

시군구에 따른 대기오염 분석

데이터 마이닝10조

20101340 김현석
20100593 박지원
20102020 서지훈

Contents

01. Intro

- 분석 배경
- 분석 목적
- 데이터 획득

02. EDA & 전처리

- 미세먼지
- 오존

02. 대기오염 분석

- 미세먼지 분석
 - 분석과정 (Linear regression, logistic regression, SVMs, Decision tree)
 - 분석결과
- 오존분석
 - 분석과정 (logistic regression, SVMs, Decision tree)
 - 분석결과

03. 분석 활용

- 기대효과
- 한계점 및 추후 개선점



Intro

겨울철 미세먼지주의보... 작아서 더 무서운 미세먼지, 혈관 타고 장기 곳곳 악영향

2021-11-30



머리카락 5분의 1 ~ 20분의 1 크기...코털 등
서 걸러지지 않아
염증반응 일으켜 조직 손상 가져오고 치매 위
험성까지 높아
먼지농도 상승하는 겨울철에는 KF94이상 마
스크 착용 권고



대기오염

대기오염은 인위적 발생원에서 배출된 물질이 생물이나 기물에 직접적으로 해를 끼칠 만큼 다량으로 대기 중에 존재하는 상태이다.

그중 대표적인 요소인 미세먼지와 오존의 발생 원인과 대기오염에 취약한 지역을 찾아보고자 한다.

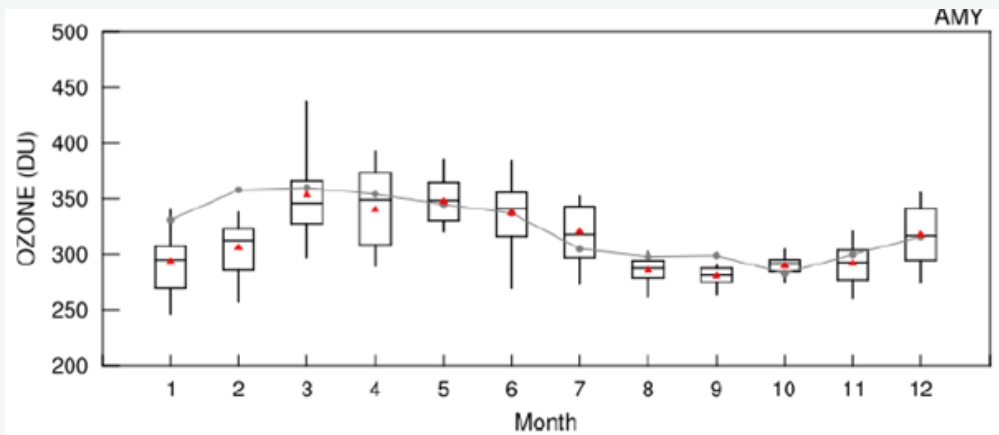
미세먼지

대기 중에 떠다니거나 흩날려 내려오는 입자상 물질로 그 크기가 $10\mu\text{m}$ 이하인 것을 미세먼지, $2.5\mu\text{m}$ 이하인 것을 초미세먼지라고 한다.

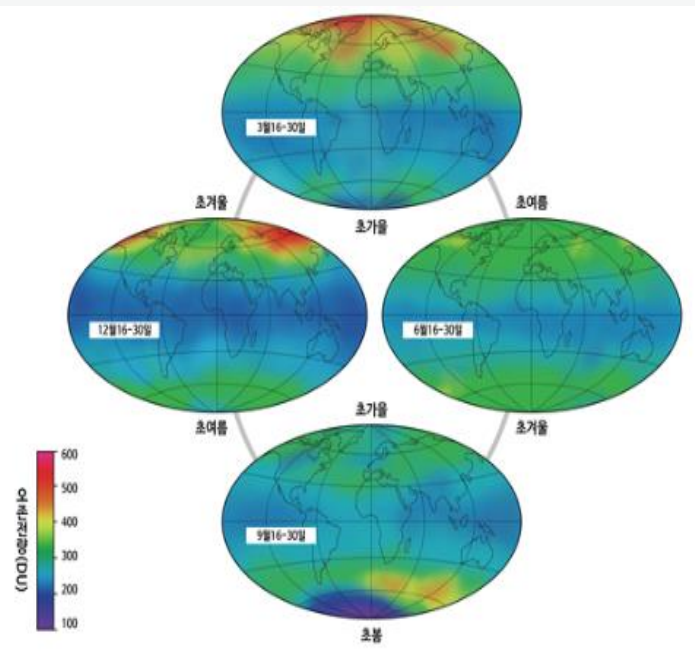
우리나라의 미세먼지는 30%~50%가 중국발 미세먼지이며 50%~70%는 국내 환경오염, 중금속-석유 석탄 등 화석연료 사용, 공장의 매연 등으로 발생한다.

미세먼지는 협심증, 뇌졸중, 심혈관질환 등을 유발하여 혈관건강을 위협하고, 후두염, 호흡곤란, 기관지염, 천식, 폐렴 등을 유발하여 호흡기의 건강을 위협한다. 그 외에 안구와 피부 등에도 영향을 주는 1급 발암물질이다.

분석배경



<안면도 기후변화감시소의 2019년 월평균 오존전량 분포>



대기오염

오존

산소 원자 3개로 이루어진 산소의 동소체로 우리가 마시는 산소(O₂)와 달리 독성을 가지고 있다.

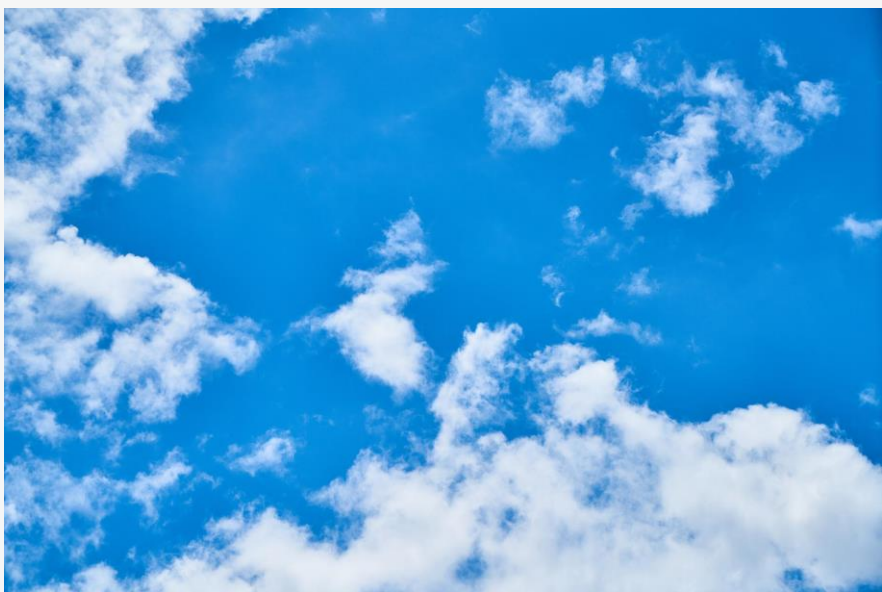
오존 또한 자동차, 공장 등에서 배출되는 질소산화물과 휘발성유기화합물질이 반응하여 생성되는 대기오염 물질이다. 도심, 공장지대에서 더욱 잘 발생하고 대기가 정체일 때 수치가 상승할 확률이 높다.

오존은 자극성과 산화력이 강해 사람의 눈과 피부를 자극하고 호흡기 질환을 유발하며 식물의 수확량감소, 건축물 부식, 스모그에 의한 대기오염 등 생태계 및 산업활동 전반에 악영향을 미친다.

오존전량은 지상 어느 위치든 그 위치 상공에 존재하는 오존 분자의 총량으로 정의되며, 위도, 경도, 계절에 따라 달라진다.

북반구의 오존전량은 봄에 뚜렷한 최댓값을 갖고 여름부터 가을까지 값이 낮아지는 계절변동을 보인다. 이는 늦가을과 겨울동안 열대지방에서 극지방으로 오존 수송이 증가하여 봄철 고위도에서 최댓값을 보인다.

분석목적



시군구에 따른 대기오염 분석

1. 여러 모델들을 통해 대기환경에 가장 영향을 많이 주는 요소를 추출할 수 있다.
2. 추출한 데이터를 바탕으로 대기오염이 심한 지역의 특징을 파악할 수 있다.
3. 최종적으로 지역별 대기오염의 해결방안을 탐색할 수 있고 해결방안을 탐색할 때 모델이 대기오염의 원인을 찾아주므로 효과적인 해결방안을 모색할 수 있다.

데이터 획득

Air korea

에어코리아란 실시간자료조회

실시간 대기정보 내일의 대기정보 시도별 대기/경보 정보

통합대기환경지수(CAI)
초미세먼지 (PM_{2.5})
미세먼지 (PM₁₀)
오존 (O₃)
이산화질소 (NO₂)
일산화탄소 (CO)
야황산가스 (SO₂)

현경도
백령도
울도
격렬비열도
외연도

한국전력기술 회사소개 사업

500MW 석탄화력발전소 사업수행

통계청 내일을 위한 정부혁신 보다는 나은 정부

국가통계포털 마이크로

통계청소개

뉴스기반통계검색
국가통계포털(KOSIS)
· 국내통계

DATA 공공데이터포털 . GO . KR

어떤 공공데이터를 찾으시나요

검색조건 ↻

분류체계
서비스유형
확장자

인기검색어



EDA & 전처리

EDA_미세먼지

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	fine_dust	O3	region	heavy_metal	traffic	num_of_factories	num_of_power_plants	num_of_people	household_waste_discharge_per_person (kg/days)	wind_speed	wind_direction	rainfall(mm)	C	avg_O3	avg_fine_dust	
2	0	1	강원도강릉시	0.9711	37.4	450	2	213321	1.3	6.295	64.02192	42.53333	14.6	0.035	38.83333	
3	0	1	강원도고성군	0.9711	38.4	81	0	26757	2.4	3.81	-27.4523	37.68333	13.2	0.035	34	
4	0	0	강원도동해시	0.9711	36	257	4	90593	1.6	6.295	64.02192	32.91667	13.5	0.027	35.75	

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 230 entries, 0 to 229

Data columns (total 15 columns):

#	Column	Non-Null Count	Dtype
0	fine_dust	230 non-null	int64
1	O3	230 non-null	int64
2	region	230 non-null	object
3	heavy_metal	230 non-null	float64
4	traffic	230 non-null	float64
5	num_of_factories	230 non-null	int64
6	num_of_power_plants	230 non-null	int64
7	num_of_people	230 non-null	int64
8	household_waste_discharge_per_person (kg/days)	230 non-null	float64
9	wind_speed	230 non-null	float64
10	wind_direction	230 non-null	float64
11	rainfall(mm)	230 non-null	float64
12	C	230 non-null	float64
13	avg_O3	230 non-null	float64
14	avg_fine_dust	230 non-null	float64

dtypes: float64(9), int64(5), object(1)

memory usage: 27.1+ KB

	fine_dust	region
0	0	강원도강릉시
1	0	강원도고성군
2	0	강원도동해시
3	0	강원도삼척시
4	0	강원도속초시
...
201	0	제주특별자치도제주시
202	0	충청남도계룡시
218	0	충청북도괴산군
219	0	충청북도단양군

	fine_dust	region
11	1	강원도철원군
14	1	강원도평창군
15	1	강원도홍천군
19	1	경기도고양시
20	1	경기도과천시
...
225	1	충청북도증평군
226	1	충청북도진천군
227	1	충청북도청원군
228	1	충청북도청주시
229	1	충청북도충주시

EDA_오존

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	fine_dust	O3	region	heavy_metal	traffic	num_of_factories	num_of_power_plants	num_of_people	household_waste_discharge_per_person (kg/days)	wind_speed	wind_direction	rainfall(mm)	C	avg_O3	avg_fine_dust	
2	0	1	강원도강릉시	0.9711	37.4	450	2	213321	1.3	6.295	64.02192	42.53333	14.6	0.035	38.83333	
3	0	1	강원도고성	0.9711	38.4	81	0	26757	2.4	3.81	-27.4523	37.68333	13.2	0.035	34	
4	0	0	강원도동해시	0.9711	36	257	4	90593	1.6	6.295	64.02192	32.91667	13.5	0.027	35.75	

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 230 entries, 0 to 229

Data columns (total 15 columns):

#	Column	Non-Null Count	Dtype
0	fine_dust	230 non-null	int64
1	O3	230 non-null	int64
2	region	230 non-null	object
3	heavy_metal	230 non-null	float64
4	traffic	230 non-null	float64
5	num_of_factories	230 non-null	int64
6	num_of_power_plants	230 non-null	int64
7	num_of_people	230 non-null	int64
8	household_waste_discharge_per_person (kg/days)	230 non-null	float64
9	wind_speed	230 non-null	float64
10	wind_direction	230 non-null	float64
11	rainfall(mm)	230 non-null	float64
12	C	230 non-null	float64
13	avg_O3	230 non-null	float64
14	avg_fine_dust	230 non-null	float64

dtypes: float64(9), int64(5), object(1)

memory usage: 27.1+ KB

	O3	region
2	0	강원도동해시
3	0	강원도삼척시
5	0	강원도양구군
7	0	강원도영월군
8	0	강원도원주시
...
225	0	충청북도증평군
226	0	충청북도진천군
227	0	충청북도청원군
228	0	충청북도청주시

	O3	region
0	1	강원도강릉시
1	1	강원도고성군
4	1	강원도속초시
6	1	강원도양양군
9	1	강원도인제군
...
199	1	전라북도진안군
200	1	제주특별자치도서귀포시
201	1	제주특별자치도제주시
209	1	충청남도서산시
215	1	충청남도청양군

전처리_미세먼지

전국 시군구 미세먼지의 AVERAGE보다 크면 1, 작으면 0 부여

X = 지역, 중금속(평균값), 교통량, 공장 수, 발전소 수, 인구 수, 1인당 생활 폐기물 배출량, 풍속, 풍향, 강수량, 기온

y = 미세먼지 농도의 평균보다 크면 1, 작으면 0(DT,SVM,logistic)

y = 미세먼지 농도(linear)

으로 X,y데이터 분리

train_test_split 이용해 train, validation, test 분리

.pop 이용해 'region' column 분리 후 list 저장

DT 모델 제외한 나머지 모델엔 StandardScaler, MinMaxScaler 이용해 scaling 진행

전처리_오존

전국 시군구 오존의 AVERAGE보다 크면 1, 작으면 0 부여

X = 지역, 중금속(평균값), 교통량, 공장 수, 발전소 수, 인구 수, 1인당 생활 폐기물 배출량, 풍속, 풍향, 강수량, 기온

Y = 오존 농도의 평균보다 크면 1, 작으면 0(DT,SVM,logistic)

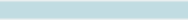
Y = 오존 농도(linear)

으로 X,y데이터 분리

train_test_split 이용해 train, validation, test 분리

.pop 이용해 'region' column 분리 후 list 저장

DT 모델 제외한 나머지 모델엔 StandardScaler, MinMaxScaler 이용해 scaling 진행



대기오염 분석_미세먼지

미세먼지 분석_Decision Tree

```
pd.DataFrame({"min_samples_leaf":sorted(msl_settings*7), "max_depth":
```

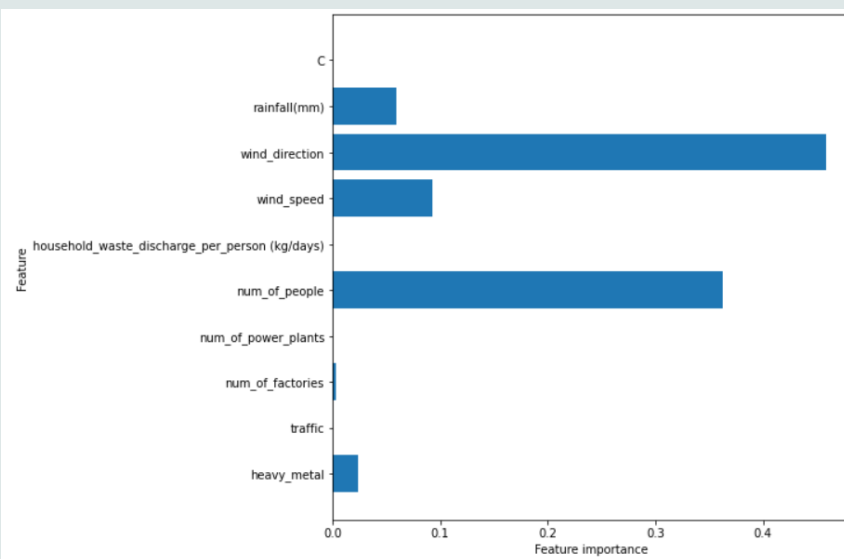
	min_samples_leaf	max_depth	training accuracy	validation accuracy
0	3	1	0.751938	0.720930
1	3	2	0.844961	0.790698
2	3	5	0.914729	0.813953
3	3	7	0.937984	0.837209
4	3	13	0.937984	0.837209
5	3	20	0.937984	0.837209
6	3	30	0.937984	0.837209
7	5	1	0.751938	0.720930
8	5	2	0.844961	0.790698
9	5	5	0.906977	0.813953
10	5	7	0.906977	0.813953

36	10	2	0.844961	0.790698
37	10	5	0.860465	0.790698
38	10	7	0.860465	0.790698
39	10	13	0.860465	0.790698
40	10	20	0.860465	0.790698
41	10	30	0.860465	0.790698
42	20	1	0.751938	0.720930

미세먼지 분석_Decision Tree

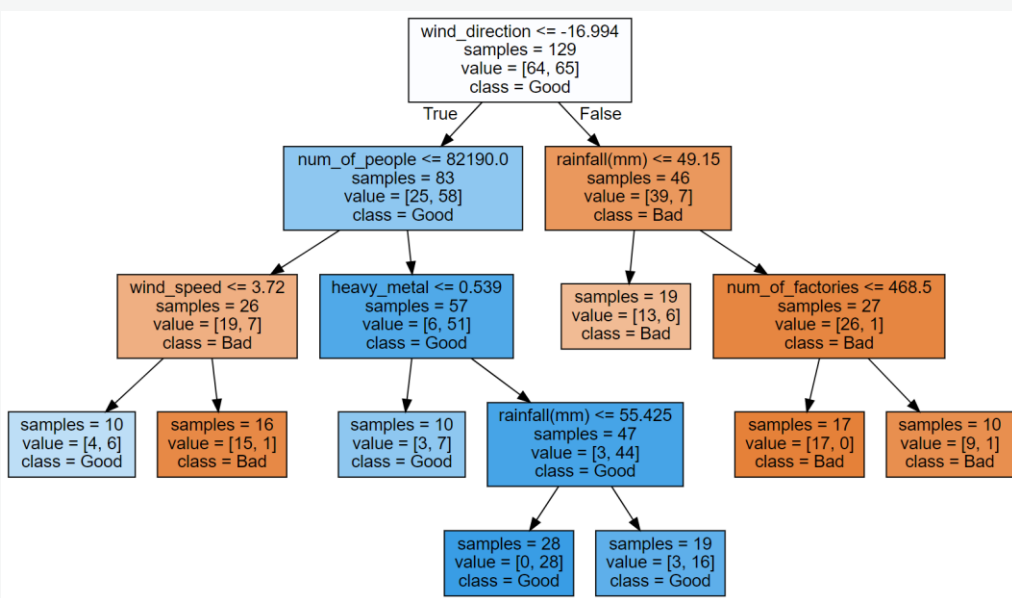
지역별 예측값 실제값 비교 (X_test)

	region	fine_dust_predict	fine_dust_value
0	전라북도정읍시	1	0
1	부산광역시연제구	0	0
2	광주광역시남 구	0	0
3	경상남도밀양시	0	0
4	경기도광주시	1	1
5	강원도홍천군	0	1
6	울산광역시울주군	0	0
7	전라북도임실군	0	1
8	경기도과천시	0	1
9	경상북도상주시	1	1
10	강원도고성군	0	0
11	경기도부천시	1	1
12	제주특별자치도제주시	0	0
13	부산광역시해운대구	0	0
14	충청북도청주시	1	1
15	인천광역시동 구	1	1
16	대구광역시남 구	0	1
17	경상북도고령군	0	0
18	인천광역시용진군	1	1
19	전라남도순천시	0	0



Feature importance :
 풍향 > 인구 수 > 풍속 > 강수량 > 중금속 > 공장수

Accuracy on training set: 0.86
 Accuracy on test set: 0.78
 [[24 5]
 [8 21]]



미세먼지 분석_SVM

StandardScaler

	C	gamma	training accuracy	validation accuracy
0	0.01	0.01	0.503876	0.511628
1	0.01	0.10	0.503876	0.511628
2	0.01	1.00	0.503876	0.511628
3	0.01	10.00	0.503876	0.511628
4	0.01	100.00	0.503876	0.511628
5	1.00	0.01	0.751938	0.906977
6	1.00	0.10	0.821705	0.883721
7	1.00	1.00	0.976744	0.697674
8	1.00	10.00	1.000000	0.511628
9	1.00	100.00	1.000000	0.488372
10	10.00	0.01	0.759690	0.930233
11	10.00	0.10	0.906977	0.837209
12	10.00	1.00	0.992248	0.697674
13	10.00	10.00	1.000000	0.488372
14	10.00	100.00	1.000000	0.488372
15	100.00	0.01	0.837209	0.883721
16	100.00	0.10	0.968992	0.720930
17	100.00	1.00	1.000000	0.674419
18	100.00	10.00	1.000000	0.488372
19	100.00	100.00	1.000000	0.488372
20	1000.00	0.01	0.868217	0.860465
21	1000.00	0.10	0.992248	0.697674
22	1000.00	1.00	1.000000	0.674419
23	1000.00	10.00	1.000000	0.488372
24	1000.00	100.00	1.000000	0.488372

MinMaxScaler

	C	gamma	training accuracy	validation accuracy
0	0.01	0.01	0.503876	0.511628
1	0.01	0.10	0.503876	0.511628
2	0.01	1.00	0.503876	0.511628
3	0.01	10.00	0.503876	0.511628
4	0.01	100.00	0.503876	0.511628
5	0.10	0.01	0.503876	0.511628
6	0.10	0.10	0.503876	0.511628
7	0.10	1.00	0.697674	0.720930
8	0.10	10.00	0.534884	0.558140
9	0.10	100.00	0.503876	0.511628
10	1.00	0.01	0.503876	0.511628

8	0.10	10.00	0.534884	0.558140
9	0.10	100.00	0.503876	0.511628
10	1.00	0.01	0.503876	0.511628
11	1.00	0.10	0.705426	0.720930
12	1.00	1.00	0.759690	0.930233
13	1.00	10.00	0.899225	0.906977
14	1.00	100.00	0.976744	0.604651
15	10.00	0.01	0.697674	0.720930
16	10.00	0.10	0.759690	0.930233
17	10.00	1.00	0.837209	0.930233
18	10.00	10.00	0.976744	0.720930
19	10.00	100.00	1.000000	0.511628
20	100.00	0.01	0.759690	0.930233

미세먼지 분석_SVM

StandardScaler

Accuracy on training set: 0.7596899224806202
Accuracy on testing set: 0.8103448275862069
[[24 5]
[6 23]]

MinMaxScaler

Accuracy on training set: 0.7596899224806202
Accuracy on testing set: 0.7586206896551724
[[20 9]
[5 24]]

StandardScaler가 MinMaxScaler보다 정확도가 더 높음

지역별 예측값 실제값 비교 (StandardScaler)

	region	fine_dust_predict	fine_dust_value
0	경상남도산청군	0	0
1	충청남도논산시	1	1
2	전라남도곡성군	0	0
3	경기도시흥시	1	1
4	경기도남양주시	1	1
5	전라남도무안군	0	1
6	대구광역시달성군	1	1
7	경상북도칠곡군	0	1
8	부산광역시동 구	0	0
9	전라남도영광군	0	1
10	전라북도진안군	0	0
11	강원도태백시	0	0
12	대전광역시중 구	1	1
13	경기도여주시	1	1
14	부산광역시동래구	0	0
15	경기도오산시	1	1
16	경상북도고령군	0	0
17	경상북도문경시	0	0
18	전라남도해남군	0	0

지역별 예측값 실제값 비교 (MinMaxScaler)

	region	fine_dust_predict	fine_dust_value
0	경상남도산청군	0	0
1	충청남도논산시	1	1
2	전라남도곡성군	0	0
3	경기도시흥시	1	1
4	경기도남양주시	1	1
5	전라남도무안군	0	1
6	대구광역시달성군	1	1
7	경상북도칠곡군	0	1
8	부산광역시동 구	0	0
9	전라남도영광군	0	1
10	전라북도진안군	1	0
11	강원도태백시	0	0
12	대전광역시중 구	1	1
13	경기도여주시	1	1
14	부산광역시동래구	0	0
15	경기도오산시	1	1
16	경상북도고령군	0	0
17	경상북도문경시	0	0
18	전라남도해남군	0	0

미세먼지 분석_Logistic Regression

C			
	training accuracy	val accuracy	
0	0.01	0.790698	0.837209
1	0.10	0.782946	0.813953
2	1.00	0.798450	0.837209
3	10.00	0.806202	0.837209
4	100.00	0.806202	0.837209
5	1000.00	0.806202	0.837209
6	10000.00	0.806202	0.837209

미세먼지 분석_Regression

Linear

결정계수: 0.4745643155100302

Logistic

Accuracy on training set: 0.6666666666666666

Accuracy on testing set: 0.7413793103448276

[[22 7]
[9 20]]

회귀계수:

[[-0. -0.014 0.001 0. 0. -0. -0.002
 -0.019 -0.016 -0.005]] 상수항: [-0.]

지역별 예측값 실제값 비교 (Linear)

	region	fine_dust_predict	fine_dust_value
0	충청남도홍성군	46.218385	46.166667
1	경상남도합천군	39.118621	30.333333
2	강원도인제군	44.895638	34.833333
3	전라북도완주군	41.593626	45.000000
4	강원도홍천군	44.599627	47.000000
5	전라북도진안군	41.399238	37.666667
6	경기도김포시	54.721885	54.666667
7	전라북도임실군	42.949791	46.333333
8	인천광역시강화군	43.270353	41.405270
9	서울특별시강북구	44.542651	51.608070
10	충청북도괴산군	43.195092	41.666667
11	경기도고양시	53.842370	48.500000
12	대구광역시달서구	45.955954	43.002590
13	전라남도영광군	35.240637	44.500000
14	충청북도음성군	46.344400	47.333333
15	서울특별시서대문구	44.861406	40.610980
16	서울특별시종로구	44.297744	46.834830
17	경기도수원시	49.768431	50.833333
18	충청북도증평군	44.328623	49.666667

지역별 예측값 실제값 비교 (Logistic)

	region	fine_dust_predict	fine_dust_value
0	전라북도정읍시	1	0
1	부산광역시연제구	0	0
2	광주광역시남 구	0	0
3	경상남도밀양시	0	0
4	경기도광주시	1	1
5	강원도홍천군	0	1
6	울산광역시울주군	1	0
7	전라북도임실군	1	1
8	경기도과천시	0	1
9	경상북도상주시	1	1
10	강원도고성군	0	0
11	경기도부천시	1	1
12	제주특별자치도제주시	0	0
13	부산광역시해운대구	1	0
14	충청북도청주시	1	1
15	인천광역시동 구	1	1
16	대구광역시남 구	0	1
17	경상북도고령군	1	0
18	인천광역시옹진군	1	1

미세먼지 분석 결과

Decision Tree 정확도 : 0.78

SVM(StandardScaler) 정확도 : 0.81

SVM(MinMaxScaler) 정확도 : 0.76

Linear Regression 정확도 : 0.47

Logistic Regression 정확도 : 0.74

SVM(StandardScaler) > Decision Tree > SVM(MinMaxScaler) >
Logistic Regression > Linear Regression 순으로 적합



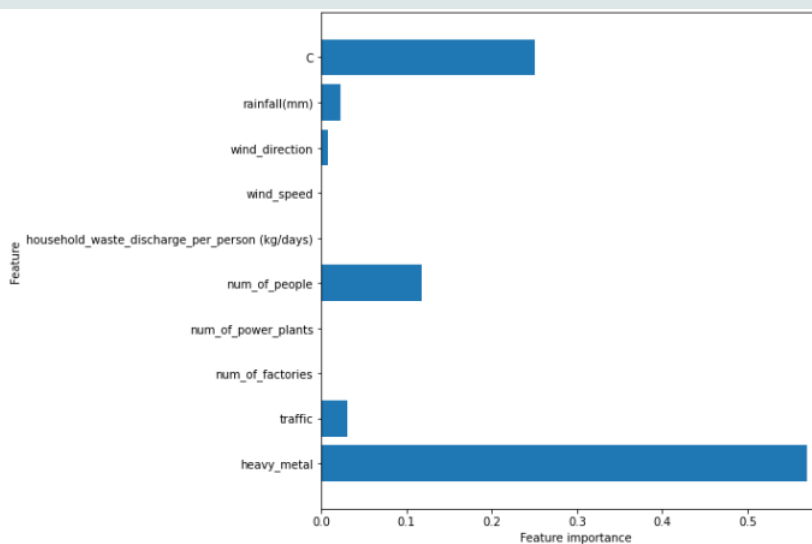
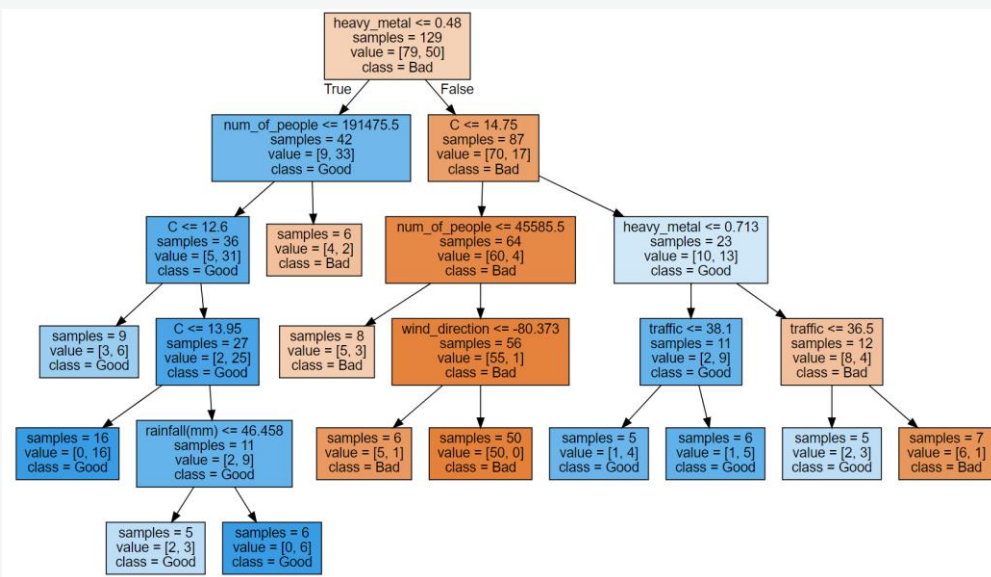
대기오염 분석_오존

오존 분석_Decision Tree

	min_samples_leaf	max_depth	training accuracy	validation accuracy
0	3	1	0.798450	0.767442
1	3	2	0.844961	0.860465
2	3	5	0.930233	0.837209
3	3	7	0.930233	0.837209
4	3	13	0.930233	0.837209
5	3	20	0.930233	0.837209
6	3	30	0.930233	0.837209
7	5	1	0.798450	0.767442
8	5	2	0.837209	0.860465
9	5	5	0.875969	0.860465
10	5	7	0.875969	0.860465

5	3	20	0.930233	0.837209
6	3	30	0.930233	0.837209
7	5	1	0.798450	0.767442
8	5	2	0.837209	0.860465
9	5	5	0.875969	0.860465
10	5	7	0.875969	0.860465
11	5	13	0.875969	0.860465
12	5	20	0.875969	0.860465
13	5	30	0.875969	0.860465
14	7	1	0.798450	0.767442
15	7	2	0.829457	0.860465

오존 분석_Decision Tree



Feature importance :
중금속 > 기온 > 인구 수 > 교통량 > 강수량 > 풍향

Accuracy on training set: 0.88
Accuracy on test set: 0.79
[[32 4]
[8 14]]

지역별 예측값 실제값 비교 (X_test)

	region	fine_dust_predict	fine_dust_value
0	경기도의왕시	0	0
1	강원도강릉시	0	1
2	경상북도성주군	1	1
3	서울특별시영등포구	0	0
4	충청북도보은군	0	0
5	충청북도제천시	0	0
6	전라남도나주시	1	1
7	경기도부천시	0	0
8	충청남도예산군	0	0
9	경기도광주시	0	0
10	강원도인제군	0	1
11	강원도양양군	0	1
12	경상북도봉화군	1	1
13	경상북도울릉군	1	1
14	충청남도아산시	0	0
15	인천광역시서구	0	0
16	대구광역시중구	0	0
17	경기도안산시	0	1
18	대구광역시남구	0	0

오존 분석_SVM

StandardSclaer

	C	gamma	training accuracy	validation accuracy
0	0.01	0.01	0.612403	0.604651
1	0.01	0.10	0.612403	0.604651
2	0.01	1.00	0.612403	0.604651
3	0.01	10.00	0.612403	0.604651
4	0.01	100.00	0.612403	0.604651
5	1.00	0.01	0.906977	0.883721
6	1.00	0.10	0.992248	0.906977
7	1.00	1.00	1.000000	0.744186
8	1.00	10.00	1.000000	0.604651
9	1.00	100.00	1.000000	0.604651
10	10.00	0.01	0.992248	0.953488

11	10.00	1.00	1.000000	0.953488
13	10.00	10.00	1.000000	0.604651
14	10.00	100.00	1.000000	0.604651
15	100.00	0.01	0.992248	0.953488
16	100.00	0.10	1.000000	0.953488
17	100.00	1.00	1.000000	0.767442
18	100.00	10.00	1.000000	0.604651
19	100.00	100.00	1.000000	0.604651
20	1000.00	0.01	1.000000	0.953488
21	1000.00	0.10	1.000000	0.953488
22	1000.00	1.00	1.000000	0.767442
23	1000.00	10.00	1.000000	0.604651
24	1000.00	100.00	1.000000	0.604651

MinMaxSclaer

	C	gamma	training accuracy	validation accuracy
0	0.01	0.01	0.612403	0.604651
1	0.01	0.10	0.612403	0.604651
2	0.01	1.00	0.612403	0.604651
3	0.01	10.00	0.612403	0.604651
4	0.01	100.00	0.612403	0.604651
5	0.10	0.01	0.612403	0.604651
6	0.10	0.10	0.612403	0.604651
7	0.10	1.00	0.651163	0.627907
8	0.10	10.00	0.612403	0.604651
9	0.10	100.00	0.612403	0.604651
10	1.00	0.01	0.612403	0.604651
12	1.00	1.00	0.976744	0.883721
13	1.00	10.00	1.000000	0.883721
14	1.00	100.00	1.000000	0.697674
15	10.00	0.01	0.813953	0.837209
16	10.00	0.10	0.976744	0.906977
17	10.00	1.00	1.000000	0.953488
18	10.00	10.00	1.000000	0.883721
19	10.00	100.00	1.000000	0.697674
20	100.00	0.01	0.961240	0.930233
21	100.00	0.10	0.992248	0.953488
22	100.00	1.00	1.000000	0.953488
23	100.00	10.00	1.000000	0.883721
24	100.00	100.00	1.000000	0.697674

오존 분석_SVM

StandardScaler

Accuracy on training set: 0.7906976744186046
Accuracy on testing set: 0.7758620689655172
[[32 4]
[9 13]]

MinMaxScaler

Accuracy on training set: 0.7906976744186046
Accuracy on testing set: 0.7586206896551724
[[31 5]
[9 13]]

StandardScaler가 MinMaxScaler보다 정확도가 더 높음

지역별 예측값 실제값 비교 (StandardScaler)

	region	fine_dust_predict	fine_dust_value
0	경기도의왕시	0	0
1	강원도강릉시	0	1
2	경상북도성주군	1	1
3	서울특별시영등포구	0	0
4	충청북도보은군	0	0
5	충청북도제천시	0	0
6	전라남도나주시	1	1
7	경기도부천시	0	0
8	충청남도예산군	0	0
9	경기도광주시	0	0
10	강원도인제군	0	1
11	강원도양양군	1	1
12	경상북도봉화군	1	1
13	경상북도울릉군	1	1
14	충청남도아산시	0	0
15	인천광역시서 구	0	0
16	대구광역시중 구	0	0
17	경기도안산시	0	1
18	대구광역시남 구	0	0

지역별 예측값 실제값 비교 (MinMaxScaler)

	region	fine_dust_predict	fine_dust_value
0	경기도의왕시	0	0
1	강원도강릉시	0	1
2	경상북도성주군	1	1
3	서울특별시영등포구	0	0
4	충청북도보은군	0	0
5	충청북도제천시	0	0
6	전라남도나주시	1	1
7	경기도부천시	0	0
8	충청남도예산군	0	0
9	경기도광주시	0	0
10	강원도인제군	0	1
11	강원도양양군	1	1
12	경상북도봉화군	1	1
13	경상북도울릉군	1	1
14	충청남도아산시	0	0
15	인천광역시서 구	0	0
16	대구광역시중 구	0	0
17	경기도안산시	0	1
18	대구광역시남 구	0	0

오존 분석_Logistic Regression

	C	training accuracy	val accuracy
0	0.01	0.759690	0.697674
1	0.10	0.813953	0.744186
2	1.00	0.829457	0.744186
3	10.00	0.829457	0.744186
4	100.00	0.829457	0.744186
5	1000.00	0.829457	0.744186
6	10000.00	0.829457	0.744186

회귀계수: [[-0. 0.005 -0.001 -0. -0. 0. 0.001
0.007 0.005 0.002]] 상수항: [0.]

Accuracy on training set: 0.5426356589147286
Accuracy on testing set: 0.6724137931034483
[[31 5]
[9 13]]

지역별 예측값 실제값 비교

	region	fine_dust_predict	fine_dust_value
0	경기도의왕시	0	0
1	강원도강릉시	0	1
2	경상북도성주군	0	1
3	서울특별시영등포구	0	0
4	충청북도보은군	0	0
5	충청북도제천시	0	0
6	전라남도나주시	1	1
7	경기도부천시	0	0
8	충청남도예산군	0	0
9	경기도광주시	0	0
10	강원도인제군	1	1
11	강원도양양군	1	1
12	경상북도봉화군	0	1
13	경상북도울릉군	1	1
14	충청남도아산시	0	0
15	인천광역시서 구	0	0
16	대구광역시중 구	1	0
17	경기도안산시	0	1
18	대구광역시남 구	1	0

오존 분석 결과

Decision Tree 정확도 : 0.79

SVM(StandardScaler) 정확도 : 0.78

SVM(MinMaxScaler) 정확도 : 0.76

Logistic Regression 정확도 : 0.67

SVM(StandardScaler) > SVM(MinMaxScaler) > Decision Tree > Logistic Regression
순으로 적합

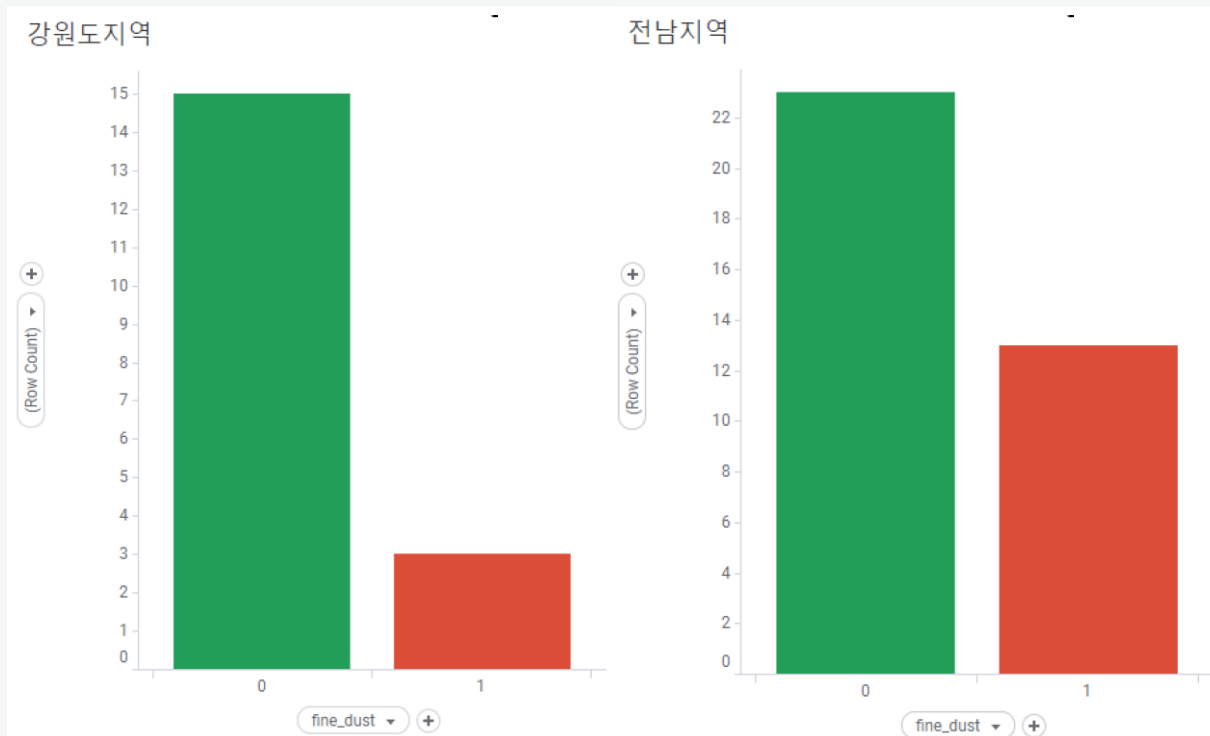


분석 활용

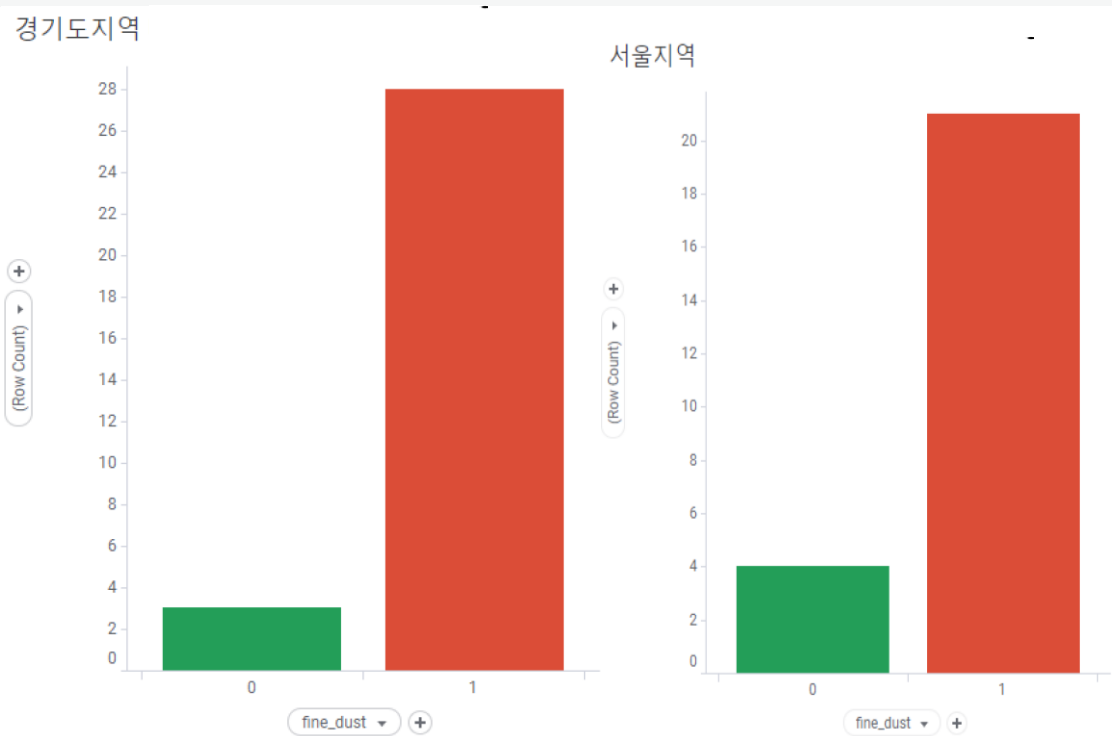
기대효과

1. feature importance가 높은 '인구수'를 보았을 때, 인구수가 많은 수도권에 가까울수록 미세먼지, 오존농도가 높은 것으로 보아 미세먼지, 오존을 줄이기 위해서는 아파트와 같은 밀집된 공간을 줄이고 다른 지역을 홍보해 인구 밀집을 분산시키는 방안이 필요

인구수 낮은 비수도권 지역



인구수 높은 수도권 지역



기대효과

2. 우리나라의 미세먼지는 북서풍의 영향을 많이 받는 것으로 보아 중국의 영향이 크기 때문에 이에 대한 대책이 필요
3. 더 많은 X변수나 새로운 y변수를 추가해 데이터분석을 한다면 또 다른 대기오염의 영향을 주는 다른 요인 또한 파악하고 개선해 나갈 수 있을 것

한계점 및 추후 개선점

1. 모델링 할 때 y target으로 continuous value를 다루기 쉽지 않아 평균값으로 대체한 점에서 데이터의 손실이 이루어질 수 있음
-> continuous value에 관한 새로운 모델 이용
1. 전국 시군구의 데이터 수가 230개라 train, test, validation 세개로 나누었을 때, test, validation 데이터가 상대적으로 적어서 충분치 못함
-> 시군구 -> 읍면동까지 넓히기

Github repo

<https://github.com/hsgalaxy-K/Dataminig>

참고자료

두산백과 – 미세먼지

<https://terms.naver.com/entry.naver?docId=1080596&cid=40942&categoryId=32412>

두산백과 – 오존

<https://terms.naver.com/entry.naver?docId=1128529&cid=40942&categoryId=32251>

오존 분석배경

http://www.climate.go.kr/home/10_wiki/index.php/%EC%84%B1%EC%B8%B5%EA%B6%8C_%EC%98%A4%EC%A1%B4

<https://blog.naver.com/momomo0505/221306912948>

미세먼지 분석배경

<https://www.yeongnam.com/web/view.php?key=20211129010003519>

감사합니다