개인의 건강보험요금 예측

2021120120 이윤경

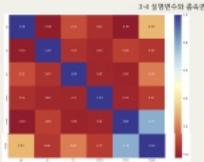
문제정의(problem definition)

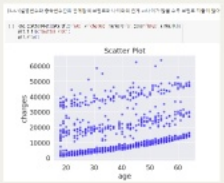
개인의 건강보험요금 예측

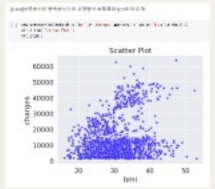
결론 선정 회귀 모델을 사용하여 개인의 건강보험 요금을 예측할 때, 모델은 주어진 독립 변수의 값에 기반하여 건강보험 요금을 계산

종속변수:건강보험료금(charges)
독립변수:예측에 사용되는 변수로, 나이(age), 성별(sex), BMI(bmi), 자녀(children), 흡연자(smoker)

문제 -> 분석->해결







개인의 건강보험요금 예측

2021120120 이윤경

문제 -> 분석->해결

당 14~27g

탕후루

나트륨
2000~3000mg

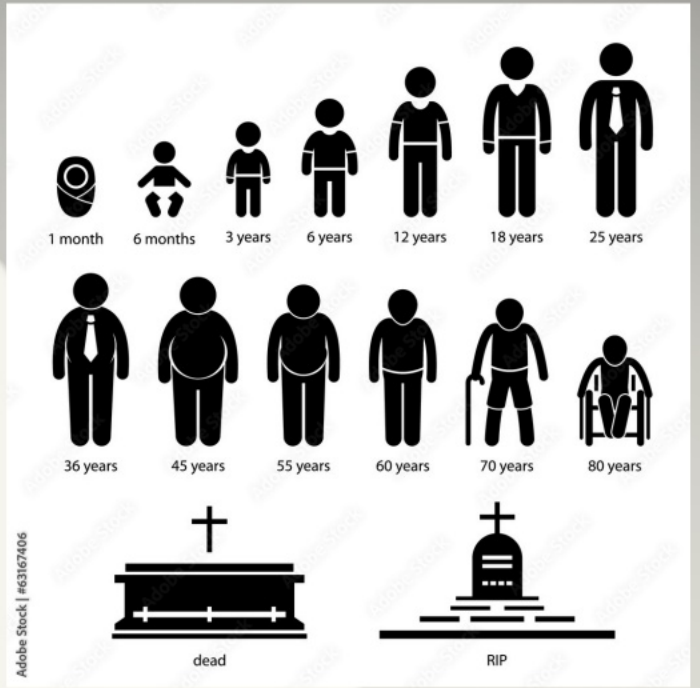
마라탕

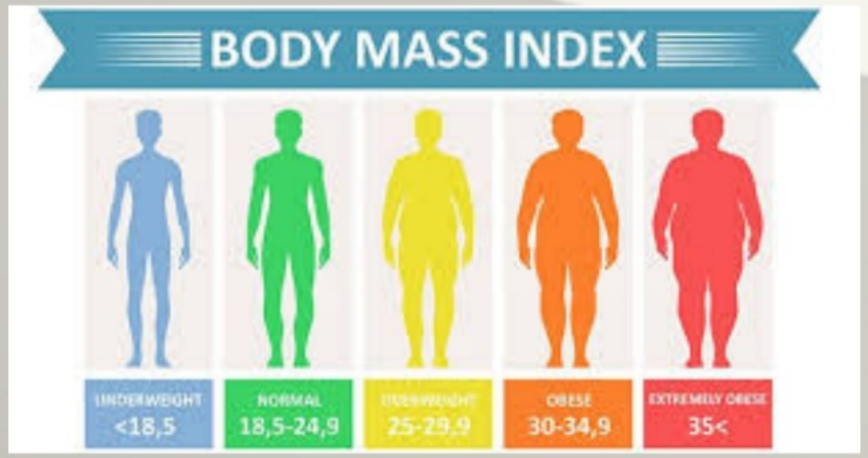
당 65g

스무디·에이드류

domain knowledge

건강보험 요금 계산 방법





문제정의(problem definition)

개인의 건강보험요금 예측

결론:선형 회귀 모델을 사용하여 개인의 건강보험 요금을 예측할 때, 모델은 주어진 독립 변수의 값에 기반하여 건강보험 요금을 계산

종속변수:건강보험요금(charges)
독립변수:예측에 사용되는 변수로, 나이(age), 성별(sex), BMI(bmi), 자녀(children), 흡연자(smoker)

2.데이터 수집및 전처리/
3-1.데이터셋기본정보파악

	age	sex	bmi	children	smoker	charges
0	19	female	27.90	0	yes	16884.9240
1	18	male	33.77	1	no	1725.5523
2	28	male	33.00	3	no	4449.4620

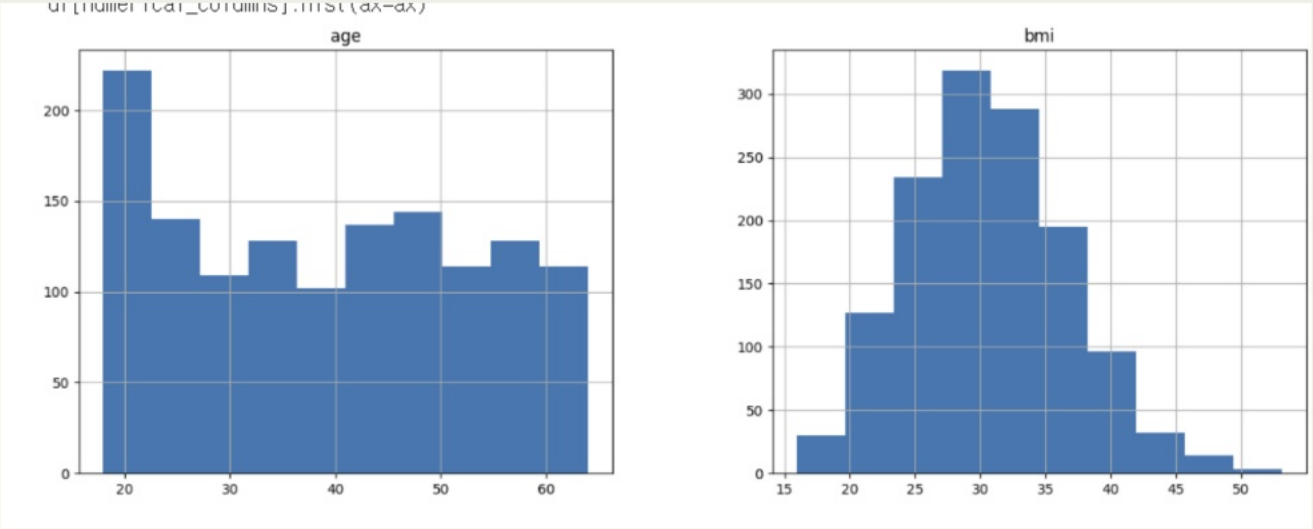
```python
import pandas as pd


# sex와 smoker 열을 수치형으로 변환
df['sex'] = df['sex'].astype('category').cat.codes
df['smoker'] = df['smoker'].astype('category').cat.codes

#수치형 열 리스트에 sex와 smoker 추가
numerical_columns = ['나이', '수입', 'sex', 'smoker']
```

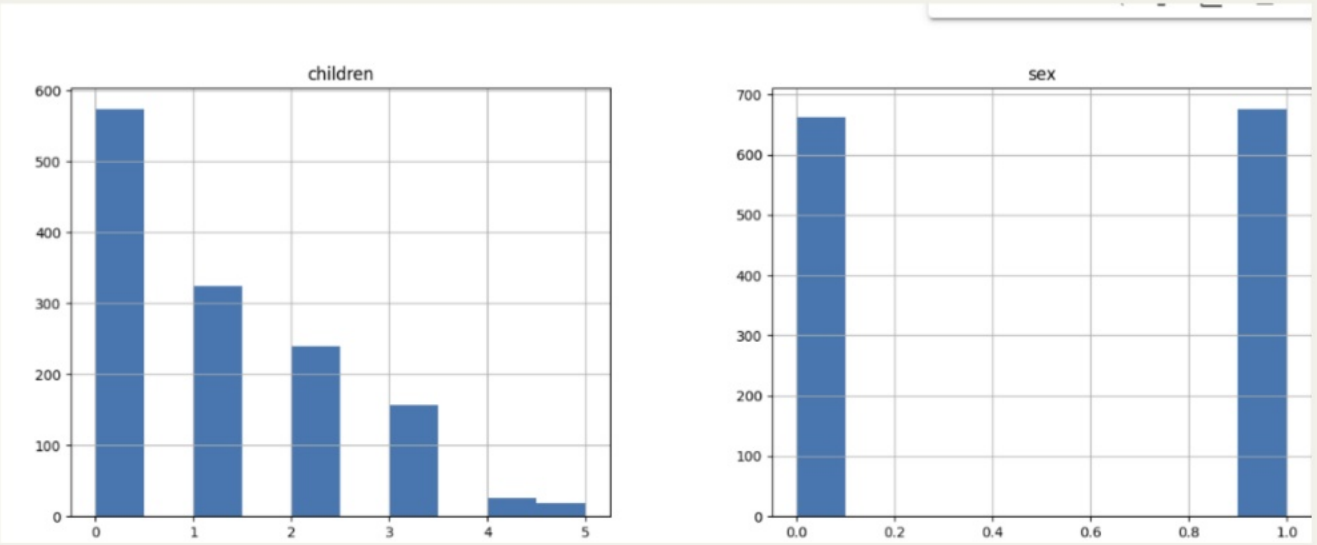
age	sex	bmi	children	smoker	charges
19	female	27.9	0	yes	16884.92
18	male	33.77	1	no	1725.552
28	male	33	3	no	4449.462
33	male	22.705	0	no	21984.47
32	male	28.88	0	no	3866.855
31	female	25.74	0	no	3756.622
46	female	33.44	1	no	8240.59
37	female	27.74	3	no	7281.506
37	male	29.83	2	no	6406.411
60	female	25.84	0	no	28923.14
25	male	26.22	0	no	2721.321
62	female	26.29	0	yes	27808.73
23	male	34.4	0	no	1826.843
56	female	39.82	0	no	11090.72
27	male	42.13	0	yes	39611.76
19	male	24.6	1	no	1837.237
52	female	30.78	1	no	10797.34
23	male	23.845	0	no	2395.172
56	male	40.3	0	no	10602.39
30	male	35.3	0	yes	36837.47

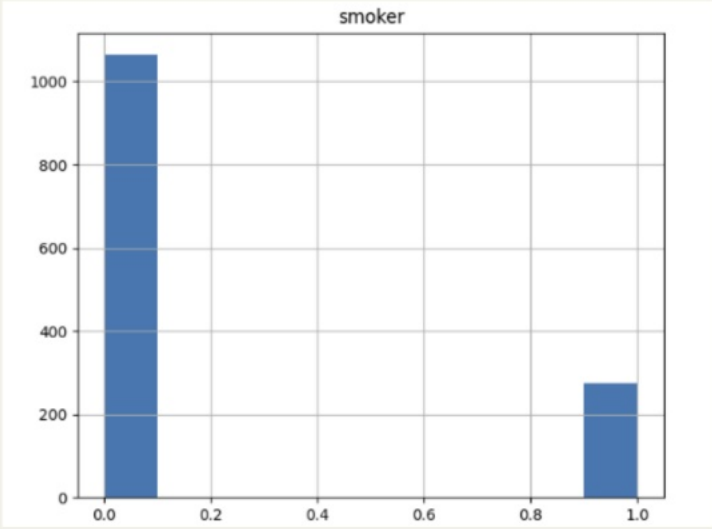
```
] df.head()
```

	age	sex	bmi	children	smoker	charges
0	19	0	27.900	0	1	16884.92400
1	18	1	33.770	1	0	1725.55230
2	28	1	33.000	3	0	4449.46200
3	33	1	22.705	0	0	21984.47061
4	32	1	28.880	0	0	3866.85520

3-2.데이터 탐색하기(독립변수 탐색)





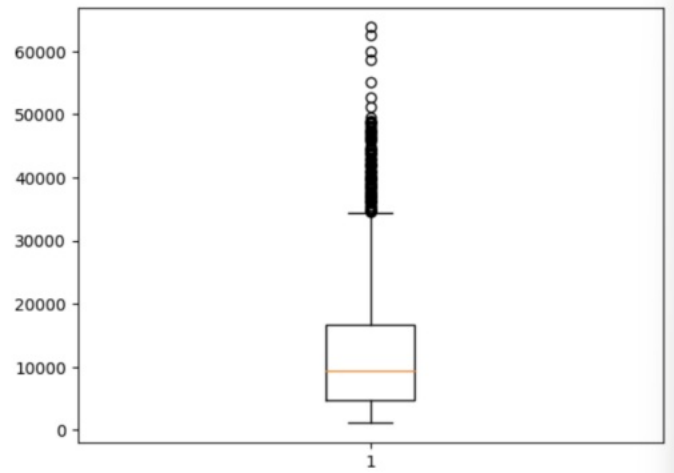


3-3종속변수탐색

```
#[기초통계량]:종속변수의 기초통계량을 살펴본다.
df['charges'].describe()
```

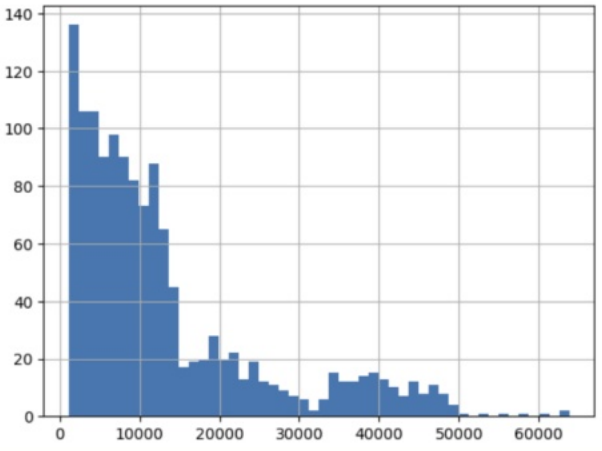
```
count     1338.000000
mean     13270.422265
std      12110.011237
min       1121.873900
25%       4740.287150
50%       9382.033000
75%      16639.912515
max      63770.428010
Name: charges, dtype: float64
```

```
#boxplot을 이용해 또 다른 시각화의 방법으로 분포를 본다.

plt.boxplot(df['charges'])

plt.show()
```

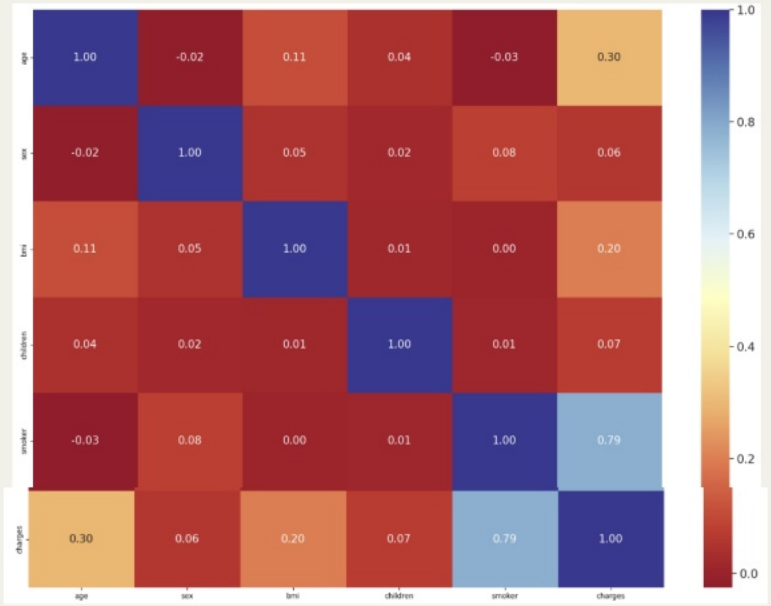



```
#시각화를 해서 살펴본다.

df['charges'].hist(bins=50)
```

```
<Axes: >
```

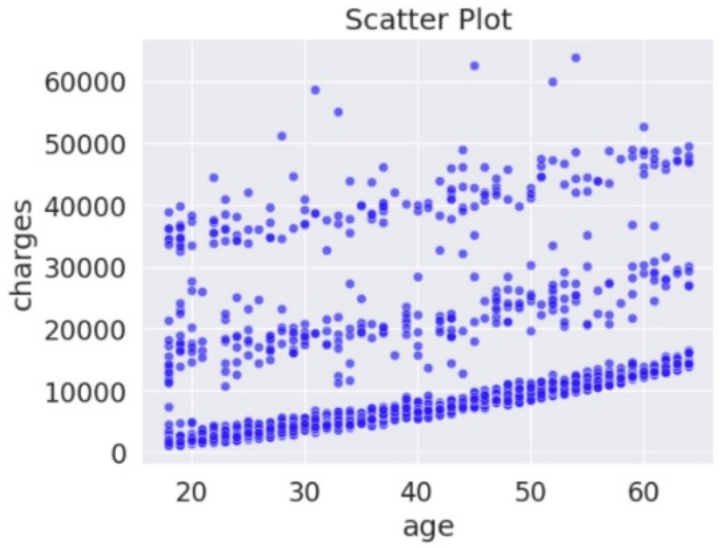


3-4 설명변수와 종속변수간의 탐색



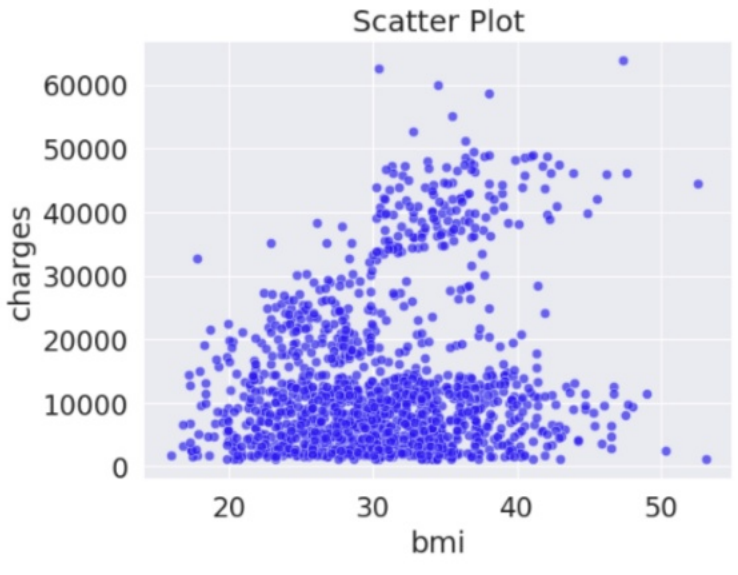
[3-4-1]설명변수와 종속변수간의 관계탐색 보험료와 나이와의 관계 =>나이가 많을 수록 보험료 지출이 많아

```
[ ]  sns.scatterplot(data=df,x='age',y='charges',markers='o',color='blue',alpha=0.6)
     plt.title('Scatter Plot')
     plt.show()
```



[3-4-3]설명변수와 종속변수간의 관계탐색 보험료와 bmi와의 관계

```
[ ]  sns.scatterplot(data=df,x='bmi',y='charges',markers='o',color='blue',alpha=0.6)
     plt.title('Scatter Plot')
     plt.show()
```



Kurtosis:	5.563	Cond. No.	1.14	

4-1.모델링하기/xtrain,ytrain나누기

```
df.head()   #Scaler 적용전, Table
```

	age	sex	bmi	children	smoker	charges
0	19	female	27.900	0	yes	16884.92400
1	18	male	33.770	1	no	1725.55230
2	28	male	33.000	3	no	4449.46200
3	33	male	22.705	0	no	21984.47061
4	32	male	28.880	0	no	3866.85520

Next steps: Generate code with df View recommended plots

[Scaler]

```
[ ]  ### 사이킷런은 파이썬에서 머신러닝 분석을 할 때 유용하게 사용할 수 있는 라이브러
     ### 여러가지 머신러닝 모듈로 구성되어있습니다.

     from sklearn.preprocessing import StandardScaler



     scaler = StandardScaler()  # 평균 0, 분산 1
     scale_columns = ['age', 'bmi', 'children', 'sex', 'smoker','charges']
     df[scale_columns] = scaler.fit_transform(df[scale_columns])
```

```
[ ]  df[scale_columns].head() # Scaler 적용 후, Table
```

	age	bmi	children	sex	smoker	charges
0	-1.438764	-0.453320	-0.908614	-1.010519	1.970587	0.298584
1	-1.509965	0.509621	-0.078767	0.989591	-0.507463	-0.953689
2	-0.797954	0.383307	1.580926	0.989591	-0.507463	-0.728675
3	-0.441948	-1.305531	-0.908614	0.989591	-0.507463	0.719843
4	-0.513149	-0.292556	-0.908614	0.989591	-0.507463	-0.776802

```
[ ]  y_train
```

```
216    -0.240782
731    -0.264757
866    -1.001941
202    -0.021330
820    -0.481146
          ...
715    -0.092805
905    -0.719197
1096    2.591451
235     0.510004
1061   -0.141770
Name: charges, Length: 1070, dtype: float64
```

```
[ ]  X_train
```

	age	bmi	children	sex	smoker
216	0.982076	-0.666578	-0.908614	-1.010519	-0.507463
731	0.982076	-1.519609	-0.078767	0.989591	-0.507463
866	-1.509965	1.087058	-0.908614	0.989591	-0.507463
202	1.480485	-1.087352	-0.908614	-1.010519	-0.507463
820	0.412467	0.498138	-0.078767	0.989591	-0.507463
...	...	...	...	...	...
715	1.480485	-0.289276	-0.908614	0.989591	-0.507463
905	-0.940356	-0.214635	0.751079	-1.010519	-0.507463
1096	0.839674	0.704834	0.751079	-1.010519	1.970587
235	0.056461	-1.385093	0.751079	-1.010519	1.970587
1061	1.266881	-0.446758	-0.078767	0.989591	-0.507463

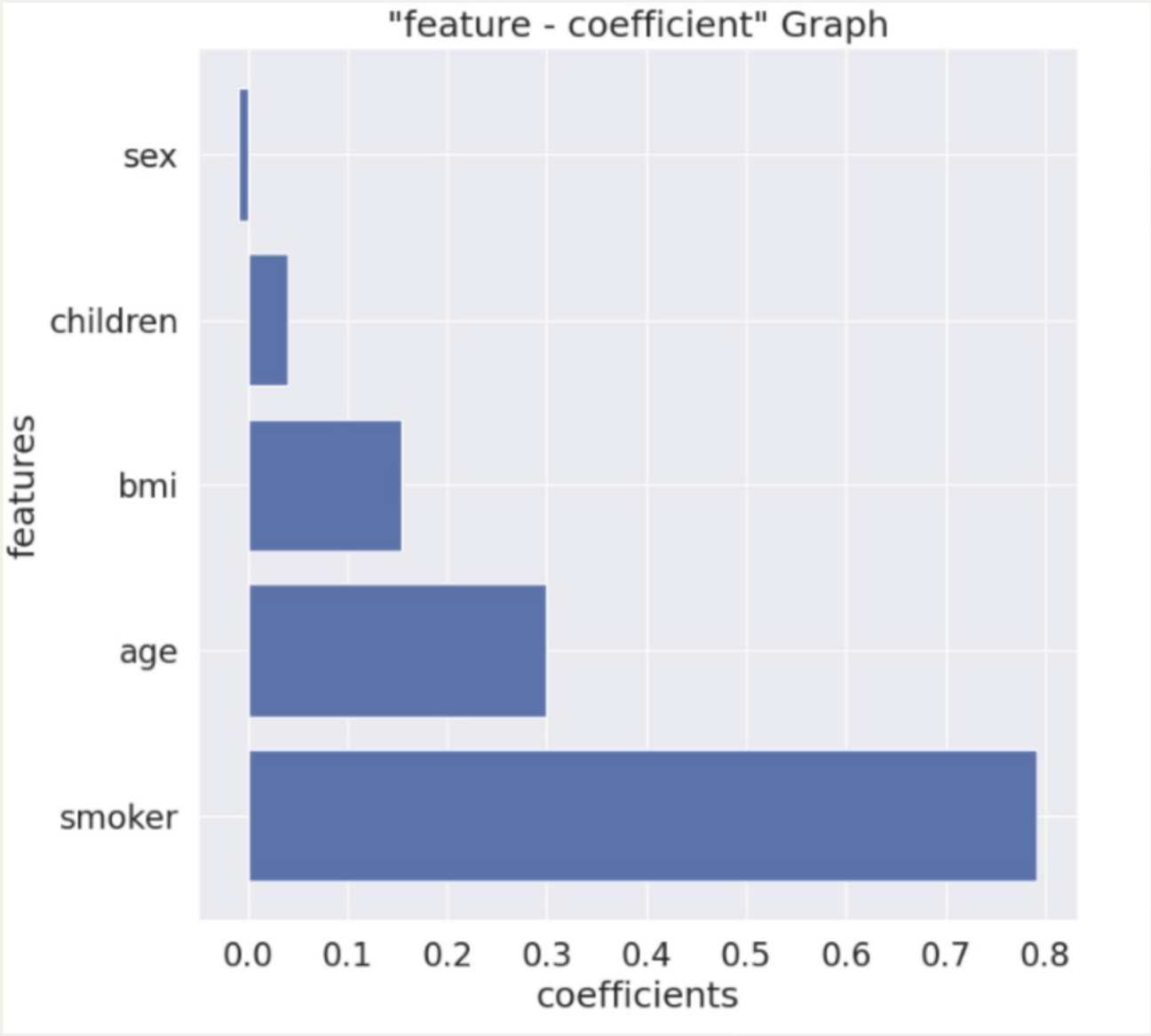
1070 rows × 5 columns

4-2.회귀모델링

```
from sklearn import linear_model

# fit regression model in training set
lr = linear_model.LinearRegression()
model = lr.fit(X_train, y_train)

# predict in test set
pred_test = lr.predict(X_test)
```

```
### https://m.blog.naver.com/tkdldjs35/2
```

	feature	coefficients
0	age	0.299486
1	bmi	0.153533
2	children	0.040430
3	sex	-0.009753
4	smoker	0.792312



4-3.모델해석

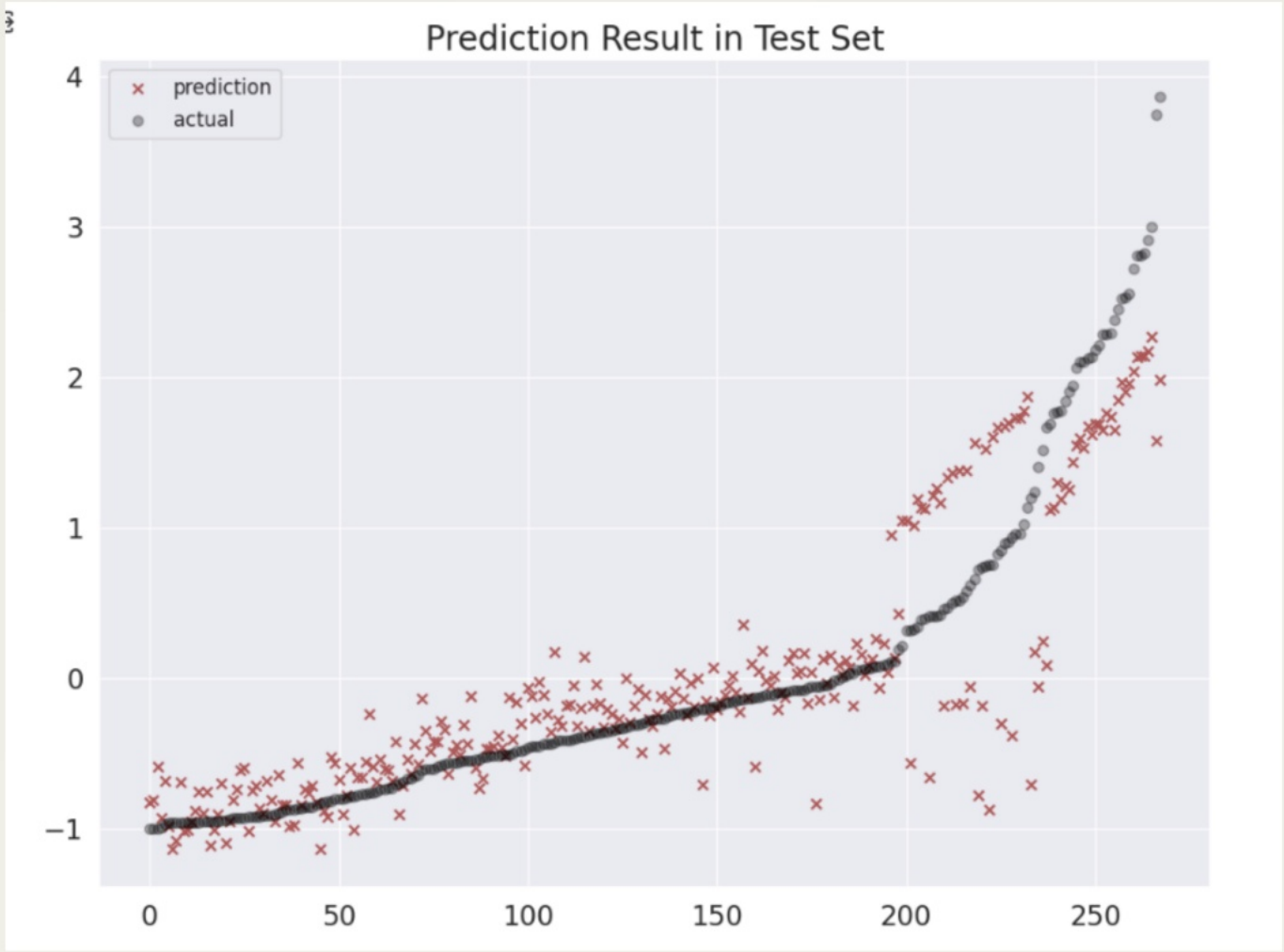
OLS Regression Results

Dep. Variable:	charges	**R-squared:**	0.747
Model:	OLS	**Adj. R-squared:**	0.745
Method:	Least Squares	**F-statistic:**	627.2
Date:	Fri, 12 Apr 2024	**Prob (F-statistic):**	3.52e-314
Time:	09:34:01	**Log-Likelihood:**	-781.16
No. Observations:	1070	**AIC:**	1574.
Df Residuals:	1064	**BIC:**	1604.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0007	0.015	-0.048	0.962	-0.031	0.029
age	0.2995	0.016	19.202	0.000	0.269	0.330
bmi	0.1535	0.015	9.942	0.000	0.123	0.184
children	0.0404	0.015	2.630	0.009	0.010	0.071
sex	-0.0098	0.015	-0.631	0.528	-0.040	0.021
smoker	0.7923	0.015	51.386	0.000	0.762	0.823

Omnibus:	239.860	**Durbin-Watson:**	1.968
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	555.395
Skew:	1.213	**Prob(JB):**	2.50e-121
Kurtosis:	5.563	**Cond. No.**	1.14

4-4.모델예측 결과 및 성능평가(예측결과 시각화)



4-4.모델성능평가

[4-4-2]모델의 성능평가(R-squred와 RMSE)

```
[ ]  print(model.score(X_train, y_train))  # training set
     print(model.score(X_test, y_test))  # test set
```

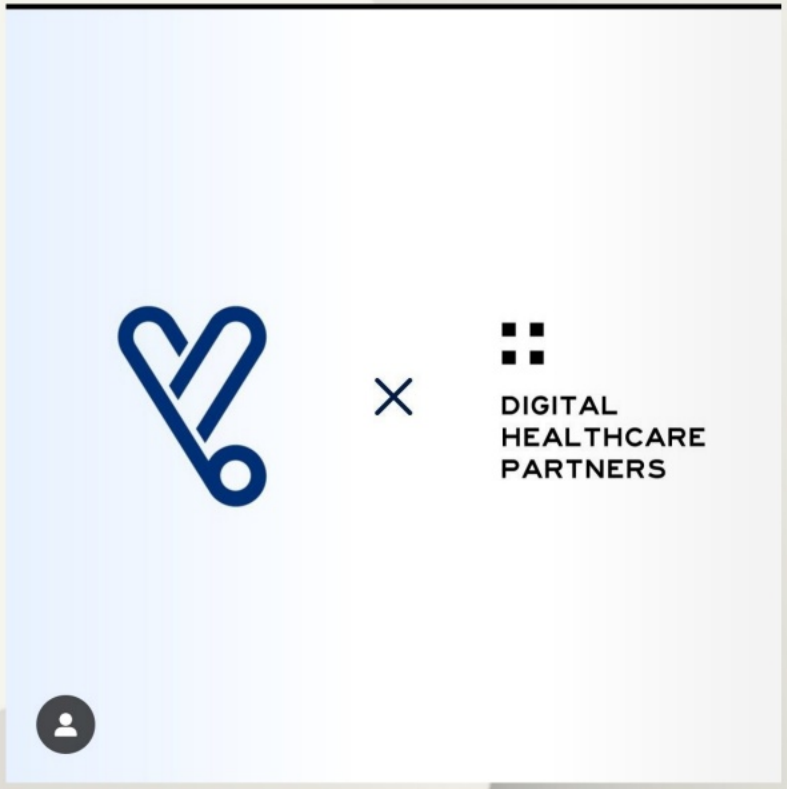
```
0.7466601137582649
0.760858175073853
```

```
[ ]  ### RMSE(Root Mean Squared Eror)
     from sklearn.metrics import mean_squared_error
     from math import sqrt

     ### training set
     pred_train = lr.predict(X_train)
     print(sqrt(mean_squared_error(y_train, pred_train)))

     ### test set
     print(sqrt(mean_squared_error(y_test, pred_test)))
```

```
0.5021358289563582
0.4935683696140902 3
```

Conclusion

$$y = 0.7923120 * \text{smoker} + 0.2994861 * \text{age} + 0.1535332 * \text{bmi} + 0.0404303 * \text{children} - 0.009753 * \text{sex}$$


×
DIGITAL
HEALTHCARE
PARTNERS


Healthcare
Partners
케어 전문 투자사, 디지털 헬스케어 파트너스
MEDILUX


Project
Make it simple and achieve without struggling
루티너리 서비스 유입과 사용 과정을 분석하여
고객 경험을 개선하는 기업 과제가 예정되어 있습니다.

개인의 건강보험요금 예측

2021120120 이윤경

문제정의(problem definition)

개인의 건강보험요금 예측

결론 선정 회귀 모델을 사용하여 개인의 건강보험 요금을 예측할 때, 모델은 주어진 독립 변수의 값에 기반하여 건강보험 요금을 계산

종속변수:건강보험료금(charges)
독립변수:예측에 사용되는 변수로, 나이(age), 성별(sex), BMI(bmi), 자녀(children), 흡연자(smoker)

문제 -> 분석->해결