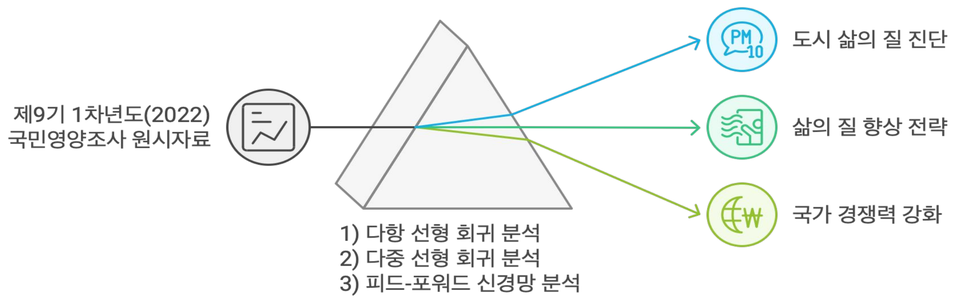
국민 건강조사를 활용한 스트레스 예측 모형

경희대학교 지리학과 19학번 김동건, 19학번 정형민
21학번 왕유민, 22학번 최지원, 22학번 서혜령

연구 배경 및 목적

- 현대 사회에서 도시 삶의 질(Urban Quality of Life)은 국가 및 도시 경쟁력의 핵심 요소로 대두됨
- 본 연구는 국민영양조사 데이터를 통해 우리나라 도시 삶의 질을 분석하고 진단하기 위해 모델을 만들고자 함
- 궁극적으로 대한민국 경쟁력 향상을 위해 도시 삶의 질 향상을 위한 실질적이고 효과적인 정책 수립에 기여하고자 함

연구방법



연구내용

1. 데이터 전처리

전처리 과정

- 결측치, 무응답, 모르겠음 제거
- 명목척도인 '성별' one-hot encoding 진행
- 학습 세트와 검증 세트를 8:2 비율로 설정
- 과적합을 방지해주기 위해 <표준화와 정규화 중 표준화 진행>

전처리 결과

Scaled Training Data:

	age	incm	edu	D_1_1	BP16_1	B01	BP_PHQ_2	N_WAT_C	L_BR_FQ	sex_F	sex_M
0	0.770492	0.666667	1.000000	0.25	0.333333	0.25	0.000000	0.100000	1.000000	0.0	1.0
1	0.672131	0.333333	0.666667	0.50	0.333333	0.50	0.000000	0.100000	1.000000	0.0	1.0
2	0.868852	0.333333	0.000000	0.00	0.500000	0.75	0.000000	0.100000	1.000000	1.0	0.0
3	0.557377	0.333333	0.666667	0.75	0.666667	0.50	0.333333	0.283333	0.000000	1.0	0.0
4	0.360656	0.333333	1.000000	0.25	0.333333	0.50	0.000000	0.100000	0.666667	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...
3636	0.508197	0.333333	1.000000	0.75	0.333333	1.00	0.000000	0.166667	0.666667	0.0	1.0
3637	0.639344	0.333333	0.666667	0.25	0.250000	0.50	0.000000	0.100000	0.666667	0.0	1.0
3638	0.786885	0.333333	0.333333	0.50	0.416667	1.00	0.000000	0.066667	1.000000	1.0	0.0
3639	0.901639	0.000000	0.000000	0.50	0.500000	1.00	0.000000	0.100000	1.000000	0.0	1.0
3640	0.000000	1.000000	0.666667	0.50	0.583333	0.75	0.000000	0.133333	1.000000	1.0	0.0

[3641 rows x 11 columns]

2. 모델학습 (다중 선형 회귀)

다중 선형 회귀 모델

- Statsmodel 라이브러리 사용
- 상수항 추가하여 분석 진행
- N_WAT_C (물 섭취량), incm (소득) 변수 유의하지 않음
- 다중 공선성 문제 확인

OLS Regression Results

Dep. Variable:	BP1	R-squared:	0.213
Model:	OLS	Adj. R-squared:	0.211
Method:	Least Squares	F-statistic:	122.9
Date:	Sat, 16 Nov 2024	Prob (F-statistic):	1.31e-227
Time:	05:47:26	Log-Likelihood:	-4500.0
No. Observations:	4552	AIC:	9022.
Df Residuals:	4541	BIC:	9093.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.9022	0.066	28.655	0.000	1.772	2.032
age	0.0079	0.001	10.168	0.000	0.006	0.009
incm	0.0023	0.009	0.260	0.795	-0.015	0.020
edu	-0.0450	0.011	-3.973	0.000	-0.067	-0.023
D_1_1	-0.1509	0.012	-12.932	0.000	-0.174	-0.128
BP16_1	0.0444	0.007	5.919	0.000	0.030	0.059
B01	-0.0296	0.010	-2.934	0.005	-0.050	-0.009
BP_PHQ_2	-0.3518	0.017	-21.094	0.000	-0.385	-0.319
N_WAT_C	0.0006	0.003	0.183	0.855	-0.006	0.007
L_BR_FQ	0.0255	0.009	2.782	0.005	0.008	0.043
sex_F	0.9410	0.034	27.584	0.000	0.874	1.008
sex_M	0.9612	0.035	27.280	0.000	0.892	1.030

Omnibus:	133.338	Durbin-Watson:	1.918
Prob(Omnibus):	0.000	Jarque-Bera (JB):	157.860
Skew:	-0.372	Prob(JB):	5.26e-35
Kurtosis:	3.527	Cond. No.	2.91e+16

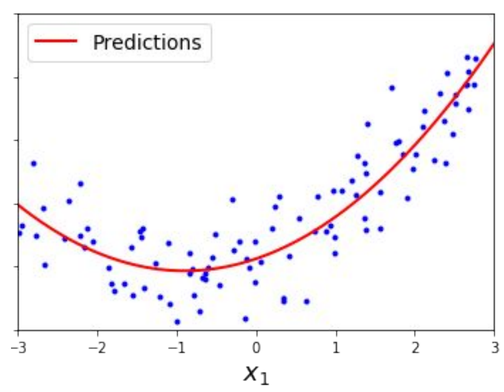
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.72e-26. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

3. 모델학습 (다항 회귀)

다항 선형 회귀 모델

- 변수들을 Polynomial transformer를 사용하여 2차항으로 전환
- 해당 변수들을 다항 선형 회귀 진행

```
array(['1', 'const', 'age', 'incm', 'edu', 'D_1_1', 'BP16_1', 'B01',
       'BP_PHQ_2', 'N_WAT_C', 'L_BR_FQ', 'sex_F', 'sex_M', 'const^2',
       'const age', 'const incm', 'const edu', 'const D_1_1',
       'const BP16_1', 'const B01', 'const BP_PHQ_2', 'const N_WAT_C',
       'const L_BR_FQ', 'const sex_F', 'const sex_M', 'age^2', 'age incm',
       'age edu', 'age D_1_1', 'age BP16_1', 'age B01', 'age BP_PHQ_2',
       'age N_WAT_C', 'age L_BR_FQ', 'age sex_F', 'age sex_M', 'incm^2',
       'incm edu', 'incm D_1_1', 'incm BP16_1', 'incm B01',
       'incm BP_PHQ_2', 'incm N_WAT_C', 'incm L_BR_FQ', 'incm sex_F',
       'incm sex_M', 'edu^2', 'edu D_1_1', 'edu BP16_1', 'edu B01',
       'edu BP_PHQ_2', 'edu N_WAT_C', 'edu L_BR_FQ', 'edu sex_F',
       'edu sex_M', 'D_1_1^2', 'D_1_1 BP16_1', 'D_1_1 B01',
       'D_1_1 BP_PHQ_2', 'D_1_1 N_WAT_C', 'D_1_1 L_BR_FQ', 'D_1_1 sex_F',
       'D_1_1 sex_M', 'BP16_1^2', 'BP16_1 B01', 'BP16_1 BP_PHQ_2',
       'BP16_1 N_WAT_C', 'BP16_1 L_BR_FQ', 'BP16_1 sex_F', 'BP16_1 sex_M',
       'B01^2', 'B01 BP_PHQ_2', 'B01 N_WAT_C', 'B01 L_BR_FQ', 'B01 sex_F',
       'B01 sex_M', 'BP_PHQ_2^2', 'BP_PHQ_2 N_WAT_C', 'BP_PHQ_2 L_BR_FQ',
       'BP_PHQ_2 sex_F', 'BP_PHQ_2 sex_M', 'N_WAT_C^2', 'N_WAT_C L_BR_FQ',
       'N_WAT_C sex_F', 'N_WAT_C sex_M', 'L_BR_FQ^2', 'L_BR_FQ sex_F',
       'L_BR_FQ sex_M', 'sex_F^2', 'sex_F sex_M', 'sex_M^2'], dtype=object)
```



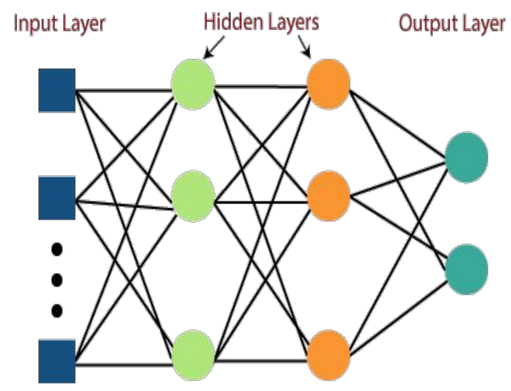
분석)

- Tensorflow Keras 고수준 api를 활용한 피드 포워드 신경망 구현
- 최적의 하이퍼 파라미터를 찾기 위해 그리드 서치 진행
- 노드 범위 32~256개, 은닉층 개수 1~3개

Activation Function	RELU
Optimizer	ADAM
Loss Function	Mean Squared Error
Metric	Mean Absolute Error

```
# 최적 하이퍼파라미터로 모델 학습
best_model = tuner.hypermodel.build(best_hps)
best_model.fit(X_train, y_train, epochs=50, validation_split=0.2)
```

```
Epoch 1/50
/usr/local/lib/python3.10/dist-packages/keras/src/layers/core/dense.py:87: UserWarning: Do not pass an `input_shape`/`input_di
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
91/91 ━━━━━━━━━━━━━━━━━━━━ 2s 6ms/step - loss: 11.6887 - mae: 1.7449 - val_loss: 0.6115 - val_mae: 0.6357
Epoch 2/50
91/91 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.5160 - mae: 0.5564 - val_loss: 0.6322 - val_mae: 0.6221
Epoch 3/50
91/91 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.4736 - mae: 0.5345 - val_loss: 0.5059 - val_mae: 0.5460
Epoch 4/50
91/91 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.4901 - mae: 0.5485 - val_loss: 0.5715 - val_mae: 0.5892
Epoch 5/50
91/91 ━━━━━━━━━━━━━━━━━━━━ 1s 4ms/step - loss: 0.4908 - mae: 0.5460 - val_loss: 0.5778 - val_mae: 0.5852
Epoch 6/50
91/91 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.4768 - mae: 0.5312 - val_loss: 0.4936 - val_mae: 0.5375
Epoch 7/50
91/91 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.4768 - mae: 0.5276 - val_loss: 0.5215 - val_mae: 0.5587
Epoch 8/50
91/91 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - loss: 0.4780 - mae: 0.5272 - val_loss: 0.4723 - val_mae: 0.5186
Epoch 9/50
91/91 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - loss: 0.4603 - mae: 0.5142 - val_loss: 0.4846 - val_mae: 0.5311
Epoch 10/50
```



- 그리드 서치 결과

Best units for layer no.1	192
Best number of layers	3
Best learning rate	0.005537024633539518

연구 결론

- 도시 삶의 질 향상을 위해서는 스트레스 관리가 필요하며, 우울감과 주관적 건강인지, 그리고 아침식사가 중요하다는 단서를 얻음
- 따라서 본 연구를 통해 우리나라 각 지자체가 구성원들의 스트레스 관리를 위해 우울감과 주관적 건강인지, 아침식사 빈도를 진단할 필요성이 있다고 판단됨
- 또한, 우울감을 해소하기 위한 대책 마련, 국민 건강 증진 도모, 아침식사 권장 캠페인 등을 제안하는 바임.

연구 의의 및 한계

- 통계적으로 변수를 찾고 선형과 비선형 관계를 모두 분석해보았음. 또한, 도시 삶의 질 향상을 위한 단서를 도출함
- 스트레스 지수를 R-squared 기준으로 매우 높은 설명력을 가공해내지 못했으며 표준화로 인한 underfitting 이 일어났을 수 있음
- 도시 삶의 질 향상을 위한 다소 추상적인 정책을 제안함

참고문헌

- 질병관리청. (2022). *국민건강영양조사 원시자료 이용지침서: 제 9기 1차년도 (2022).* 질병관리청. https://knhanes.kdca.go.kr