2020년 K-water 대국민 빅데이터 공모전 수행 결과보고서

제 목	딥러닝 기술을 활용한 대기오염 예측모델 생성 -화력발전소 인근 대기오염측정값을 중심으로-			
공모분야	서비스 개발		융합데이터	O
성 명	팀 장	(성 명)박강현	(연락처)010-8806-6392	
		(소 속) 한국지질자원연구원 석사과정연구생		
	팀 원	(성 명)김대열	(소속) 한국지질자원연구원 인턴연구원	

Ⅰ. 과제 목표

2016년 초부터 대한민국의 미세먼지 문제가 심각해지고 지속화되자 환경기준이 선진국 수준으로 강화되고 비상저감조치가 시행되었다. 미세먼지 예보가 서비스되면서 국민들의 건강에 대한 관심이 화두가 되었고, 이에 따라 필터 마스크. 공기정화식물 등에 관심이 폭증하고 공기청정기 시장이 가파르게 성장하였다.

국내 미세먼지의 주요 발생원으로는 화력발전소가 지목받고 있다. 국내전력 생산의 40.4%가(2019년) 화력발전소에서 생산되고 석탄을 에너지원으로 사용하기 때문에 미세먼지와 대기오염물질을 많이 발생시킨다.

화력발전소 가동 과정에서 나오는 대기 오염물질들은 인체에 해로운 영향을 미치고 질병의 원인이 된다. 특히 수도권 지역은 국토 면적의 11.8%를 차지하지만 인구밀도가 50.16%에 달해 전력수요가 많은데, 전력량 확보를 위해 공간의 제약을 적게 받는 화력발전소가 주로 집중되어 있어 수많은 인구가 대기오염에 노출되어있다. [그림3 참고]

통계청「2019 한국의 사회동향」에 따르면 일평균 초미세먼지 농도가 '매우 나쁨' 빈도와 미세먼지 주의보.경보 발령 횟수가 점점 증가

하는 추세에 있어서 이로 인한 만성.급성 질환자도 증가할 가능성이 우려되는 상황이다.

본 과제에서는 먼저 화력발전소가 밀집된 지역을 선정하고 측정소에서 측정한 대기오염물질 빅데이터를 GIS 프로그램으로 가시화 및 데이터베이스화 하였다.

다음으로 앞서 구축된 데이터셋을 기반으로하여 환경빅데이터 플랫폼에서 제공하는 화력발전소 별 전국 평균 미세먼지 대비 목표 화력발전소의 월평균 미세먼지량의 비율이 가장 큰 화력발전소(일산열병합발전소)를 참조하여 연구 목표로 삼는다.

마지막으로 목표가 되는 발전소에서 가장 가까운 도심지역 측정소의 일단위 대기오염물질의 데이터를 정제한 후, 딥러닝 기술을 중심으로 분석하여 현재 에어코리아 공개되어있는 기존 2020년 5월까지의 데이터를 기준으로 2020년 6월의 30일 간의 미세먼지 예측모델을 만들어보고자 한다.

II. 주요 내용

- 환경빅데이터플랫폼에서「화력발전소 인근 질병정보」데이터를 다운받은 후, 발전소의 위경도 좌표를 Excel 중복값 제거를 통해 48개의 화력발전소 위치를 추출하여 CSV파일로 저장한다.

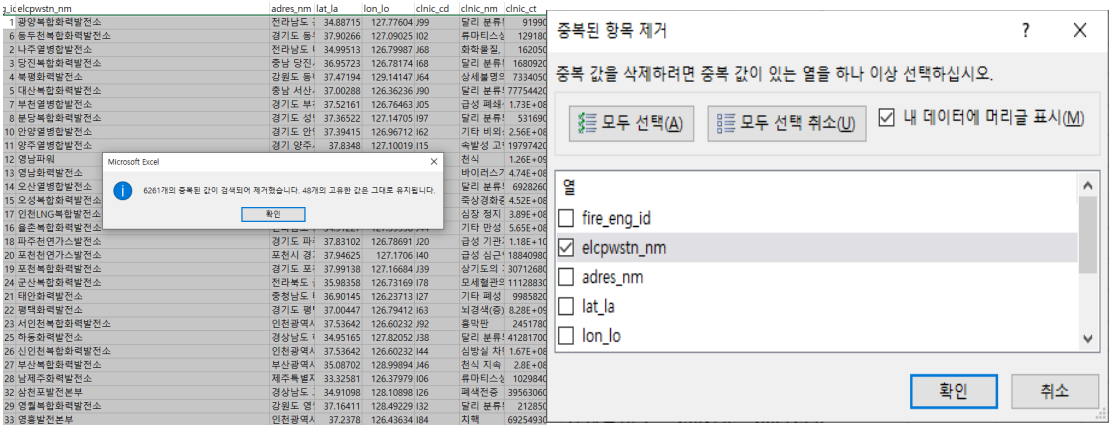


그림 1 중복값 제거로 화력발전소 리스트 확보

- CSV형식의 화력발전소의 경위도 좌표를 SHP파일로 변환하기 위해 GeoCoder-Xr 프로그램을 사용하여 공간참조를 한다. 국가공간정보포털에서「행정구역 SHP」파일을 다운받아 화력발전소 경위도 좌표 SHP 파일을 중첩시킨다.

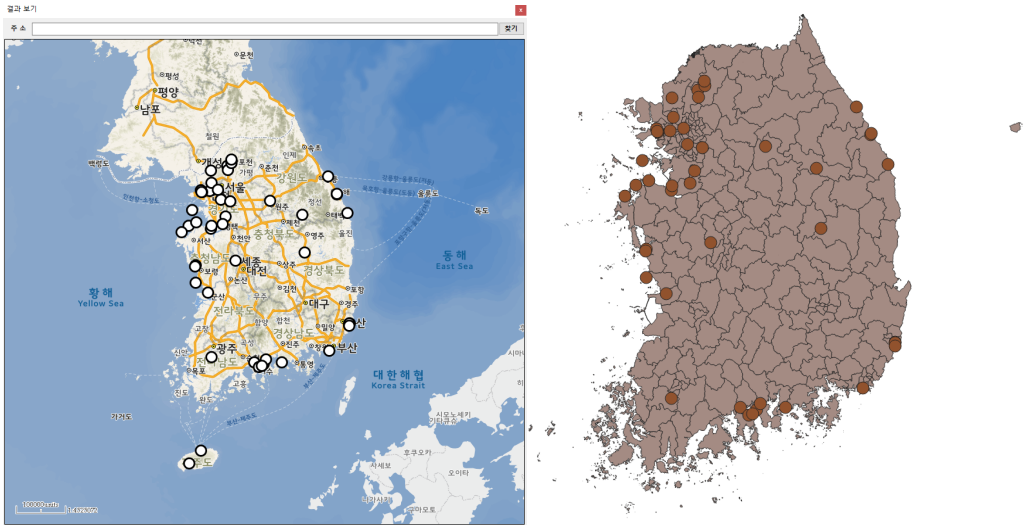


그림 2 좌표값 공간 참조를 통한 화력발전소 SHP파일 생성

- 핫스팟 분석(Hot Spot Analysis)를 통해 화력발전소의 분포 패턴을 분석하여 화력발전소 입지가 편중된 지역을 확인한다.

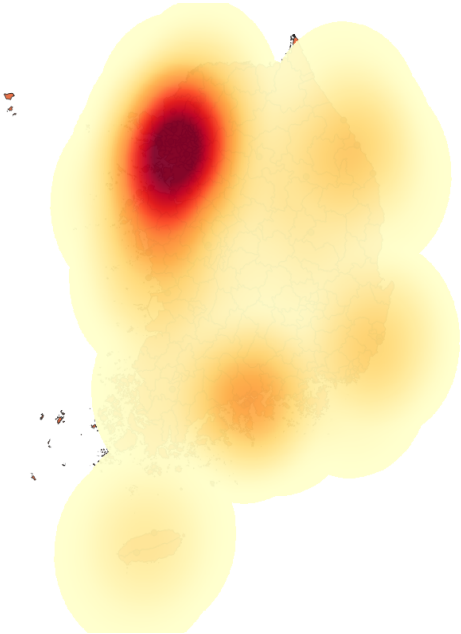


그림 3 화력발전소 핫스팟 분석

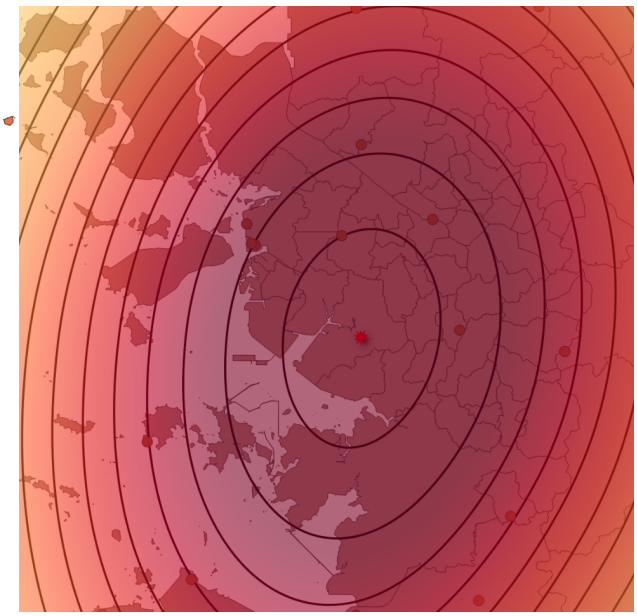
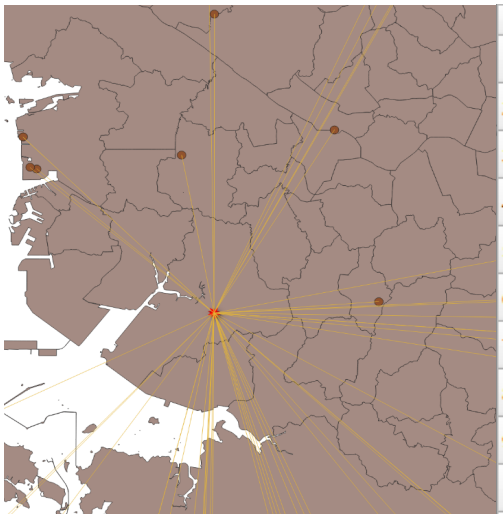


그림 4 등고선 변환을 통한 중심점 생성

- 핫스팟 분석으로 생성한 열지도를 등고선 지도로 벡터라이징 한 후, 중심점 분석을 통해 화력발전소의 분포 밀도가 가장 높은 지역의 포인트 좌표를 생성한다. 최근린 거리로 허브라인을 생성하여 열지도 중심 포인트에서 가장 가까운 화력발전소 10개를 추출한다.



	HubDist	elcpwstn_n
1	0.1425710290...	부천열병합발전소
2	0.1684527028...	안양열병합발전소
3	0.2021476336...	서울화력발전소
4	0.2212104511...	인천LNG복합발전소
5	0.2276389267...	인천복합화력발전소
6	0.2493039018...	신인천복합화력발전소
7	0.2493039018...	서인천복합화력발전소
8	0.2620786018...	일산열병합발전소
9	0.3484962691...	분당복합화력발전소
10	0.3484962691...	분당발전본부

그림 5 연구대상 화력발전소 선택

- 환경빅데이터포털의 「화력발전소 인근 미세먼지 정보_200305」를 분석하여 얻은 자료를 기반으로 화력발전소 미세먼지 비율이 가장 높은 일산열병합발전소를 예제로 앞으로의 과제를 진행한다. 화력발전소 인근의 대기오염물질 데이터를 확보하기 위해 에어코리아 홈페이지의 최종확정 측정자료 조회 탭에서 일산열병합발전소의 도로명 주소값을 입력 후, 가장 가까운 도시대기 측정소의 측정소명을 확인한다.

그림 6 화력발전소와 가장 가까운 측정소 선택

- 확정자료 다운로드 탭에서 「2010-2019년간 전국 측정소의 시간 단위 측정값」 자료를 다운받고, EXCEL의 필터링 기능을 통해 측정소의 시간별 대기오염물질 데이터 값을 확보한다.

1	지역	측정소명	측정소코드	측정일시	SO2	CO	O3	NO2	PM10
190082	경기 고양시	식사동	131382	2018010101	0.005	0.5	0.022	0.017	42
190083	경기 고양시	식사동	131382	2018010102	0.005	0.4	0.021	0.018	38
190084	경기 고양시	식사동	131382	2018010103	0.006	0.5	0.014	0.021	41
190085	경기 고양시	식사동	131382	2018010104	0.005	0.5	0.019	0.019	41
190086	경기 고양시	식사동	131382	2018010105	0.005	0.5	0.018	0.017	41
190087	경기 고양시	식사동	131382	2018010106	0.005	0.5	0.019	0.017	33
190088	경기 고양시	식사동	131382	2018010107	0.005	0.5	0.013	0.023	31
190089	경기 고양시	식사동	131382	2018010108	0.005	0.5	0.009	0.029	32
190090	경기 고양시	식사동	131382	2018010109	0.006	0.6	0.006	0.031	34
190091	경기 고양시	식사동	131382	2018010110	0.007	0.7	0.007	0.031	34
190092	경기 고양시	식사동	131382	2018010111	0.005	0.5	0.021	0.02	32
190093	경기 고양시	식사동	131382	2018010112	0.006	0.5	0.025	0.02	38
190094	경기 고양시	식사동	131382	2018010113	0.005	0.4	0.031	0.015	42
190095	경기 고양시	식사동	131382	2018010114	0.004	0.3	0.035	0.01	36
190096	경기 고양시	식사동	131382	2018010115	0.004	0.4	0.038	0.008	34
190097	경기 고양시	식사동	131382	2018010116	0.005	0.4	0.038	0.01	37
190098	경기 고양시	식사동	131382	2018010117	0.004	0.4	0.035	0.011	44
190099	경기 고양시	식사동	131382	2018010118	0.005	0.5	0.027	0.021	44
190100	경기 고양시	식사동	131382	2018010119	0.005	0.5	0.018	0.028	45
190101	경기 고양시	식사동	131382	2018010120	0.006	0.5	0.012	0.033	46
190102	경기 고양시	식사동	131382	2018010121	0.007	0.6	0.004	0.043	40
190103	경기 고양시	식사동	131382	2018010122	0.007	0.6	0.003	0.042	46
190104	경기 고양시	식사동	131382	2018010123	0.006	0.6	0.008	0.035	48

그림 7 Excel 필터링 기능을 통한 데이터 정렬과 병합

Ⅲ. 활용데이터 및 수행내용

- 수집된 화력발전소의 데이터 중 월평균 전국 미세먼지량 대비 위치한 미세먼지량이 가장 높았던 “일산열병합발전소”를 기준으로 하여, 해당 발전소와 가장 가까운 곳에 위치한 도심지역 미세먼지 측정소인 식사동 측정소에서 수집된 2014년부터 2020년 5월까지의 일간 미세먼지 데이터를 사용하였다.

날짜	PM10	PM2.5	오존	이산화질소	일산화탄소	아황산가스
2014-01-01	100.0	NaN	0.017	0.023	0.6	0.009
2014-01-02	48.0	NaN	0.016	0.022	0.5	0.007
2014-01-03	63.0	NaN	0.006	0.033	0.7	0.008
2014-01-04	37.0	NaN	0.018	0.016	0.4	0.005
2014-01-05	43.0	NaN	0.018	0.016	0.5	0.005
...	...	...	...	...	...	...
2020-05-27	36.0	10.0	0.040	0.012	0.3	0.003
2020-05-28	48.0	16.0	0.048	0.014	0.3	0.003
2020-05-29	49.0	16.0	0.034	0.015	0.3	0.003
2020-05-30	58.0	22.0	0.056	0.014	0.4	0.003
2020-05-31	57.0	25.0	0.055	0.010	0.3	0.003

2229 rows × 6 columns

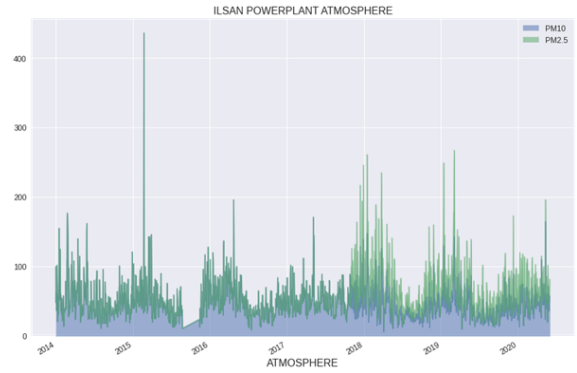


그림 8 정제된 데이터셋의 데이터프레임과 그래프

- PM2.5의 경우에는 2017년 말부터 수집되기 시작한 자료로, 데이터양에 따른 딥러닝 학습의 효율을 고려하여 대상에서 제외하였고, 데이터양이 결측값을 제외하고도 2229개인 PM10 데이터를 유용하였다.

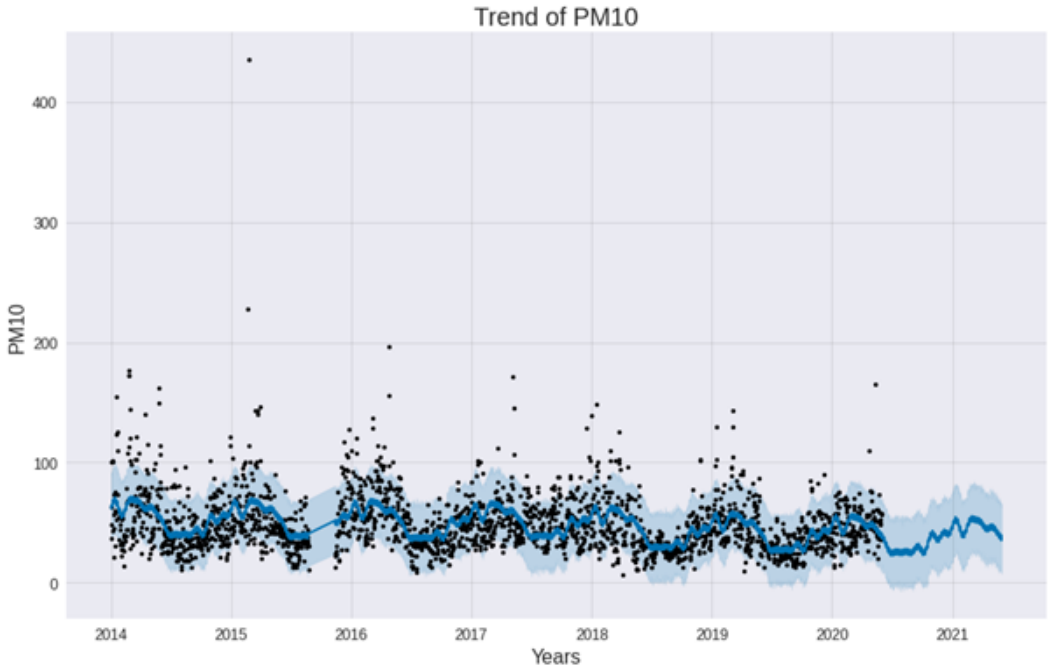


그림 9 미세먼지 데이터의 분포와 추세 그래프

- 검은색 점은 PM10의 실제적인 데이터, 파란색 굵은 선은 최소제곱법을 통해 연산한 추세선을 나타낸다. 이후 그간의 추세를 통하여 2020년 5월부터 2021년 5월까지의 추세를 그려내었다.

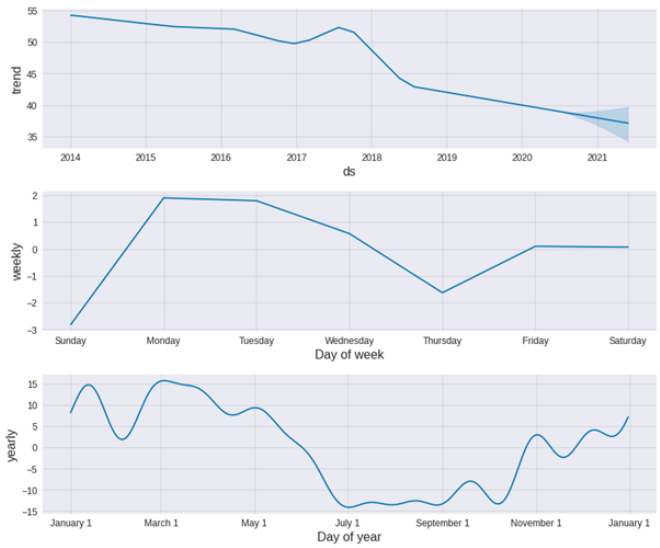
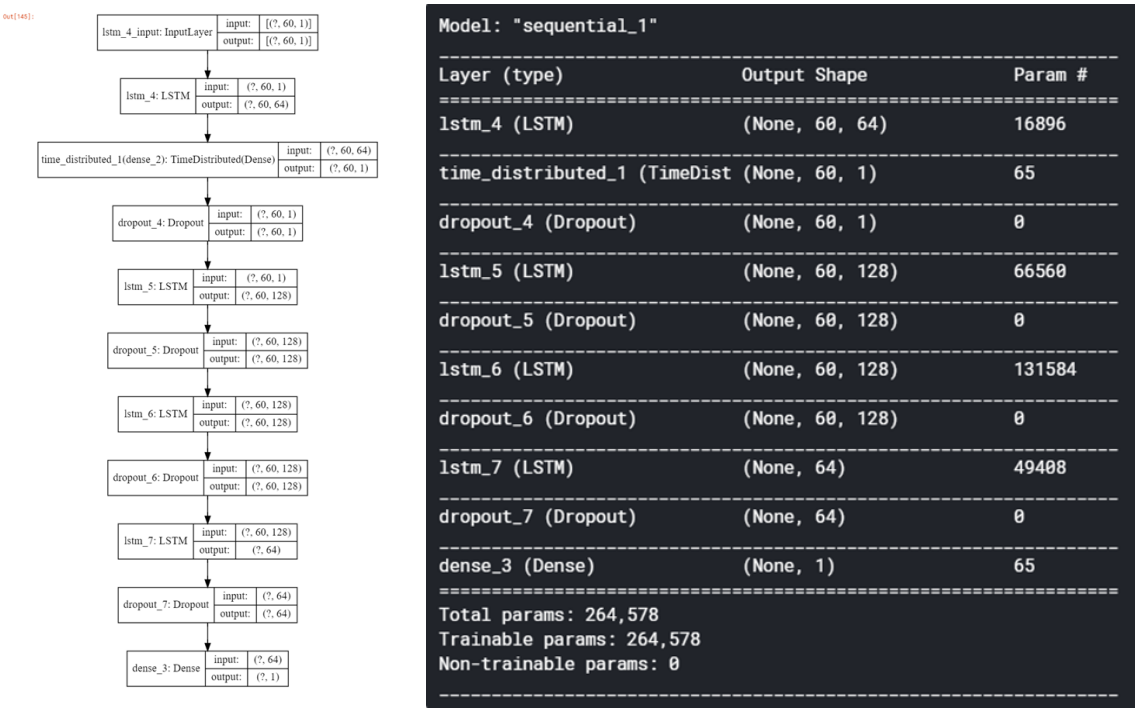


그림 10 연도별, 주별, 월별 미세먼지 변화양상

- 본 연구에서는 데이터에 대한 형태와 추세에 대한 분석을 토대로 딥러닝 모델을 통한 예측분석을 실행하였다. RNN(Recurrent Neural Networks, 순환신경망) 딥러닝 모델이 가지는, 시계열이 클 경우 제대로 된 학습이 이루어지지 않는 문제를 개선하기 위해 변형 모델 LSTM(Long Short Term Memory)을 채택하였다.




```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_4 (LSTM)                (None, 60, 64)            16896
_________________________________________________________________
time_distributed_1 (TimeDist (None, 60, 1)             65
_________________________________________________________________
dropout_4 (Dropout)          (None, 60, 1)             0
_________________________________________________________________
lstm_5 (LSTM)                (None, 60, 128)           66560
_________________________________________________________________
dropout_5 (Dropout)          (None, 60, 128)           0
_________________________________________________________________
lstm_6 (LSTM)                (None, 60, 128)           131584
_________________________________________________________________
dropout_6 (Dropout)          (None, 60, 128)           0
_________________________________________________________________
lstm_7 (LSTM)                (None, 64)                49408
_________________________________________________________________
dropout_7 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 65
=================================================================
Total params: 264,578
Trainable params: 264,578
Non-trainable params: 0
_________________________________________________________________
```

그림 11 학습 모델 구조도와 구조도표

- 그림 11은 해당 연구에서 사용된 Model의 구조도를 나타낸다. 모델 계층은 11층으로 구성하였다.

```
Epoch 43/50
68/68 [==============================] - 10s 144ms/step - loss: 0.0020 - mean_squared_error: 0.0020
Epoch 44/50
68/68 [==============================] - 11s 155ms/step - loss: 0.0019 - mean_squared_error: 0.0019
Epoch 45/50
68/68 [==============================] - 10s 154ms/step - loss: 0.0022 - mean_squared_error: 0.0022
Epoch 46/50
68/68 [==============================] - 10s 145ms/step - loss: 0.0019 - mean_squared_error: 0.0019
Epoch 47/50
68/68 [==============================] - 11s 156ms/step - loss: 0.0019 - mean_squared_error: 0.0019
Epoch 48/50
68/68 [==============================] - 10s 145ms/step - loss: 0.0019 - mean_squared_error: 0.0019
Epoch 49/50
68/68 [==============================] - 10s 144ms/step - loss: 0.0021 - mean_squared_error: 0.0021
Epoch 50/50
68/68 [==============================] - 10s 144ms/step - loss: 0.0018 - mean_squared_error: 0.0018
```

그림 12 모델 학습과정

- 그림 12는 모델의 학습과정을 나타내며, 반복적인 실험을 통한 경험식으로, Dense층을 삽입하고 계층을 11층까지 늘렸으며, 학습 반복 횟수는 약 50회로 조정하였다. 최종적으로 손실값이 0.0018을 가지는 모델 및 변수값을 선정하여 예측에 사용하였다.

그림 13 모델 검증결과

- 그림 13은 2020년 1월부터 5월까지의 데이터를 기반으로, 약 152일 분량의 데이터를 예측한 결과를 그래프로 도시화한 것이다.
모델이 전체적으로 실제값과 유사한 방향으로 예측을 진행하는 것을

알 수 있으며, 모델이 데이터를 의도에 맞게 학습하였는지 확인할 수 있었다.

그림 14-1,2 모델의 미세먼지 예측 결과

- 그림 14는 기존 학습된 모델을 토대로 실제적인 모델의 미래 미세먼지 예측의 결과를 그래프로 도시한 것이다.

모델의 미래 예측 방식을 다르게 하여, 두 종류의 결과값을 나타내었다. 그림 14-1의 경우 향후 30일간의 미세먼지 데이터를 모델에게 한 번에 예측하게 한 결과값이며, 그림 14-2는 기존 학습된 일별 데이터(T)를 기반으로 다음의 예측값(t)을 계산하고, 예측값(t)과 기존 학습된 데이터(T)를 결합한 후 그 다음날의 예측값을 도출하는, 이 과정을 지정된 날짜까지 반복하는 시퀀스 이기 때문이다.

(T-1+t -> t+1 ====> T-2+t+[t+1] ====> 반복)

Ⅳ. 결과 및 기대효과

최근 대두되고 있는 인공지능을 이용한 미래가치의 예측은 기존 인터넷 상에 퍼져있는 방대한 양의 데이터에 대해 다른 시각으로 접근하거나 새로운 가치를 창출해낼 수 있는 도구이다.

본 연구에서는 환경 빅데이터 플랫폼에 존재하는 화력발전소 명과 주소, 좌표를 데이터베이스화 하여, 1차적으로 오픈 플랫폼인 QGIS프로그램을 기반으로, GIS 기법을 이용한 전국 화력발전소 데이터를 시각화하고 데이터셋을 분리 및 정제하였다. 2차적으로는 기존의 방대한 시계열 데이터에 대한 한계를 가지고 있는 RNN 모델의 개선버전인 LSTM을 이용하여 기존 GIS 데이터 내에서 파생된 특정 화력발전소 인근의 미세먼지 데이터를 학습시켜 선정지역의 미세먼지량을 예측하였다.

2020년 5월까지의 데이터를 예측한 바에 대한 검증 결과 모델은 의도에 적합한 학습 결과를 내보였으며, 이를 기반으로 6월의 가상 예측 결과를 도출했다.

아직 딥러닝 인공신경망은 아직까지도 불완전한 요소가 많으나, 이미지 분류모델인 CNN의 등장 이후 폭발적인 성장세를 보이는 중이다. 사회 전 분야에서 딥러닝을 활용한 데이터 분석이 이루어지고 있고, 딥러닝의 가능성에 투자가 이루어지고 있다.

에너지경제연구원의 「세계 에너지수급 현황(2016년) 및 구조변화 분석」에 의하면 세계 에너지생산에서 화석에너지의 비중은 81.1%에 해당한다. 환경 문제에 대한 관심이 증가함에 따라, 환경과 결합한 빅데이터 활용 및 분석에 대한 수요가 증대될 것으로 예상된다.

본 연구에서 진행된 환경 빅데이터를 이용한 GIS 및 인공지능 모델의 개발 및 적용 프로세스는 대기 뿐만 아닌 수자원, 식생, 지질 등의 다양한 환경 분야에 대한 다양한 가치를 창출해 낼 수 있을 것이다.