AUGUST . 02 / 2024

Project Presentation @ Yonsei IT Future Education Center (No. 702)

Analysis of the local extinction crisis based on data and machine learning **techniques**

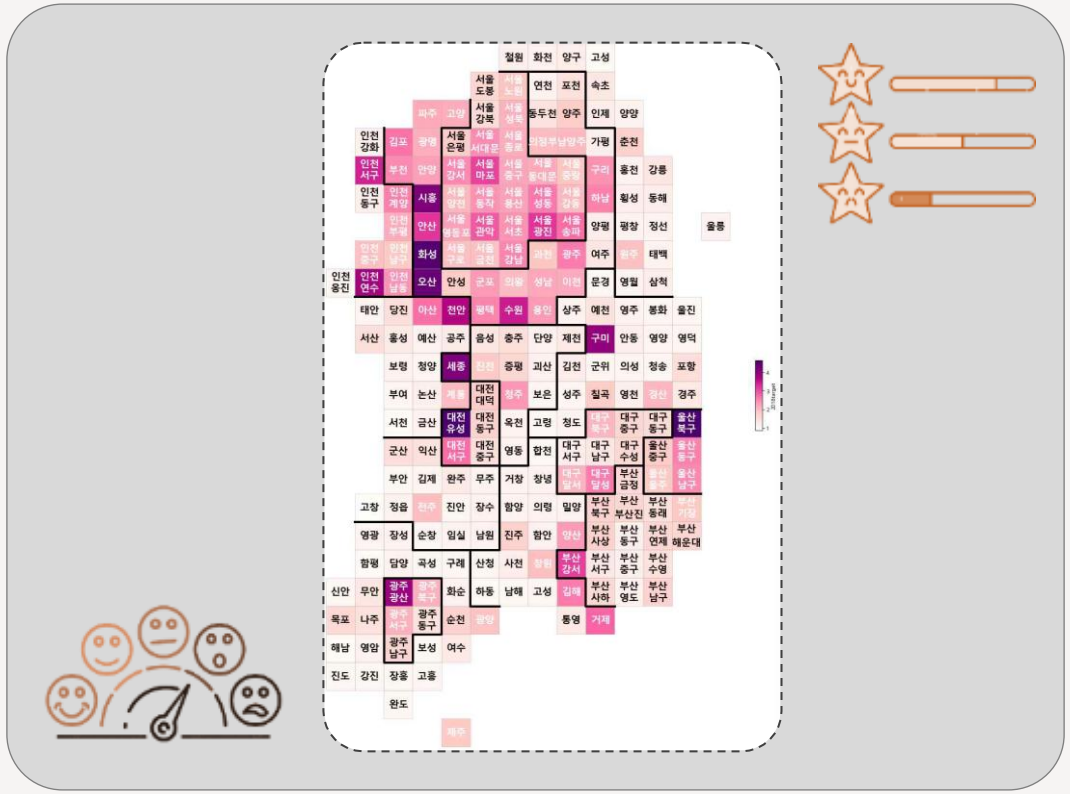
Contents

Contents

I. Project Overview

ocal extinction c

① Visualization of extinction risk inde ② Prediction of areas at risk of extinction in th



Source : Yonsei

Source : KBS

(Source) DAPT Desk Research

Expected effect

Areas at risk of extinction based on variables

Find out and **predict areas at risk of extinction in the future**

(Source) DAPT Desk Research

" The disappearance **of areas and villages** where people have lived and are still living ."

low birth rate , aging population and population movement to large cities

De-

squatter phenomenon

Fat loss ?

Former Chairman of the Japan Creation Association

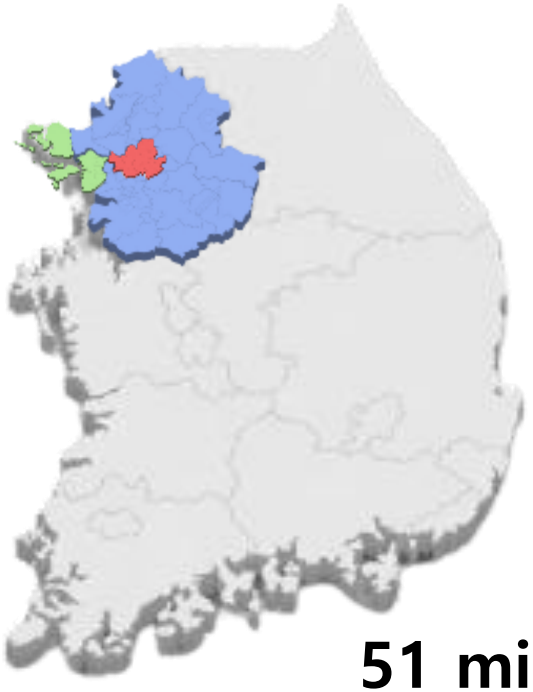
Masuda Hiroya

First time use

(Source) Hong Seon-mi , "[Project 2] The Current Status of Local Extinction and Mu

I. Project Overview – Reasons for Selection

Nationwide



51 million

100% population share

100%

Seoul Metropolitan City

gyeonggi-do

Incheon Metropolitan City

metro polita n area

13.4 million

3 million

9.3 million

25.7 million

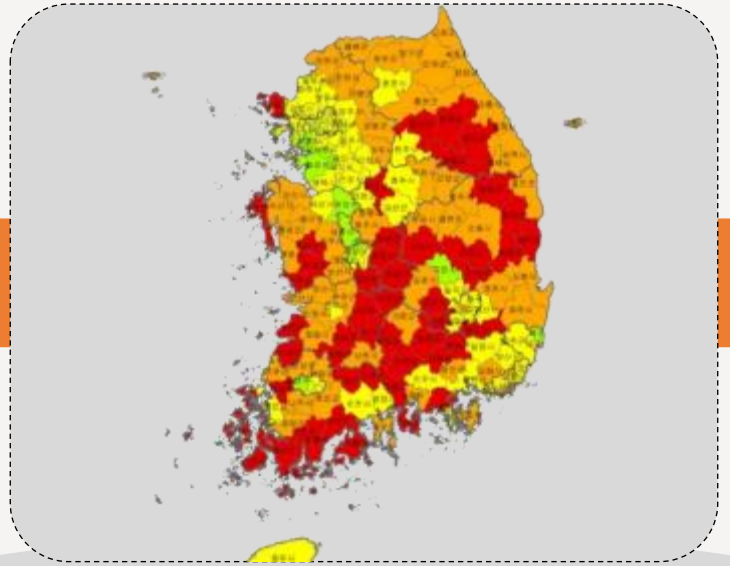
51% of the population

12%

(Source) DAPT Desk Research

Why fat loss ?

Good



Serious

" The society is seriously aware of the problems of regional extinction and local extinction ,
Accordingly , we hope to help **resolve the regional extinction crisis** ."

(Source) National Employment Information Service , Local Extinction 2024: The Extinction Risk Spreading to Metropolitan Cities , June 28, 2024

I. Project Overview – Reasons for Selection (2)

뉴스광장(창원)

지방소멸 대응 급한데…통계는 '제 각각'

입력 2024.07.02 (07:42) | 수정 2024.07.02 (08:51)

【뉴스초점】 지방소멸 시대 '모범해법'으로 … 진천군 충북 첫 인구 전담 부서 신설

입력 2024.07.07 10:08 | 수정 2024.07.07 10:10 | 댓글 0

buy episode this choux

함양군 지방소멸대응 전담추진단 회의 개최

기사승인 2024.07.09 10:50:27

홈 > 사회 > 일반사회

지방소멸 가속화 충북도 위험하다

청주시도 10년 후 소멸 위험단계 진입 전망

2022.09.26 21:03:43

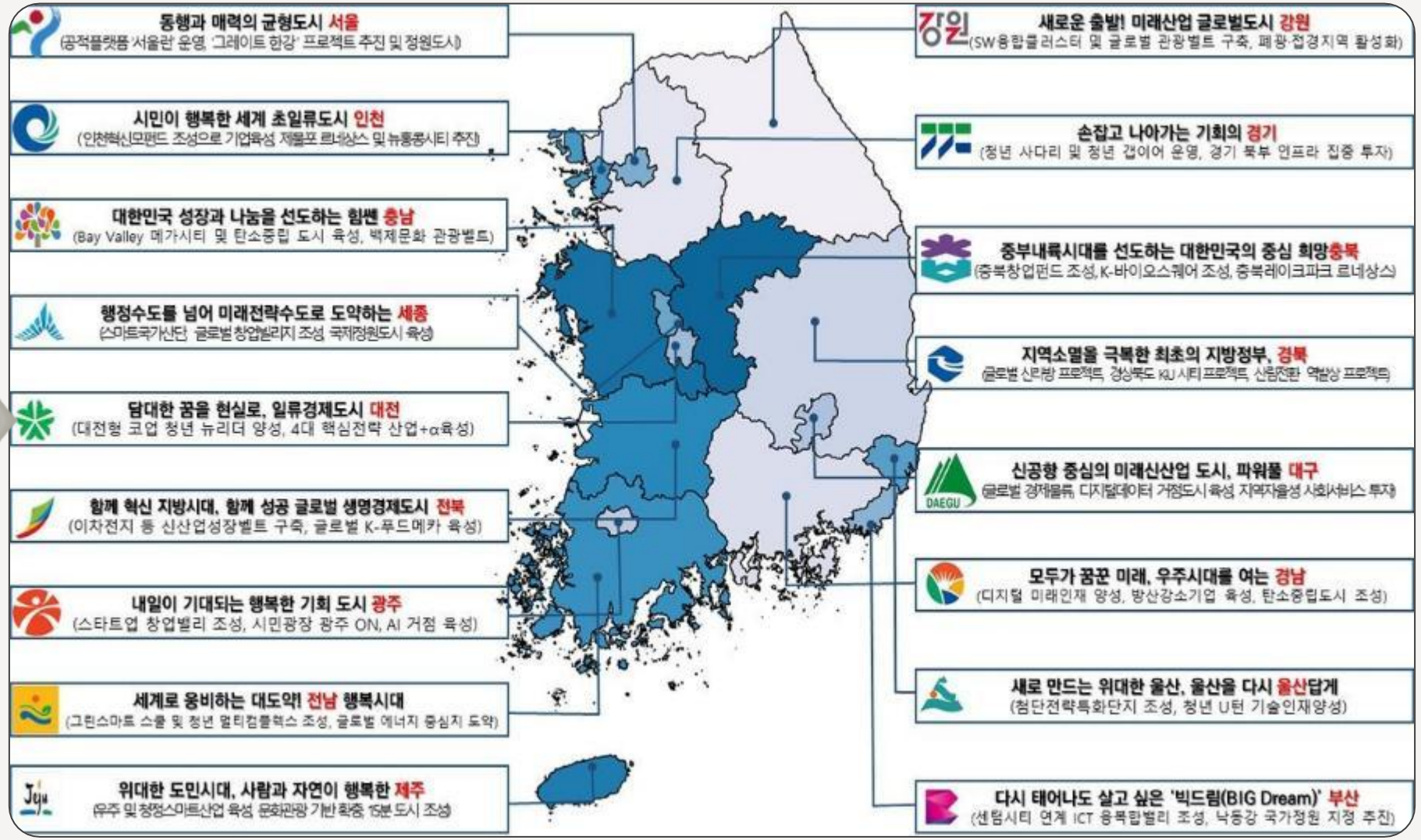
(Source) Dedicated team Desk Research

I. Project Overview – Analysis Plan – Target (1)

- **Analysis target : Nationwide (17 cities and provinces) / 229 cities , counties, and districts)**

Existing : Chungcheong region

Change : Nationwide



(Source) Local Government Committee

Just by improving the infrastructure in the Chungcheong region There are limits to analysis !

Sejong

The population continues to in flow , and infrastructure also c ontinues to increase.

electrification

Continuous population outflow , no decrease in infrastructure

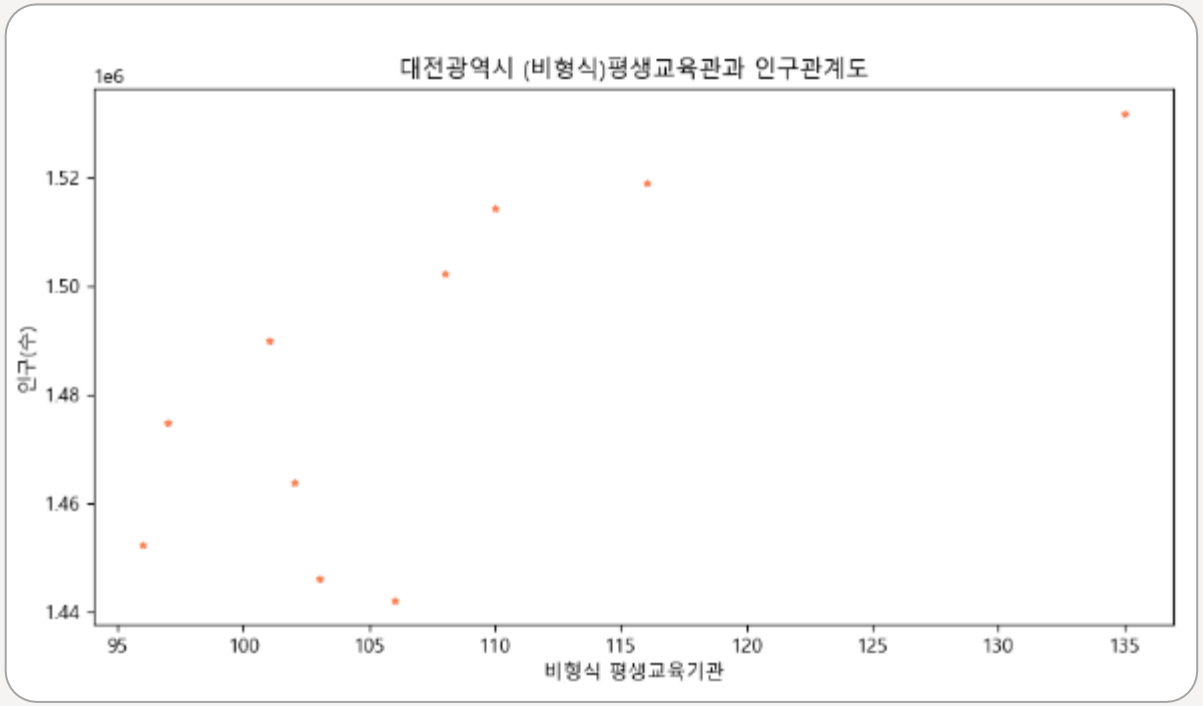
< Analysis Results >

" We face limitations in accurately analyzing population inflow and outflow and local extinction **based on a small number of regional samples and infrastructure alone** ."

" To increase the reliability and accuracy of the analysis, **the analysis target was expanded nationwide** ."

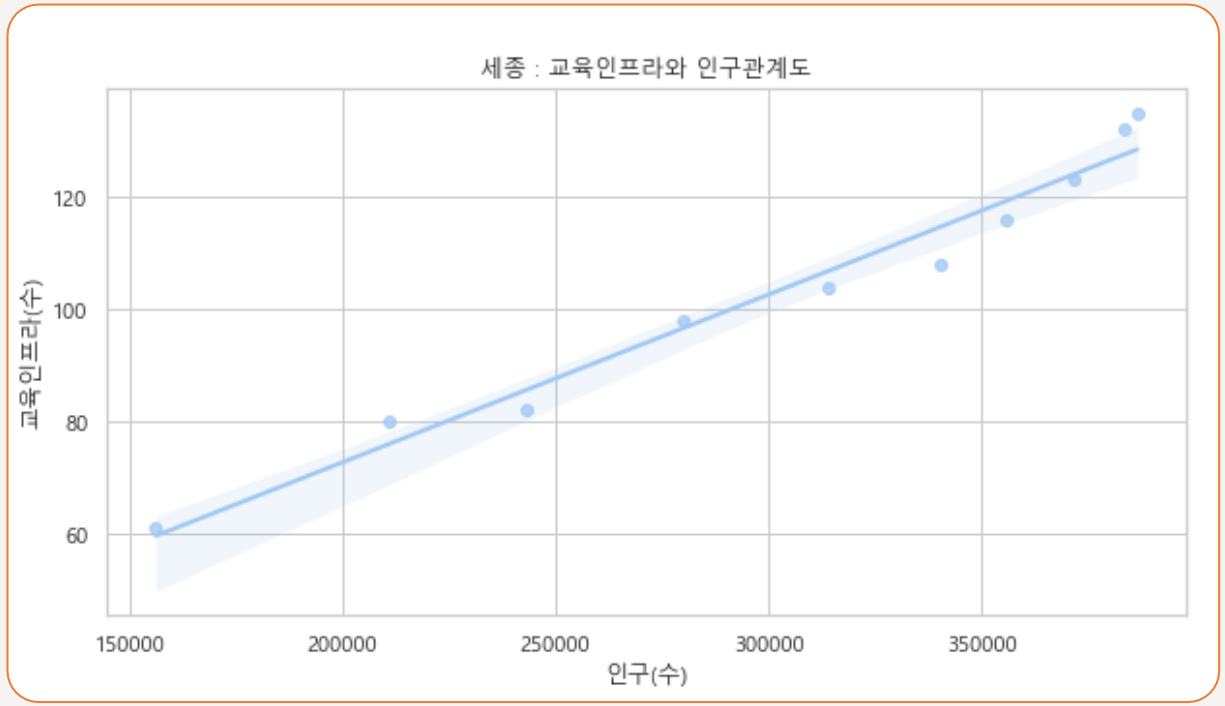
(Source) Team 2 Desk Research

electrification



No significant relationship appears

Sejong



A positive correlation appears

(Source) Team 2 Desk Research

II. Introduction of team members and their roles

Data Analytics Professional Team

Seong

Data crawling
Data preprocessing
Data collection
presentation

Shin

DBMS
Data preprocessing
Data collection
Data Anaylsis

Park

Planning
Building
Data preprocessing
Data collection

Lee

Backend construction
Building
Data visualization
Data analysis

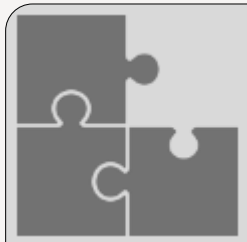
Oh

Backend construction
Data preprocessing
Data visualization
Data collection

III. Development Process

Self-analysis of the pilot project

Initial goal : Let's do something people have never tried before !



1) Since we started from scratch , the project completion rate was low.

2) To supplement this, the RCF technique was used.

(Source) Vent Florvia , Dan Gardner , and Park Young-jun , "Project

the RCF technique

Goal improvem

During the planning stage of the project, I thought about how to proceed with the project in the future .

Decided to carry out the project based on **the RCF technique**

the RCF technique ?

Time saving & completeness

The RCF technique refers to similar projects that have been carried out in the past to identify potential problems. (Time cost/ Pre-removal of variables with missing values /Result prediction) is prevented , and based on this, **time is saved and completion is improved. target**

Reference

「지방정부연구」 제24권 제4호(2021 겨울): 443-476

The Korean Journal of Local Government Studies, Vol.24 No.4 (2021 Winter)

http://dx.doi.org/10.20484/klog.24.4.18

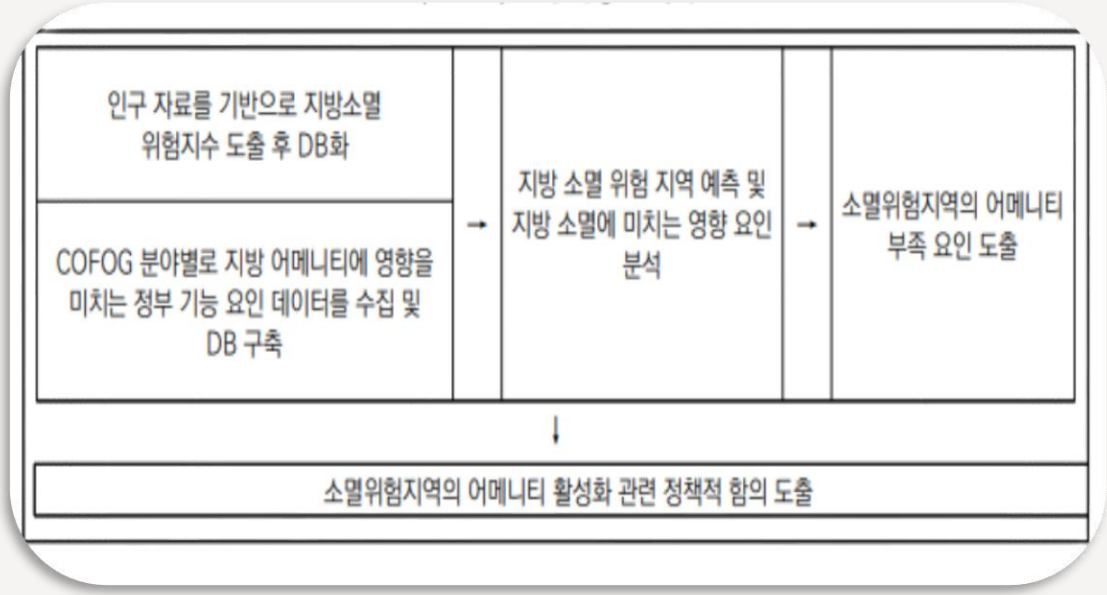
한국 지방소멸 요인과 극복 방안에 관한 연구:

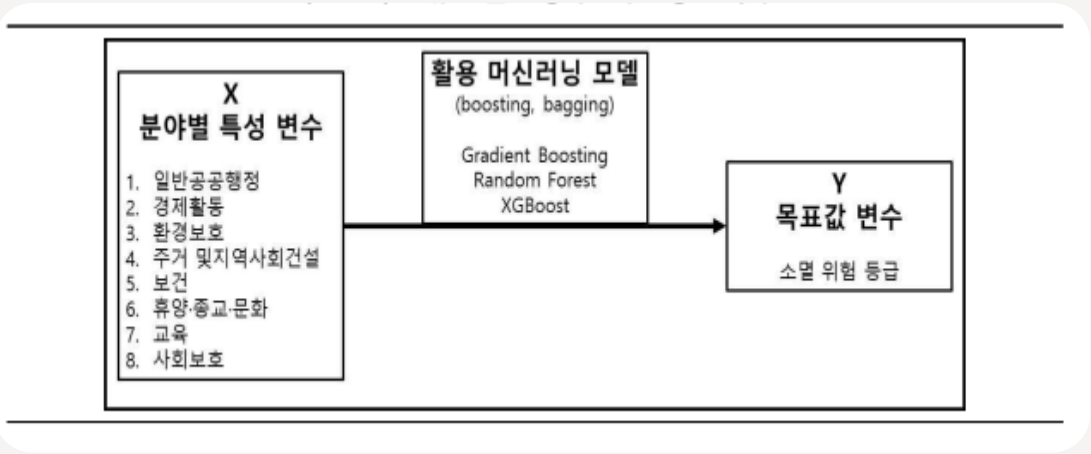
머신러닝 방법을 통한 탐색*

유 한 별**

탁 근 주***

문 정 승****





Existing research and Differences

Limitations of existing research

Due to insufficient variable data size , we utilize

Study : 2015-2017 (3 years of data) / Validation : 2018 (1 year)

Overcoming New Research Limitations

Data-centric by increasing data size Improvement in the way

2015~2020 (6- year data) / Verification : 2021 (1- year)

Previous research

GBM

Among the models used i n previous studies, **Highes t performing model**

Latest research

LGBM

The model with the highest predictive ra te in the recent mach ine learning field

Prediction rate I raised it higher .

(Source) DAPT Desk Research

III. Development Process – Dependent Variables

Existing extinction risk index

Calculation formula

$$\frac{\text{가임여성인구}}{\text{노인인구}}$$

Scaling technique
(doesn't exist)

(limit point)
Local extinction and population migration

Absence of variables explaining the relationship

VS

New extinction risk index

Calculation formula

$$\text{기존소멸위험지수}^2 + Log(\frac{\text{전입인구}}{\text{전출인구}})$$

Added scaling technique
(Min-Max Scaling + Robust Scaling +Standard Scaling)

(Improvement points)
Local extinction and population migration

Reflect variables that explain relationships

III. Development Process – Independent Variables

Unselected independent va

- air quali
- National Def
- Recreational
- public order an d safety

Reason for

air quality
The final decision was made to remove missing values because there were too many missing values to replace them , and the data **was inconsistent** from un it to unit by year

National Defense
excluded the indicator from the characteristic variables because **it targeted local governments** as a characteristic of expenditure at the national level.

Recreation, Culture, Religion & Public Order and Safety
Couldn't add more due to **lack of time**

Selection variable factors

- econom
- Transport
- educati
- environ
- Health
- Housing A
- social prot

COFOG: “Classification of the Functions of Government”

① An indicator with

By function according to the g overnment function classificati on

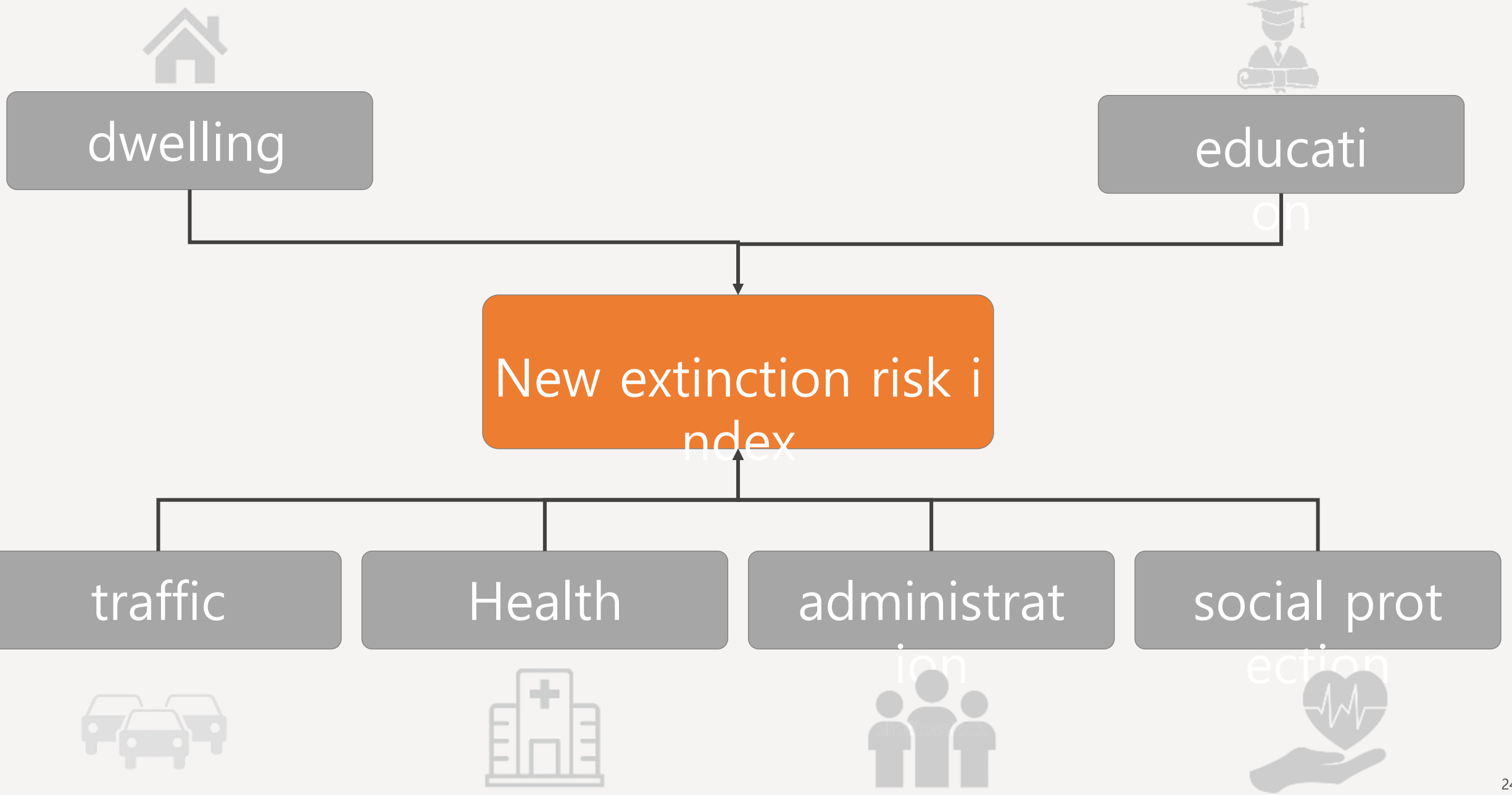
Categorized characteristics

Elements

, leisure culture and religion , pu blic order and safety factors fro m

the 7 elements

III. Development Process – Independent and Dependent Variables



III. Development Process - 2. Data Preprocessing – Check

) Data preprocessing (integrity)

Existing problems

(Multiple missing values)

(Case 1)

- Daegu Metropolitan City : Previously part of Gyeongsangbuk-do , but incorporated into Daegu Metropolitan City.
- Sejong Special Self-Governing City : Not long ago it became

(Case 2)

' Number of private academies for school subjects '

There are many missing values (0 or -)

Solution

1) In a larger category, the city-level data was divided by the number of districts and replaced with the district average data value to increase data integrity by configuring it to minimize duplicate data.

2) 'Integrity constraints' In accordance with the principle of null integrity, a variable (number of students per private academy) was added

arbitrarily to ensure that there are no null val

Data collection

Case ① Crawling (Automatically collected)

통계청 API (Data URL)

+

crawling (URL -based collection)

Case ② Research (direct collection)

통계청 directly Download

+

Excel file Download

Data preprocessing (Data organization)

Case ① Summary by year

City/county /district	year	Variable 1
Suwon City	2015	10
Seongnam City	2015	20
Yongin City	2015	30
Anyang City	2015	40
Gwacheon City	2015	50

Case ② Region name update

Existing region name	New region name
Gunwi-gun, Gyeongsangbuk-do	Gunwi-gun, Daegu Metropolitan City
Nam-gu, Incheon Metropolitan City	Michuhol-gu,

경상북도 → DAEGU 대구광역시

Data preprocessing (Check missing values)

Case ① Partial missing values (City, county , district Replace missing values with trial mean)

City/county /district	2015	2016
Suwon City	1	2
Seongnam City	2	4
Yongin City	3	6
Anyang City	4	8
Gwacheon C	5	5
City/county /district	**2015**	**2016**
Gyeonggi-do average	5	5

Case ② Missing values for an entire specific year (Replace data with the most r

City/county/district	2015	2016	2017	2018
Suwon City	2	2	1	3
Seongnam City	4	4	2	9
City/county /district	**2015**		**2016**	
Suwon City	8	8	4	2 5
g City				

Data preprocessing (Scaling)

Case ① min-max scaling

A technique for calcu lating the average ba sed on

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

Case ② Robust scaling

A technique for calcu lating the average ba sed on

$$\frac{x_i - median(x)}{Q_3 - Q_1}$$

Case ③ standard scaling

o that the ranges of all features are simila r.

Matching technique

$$Z = \frac{x - \mu}{\sigma}$$

Data preprocessing (Scaling)

Standard Scaler:

OLS Regression Results

Dep. Variable:	소멸위험등급		R-squared:		0.346	
Model:	OLS		Adj. R-squared:		0.329	
Method:	Least Squares		F-statistic:		20.18	
Date:	Thu, 01 Aug 2024		Prob (F-statistic):		4.34e-115	
Time:	15:21:16		Log-Likelihood:		-2112.9	
No. Observations:	1603		AIC:		4310.	
Df Residuals:	1561		BIC:		4536.	
Df Model:	41					
Covariance Type:	nonrobust					

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.4997	0.023	109.242	0.000	2.455	2.545
x1	0.0814	0.056	1.456	0.146	-0.028	0.191
x2	0.2017	0.157	1.285	0.199	-0.106	0.510
x3	0.4873	0.117	4.166	0.000	0.258	0.717
x4	0.2949	0.077	3.845	0.000	0.144	0.445
x5	0.1529	0.057	2.671	0.008	0.041	0.265
x6	-0.4466	0.155	-2.876	0.004	-0.751	-0.142
x7	-0.3884	0.100	-3.874	0.000	-0.585	-0.192
x8	-0.0993	0.076	-1.313	0.189	-0.248	0.049
x9	-0.4006	0.088	-4.573	0.000	-0.572	-0.229
x10	0.2636	0.092	2.873	0.004	0.084	0.444
x11	-0.0073	0.032	-0.225	0.822	-0.071	0.056
x12	0.0237	0.101	0.234	0.815	-0.175	0.222
x13	0.0603	0.121	0.500	0.617	-0.176	0.297
x14	-0.1394	0.042	-3.309	0.001	-0.222	-0.057
x15	-0.3058	0.058	-5.267	0.000	-0.420	-0.192
x16	-0.0342	0.089	-0.386	0.699	-0.208	0.139
x17	0.1089	0.123	0.887	0.375	-0.132	0.350
x18	0.1160	0.034	3.395	0.001	0.049	0.183
x19	0.2305	0.124	1.859	0.063	-0.013	0.474
x20	-0.0402	0.032	-1.241	0.215	-0.104	0.023
x21	0.1901	0.067	2.854	0.004	0.059	0.321
x22	0.0375	0.029	1.309	0.191	-0.019	0.094
x23	0.0062	0.033	0.188	0.851	-0.059	0.071
x24	-0.0190	0.026	-0.743	0.457	-0.069	0.031
x25	0.1202	0.027	4.400	0.000	0.067	0.174
x26	-6.0034	0.871	-6.893	0.000	-7.712	-4.295
x27	0.0481	0.045	1.075	0.283	-0.040	0.136
x28	-0.0182	0.036	-0.502	0.616	-0.089	0.053
x29	-0.0398	0.042	-0.943	0.346	-0.123	0.043
x30	-0.1148	0.028	-4.102	0.000	-0.170	-0.060
x31	0.0919	0.038	2.430	0.015	0.018	0.166
x32	-0.1516	0.034	-4.458	0.000	-0.218	-0.085
x33	-0.0204	0.024	-0.856	0.392	-0.067	0.026
x34	-0.0543	0.034	-1.605	0.109	-0.121	0.012
x35	-1.2348	0.216	-5.723	0.000	-1.658	-0.812
x36	-0.0983	0.115	-0.857	0.392	-0.323	0.127
x37	-0.1445	0.037	-3.940	0.000	-0.216	-0.073
x38	0.2804	0.031	9.073	0.000	0.220	0.341
x39	0.1330	0.028	4.754	0.000	0.078	0.188
x40	7.1802	0.903	7.950	0.000	5.409	8.952
x41	-0.0203	0.025	-0.809	0.419	-0.070	0.029

Omnibus:	41.333	Durbin-Watson:	1.700
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23.784
Skew:	0.127	Prob(JB):	6.84e-06
Kurtosis:	2.460	Cond. No.	212.

Standard Scaling

Coefficient of d

etermination :

Data preprocessing (Scaling)

Robust Scaler:

OLS Regression Results

Dep. Variable:	소멸위험등급		R-squared:		0.346	
Model:	OLS		Adj. R-squared:		0.329	
Method:	Least Squares		F-statistic:		20.18	
Date:	Thu, 01 Aug 2024		Prob (F-statistic):		4.34e-115	
Time:	15:21:16		Log-Likelihood:		-2112.9	
No. Observations:	1603		AIC:		4310.	
Df Residuals:	1561		BIC:		4536.	
Df Model:	41					
Covariance Type:	nonrobust					
	coef	std err	t	P>|t|	[0.025	0.975]
const	2.5092	0.052	47.943	0.000	2.407	2.612
x1	0.0911	0.063	1.456	0.146	-0.032	0.214
x2	0.3816	0.297	1.285	0.199	-0.201	0.964
x3	0.7506	0.180	4.166	0.000	0.397	1.104
x4	0.4564	0.119	3.845	0.000	0.224	0.689
x5	0.2259	0.085	2.671	0.008	0.060	0.392
x6	-0.7081	0.246	-2.876	0.004	-1.191	-0.225
x7	-0.5312	0.137	-3.874	0.000	-0.800	-0.262
x8	-0.1330	0.101	-1.313	0.189	-0.332	0.066
x9	-0.2459	0.054	-4.573	0.000	-0.351	-0.140
x10	0.1493	0.052	2.873	0.004	0.047	0.251
x11	-0.0013	0.006	-0.225	0.822	-0.013	0.010
x12	0.0256	0.109	0.234	0.815	-0.189	0.240
x13	0.0675	0.135	0.500	0.617	-0.197	0.332
x14	-0.1626	0.049	-3.309	0.001	-0.259	-0.066
x15	-0.3452	0.066	-5.267	0.000	-0.474	-0.217
x16	-0.0352	0.091	-0.386	0.699	-0.214	0.143
x17	0.1297	0.146	0.887	0.375	-0.157	0.416
x18	0.0233	0.007	3.395	0.001	0.010	0.037
x19	0.2558	0.138	1.859	0.063	-0.014	0.526
x20	-0.0464	0.037	-1.241	0.215	-0.120	0.027
x21	0.2261	0.079	2.854	0.004	0.071	0.381
x22	0.0503	0.038	1.309	0.191	-0.025	0.126
x23	0.0084	0.045	0.188	0.851	-0.080	0.096
x24	-0.0256	0.034	-0.743	0.457	-0.093	0.042
x25	0.1562	0.036	4.400	0.000	0.087	0.226
x26	-7.7056	1.118	-6.893	0.000	-9.898	-5.513
x27	0.0187	0.017	1.075	0.283	-0.015	0.053
x28	-0.0273	0.054	-0.502	0.616	-0.134	0.079
x29	-0.0396	0.042	-0.943	0.346	-0.122	0.043
x30	-0.0743	0.018	-4.102	0.000	-0.110	-0.039
x31	0.1319	0.054	2.430	0.015	0.025	0.238
x32	-0.1505	0.034	-4.458	0.000	-0.217	-0.084
x33	-0.0059	0.007	-0.856	0.392	-0.020	0.008
x34	-0.0413	0.026	-1.605	0.109	-0.092	0.009
x35	-1.5634	0.273	-5.723	0.000	-2.099	-1.028
x36	-0.1170	0.137	-0.857	0.392	-0.385	0.151
x37	-0.1906	0.048	-3.940	0.000	-0.286	-0.096
x38	0.3033	0.033	9.073	0.000	0.238	0.369
x39	0.1052	0.022	4.754	0.000	0.062	0.149
x40	9.3599	1.177	7.950	0.000	7.051	11.669
x41	-0.0193	0.024	-0.809	0.419	-0.066	0.027

Omnibus:	41.333	Durbin-Watson:	1.700
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23.784
Skew:	0.127	Prob(JB):	6.84e-06
Kurtosis:	2.460	Cond. No.	385.

Robust Scaling

Coefficient of d

etermination :

Data preprocessing (Scaling)

MinMax Scaler:

OLS Regression Results

Dep. Variable:	소멸위험등급	R-squared:				0.346
Model:	OLS	Adj. R-squared:				0.329
Method:	Least Squares	F-statistic:				20.18
Date:	Thu, 01 Aug 2024	Prob (F-statistic):				4.34e-115
Time:	15:21:16	Log-Likelihood:				-2112.9
No. Observations:	1603	AIC:				4310.
Df Residuals:	1561	BIC:				4536.
Df Model:	41					
Covariance Type:	nonrobust					

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.8132	0.376	4.821	0.000	1.075	2.551
x1	0.4250	0.292	1.456	0.146	-0.148	0.998
x2	0.8176	0.637	1.285	0.199	-0.431	2.066
x3	2.1017	0.505	4.166	0.000	1.112	3.091
x4	1.8255	0.475	3.845	0.000	0.894	2.757
x5	0.7529	0.282	2.671	0.008	0.200	1.306
x6	-1.7703	0.616	-2.876	0.004	-2.978	-0.563
x7	-2.2006	0.568	-3.874	0.000	-3.315	-1.086
x8	-0.6755	0.514	-1.313	0.189	-1.685	0.334
x9	-1.6695	0.365	-4.573	0.000	-2.386	-0.953
x10	1.4313	0.498	2.873	0.004	0.454	2.409
x11	-0.1131	0.502	-0.225	0.822	-1.098	0.872
x12	0.1486	0.635	0.234	0.815	-1.097	1.394
x13	0.3284	0.657	0.500	0.617	-0.961	1.618
x14	-0.8941	0.270	-3.309	0.001	-1.424	-0.364
x15	-1.8553	0.352	-5.267	0.000	-2.546	-1.164
x16	-0.3520	0.911	-0.386	0.699	-2.139	1.435
x17	0.7317	0.825	0.887	0.375	-0.886	2.349
x18	1.4812	0.436	3.395	0.001	0.625	2.337
x19	1.2976	0.698	1.859	0.063	-0.071	2.667
x20	-0.2901	0.234	-1.241	0.215	-0.749	0.168
x21	1.2652	0.443	2.854	0.004	0.396	2.135
x22	0.2594	0.198	1.309	0.191	-0.129	0.648
x23	0.0424	0.225	0.188	0.851	-0.399	0.484
x24	-0.1160	0.156	-0.743	0.457	-0.422	0.190
x25	0.7956	0.181	4.400	0.000	0.441	1.150
x26	-32.8079	4.760	-6.893	0.000	-42.144	-23.472
x27	0.3188	0.297	1.075	0.283	-0.263	0.900
x28	-0.0801	0.160	-0.502	0.616	-0.393	0.233
x29	-0.2891	0.307	-0.943	0.346	-0.891	0.313
x30	-1.9316	0.471	-4.102	0.000	-2.855	-1.008
x31	0.5299	0.218	2.430	0.015	0.102	0.958
x32	-0.9571	0.215	-4.458	0.000	-1.378	-0.536
x33	-0.2443	0.285	-0.856	0.392	-0.804	0.316
x34	-0.3489	0.217	-1.605	0.109	-0.775	0.077
x35	-6.8086	1.190	-5.723	0.000	-9.142	-4.475
x36	-0.7233	0.844	-0.857	0.392	-2.379	0.933
x37	-1.0437	0.265	-3.940	0.000	-1.563	-0.524
x38	2.2144	0.244	9.073	0.000	1.736	2.693
x39	2.3583	0.496	4.754	0.000	1.385	3.331
x40	38.8799	4.890	7.950	0.000	29.287	48.472
x41	-0.2191	0.271	-0.809	0.419	-0.750	0.312

Omnibus:	41.333	Durbin-Watson:	1.700
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23.784
Skew:	0.127	Prob(JB):	6.84e-06
Kurtosis:	2.460	Cond. No.	805.

Min-Max Scaling

Coefficient of determinati

on : 0.346

Data preprocessing (Scaling)

OLS Regression Results 회귀분석 결과

Dep. Variable:	소멸위험등급	R-squared (uncentered):	0.891
Model:	OLS	Adj. R-squared (uncentered):	0.889
Method:	Least Squares	F-statistic:	312.9 F-statistic은 모델의 설명력이 통계적으로 유의미한지 검정
Date:	Wed, 31 Jul 2024	Prob (F-statistic):	0.00 매우 낮은 p-value으로 모델이 유의미하다는 것을 강하게 시사
Time:	09:31:13	Log-Likelihood:	-2109.6
No. Observations:	1603	AIC:	4301.
Df Residuals:	1562	BIC:	4522.
Df Model:	41		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
고원_1인당_학생수_유치원	0.0260	0.021	1.247	0.213	-0.015	0.067
고원_1인당_학생수_초등학교	0.0399	0.043	0.935	0.350	-0.044	0.124
고원_1인당_학생수_중학교	0.1536	0.036	4.279	0.000	0.083	0.224
고원_1인당_학생수_고등학교	0.1153	0.030	3.901	0.000	0.057	0.173
유치원_학급당 학생 수 (명)	0.0423	0.014	2.993	0.003	0.015	0.070
초등학교_학급당 학생 수 (명)	-0.0796	0.031	-2.577	0.010	-0.140	-0.019
중학교_학급당 학생 수 (명)	-0.0776	0.019	-3.982	0.000	-0.116	-0.039
고등학교_학급당 학생 수 (명)	-0.0222	0.016	-1.431	0.153	-0.053	0.008
학교교과교습학원 (개)	-0.0005	0.000	-5.155	0.000	-0.001	-0.000
평생직업교육학원 (개)	0.0035	0.001	3.498	0.000	0.002	0.006
사설학원당 학생수 (명)	-0.0001	0.000	-0.744	0.457	-0.001	0.000
유치원생 수	-3.901e-06	3e-05	-0.130	0.897	-6.28e-05	5.5e-05
초등학생수	9.568e-07	9.49e-06	0.101	0.920	-1.76e-05	1.96e-05
종합병원	-0.0689	0.025	-2.783	0.005	-0.117	-0.020
병원	-0.0387	0.008	-4.661	0.000	-0.055	-0.022
의원	-0.0005	0.001	-1.018	0.309	-0.002	0.000
치과병(의)원	0.0011	0.001	0.774	0.439	-0.002	0.004
한방병원	0.0119	0.003	3.486	0.001	0.005	0.019
한의원	0.0035	0.002	1.951	0.051	-1.87e-05	0.007
인구 천명당 의료기관병상수(개)	-0.0038	0.003	-1.158	0.247	-0.010	0.003
총병상수 (개)	5.532e-05	2.14e-05	2.584	0.010	1.33e-05	9.73e-05
고위험음주율	0.0092	0.009	1.071	0.284	-0.008	0.026
비만율	0.0019	0.008	0.233	0.816	-0.014	0.018
EQ_5D(건강상태 표준화)	-2.1779	0.706	-3.086	0.002	-3.562	-0.794
주관적건강수준인지율	0.0162	0.004	4.558	0.000	0.009	0.023
건강보험 적용인구 현황	-2.363e-05	4.17e-06	-5.667	0.000	-3.18e-05	-1.54e-05
문화관광형태영역	0.0040	0.003	1.145	0.252	-0.003	0.011
교통안전영역	-0.0025	0.004	-0.559	0.576	-0.011	0.006
보행행태영역	-0.0133	0.012	-1.079	0.281	-0.037	0.011
1인당 자동차등록대수	-0.7825	0.181	-4.315	0.000	-1.138	-0.427
도시지역면적	-6.21e-10	3.63e-10	-1.709	0.088	-1.33e-09	9.19e-11
주택 수	-1.651e-05	3.15e-06	-5.239	0.000	-2.27e-05	-1.03e-05
출생아수	-9.623e-05	7.13e-05	-1.349	0.177	-0.000	4.37e-05
합계출산율	-0.5015	0.123	-4.086	0.000	-0.742	-0.261
남녀성비	0.0529	0.006	9.562	0.000	0.042	0.064
인구증가율	0.0503	0.011	4.543	0.000	0.029	0.072
주민등록인구	3.284e-05	4.09e-06	8.034	0.000	2.48e-05	4.09e-05
가구수	-9.34e-06	3.48e-06	-2.686	0.007	-1.62e-05	-2.52e-06
일반공공 공행정예산비중	-0.0052	0.007	-0.772	0.440	-0.019	0.008
하수도보급률	0.0056	0.002	2.537	0.011	0.001	0.010
상수도보급률	-0.0128	0.003	-4.474	0.000	-0.018	-0.007

Omnibus:	34.036	Durbin-Watson:	1. 자기상관을 측정 값이 2에 가까울수록 자기상관이 없음을 의미
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21.556
Skew:	0.140	Prob(JB):	2.09e-05
Kurtosis:	2.506	Cond. No. 다중공선성	3.72e+09

Unscaled

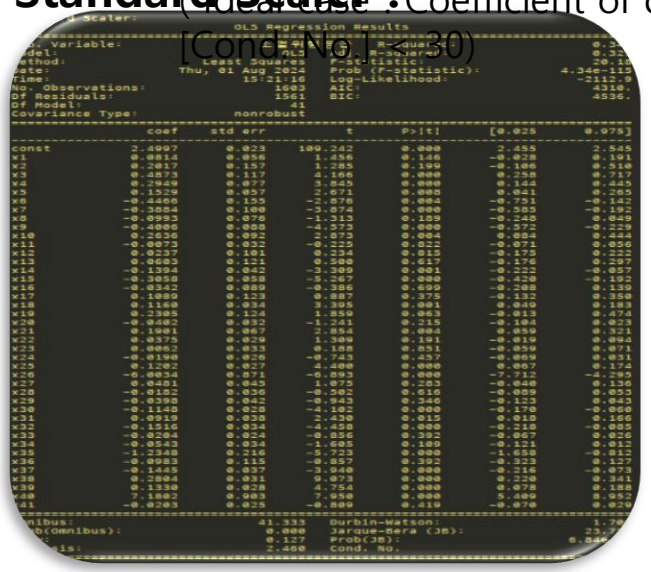
Coefficient of determinati

on : 0.805

Multiple regression analysis results (1) [Independent variable : cofog variable / Dependent variable : fat extinction index]

(Ideal case : Coefficient of determination [R-squared]: explanatory power of independent variables (close to 1) / Multicollinearity [Cond. No.] < 30)

Standard Scaler :



Coefficient of determination : 0.346
Multicollinearity : 212

Robust Scaler :

Coefficient of determination : 0.346
Multicollinearity : 385

Unscaled :

Coefficient of determination : 0.891
Multicollinearity : 3,720,000,000

Minmax Scaler :

Coefficient of determination : 0.346
Multicollinearity : 805

Results Interpretation : Although multicollinearity was partially resolved, we decided to apply

Data preprocessing (Multicollinearity)

① In regression models, logistic models, etc., variables with high correlation between variables

(VIF value If the number is 10 or more, **consider multicollinearity and appropriately** exclude

② Classification models utilizing machine learning have the potential to increase explanatory power by reflecting variables with high multicollinearity . **In this project, we also considered multicollinearity, but decided to use it in a way that reflects the variables presented above while comparing the model performance as much as possible.**

Multiple regression analysis results (2) - Normality test , VIF test , QQ plot

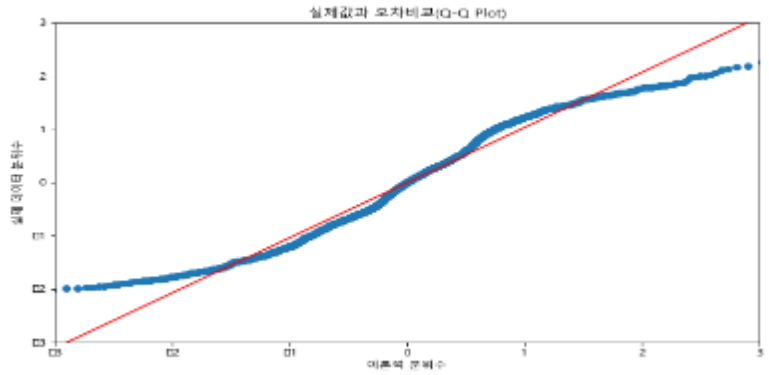
Variables affecting the increase in the level of fat loss (41 variables)

Based on the regression a nalysis results ,
VIF test was performed on variables with a P-value of 0.05 or less and a regressi on coefficient of > 0.

The VIF test is a measure of th e correlation between indepen dent variables .
measure

지방소멸 위험도를 낮추어 주는 변수	회귀계수	유의확률
교원_1인당_학생수_중학교	0.153609	1.99112e-05
교원_1인당_학생수_고등학교	0.115309	9.98662e-05
유치원_학급당 학생 수 (명)	0.0423004	0.00280763
평생직업 교육학원 (개)	0.00353915	0.00048198
한방병원	0.0119425	0.000504231
총병상수 (개)	5.53224e-05	0.00986624
주관적건강수준인지율	0.0162483	5.57823e-06
남녀성비	0.0529228	4.31532e-21
인구증가율	0.0503201	5.97605e-06
주민등록인구	3.28428e-05	1.84577e-15
하수도보급률	0.00558597	0.0112909

11 variables affecting the increase in the level of local extinction



Results : Based on the results of multiple regression analysis, the VIF test was performed on 11 variables that affect the dependent variable, the fat extinction grade. As a result, multicollinearity was alleviated, but the err or was still large, so the conclusion was to apply all variables to the classification model .

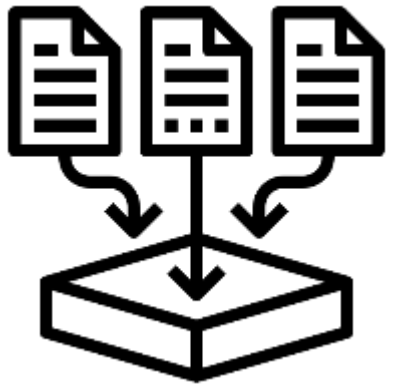
III. Development Process - 3. DBMS Construction - Procedure

Step 1 : Preprocessing

Uniform row / column items

Step 2: Loading data

DBMS construction



Step 3: Data Utilization

Enter SQL

III. Development Process - 3. DBMS Construction - Table Creation

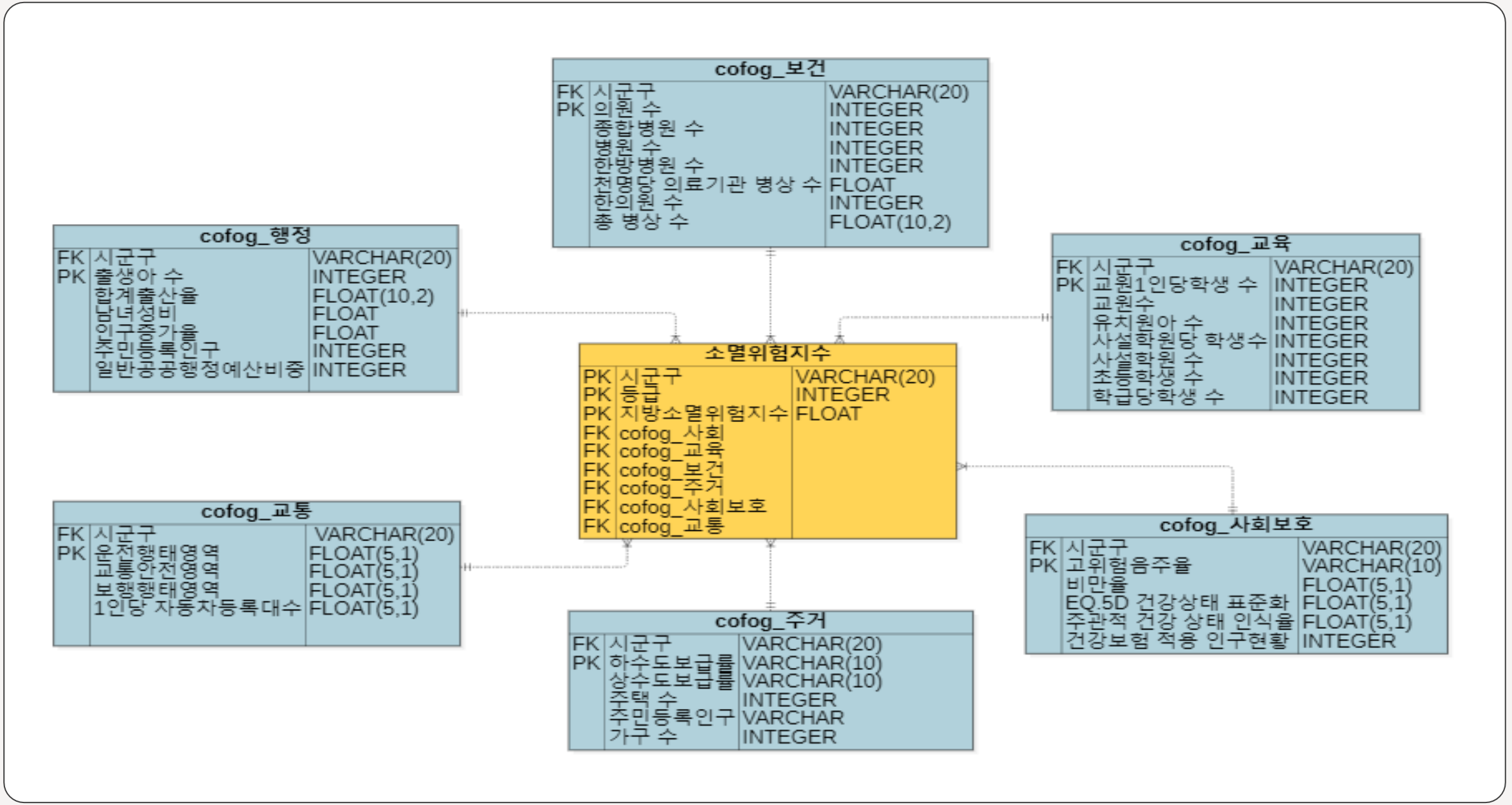

```
-- 보건 테이블 생성
-- 테이블 생성
CREATE TABLE cofog_health (
    sigun_gu VARCHAR(20),
    general_hospital INT(10) NOT NULL,
    hospital INT(10) NOT NULL,
    clinic INT(10) NOT NULL,
    dental_clinic INT(10) NOT NULL,
    medicine_hospital INT(10) NOT NULL,
    koreanmedical_clinic int(10) not null,
    beds_per_thousandpeople FLOAT(5,1) NOT NULL,  -- 소수점 자리 포함
    total_beds FLOAT(10,1) NOT NULL  -- 총 침대 수, FLOAT으로 수정
);

-- CSV 데이터 로드
LOAD DATA INFILE 'C:/MySQL/8.4/Data/Uploads/extinct/health.csv'
INTO TABLE cofog_health
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';
```

```
use extinction;
CREATE TABLE extinct (
    sigun_gu VARCHAR(20),          -- 지역 이름
    extinction_index FLOAT(25) NOT NULL,  -- 소멸 지수
    extinction_grade VARCHAR(3) NOT NULL  -- 소멸 등급
);
show variables like 'secure_file_priv';
load data infile 'C:/MySQL/8.4/Data/Uploads/extinct/merged.csv' into table extinct fields terminated by ',';
```

```
-- 교육테이블 생성
CREATE TABLE cofog_education (
    sigun_gu_edu VARCHAR(20),          -- 지역 이름
    student_per_teacher_kin INT(10) NOT NULL,  -- 교사 당 유치원 학생 수
    teacher_per_student_pri INT(10) NOT NULL,  -- 교사 당 초등학생 수
    teacher_per_student_mid INT(10) NOT NULL,  -- 교사 당 중학생 수
    teacher_per_student_high INT(10) NOT NULL, -- 교사 당 고등학생 수
    student_per_class_kin INT(10) NOT NULL,    -- 학급 당 유치원 학생 수
    student_per_class_pri INT(10) NOT NULL,    -- 학급 당 초등학생 수
    student_per_class_mid INT(10) NOT NULL,    -- 학급 당 중학생 수
    student_per_class_high INT(10) NOT NULL,   -- 학급 당 고등학생 수
    extra_curr_school INT(10) NOT NULL,        -- 학교 부속 특별 활동 프로그램 수
    extra_curr_lifelong INT(10),               -- 평생 교육 관련 프로그램 수
    student_per_extra_curr INT(10) NOT NULL,   -- 사설학원 당 학생 수
    kin_stu INT(10) NOT NULL,                  -- 유치원 학생 수
    pri_stu INT(10) NOT NULL                   -- 초등학생 수
);

-- CSV 데이터 로드
LOAD DATA INFILE 'C:/MySQL/8.4/Data/Uploads/extinct/education.csv'
INTO TABLE education
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;  -- 첫 번째 행이 헤더라면 무시
```

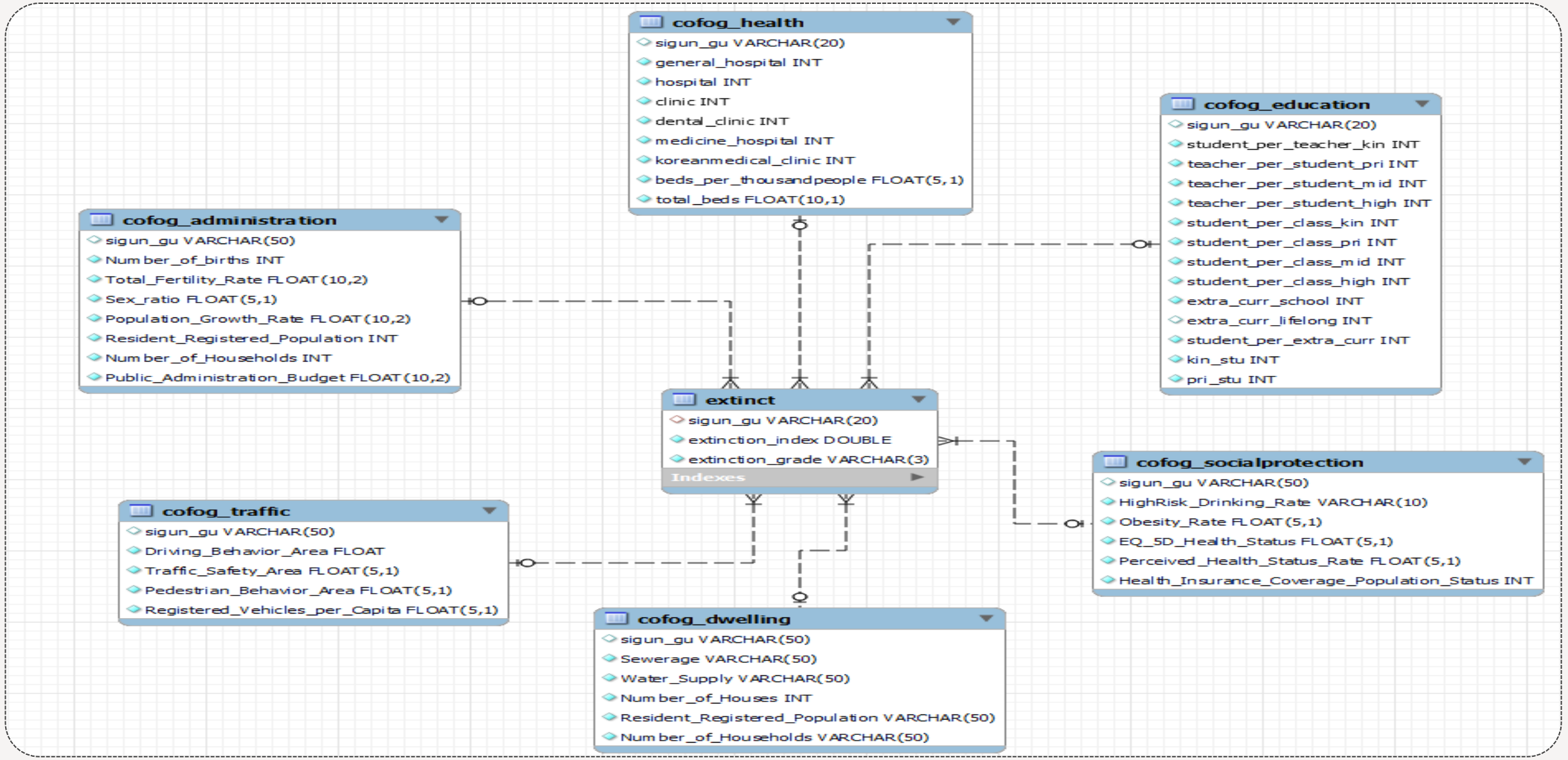
Source : MySQL by Shin Jeong-yoon

III. Development Process - 3. DBMS Construction - Learning Relational Data Structures (1)



Source : StarUML (ERD) by Shin Jeon

III. Development Process - 3. DBMS Construction - Learning Relational Data Structures (2)



Source : MySQL (ERD) by Shin

III. Development Process - 3. DBMS Construction - Extracting Specific Datasets


```sql
-- coef>0.05 & P-value < 0.05 인 변수 11개 리스트 DBMS에서 추출:
CREATE VIEW integrated_data AS
SELECT
    e.sigun_gu,
    e.teacher_per_student_mid,
    e.teacher_per_student_high,
    e.student_per_class_kin,
    e.extra_curr_lifelong,
    h.medicine_hospital,
    h.total_beds,
    s.Perceived_Health_Status_Rate,
    a.Sex_Ratio,
    a.Population_Growth_Rate,
    a.Resident_Registered_Population,
    d.Sewerage,
    ext.extinction_point
FROM
    cofog_education e
LEFT JOIN
    extinct ext ON e.sigun_gu = ext.sigun_gu
LEFT JOIN
    cofog_health h ON e.sigun_gu = h.sigun_gu
LEFT JOIN
    cofog_socialprotection s ON e.sigun_gu = s.sigun_gu
LEFT JOIN
    cofog_administration a ON e.sigun_gu = a.sigun_gu
LEFT JOIN
    cofog_dwelling d ON e.sigun_gu = d.sigun_gu
WHERE
    e.sigun_gu LIKE '%2015%'
    OR e.sigun_gu LIKE '%2016%'
    OR e.sigun_gu LIKE '%2017%'
    OR e.sigun_gu LIKE '%2018%'
    OR e.sigun_gu LIKE '%2019%'
    OR e.sigun_gu LIKE '%2020%'
    OR e.sigun_gu LIKE '%2021%';
```

(Method 1) Extracting MySQL dataset

You can extract data files in CSV or EXCEL format based on various combinations of query statements

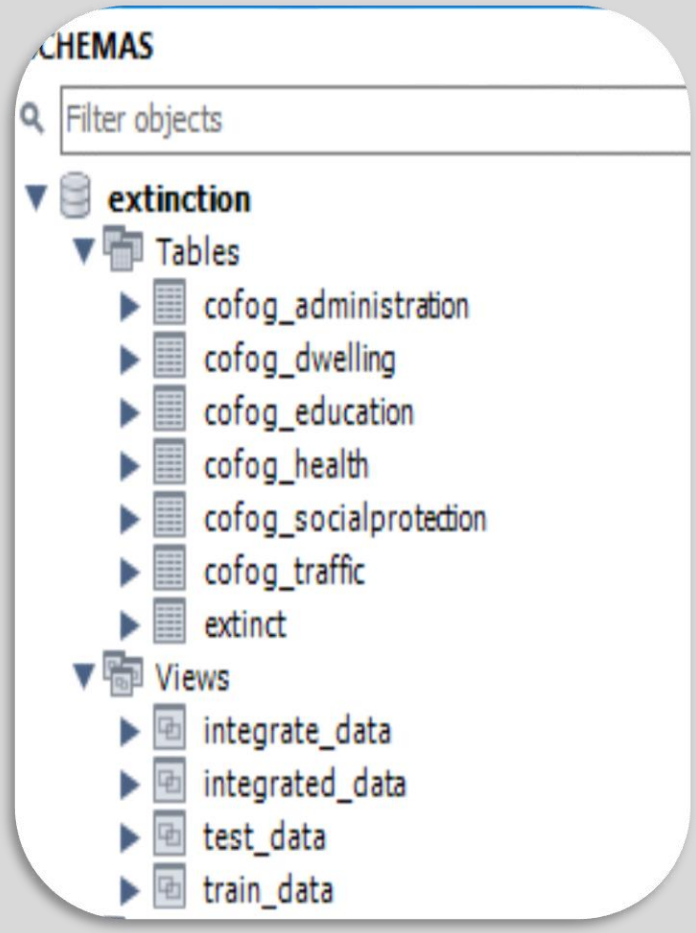
(Method 2) pymysql Library → Extract dataset

After obtaining the author's authority from the DB built in Python You can read data into a dataframe in Python by writing a query statement .

III. Development Process - 3. DBMS Construction - Utilization Plan

DBMS utilization plan

Defunct DBMS



example)

sigun_gu	teacher_per_student_	teacher_per_student_	student_per_class_ki	extra_curr_lifelong
세종특별자치시_2015	8	9	16	7
경기도 수원시_2015	15	14	22	158
경기도 성남시_2015	14	12	22	103
경기도 안양시_2015	14	12	21	63
경기도 부천시_2015	15	11	19	100

Main DB
(IP Address)

Main dish | Medical t | Woojin | Jeong Yo | Seungpil

the IP address of

III. Development Process - 3. DBMS Construction - Specific Case

Data Generation (By MySQL)

```
-- coef>0.05 & P-value < 0.05 인 변수 11개 리스트 DBMS에서 추출:
CREATE VIEW integrated_data AS
SELECT
    e.sigun_gu,
    e.teacher_per_student_mid,
    e.teacher_per_student_high,
    e.student_per_class_kin,
    e.extra_curr_lifelong,
    h.medicine_hospital,
    h.total_beds,
    s.Perceived_Health_Status_Rate,
    a.Sex_Ratio,
    a.Population_Growth_Rate,
    a.Resident_Registered_Population,
    d.Sewerage,
    ext.extinction_point
FROM
    cofog_education e
LEFT JOIN
    extinct ext ON e.sigun_gu = ext.sigun_gu
LEFT JOIN
    cofog_health h ON e.sigun_gu = h.sigun_gu
LEFT JOIN
    cofog_socialprotection s ON e.sigun_gu = s.sigun_gu
LEFT JOIN
    cofog_administration a ON e.sigun_gu = a.sigun_gu
LEFT JOIN
    cofog_dwelling d ON e.sigun_gu = d.sigun_gu
WHERE
    e.sigun_gu LIKE '%2015%'
    OR e.sigun_gu LIKE '%2016%'
    OR e.sigun_gu LIKE '%2017%'
    OR e.sigun_gu LIKE '%2018%'
    OR e.sigun_gu LIKE '%2019%'
    OR e.sigun_gu LIKE '%2020%'
    OR e.sigun_gu LIKE '%2021%';
```

Importing Data (By MySQL / Python)

```
import pandas as pd
from sqlalchemy import create_engine

#%%

# 데이터베이스 연결 URL 형식
# DATABASE_URL = 'mysql+pymysql://username:password@localhost:3306/database'
DATABASE_URL = 'mysql+pymysql://shin:1234@192.168.71.233:3306/extinction'
engine = create_engine(DATABASE_URL)

#%%

# SQL 쿼리 정의
sql = "SELECT * FROM extinction.integrated_data"

data = pd.read_sql_query(sql, engine)

print(data)
```

Sharing Data with Team Members (By Python)

Index	sigun_gu	_per_stude	_per_stude	nt_per_clas	a_curr_lifel	dicine_hosp	total_beds	_Health_St	Sex_Ratio	tion_Growt	Registered_P	Sewerage	inction_po
0	세종특별자치시_2015	8	9	16	7	0	1229	43.1	100.6	34.94	210884	89.5	4
1	경기도 수원시_2015	15	14	22	158	2	10526	43.8	101.4	1.06	1184624	99.0	4
2	경기도 성남시_2015	14	12	22	103	2	9131	50.1	98.8	-0.26	971424	99.5	4
3	경기도 안양시_2015	14	12	21	63	1	4785	45.6	99	-0.47	597789	100.0	4
4	경기도 부천시_2015	15	11	19	100	7	11868	45.5	99.9	-0.64	848987	100.0	4
5	경기도 광명시_2015	14	12	22	15	0	2076	46.8	98.5	-0.88	344978	98.5	4
6	경기도 평택시_2015	15	14	16	35	0	4171	41.5	103.9	2.67	460532	85.3	4
7	경기도 안산시_2015	16	13	20	80	10	9492	37.9	105.1	-1.05	697885	98.9	4
8	경기도 과천시_2015	16	12	19	1	0	18	52	94.5	-1.74	68946	98.5	4
9	경기도 오산시_2015	16	14	21	8	0	1987	46.1	105.1	-0.42	206828	98.9	4

III. Development Process – cofog variable

(Time) Total 2015-2021 / (Category) **cofog variable** / (Index) Improvement extinction risk index

Index	소멸위험지수	1당_확생수	당_확생수_	1당_확생수	당_확생수_	확급당 확산	확급당 확-	확급당 확산	확급당 확-	과 교습확	인 교육확	원당 확생	유시원생 수	초등확생 수	종합병원	병원	의원	내과병(의)원	한방병원	한의원	3 의료기관	생활수 (기	1수도보급률	1수도보급률
0	0.623911	11	12	8	9	16	18	19	15	162	7	36	1356	3570	4	5	5	104	0	52	14	3013	99	100
1	0.1875	14	17	15	14	22	26	30	32	1800	158	73	20400	69880	0	0	0	8	0	5	0.3	8	99.5	99.9
2	0.708359	13	17	14	12	22	25	30	29	1628	103	55	11510	48568	1	2	2	50	0	22	16.8	1578	100	100
3	0.230535	13	17	14	12	21	25	31	31	1045	63	55	7555	31499	1	1	1	26	0	13	6	424	100	100
4	0.842704	14	17	15	11	19	25	30	27	1167	100	75	11582	42474	2	0	0	58	0	26	12	987	98.5	100
5	0.585238	15	18	14	12	22	27	30	29	584	15	79	4592	20710	0	3	5	6	0	5	8.2	197	85.5	97.8
6	0.203023	13	16	15	14	16	24	30	32	612	35	70	7631	27812	0	0	0	6	0	5	1.1	30	98.9	99.8
7	0.254017	14	17	16	13	20	25	30	30	982	80	84	9801	40017	1	1	1	12	0	6	9.7	390	98.5	99.7
8	1.7154	14	17	16	12	19	25	33	27	69	1	115	684	3665	2	15	15	179	3	95	14.5	4817	98.9	100
9	0.639372	15	17	16	14	21	26	32	34	240	8	112	6259	15598	0	1	1	7	0	5	5.7	191	94.5	99.9
10	0.221961	15	17	15	15	20	25	29	33	545	29	94	5200	24094	0	2	2	18	0	5	8.8	343	99.8	100
11	0.589004	15	16	15	14	23	26	31	32	371	29	86	4396	16082	0	1	1	18	0	7	9.1	445	99.3	99.9
12	1.09943	13	16	15	12	23	24	29	27	177	4	78	2020	7842	2	7	7	140	3	78	15.4	4272	88	98
13	0.938585	15	17	15	11	20	25	27	26	162	5	90	2186	7861	1	0	0	25	0	10	14.3	680	92.9	98.5
14	0.260275	15	18	16	14	21	26	32	33	1645	45	77	18159	67703	0	0	0	15	0	9	4.4	190	90	94
15	0.32034	13	15	14	12	17	23	28	29	336	27	73	2718	12787	1	2	2	30	0	15	11.9	840	77.5	91.2
16	0.357075	12	13	13	14	16	20	27	29	1322	15	13	2784	10861	0	0	0	7	0	4	1.9	50	85.9	92.7
17	0.322423	15	17	15	13	19	24	29	28	493	21	82	8285	22951	0	2	2	12	0	7	5.4	240	82.9	96.4
18	0.425847	16	17	16	16	20	25	32	36	910	29	81	13381	44842	0	4	21	16	2	15	10.7	1222	94.3	88.9
19	2.60021	14	19	16	14	21	27	32	31	359	9	81	2838	10940	5	21	520	303	2	255	12.5	12790	79.9	85.6
20	1.93064	13	12	13	11	17	18	27	23	196	12	47	1471	5699	0	0	33	18	0	13	0.3	18	80.3	65.6
21	2.52599	11	12	11	9	14	19	24	24	78	2	89	680	4983	1	3	191	108	0	79	6	2076	98.7	90.6
22	2.61546	13	16	15	13	19	25	30	32	620	41	73	5473	22595	0	3	97	60	0	56	4.6	1424	97.8	99.3
23	2.67348	12	16	14	12	19	23	27	30	92	8	77	938	5358	1	9	120	79	0	56	13	2427	91	99.6
24	3.8975	14	18	15	14	22	26	31	32	1772	99	66	14703	56342	2	3	154	86	0	59	8.7	2507	100	99.9
25	2.38472	14	17	16	13	23	25	32	32	309	13	72	2533	10049	2	4	150	101	2	75	8.4	2927	96.2	100
26	2.36217	14	18	15	13	19	25	30	31	783	26	94	10604	41703	3	11	273	166	0	125	8.5	5570	93.2	98.4
27	0.841248	14	16	16	14	19	24	31	31	428	24	107	7711	27217	0	1	58	25	0	16	26	2549	91.1	97.4
28	2.93332	12	16	16	15	16	23	32	34	199	4	120	2924	13265	5	22	481	269	7	204	14	11068	63.6	75.9
29	2.41735	12	13	12	12	14	19	25	26	128	8	120	1427	7757	6	16	670	423	2	312	9.4	9131	85.5	94.3

RandomizedSearchCV Finding optimal settings through m

```
random_search = RandomizedSearchCV(model, param_distributions=param_dist, n_iter=100, cv=5,
```

Number of searches : A total

There are several hyperparameters that have a significant impact on the performance of classification models (GBM/LGBM) .

So why **RandomizedSearchCV Is it ?**

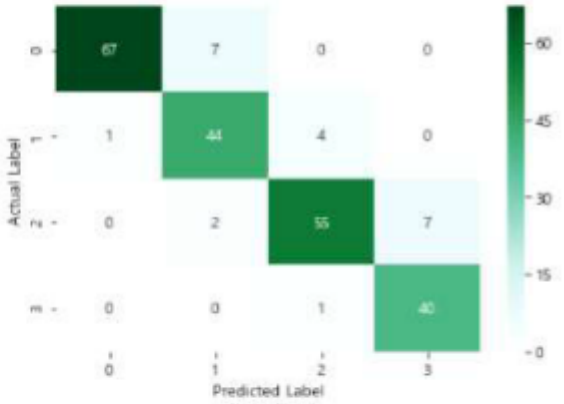
1. Reduce computational costs
Since only a few combinations are randomly tried, the computational cost can be greatly reduced .

2. Improving model performance
more accurate predictions , lower error rates , and better generalization performance .

3. Preventing overfitting
Evaluate model hyperparameters via K- fold cross-validation (different combinations of train / test sets)

Algorithm 1: Gradient Boosting Classifier Model (existing)

〈표 3〉 Gradient Boosting Classifier 결과

RandomizedSearchCV 파라미터 최적화를 이용한 stratifed K-fold 결과	모델을 통해 시험자료를 예측하고 F1 score 측정한 결과
	F1 Cross_validate 0.8915582178411123 F1 Macro: 0.8905957255158861 F1 Micro: 0.8947368421052632 F1 Weighted: 0.8956295791887187
{'subsample': 0.95, 'n_estimators': 608, 'min_samples_split': 0.075, 'min_samples_leaf': 60, 'max_features': 'sqrt', 'max_depth': 15, 'loss': 'deviance', 'learning_rate': 0.025, 'criterion': 'friedman_mse'}	
최적화된 파라미터를 반영한 GBC모델	voting 결과
GradientBoostingClassifier(learning_rate=0.2, max_depth=30, max_features='log2', min_samples_leaf=15, min_samples_split=0.1, n_estimators=773)	Validation score of a single GBC Classifier: 0.8828 Validation score of a VotingClassifier on 3 GBC with hard voting strategy: 0.8871 Validation score of a VotingClassifier on 3 GBC with soft voting strategy: 0.8871

(example)

F1 Score

Improving F1 Score by Modifying Paramete

(Usage tool)

(Source) p21, A Study on the Factors of Local Government Extinction in Korea and Solutions to It : Exploration thro

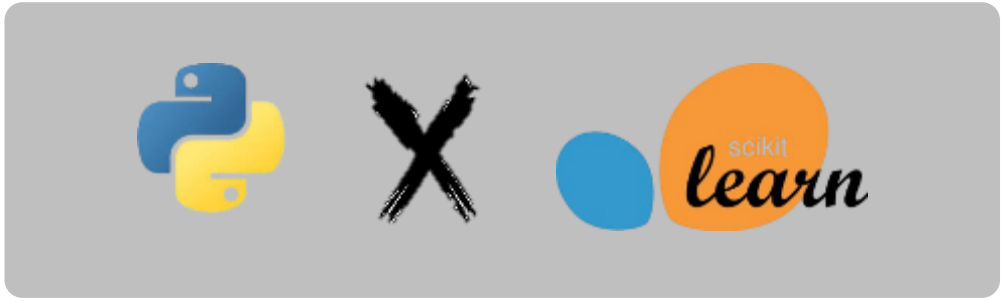
Algorithm 2: Light Gradient Boosting Classifier Model (new)

```python
#%%
import lightgbm as lgb
from sklearn.metrics import accuracy_score

# 최적의 하이퍼파라미터를 사용하여 LightGBM 모델 초기화
best_params = rand_search_lgb.best_params_
lgb_clf_best = lgb.LGBMClassifier(
    num_leaves=best_params['num_leaves'],
    learning_rate=best_params['learning_rate'],
    n_estimators=best_params['n_estimators'],
    max_depth=best_params['max_depth'],
    min_child_samples=best_params['min_child_samples'],
    subsample=best_params['subsample'],
    colsample_bytree=best_params['colsample_bytree'],
    random_state=1
)

# 최적의 파라미터로 모델 학습
lgb_clf_best.fit(X_train, y_train.values.ravel())

# 테스트 데이터에서 예측 수행
y_pred_lgb = lgb_clf_best.predict(X_test)

# 정확도 평가
accuracy = accuracy_score(y_test, y_pred_lgb)
print(f"LightGBM Accuracy with optimized parameters: {accuracy:.4f}")

# 예측 결과의 혼동 행렬과 분류 리포트 출력 (선택 사항)
from sklearn.metrics import confusion_matrix, classification_report

conf_matrix = confusion_matrix(y_test, y_pred_lgb)
print("Confusion Matrix:")
print(conf_matrix)

class_report = classification_report(y_test, y_pred_lgb)
print("Classification Report:")
print(class_report)
```

(example)

F1 Score

provement of error rate through parameter mod
and improve classification accuracy .



(Usage tool)

machine learning Python code

```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, RandomizedSearchCV
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import mean_squared_error

base_dir = "your_link"

desired_path = "your_link"

df_learn = pd.read_excel(desired_path)

# 소멸위험등급을 수치형으로 변환
grade_mapping = {'A': 1, 'B': 2, 'C': 3, 'D': 4}
df_learn['소멸위험등급'] = df_learn['소멸위험등급'].map(grade_mapping)

df_learn.replace('-', 0, inplace=True)

# 특징 변수(X)와 타겟 변수(y) 분리
X = df_learn[['소멸위험지수', '교원_1인당_학생수_유치원', '교원_1인당_학생수_초등학교', '교원_1인당_학생수_중학교', '교원_1인당_학생수_고등학교', '유치원_학급당 학
y = df_learn['소멸위험등급']

# 데이터셋 분할
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Gradient Boosting Regressor 모델 초기화
model = GradientBoostingClassifier()

# 하이퍼파라미터 범위 설정
param_dist = {
    'n_estimators': np.arange(50, 500, 50),
    'learning_rate': np.linspace(0.01, 0.2, 20),
    'max_depth': np.arange(3, 10, 1),
    'min_samples_split': np.arange(2, 10, 1),
    'min_samples_leaf': np.arange(1, 10, 1),
    'subsample': np.linspace(0.5, 1.0, 6)
}

# RandomizedSearchCV 설정
random_search = RandomizedSearchCV(model, param_distributions=param_dist, n_iter=100, cv=5, scoring='neg_mean_squared_error', random_state=42, n_jobs=-1, e

# 모델 학습
random_search.fit(X_train, y_train)

# 최적의 하이퍼파라미터 출력
print("Best hyperparameters:", random_search.best_params_)

# 최적의 모델로 예측 및 평가
best_model = random_search.best_estimator_
y_pred = best_model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)
```

Model

GradientBoostingClassifierModel

Optimal parameter values

Using the RandomizedSerchCV mo

Apply optimal parameter values to GBM

III. Development Process – Optimal Hyperparameters Settings

Predicted value Actual value

y_pred - NumPy object array

	0
0	4
1	2
2	4
3	4
4	1
5	4
6	1
7	4
8	4
9	1
10	3
11	1
12	3
13	3

y_test - Series

Index	ᄂ멸위험등ᄀ
528	4
1145	2
168	4
135	4
1330	1
940	4
1132	1
237	4
481	4
1034	1
561	3
543	1
374	3
1572	3

Optimal parameter values

Number of Trees	learning rate	node split sample
100	0.05	5

leaf node	Regression estimators	Error rate
(min) 9	8	0.0

```
Best hyperparameters: {'subsample': 0.9, 'n_estimators': 100, 'min_samples_split': 5,
'min_samples_leaf': 9, 'max_depth': 8, 'learning_rate': 0.05}
Mean Squared Error: 0.0
```

IV. Conclusion and Future Directions

IV. Conclusion – Reattempting Scaling by Variable

1) Standard Scaler

```
Test Data 예측 정확도: 0.8343
Confusion Matrix:
[[83  5  0  0]
 [ 7 65 10  0]
 [ 0  9 57 14]
 [ 0  1 11 82]]

In [16]: runcell(6, 'D:/Workspace/Local_Extinction/codes/
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.94      0.93        88
           1       0.81      0.79      0.80        82
           2       0.73      0.71      0.72        80
           3       0.85      0.87      0.86        94

    accuracy                           0.83       344
   macro avg       0.83      0.83      0.83       344
weighted avg       0.83      0.83      0.83       344


In [17]: runcell(7, 'D:/Workspace/Local_Extinction/codes/
2021년 Data 예측 정확도: 0.8696
```

2) Robust Scaler

```
Test Data 예측 정확도: 0.8401
Confusion Matrix:
[[81  7  0  0]
 [ 8 66  8  0]
 [ 0  8 60 12]
 [ 0  1 11 82]]

In [25]: runcell(6, 'D:/Workspace/Local_Extinction/codes/
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.92      0.92        88
           1       0.80      0.80      0.80        82
           2       0.76      0.75      0.75        80
           3       0.87      0.87      0.87        94

    accuracy                           0.84       344
   macro avg       0.84      0.84      0.84       344
weighted avg       0.84      0.84      0.84       344


In [26]: runcell(7, 'D:/Workspace/Local_Extinction/codes/
2021년 Data 예측 정확도: 0.8609
```

3) Min-Max Scaler

```
Test Data 예측 정확도: 0.8517
Confusion Matrix:
[[83  5  0  0]
 [ 6 69  7  0]
 [ 0  8 58 14]
 [ 0  0 11 83]]

In [13]: runcell(6, 'D:/Workspace/Local_Extinction/codes/
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.94      0.94        88
           1       0.84      0.84      0.84        82
           2       0.76      0.72      0.74        80
           3       0.86      0.88      0.87        94

    accuracy                           0.85       344
   macro avg       0.85      0.85      0.85       344
weighted avg       0.85      0.85      0.85       344


In [14]: runcell(7, 'D:/Workspace/Local_Extinction/codes/
2021년 Data 예측 정확도: 0.8565
```

4) Unscaled classification accuracy is the highest.

```
Test Data 예측 정확도: 0.8547
Confusion Matrix:
[[83  5  0  0]
 [ 4 70  7  1]
 [ 0  6 59 15]
 [ 0  0 12 82]]

In [12]: runcell(5, 'D:/Workspace/Local_Extinction/codes/
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.94      0.95        88
           1       0.86      0.85      0.86        82
           2       0.76      0.74      0.75        80
           3       0.84      0.87      0.85        94

    accuracy                           0.85       344
   macro avg       0.85      0.85      0.85       344
weighted avg       0.85      0.85      0.85       344


In [13]: runcell(6, 'D:/Workspace/Local_Extinction/codes/
2021년 Data 예측 정확도: 0.8652
```

IV. Conclusion – LGBM – Confusion Matrix, Classification Report

(Parameter Optimization) LightGBM Accuracy : 0.8652

optimized parameters to the LGBM model

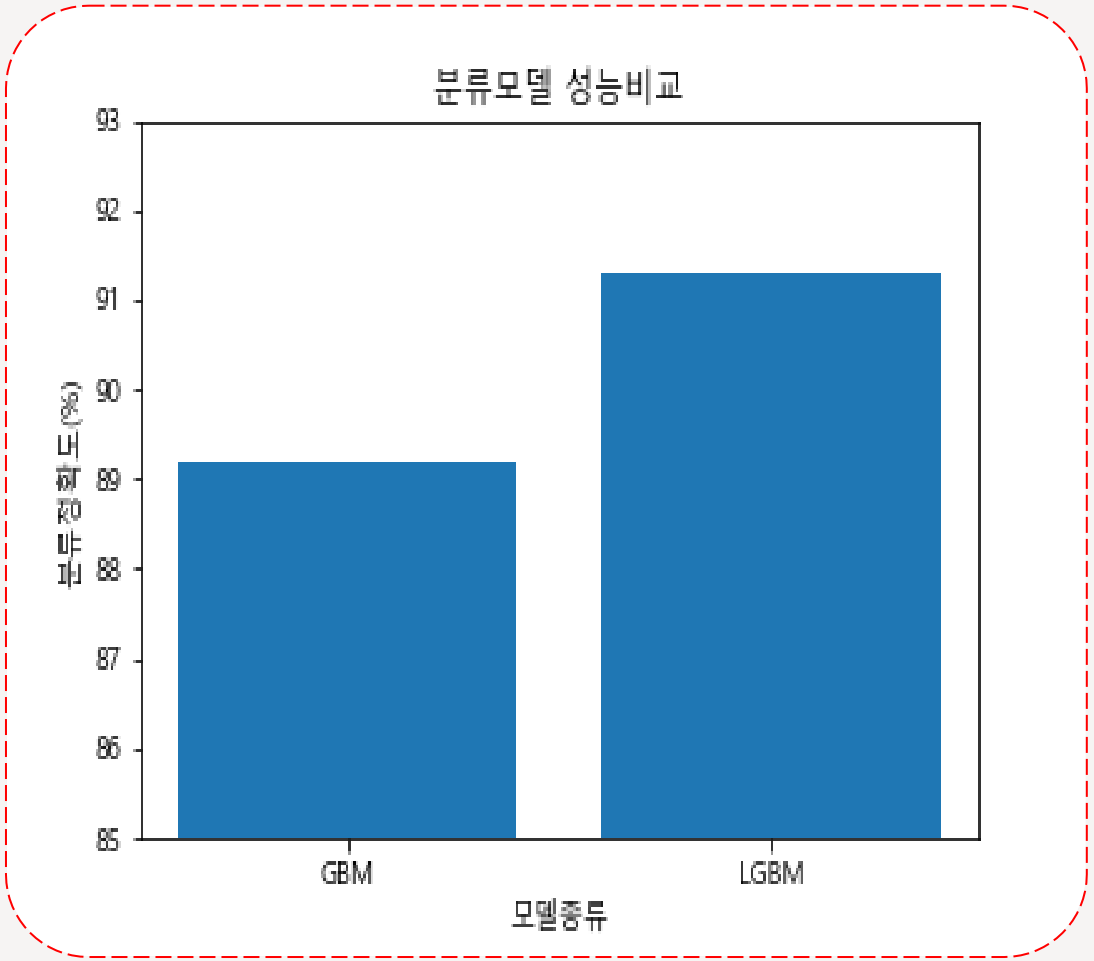
```
#최
import lightgbm as lgb
from sklearn.metrics import accuracy_score

# 최적의 하이퍼파라미터를 사용하여 LightGBM 모델 초기화
best_params = rand_search_lgb.best_params_
lgb_clf_best = lgb.LGBMClassifier(
    num_leaves=best_params['num_leaves'],
    learning_rate=best_params['learning_rate'],
    n_estimators=best_params['n_estimators'],
    max_depth=best_params['max_depth'],
    min_child_samples=best_params['min_child_samples'],
    subsample=best_params['subsample'],
    colsample_bytree=best_params['colsample_bytree'],
    random_state=1
)

# 최적의 파라미터로 모델 학습
lgb_clf_best.fit(X_train, y_train.values.ravel())

# 테스트 데이터에서 예측 수행
y_pred_lgb = lgb_clf_best.predict(X_test)

# 정확도 평가
accuracy = accuracy_score(y_test, y_pred_lgb)
print(f"LightGBM Accuracy with optimized parameters: {accuracy:.4f}")

# 예측 결과의 혼동 행렬과 분류 리포트 출력 (선택 사항)
from sklearn.metrics import confusion_matrix, classification_report

conf_matrix = confusion_matrix(y_test, y_pred_lgb)
print("Confusion Matrix:")
print(conf_matrix)

class_report = classification_report(y_test, y_pred_lgb)
print("Classification Report:")
print(class_report)
```

Applying

```
Test Data 예측 정확도: 0.8547
Confusion Matrix:
[[83  5  0  0]
 [ 4 70  7  1]
 [ 0  6 59 15]
 [ 0  0 12 82]]

In [12]: runcell(5, 'D:/Workspace/Local_Extinction/codes/
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.94      0.95        88
           1       0.86      0.85      0.86        82
           2       0.76      0.74      0.75        80
           3       0.84      0.87      0.85        94

    accuracy                           0.85       344
   macro avg       0.85      0.85      0.85       344
weighted avg       0.85      0.85      0.85       344


In [13]: runcell(6, 'D:/Workspace/Local_Extinction/codes/
2021년 Data 예측 정확도: 0.8652
```

IV. Conclusion – New Extinction Risk Index Grades - LGBM / GBM (2)

(Final result) LGBM / GBM



LGBM outperforms GBM (Classification accuracy) is better

Our team is , We improved the prediction accuracy by about 2% by using the LGBM model compared to the GBM used in the reference project !

(Final Results) Future Development Direction

In the future, our team will investigate the variables that affect fat loss.
We plan to **provide insights into policy directions** through **importance analysis .**










End of Page

- thank you -