기상요소 분석을 통한 초미세먼지 예측 모델 구현

1. 이호은(19951205)
2. ghdms1205@naver.com
3. 이호은(19951205)/유동관(19920928)/임연수(19951019)

2018년 상반기 사람들에게 가장 뜨거웠던 이슈는 무엇일까? 아래는 구글코리아에서 발표한 2018년 상반기 인기 검색어 종합 순위이다.

2018년 상반기 한국 인기 검색어 종합 순위			
1	미세먼지	11	윤식당2
2	신과 함께	12	곤지암
3	하트시그널 시즌2	13	안희정
4	블랙 팬서	14	나의 아저씨
5	NBA	15	일곱 개의 대죄
6	나 혼자 산다	16	가상화폐
7	조민기	17	어벤져스: 인피니티 워
8	외모지상주의	18	배틀그라운드
9	평창 동계올림픽	19	야생의 땅: 듀랑고
10	토르: 라그나로크	20	천애명월도

사진 출처: https://news.joins.com/article/22766496

국내 여러 사람을 울고 웃게 했던 비트코인? 축구 팬들의 마음을 뒤흔들어 놓았던 러시아 월드컵? 올해 상반기 사람들의 가장 뜨거웠던 이슈는 다름 아닌 '**미세먼지**'였다. 최근 들어 미세먼지의 심각성이 대두되면서 마스크를 끼고 다니는 사람들이 부쩍 늘어났다. 우리는 이러한 관심이 매해 반복되리라 판단하고, 다양한 분석이 필요할 것으로 생각하였다.

이미 미세먼지에 대한 다양한 요인이 있다. 중국에서 넘어온 먼지가 문제가 된다는 설이 존재하고, 국내 무분별한 개발, 심지어 고등어가 문제가 된다는 설까지도 존재한다. 하지만 현재까지 과학적으로 정확한 이유가 밝혀지지는 않았다. 이에 우리는 국내 **기상요소**들을 분석하여 어떤 요소가 미세먼지 농도에 영향을 미치는지 확인해보기로 하였다.

먼지의 지름이 2.5㎛ 이하에 속하는 이른바 **초미세먼지**는 호흡기 가장 깊은 곳까지 침투할 수 있어 폐암의 원인이 될 수 있다. 질병관리본부에 따르면 초미세먼지 농도가 10㎍/㎥ 높아질 때마다 폐암 발생률은 9%씩 증가한다고 한다. 따라서 우리는 세부적으로 초미세먼지에 초점을 맞추어 데이터 분석을 진행하기로 했다.

분석에는 한국 통계청과 한국환경공단 에어코리아에서 제공하는 2016~2018년도 데이터를 이용하였다.

- 데이터 출처 :
http://sts.kma.go.kr/jsp/home/contents/statistics/newStatisticsSearch.do?menu=SFC&MNU=MNU
http://www.airkorea.or.kr/realSearch
- 2016년 1월 1일 00시 ~ 2018년 7월 23일 23시 서울 종로구(기상청 위치)에서의 관측값 (1시간 단위)
- 21793개의 날짜, 시간별 데이터

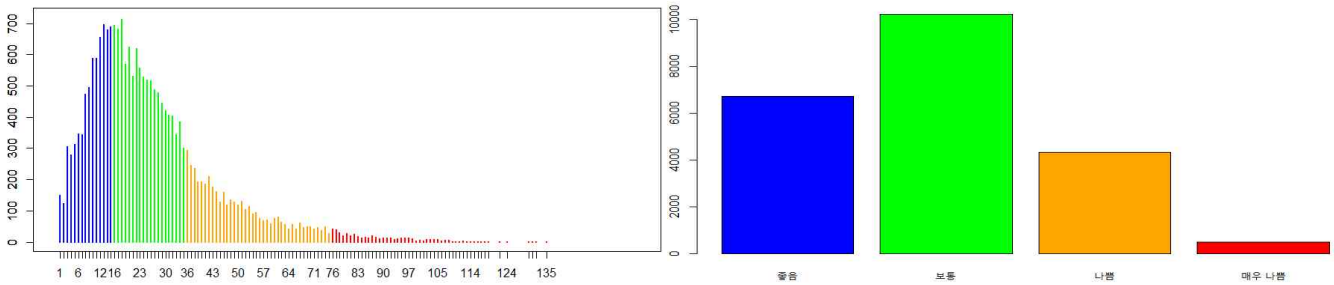
변수 이름과 설명

1. 기온(℃) : 대기의 온도를 말하며, 일반적으로 지면으로부터 1.2~1.5 m 정도 높이의 온도를 말한다.
2. 이슬점 온도(℃) : 습한 공기를 냉각시켜 가면 공기 중의 수증기는 어느 온도에서 응결하기 시작하여 이슬을 맺는데, 이때 온도를 이슬점 온도(노점,露點)라고 한다.
3. 시간 강수량 : 강수량이란 비나 눈, 우박 등과 같이 내린 강수의 양을 말한다. 어느 기간에 내린 강수가 땅 위를 흘러가거나 스며들지 않고, 땅 표면에 괴어 있다는 가정 아래 그 괸 물의 깊이를 측정한다. 눈·싸락눈 등 강수가 얼음이면 이것을 녹인 물의 깊이를 측정한다.
4. 풍향(deg) : 바람이 불어오는 방향. 보통 북 · 북동 · 동 · 남동 · 남 · 남서 · 서 · 북서의 8방위, 더욱 상세하게는 그 중간을 포함한 16방위로 나타낸다. 북의 방향을 0, 동을 90, 남을 180 라는 식으로, 북의 방향을 기준으로 하고, 시계방향으로 10마다 각도로 나타내기도 한다.
5. 풍속(m/s) : 대기의 수평적인 흐름의 속도를 풍속이라고 한다. 풍속은 지면으로부터 높이에 따라 다르므로 지상 10m에서 풍속을 표준으로 한다.
6. 해면 기압(hPa) : 어떤 관측지점에서 관측된 관측소 기압에 해면경정을 하여 얻은 평균해수면 높이에서의 기압을 말한다.
7. 수증기압(hPa) : 대기 중에 포함된 수증기만의 압력을 수증기압 또는 수증기장력이라고 한다.
8. 습도(%) : 수증기 함유량에 관한 대기의 상태를 말한다. 일반적으로 습도라 하면 상대습도를 뜻하며, 이 밖에 혼합비, 절대습도, 수증기압, 비습, 이슬점, 온도 등의 습도 표현방법이 있다.
9. 전운량(1/10) : 하늘을 덮고 있는 구름의 양의 비율을 말한다. 하늘 전체(눈에 보이는 범위)의 몇 %쯤이 구름으로 덮여 있는지에 따라 0부터 10까지의 계급으로 나누고 있다.
10. 일사(MJ/m²) : 지표면에 도달한 태양복사에너지를 말한다. 태양복사는 대기를 통과하는 동안에 공기 분자·먼지·수증기 등에 의하여 줄어지는데, 대기 중의 어느 한 점 또는 지표의 어느 한 점에서 받는 태양복사를 가리킨다.
11. 일조(hr) : 태양 광선이 구름이나 안개로 가려지지 않고 땅 위를 비추는 것을 말하며, 실제로 비친 시간을 일조시간이라고 한다. 일조시간은 보통 1일이나 한 달 동안에 비친 총시간 수로 나타낸다.
12. PM2.5(㎍/㎥) : 입자의 크기가 2.5㎛ 미만인 먼지를 말한다. 이것을 초미세먼지라고 한다.

초미세먼지 데이터 분석

기상청에 따르면 초미세먼지 농도를 0~15는 좋음, 16~35는 보통, 36~75는 나쁨, 76 이상부터는 매우 나쁨으로 분류한다. 보통에 위치하는 데이터가 가장 많고 매우 나쁨은 아주 적은 양만 차지한다.

아래의 표는 초미세먼지의 농도의 분포표이다.



정도	좋음	보통	나쁨	매우 나쁨
일수	6731	10227	4326	509
(%)	30.9	46.9	19.9	2.3

초미세먼지의 농도가 매우 좋을 때와 매우 나쁠 때의 대기요소 차이

```
print(df.isnull().sum()/len(df))
```

```
temp      0.000178
esul      0.000535
rain      0.000000
windDir   0.001961
windSpeed 0.001961
sea       0.000178
press     0.000535
humid     0.000936
cloud     0.219662
lsa       0.451981
lzo       0.451803
PM2.5     0.028789
dtype: float64
```

왼쪽의 표는 각 요소의 NULL 값을 조사한 결과이다. 일사와 일조의 경우 약 45%의 데이터가 존재하지 않고 전운량의 경우에도 약 20%의 데이터가 존재하지 않으므로 이러한 데이터들을 모두 삭제할 경우 절반에 가까운 데이터를 삭제해야 하므로 해당 요소들을 제외하고 결측값을 없앤 뒤 21719개의 데이터로 진행하였다.

초미세먼지의 농도가 좋을 때와 나쁠 때 대기요소가 어떻게 다른지 알아보기 위해 일부 표본의 데이터를 비교하여 보았다. 초미세먼지의 농도가 1로 아주 좋을 때와 130 이상으로 몹시 나쁠 때의 데이터 5개를 비교한 결과는 아래와 같다.

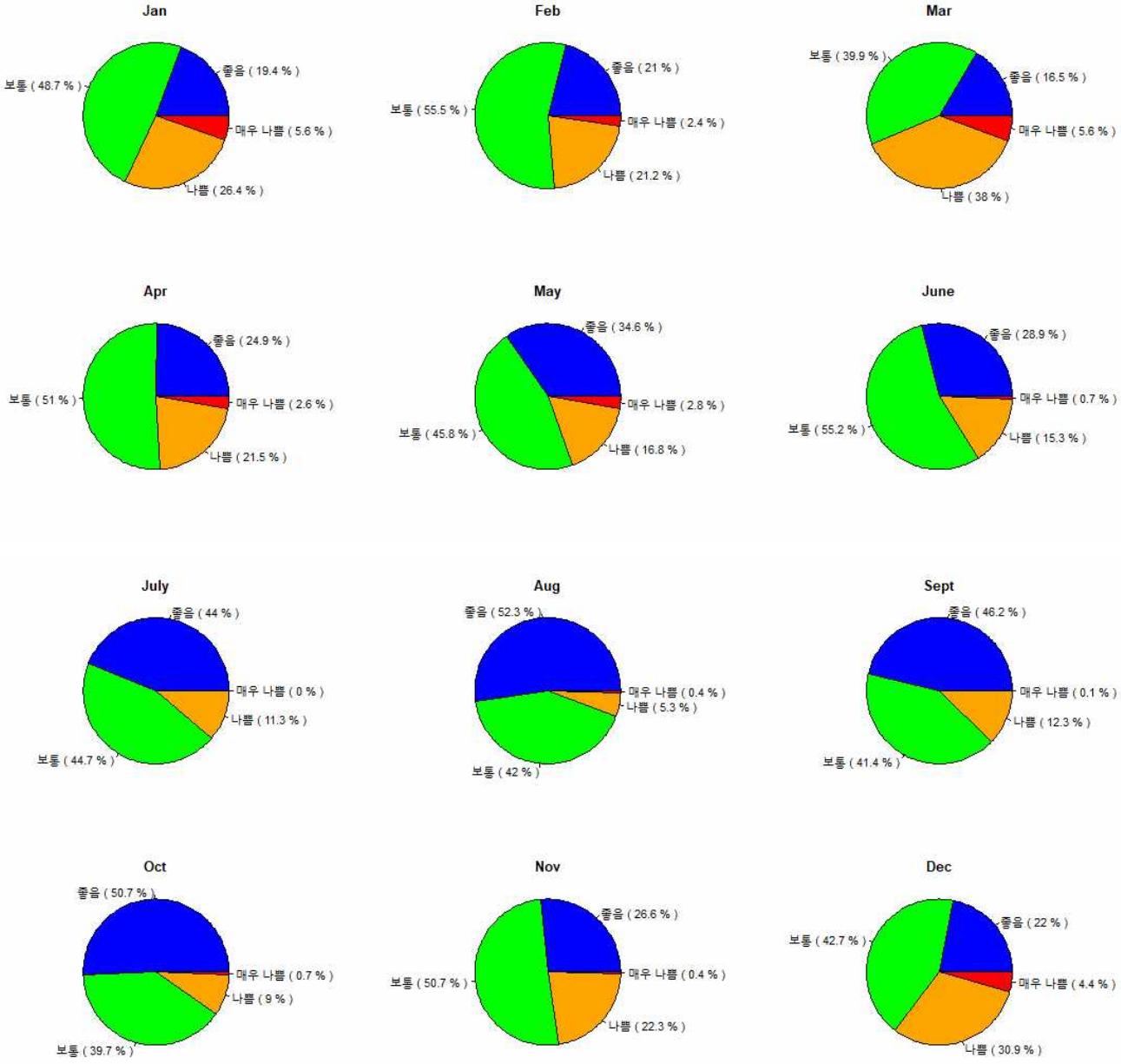
	일시	기온	이슬점	시간 강수량	풍향	풍속	해면 기압	증기압	습도	PM2.5
좋음 (PM2.5 = 1)	2016-02-15 10	-7.8	-22.0	0.0	340	3.8	1023.8	1.1	31	1
	2016-02-13 06	10.1	10.0	1.7	140	2.4	1012.5	12.3	100	1
	2016-05-24 09	16.4	15.4	4.0	250	3.2	1006.4	17.5	94	1
	2016-02-21 00	-2.4	-18.4	0.0	340	3.6	1026.0	1.4	28	1
	2016-07-01 17	23.7	21.9	1.0	200	3.1	1002.0	26.3	90	1
나쁨 및 매우 나쁨 (PM2.5 >= 130)	2016-02-09 00	1.2	-4.8	0	320	4.5	1014.4	4.3	64	131
	2016-04-09 16	13.9	9.9	0	290	2.1	1012.0	12.2	77	131
	2016-06-24 04	22.7	20.9	0	160	1.8	1002.1	24.7	90	130
	2016-08-10 07	26.2	23.0	0	200	0.7	1010.0	28.1	83	135
	2018-01-20 21	2.1	-7.7	0	340	2.8	1023.5	3.4	48	132

이 표에서 기상요소들을 비교하였을 때 강수량이 가장 눈에 띄었다. 초미세먼지 농도가 좋을 때는 강수량이 0보다 큰 경우가 잦았던 반면, 나쁠 때는 모두 강수량이 0이었다. 그 이외의 경우는 두드러지게 나타나는 특징이 보이지 않으므로 다양한 그래프를 통해 더 자세히 알아보겠다.

시간에 따라 초미세먼지의 농도에 어떤 변화가 있을까?

월별 및 시간대별로 비율을 분석하기 위해 R의 pie 함수를 활용하였다.

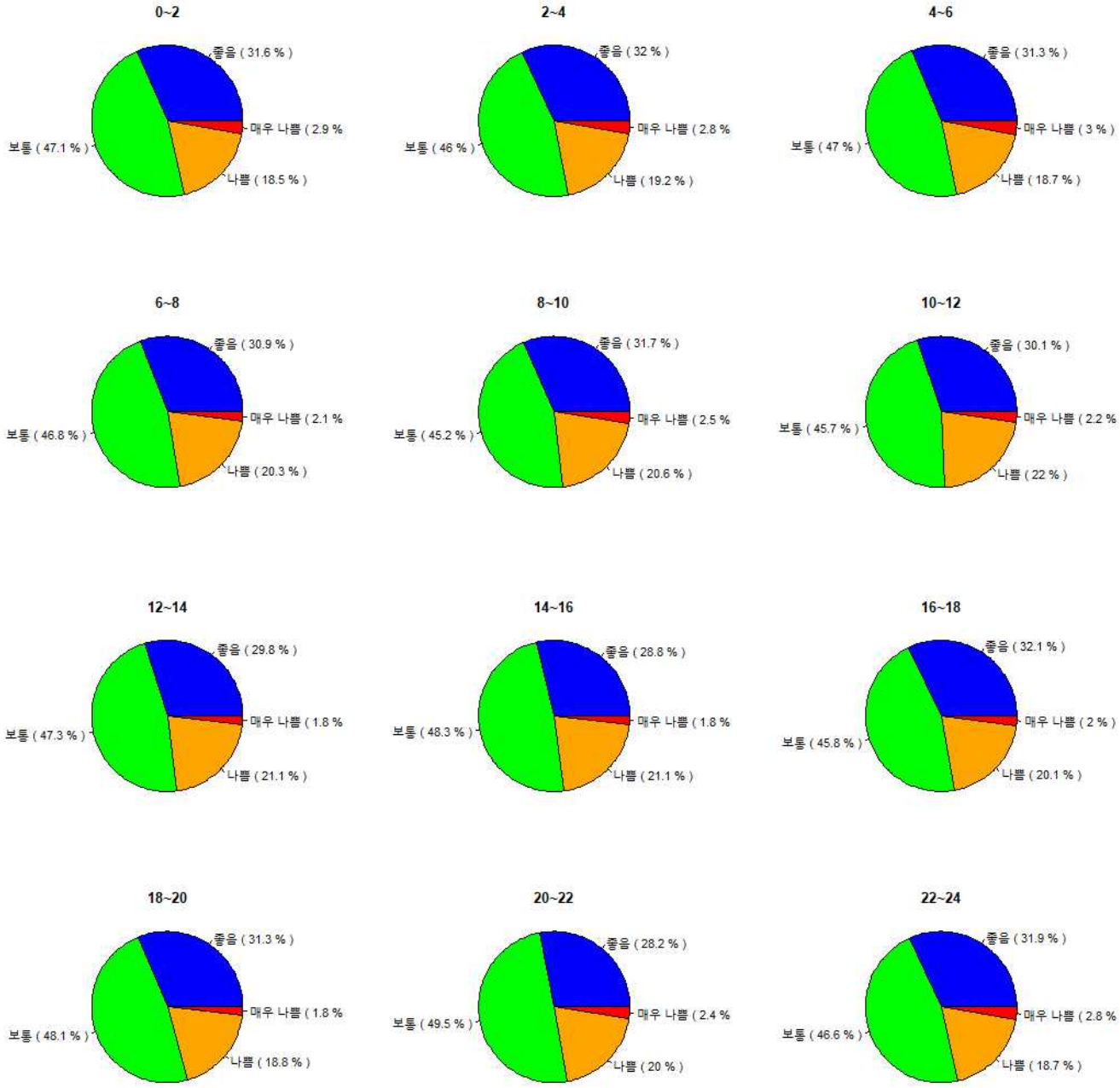
1) - 데이터 : 1월부터 12월까지 12개로 분류
- 척도 : PM2.5의 값에 따라 좋음, 보통, 나쁨, 매우 나쁨 4가지 등급으로 분류



6월부터 10월까지는 대체로 초미세먼지 농도가 좋지만 11월부터 5월까지는 비교적 나쁜 것을 알 수 있다. 즉, 여름과 가을, 겨울과 봄으로 비교가 가능한 것을 볼 수 있고, 초미세먼지가 기온, 기압, 날씨 등과 영향이 있다고 판단하고 더 나아갈 수 있었다.

2)　　- 데이터 : 0시부터 23시까지 2시간 단위로 12개로 분류

- 척도 : PM2.5의 값에 따라 좋음, 보통, 나쁨, 매우 나쁨 4가지 등급으로 분류

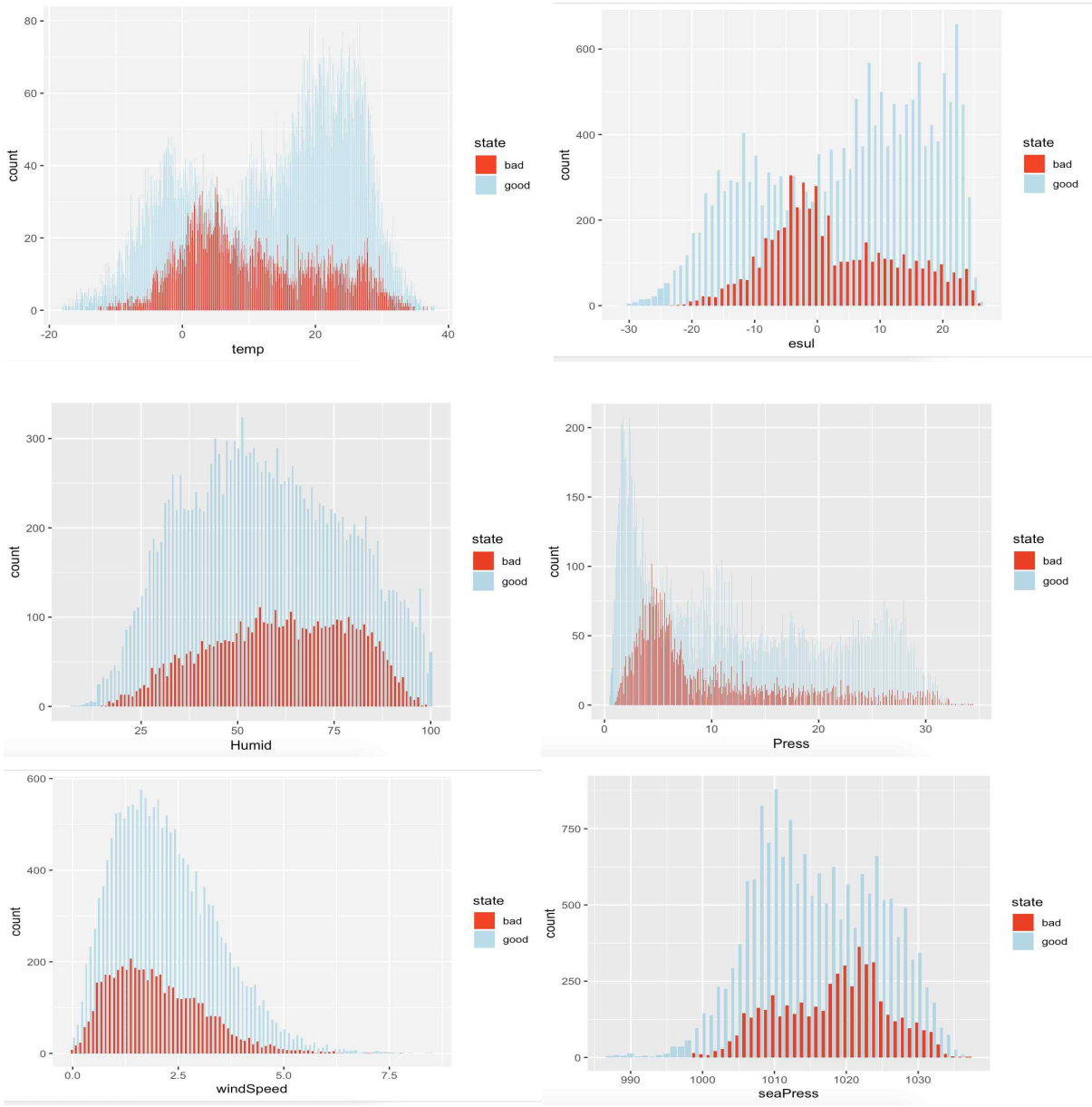


비교 결과 나쁨과 매우 나쁨의 비율이 가장 작은 시간대는 18~20시로 20.6%였고, 가장 큰 시간대는 10~12시로 24.2%였다. 최대와 최소의 차이가 크지 않은 걸 보았을 때 시간대와 초미세먼지는 큰 상관관계가 있지 않을 것 같다고 판단하였다.

기상요소 개별적으로 초미세먼지와 어떤 상관관계가 있을까?

쉬운 비교를 위해 데이터를 두 집단으로 나눈 다음 R의 ggplot 함수와 geom_bar, scale_fill_manual 함수를 이용하여 기온, 이슬점, 습도, 증기압, 풍속, 해면 기압의 그래프를 그려보았다.

- 초미세먼지 좋음 : PM2.5 수치가 35 이하인 경우
- 초미세먼지 나쁨 : PM2.5 수치가 35 초과인 경우



두 집단의 데이터 비율이 3:1 정도인 것을 고려해 보았을 때, 그래프의 크기 차이보다는 모양에 중점을 두고 차이를 판단하는 것이 옳다고 판단하였다. 기온과 이슬점인 경우 차이가 잘 보이고, 습도와 해면 기압에서도 어느 정도 차이가 드러나는 것으로 보아 초미세먼지 농도의 구분 기준으로 적합해 보인다.

초미세먼지 농도의 구분하는 데 큰 영향을 미치는 기상요소는 무엇인가?

먼저 두 가지로 분류했던 집단의 각 평균을 이용하여 그 차이를 변수별로 비교하여 보았다.

	좋은 초미세먼지 집단 평균	나쁜 초미세먼지 집단 평균	평균의 차이
기온	14.28	10.68	3.6
이슬점	5.015	2.93	2.085
시간 강수량	0.2674	0.0659	0.1015
풍속	2.343	2.014	0.329
해면 기압	1015	1018.4	-3.4
증기압	11.97	9.439	2.531
습도	57.44	61.63	-4.19

평균의 차이가 가장 큰 변수는 습도이고, 가장 작은 변수는 시간 강수량이다. 그러나 각 변수의 측정단위가 다르므로 아직 어떤 변수가 초미세먼지 농도에 영향을 준다고 단언할 수는 없다.

특정 변수가 두 집단을 얼마나 잘 나누는지를 더 정교하게 판단하기 위해 수치의 차이를 이용하기보다는 부등호를 이용하여 상대적으로 비교하는 방법을 생각해 보았다. 일관된 부등호 방향을 가지고 있다면 좋은 기준이라고 할 수 있겠다. 좋은 집단과 나쁜 집단의 쌍을 만들어 각각의 기상요소 수치를 비교하였다.

	good >= bad	good < bad	good >= bad 비율	good < bad 비율
기온	18465717	11918114	60.8%	39.2%
이슬점	17573400	12810431	57.8%	42.2%
시간 강수량	29553774	830057	97.2%	2.8%
풍속	17879495	12504336	58.8%	41.2%
해면 기압	11417459	18966372	37.6%	62.4%
증기압	17602200	12781631	58.0%	42.0%
습도	13505458	16878373	44.4%	55.6%

두 그룹에 대해 가장 부등호가 일관되게 나타난 변수는 시간 강수량이다. 단 2.8%만이 좋은 집단보다 나쁜 집단의 강수량이 높았으며, 이는 임의의 데이터의 초미세먼지 농도를 강수량으로 판단하면 대부분은 맞출 수 있다는 것을 의미한다. 그리고 해면 기압과 기온이 각각 62.4%, 60.8%로 어느 정도 일관된 부등호를 갖고 있었다.

앞에서 시간 강수량의 평균 차이는 0.1015로 가장 작은 차이를 보였으나 실제 초미세먼지 농도를 구분하는 데는 가장 중요한 변수라는 것을 알 수 있다. 추가로 이슬점의 경우 히스토그램으로 보았을 때와 달리 비율의 차이가 비교적 작게 나타났다.

초미세먼지 농도 등급 판별 모형

1) 전체 data로 4개의 class로 구별

- training data : 전체 data의 70%
- test data : 전체 data의 30%

지금까지 조사한 결과를 참고하였을 때 초미세먼지의 농도를 평가하기 위한 가장 좋은 기상요소는 시간 강수량이다. 그렇다면 여러 기상요소를 모두 고려한 구분 기준을 만들어 보기로 한다.

먼저 초미세먼지 농도 수치를 예측하는 Linear Regression model을 만들어 보았다. training 데이터로 학습시킨 모델의 예측값과 test 데이터가 평균적으로 두 배 정도 차이가 났기 때문에 이 모델을 사용하여 초미세먼지의 정확한 수치를 예측하는 것은 불가능하다고 판단했다.

문제를 단순화하여 기상청에서 정의한 초미세먼지 농도의 4가지 등급을 예측하는 모델을 제작했다. 첫 번째로 Softmax Cross Entropy 모델을 만들었다. 먼저 좋음부터 매우 나쁨까지의 등급을 0~3으로 정의하였다. training 데이터를 이용해서 모델을 학습시킨 후 test 데이터를 사용해 검증하였다. 학습결과 이 모델은 약 49.3%의 정확도를 보였다.

```
Step:  1800     Loss: 1.075     Acc: 49.32%
Step:  1900     Loss: 1.075     Acc: 49.30%
Step:  2000     Loss: 1.075     Acc: 49.28%
```

두 번째로 xavier를 사용한 nn Softmax cross entropy 모델을 사용하여 초미세먼지 농도의 4가지 등급을 구별했다. 이 모델로 시험하였을 때 약 54.7%의 정확도를 기록하였다.

```
Learning Finished!
Accuracy: 0.5471455
```

2) 초미세먼지 농도가 좋음, 나쁨, 매우 나쁜 집단만 추출하여 2개의 class로 구별

- training data : 추출한 data의 70%
- test data : 추출한 data의 30%

앞서 사용한 모델들 모두 너무 낮은 정확도를 기록하였으므로 이들을 좋은 모델로 평가할 수 없다. 위의 모델들에서 만족할만한 결과를 얻어내지 못했기 때문에 문제를 더 간단히 해 정확도를 높여보기로 한다.

좋음 등급의 초미세먼지 집단과 나쁨과 매우 나쁨 등급의 초미세먼지 집단을 데이터로 사용하여 초미세먼지 농도를 좋음과 나쁨으로 구별하는 모델들을 만들었다. 첫 번째로 Logistic Regression 모델을 만들었다. 그 결과 평균적으로 67.6% 정도의 정확도를 기록하였다.

```
Accuracy:  0.67640257
```

두 번째로 xavier를 사용한 nn Softmax Cross Entropy 모델을 사용하였다. 앞에서와 마찬가지로 데이터를 구분하고 모델을 학습시켜 시험한 결과 평균적으로 78% 정도의 정확도를 기록하였다.

```
Learning Finished!
Accuracy: 0.7802198
```

Neural Network 모델을 이용하면 예측력이 얼마나 될까?

R에서 제공하는 기계 학습의 대표적인 모델인 신경망 모델(nnet 패키지, NeuralNetTools 패키지)을 활용하여 예측력을 검사해보았다. PM2.5 값이 35 이하인 그룹과 36 이상인 그룹으로 이분화하여 classification으로 적용해보았다.

```
samp <- c(sample(1:4343,3040), sample(4344:8686,3040), sample(8687:13030,3041),
          sample(13031:17374,3041), sample(17375:21719,3041))
pm.nn <- nnet(PM2.5 ~ temp+esul+rain+windSpeed+Press+Humid+seaPress,
              data = data, subset = samp,
              size = 8, rang = 0.1, decay = 5e-04, maxit = 2000)
```

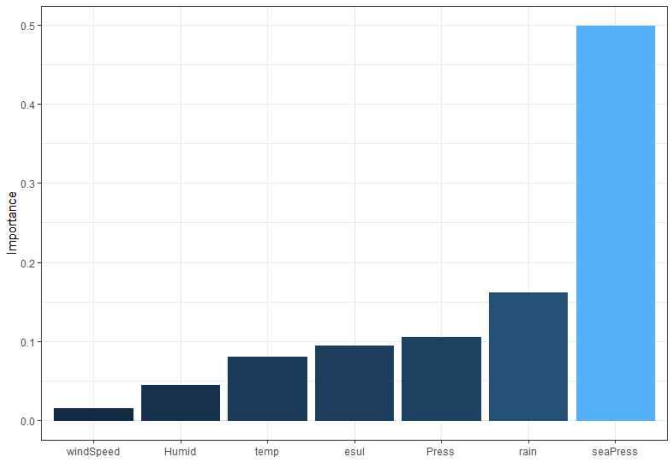
이는 nnet함수를 사용한 부분의 코드인데, training set을 전체 data의 70%로 설정하였고 구간을 5개로 나눠 고르게 표본을 추출하였다. 그리고 모델링에서 모든 요소를 넣었으며 신뢰도와 over-fitting을 조절하기 위해 rang 값과 decay 값을 조절하였다.

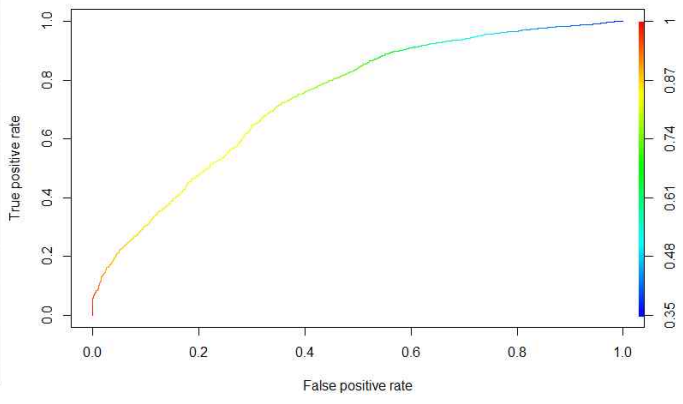
그리고 나머지 30%의 test set으로 예측을 한 결과 중 예측력이 높게 나온 3가지를 뽑아보았다. 경우에 따른 사건의 수를 알아보기 위해 predict 함수와 table 함수를 이용하였고 정보의 시각화를 위해 garson 함수와 ROCR 패키지의 prediction 함수와 performance 함수를 이용하였다.

1) 정확도 0.8012584

	0	1
0	**374**	1063
1	232	**4847**

실제 test set과 예측값을 비교했을 때, 맞은 경우는 5221개이고, 나쁜 초미세먼지로 예측했지만, 실제 좋은 농도인 경우가 1063개, 반대의 경우가 232개로 나타났다.



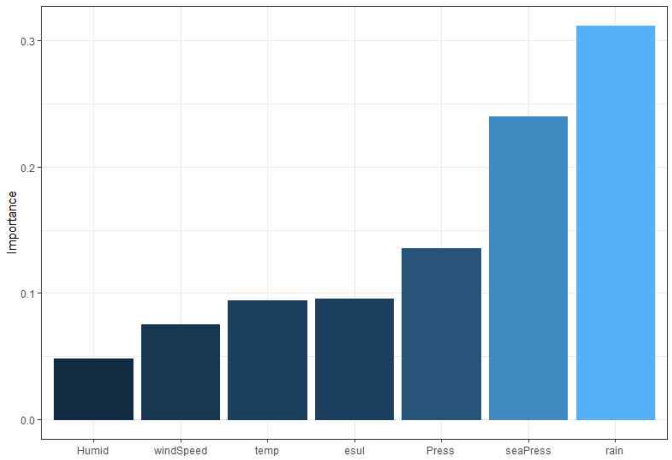


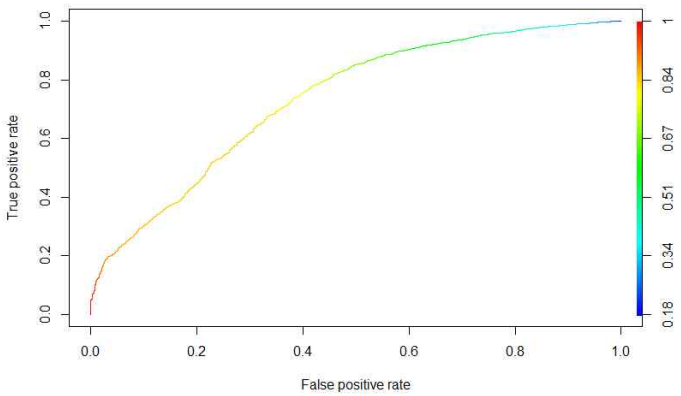
이 경우 변수의 중요도가 높은 순으로 나열하면 해면 기압 - 시간 강수량 - 증기압 - 이슬점 - 기온 - 습도 - 풍속이 된다. 그리고 plot 함수로 비율에 대한 그래프를 출력해보니 기울기가 1인 직선 그래프와 차이가 있는 걸 볼 수 있다.

2) 정확도 0.8025165

	0	1
0	**337**	1071
1	216	**4893**

이 경우 1)보다 0.13% 정도 높은 예측력을 보인다. (0, 1)의 경우가 8번 더 많지만 (1, 0)의 경우에서 16번의 감소가 보이는 걸 알 수 있다.



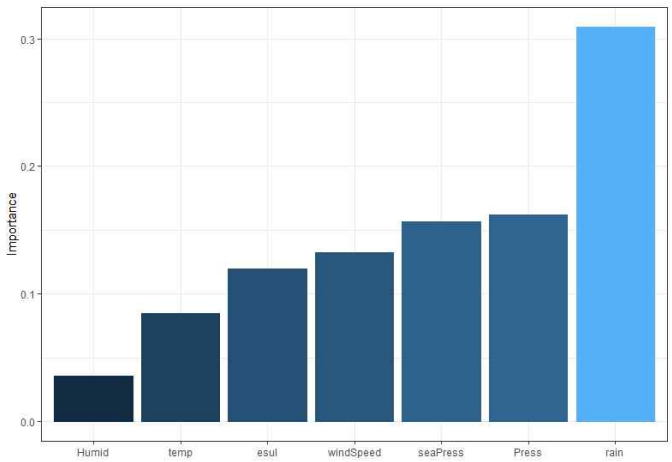


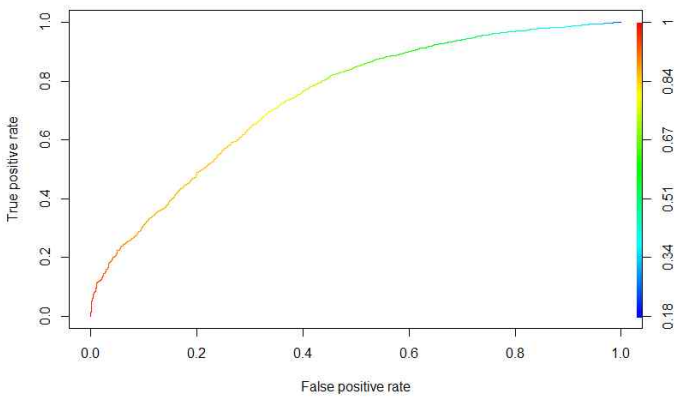
이 경우 변수의 중요도가 높은 순서는 시간 강수량 - 해면 기압 - 증기압 - 이슬점 - 기온 - 풍속 - 습도이다. 마찬가지로 비율의 그래프는 비슷한 모양을 띠고 있는 것을 볼 수 있다.

3) 정확도 0.8035605

	0	1
0	**371**	1050
1	230	**4865**

이 경우 2)보다 0.10% 정도 높은 예측력을 보인다. (0, 1)의 경우 21번 더 적고 (1, 0)의 경우 14번의 증가가 보이는 걸 알 수 있다.





이 경우 변수의 중요도가 높은 순서는 시간 강수량 - 증기압 - 해면 기압 - 풍속 - 이슬점 - 기온 - 습도이다. 마찬가지로 비율의 그래프는 비슷한 모양을 띠고 있는 것을 볼 수 있다.

확률	1	2	3	4	5	6	7
0.8012584	해면 기압	시간 강수량	증기압	이슬점	기온	습도	풍속
0.8025165	시간 강수량	해면 기압	증기압	이슬점	기온	풍속	습도
0.8035605	시간 강수량	증기압	해면 기압	풍속	이슬점	기온	습도

위의 3가지 경우에서 변수 중요도를 합친 표이다. 대체로 시간 강수량, 해면 기압, 증기압의 경우 중요도가 높았고, 반대로 습도와 기온은 중요도가 비교적 낮은 것을 알 수 있다.
결과적으로 nnet를 활용하여 예측력을 80% 정도로 모델링을 할 수 있었다.

결론 및 한계점

모델	Logistic Regression	random forest	Softmax cross entropy	Neural Network
예측률	0.68	0.45	0.78	0.80

Logistic Regression과 random forest를 활용하여 구현했을 때에는 각각 0.68, 0.45 정도의 좋지 않은 결과가 나타났다. 그래서 더 다양한 모델을 해보기로 하였고, 그 결과 Softmax cross entropy 모델과 Neural Network 모델에서 0.8 내외의 준수한 결과를 낼 수 있었다. 이는 데이터셋에 있는 기상 요소만으로 정확한 PM2.5의 수치를 예상하기엔 무리가 있지만, 좋음과 나쁨 두 분류 중 하나로 판단할 때에는 어느 정도 높은 신뢰율이 있다는 것을 의미한다.

한계점으로는 전체적인 예측률은 높았지만, 앞서 말했듯이 자연적인 이유로 초미세먼지가 좋은 집단과 나쁜 집단의 비율이 3:1 정도여서 모델이 나쁜 초미세먼지로 예측했지만 실제로는 좋은 초미세먼지 농도일 경우가 많았다는 것이다. 또한, 앞서 시간 강수량의 부등호가 매우 일관되게 나타났지만 대부분 값이 0이어서 하나만으로 모델을 만들기에 어려움이 있어서 다른 요소도 포함하게 되었다.

2018-07-24 00시 ~ 2018-07-26 09시 데이터에 적용해보자

	Softmax cross entropy	Neural Network
데이터 총 개수	57	57
맞게 예상한 개수 (정확도)	49 (0.8596491)	48 (0.8421053)

새로운 데이터 총 57개에 구현했던 모델을 적용해본 결과, Softmax cross entropy 모델로는 57개 중 49개를 맞게 예상해서 0.8596491의 정확도가 나왔고 Neural Network 모델로는 57개 중 48개를 맞게 예상해서 0.8421053의 정확도가 나왔다.