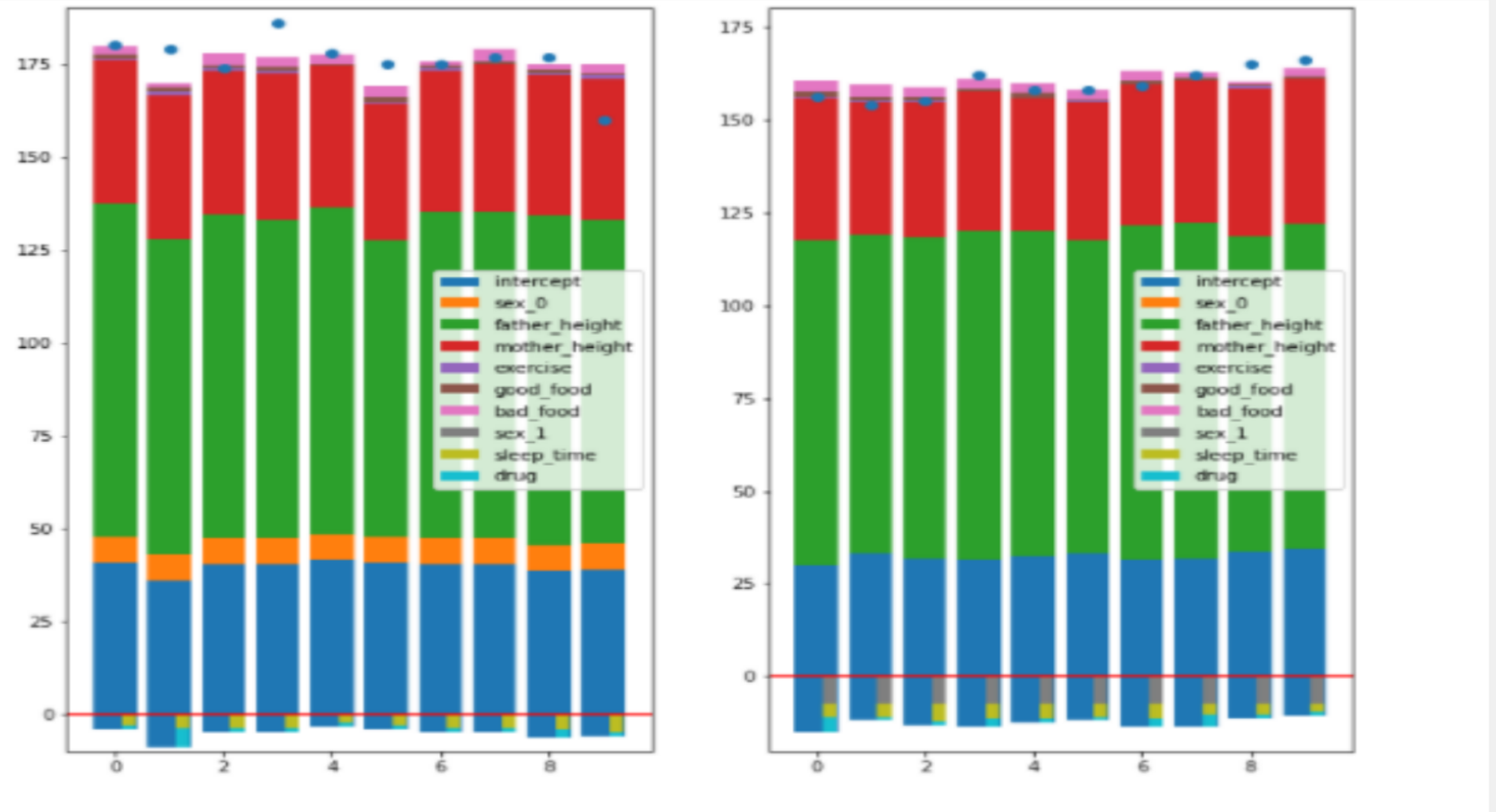
다음으로 환경적요인을 이용한 키 예측
Linear Regression을 이용
x 값은 성별과 운동, 취침시간, 음식, 약
y 값은 본인 키
학습시킨 후, 정확도 비교

```
fig, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 5))
ax1 = ax[0]
ax2 = ax[1]

plot_men = height.loc[height["sex_0"] == 1].sample(10)
plot_women = height.loc[height["sex_1"] == 1].sample(10)


ax1 = axe_make(ax1, plot_men, ['sex_0', 'sex_1', 'exercise', 'sleep_time', 'good_food', 'bad_fo
od', 'drug'], lin_reg)
ax2 = axe_make(ax2, plot_women, ['sex_0', 'sex_1', 'exercise', 'sleep_time', 'good_food', 'bad_
food', 'drug'], lin_reg)


ax1.set_ylim(-10, 190)
ax2.set_ylim(-20, 180)

plt.show()
```

위에서 한 동일한 방법으로 함수를 이용하여 그래프를 나타냈다.



결과 그래프를 보면 환경적 요인보다는 유전적 요인의 영향이 압도적으로 많다는것을 알 수 있다.



위와 같은 방법으로 전체를 넣고 그래프를 만들어보면 유전적 요인이 가장 영향력이 높다는 것을 알 수 있다.

- 키는 유전과 가장 큰 영향을 받는다는 것을 알 수 있었다.

- 환경적인 요인에서는 운동에 영향을 많이 받는다는 것을 알 수 있었다.

- 약 복용량과 키는 음의 상관관계를 보이는데 아마도 스트레스의 영향이 있거나, 키가 잘 크지 않는 아이일수록 더 약을 많이 먹기 때문이 아닐까 생각해볼 수 있었다.

- 수면 시간이나 우유 등 키크는 음식은 키의 별로 영향을 미치지 않는다는 것도 알 수 있었다.