미세먼지 예측을 위한 기상·지리 요인 기반 회귀 모델 성능 분석

32212293 신명섭 32212477 안준현 32214250 정형윤

목차

01 연구 배경 및 목적

미세먼지란?

초미세먼지(PM2.5)는 입경 2.5μm 이하의 대기 중 부유입자로, 자동차, 산업시설, 난방 등에서 발생하며 중국 등 국외 오염원이 편서풍을 타고 유입되기도 합니다. 입자가 매우 작아 폐 깊숙이 침투할 수 있어 호흡기 및 심혈관계 질환을 유발하며, WHO는 이를 조기사망의 주요 원인으로 보고하고 있습니다. PM2.5 농도는 기상 요인뿐만 아니라 지역의 고도나 해안 인접성 같은 지리적 특성에도 영향을 받을 수 있어, 본 연구에서는 기상 및 지리 요인을 결합한 예측 모델을 통해 변수별 영향력을 분석하고자 하였습니다.

02 분석 목표

1. 변수 조합에 따른 미세먼지 예측 비교 (A군 / B군)

A군: 기상 요소 (평균기온, 습도, 풍속)

B군: A군 + 지리 정보 (고도, 해안지역 여부)

→ 변수 조합에 따라 예측 정확도에 차이가 발생하는지 분석

2. 예측 모델별 성능 비교 분석

선형회귀: 변수 간 선형관계를 기반으로 한 예측

결정트리: 조건 분기 구조를 활용한 규칙 기반 예측

KNN: 거리 기반 최근접 이웃 알고리즘을 활용한 비모수 예측

→ 각 모델의 성능(RMSE, MAE 기준)을 정량적으로 비교

3. 변수 중요도 및 영향력 시각화

모델 기반 변수 중요도 분석 (결정트리 중심)

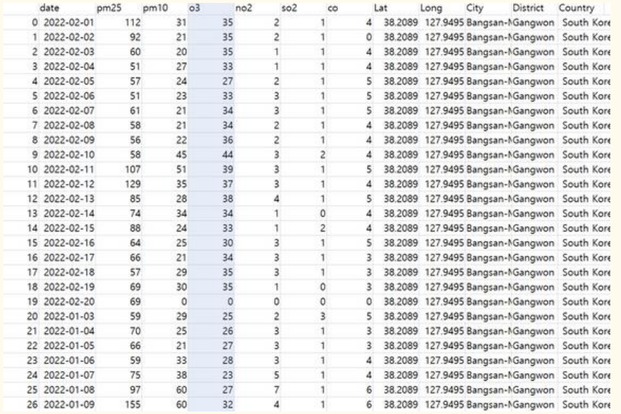
RMSE/MAE를 통한 예측 정확도 비교

산점도, 상관분석을 통해 변수 간 관계 구조 시각화

03 데이터 소개 및 수집 방법

데이터 1 대기오염 데이터(south-korean-pollution-data.csv)

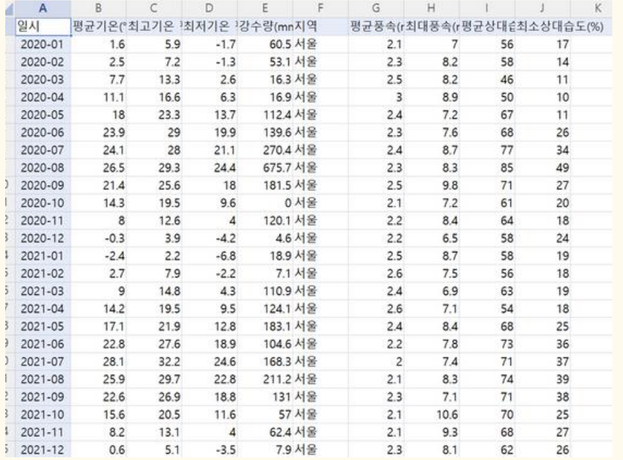
구성: 전국 10개 지역의 일별 PM2.5, PM10, O3, NO2, SO2, CO 등의 대기오염 수치
기간: 2020년 1월 ~ 2022년 12월
출처: kaggle



	date	pm25	pm10	o3	no2	so2	co	Lat	Long	City	District	Country
0	2022-02-01	112	31	35	2	1	4	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
1	2022-02-02	92	21	35	2	1	0	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
2	2022-02-03	60	20	35	1	1	4	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
3	2022-02-04	51	27	33	1	1	4	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
4	2022-02-05	57	24	27	2	1	5	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
5	2022-02-06	51	23	33	3	1	5	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
6	2022-02-07	61	21	34	3	1	5	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
7	2022-02-08	58	21	34	2	1	4	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
8	2022-02-09	56	22	36	2	1	4	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
9	2022-02-10	58	45	44	3	2	4	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
10	2022-02-11	107	51	39	3	1	5	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
11	2022-02-12	129	35	37	3	1	4	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
12	2022-02-13	85	28	38	4	1	5	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
13	2022-02-14	74	34	34	1	0	4	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
14	2022-02-15	88	24	33	1	2	4	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
15	2022-02-16	64	25	30	3	1	5	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
16	2022-02-17	66	21	34	3	1	3	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
17	2022-02-18	57	29	35	3	1	3	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
18	2022-02-19	69	30	35	1	0	3	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
19	2022-02-20	69	0	0	0	0	0	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
20	2022-01-03	59	29	25	2	3	5	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
21	2022-01-04	70	25	26	3	1	3	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
22	2022-01-05	66	21	27	3	1	3	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
23	2022-01-06	59	33	28	3	1	4	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
24	2022-01-07	75	38	23	5	1	4	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
25	2022-01-08	97	60	27	7	1	6	38.2089	127.9495	Bangsan-h	Gangwon	South Kore
26	2022-01-09	155	60	32	4	1	6	38.2089	127.9495	Bangsan-h	Gangwon	South Kore

데이터 2 기상 정보 데이터(기상정보(20~22).csv)

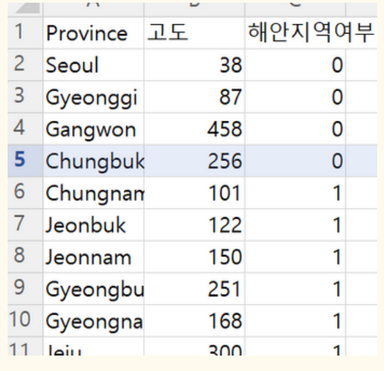
구성: 10개 도 단위 지역의 월별 평균기온(℃), 풍속(m/s), 습도(%), 지역, 일시(연도-월), 평균기온, 평균풍속, 평균상대습도
기간: 2020년 1월 ~ 2022년 12월
출처: 기상청



	A	B	C	D	E	F	G	H	I	J	K
	일시	평균기온(℃	최고기온	최저기온	강수량(mm	지역	평균풍속(r	최대풍속(r	평균상대	최소상대	습도(%)
	2020-01	1.6	5.9	-1.7	60.5	서울	2.1	7	56	17	
	2020-02	2.5	7.2	-1.3	53.1	서울	2.3	8.2	58	14	
	2020-03	7.7	13.3	2.6	16.3	서울	2.5	8.2	46	11	
	2020-04	11.1	16.6	6.3	16.9	서울	3	8.9	50	10	
	2020-05	18	23.3	13.7	112.4	서울	2.4	7.2	67	11	
	2020-06	23.9	29	19.9	139.6	서울	2.3	7.6	68	26	
	2020-07	24.1	28	21.1	270.4	서울	2.4	8.7	77	34	
	2020-08	26.5	29.3	24.4	675.7	서울	2.3	8.3	85	49	
	2020-09	21.4	25.6	18	181.5	서울	2.5	9.8	71	27	
	2020-10	14.3	19.5	9.6	0	서울	2.1	7.2	61	20	
	2020-11	8	12.6	4	120.1	서울	2.2	8.4	64	18	
	2020-12	-0.3	3.9	-4.2	4.6	서울	2.2	6.5	58	24	
	2021-01	-2.4	2.2	-6.8	18.9	서울	2.5	8.7	58	19	
	2021-02	2.7	7.9	-2.2	7.1	서울	2.6	7.5	56	18	
	2021-03	9	14.8	4.3	110.9	서울	2.4	6.9	63	19	
	2021-04	14.2	19.5	9.5	124.1	서울	2.6	7.1	54	18	
	2021-05	17.1	21.9	12.8	183.1	서울	2.4	8.4	68	25	
	2021-06	22.8	27.6	18.9	104.6	서울	2.2	7.8	73	36	
	2021-07	28.1	32.2	24.6	168.3	서울	2	7.4	71	37	
	2021-08	25.9	29.7	22.8	211.2	서울	2.1	8.3	74	39	
	2021-09	22.6	26.9	18.8	131	서울	2.3	7.1	71	38	
	2021-10	15.6	20.5	11.6	57	서울	2.1	10.6	70	25	
	2021-11	8.2	13.1	4	62.4	서울	2.1	9.3	68	27	
	2021-12	0.6	5.1	-3.5	7.9	서울	2.3	8.1	62	26	

데이터 3 고도 및 해안지역 여부 데이터(고도와 해안 정보.csv)

구성: 지역별 고도 정보(m)와 해안지역 여부(0/1), Province (지역), 고도, 해안지역 여부
출처: 공공데이터포털
특징: 해안지역 여부 이진변수 처리 내륙(0), 해안(1)



	Province	고도	해안지역여부
2	Seoul	38	0
3	Gyeonggi	87	0
4	Gangwon	458	0
5	Chungbuk	256	0
6	Chungnam	101	1
7	Jeonbuk	122	1
8	Jeonnam	150	1
9	Gyeongbu	251	1
10	Gyeongna	168	1
11	Jeju	300	1

04 데이터 전처리

1. 데이터 불러오기 및 병합

- 3개 파일 불러옴: 대기오염 데이터, 기상 정보, 고도·해안 여부
- 날짜 기준 통합을 위해 pollution과 weather 데이터에 month 파생

(일별 PM2.5 → 월평균으로 집계)

- District 이름 통일 (한글 → 영문으로 변환)

2. 변수 선택 및 정리

- 대기오염 데이터에서 PM2.5와 지역 정보만 선택(종속 변수: PM2.5)
- 기상 데이터에서는 월별 평균기온, 풍속, 습도 추출
- 고도와 해안 여부 데이터는 지역별로 병합 처리→ 총 5개 독립 변수 구성

3. 결측치 처리

- PM2.5, 기상, 지리 데이터 모두 결합 후 na.omit() 함수를 통해 결측 행 제거

4. 실험군 분리

- A군: 기상 변수만 사용 (평균기온, 풍속, 습도 → 3개 변수)
- B군: 기상 변수 + 지리 변수(고도, 해안 여부 포함 → 5개 변수)

  → 지리 정보 포함 여부가 예측 성능에 미치는 영향 평가

5. 데이터 분할 (Train/Test)

- 전체 데이터를 8:2 비율로 무작위 분할
- createDataPartition() 함수로 학습용/평가용 데이터 구분
- KNN 모델은 거리 기반 특성을 고려해 min-max 정규화 사전 적용

```r
# 월 정보 파생
pollution <- pollution %>% mutate(month = format(date, "%Y-%m"))


# 한글 District 열로 영문으로 바꾸기
weather <- weather %>%
  rename(District = 지역) %>%
  mutate(District = recode(District,
                           "서울" = "Seoul", "경기도" = "Gyeonggi", "강원도" = "Gangwon",
                           "충청북도" = "Chungbuk", "충청남도" = "Chungnam",
                           "전라북도" = "Jeonbuk", "전라남도" = "Jeonnam",
                           "경상북도" = "Gyeongbuk", "경상남도" = "Gyeongnam"
  ))

# 월 정보 생성
weather <- weather %>%
  mutate(date = as.Date(paste0(일시, "-01")),
         month = format(date, "%Y-%m")) %>%
  select(District, month,
         평균기온 = `평균기온(℃)`,
         풍속 = `평균풍속(m/s)`,
         습도 = `평균상대습도(%)`)


# 병합 수행
poll_weather <- pollution %>%
  select(-starts_with("평균기온"), -starts_with("풍속"), -starts_with("습도"),
         -starts_with("고도"), -starts_with("해안지역여부")) %>%
  left_join(weather, by = c("District", "month")) %>%
  left_join(geo, by = c("District" = "Province"))

glimpse(poll_weather)
names(poll_weather)

# ======================
# [5] 모델 학습 및 평가 - A군
# ======================
set.seed(42)
idx_A <- createDataPartition(df_A$pm25, p = 0.8, list = FALSE)
train_A <- df_A[idx_A, ]; test_A <- df_A[-idx_A, ]

lm_A <- lm(pm25 ~ ., data = train_A)
pred_lm_A <- predict(lm_A, test_A)
rmse_lm_A <- rmse(test_A$pm25, pred_lm_A)

tree_A <- rpart(pm25 ~ ., data = train_A)
pred_tree_A <- predict(tree_A, test_A)
rmse_tree_A <- rmse(test_A$pm25, pred_tree_A)

knn_A <- knn.reg(train = train_A[, -1], test = test_A[, -1], y = train_A$pm25, k = 5)
rmse_knn_A <- rmse(test_A$pm25, knn_A$pred)

# ======================
# [6] 모델 학습 및 평가 - B군
# ======================
set.seed(42)
idx_B <- createDataPartition(df_B$pm25, p = 0.8, list = FALSE)
train_B <- df_B[idx_B, ]; test_B <- df_B[-idx_B, ]

lm_B <- lm(pm25 ~ ., data = train_B)
pred_lm_B <- predict(lm_B, test_B)
rmse_lm_B <- rmse(test_B$pm25, pred_lm_B)

tree_B <- rpart(pm25 ~ ., data = train_B)
pred_tree_B <- predict(tree_B, test_B)
rmse_tree_B <- rmse(test_B$pm25, pred_tree_B)

knn_B <- knn.reg(train = train_B[, -1], test = test_B[, -1], y = train_B$pm25, k = 5)
rmse_knn_B <- rmse(test_B$pm25, knn_B$pred)
```

05 모델별 RMSE 성능 비교

- **선형회귀**

RMSE: A군 28.1 → B군 27.7 MAE: A군 21.7 → B군 21.4
(지리 변수 추가시 예측 성능 소폭 개선)

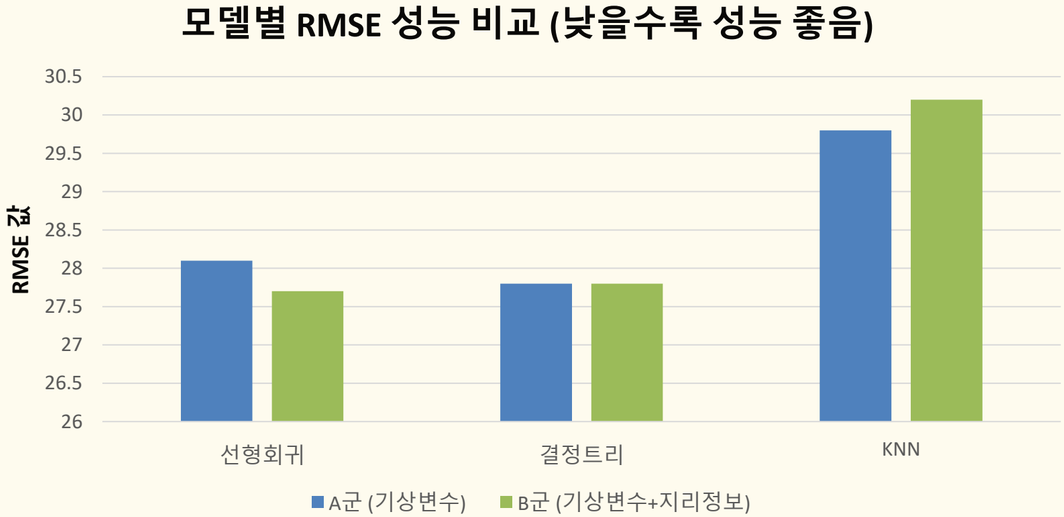
- **결정트리**

RMSE: A군 27.8 = B군 27.8, MAE: A군 21.5 = B군 21.5
(지리 정보의 영향 거의 없음 성능 지리 변수 추가 효과 미미)

- **KNN(정규화 포함)**

RMSE: A 30.1 → B 30.3, MAE: A 22.9 → B 23.1
(모든 지표에서 B군 성능 하락, 거리 기반 모델은 고도·해안 여부와 같은 변수에 민감)

- **지리 정보의 효과는 모델에 따라 다르게 작용**
- **선형회귀에서는 유의미한 개선이 있었으나, KNN에서는 오히려 예측 성능이 악화**




```
# A tibble: 3 × 5
  모델       A_RMSE A_MAE B_RMSE B_MAE
  <chr>       <dbl> <dbl>  <dbl> <dbl>
1 선형회귀     28.1  21.7   27.7  21.4
2 결정트리     27.8  21.5   27.8  21.5
3 KNN          30.1  22.9   30.3  23.1
```

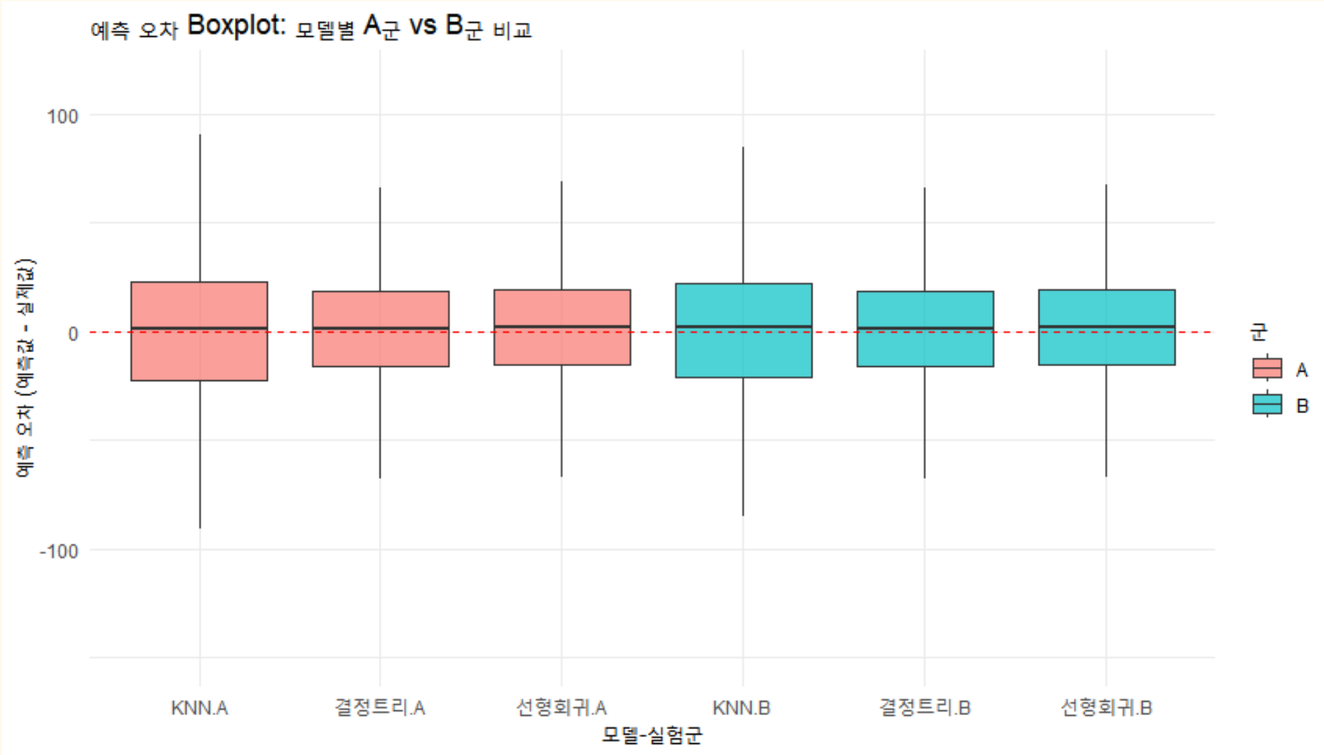
06 모델별 예측 오차 분포 비교 (RMSE 보조 분석)

모델별 예측 안정성 및 중심성 평가

선형회귀: 예측 오차 분산이 작고, 오차의 중심이 0에 가까워 비교적 안정적인 예측을 보임

결정트리: 선형회귀와 유사하게 오차 분산이 작고, 중심이 0 부근에 위치함

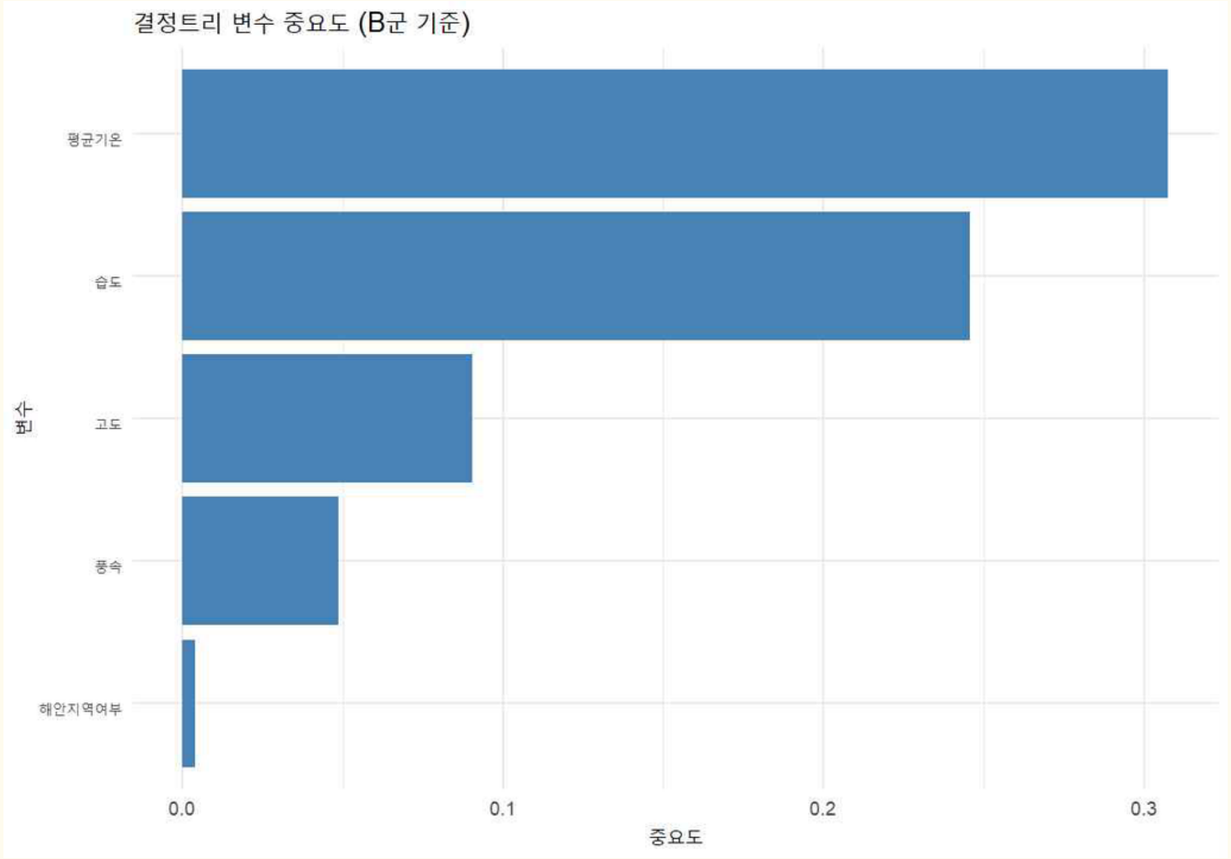
KNN: 예측 오차의 분산이 크고, 중심이 0에서 벗어난 경우 존재
(일관성이 낮고 예측 편향 발생 가능성 존재, 특히 거리 기반 특성상 고도·해안 변수의 영향력이 왜곡되어 나타날 수 있음)



07 변수 중요도 분석(결정트리 기준)

PM2.5 예측에 기여한 주요 기상·지리 요인 확인

- **평균기온**: 매우 높음 – PM2.5 예측에 가장 크게 기여한 변수
- **습도**: 매우 높음 – 평균기온 다음으로 중요한 예측 변수
- **고도**: 중간 – 상대적으로 낮지만, 지형적 특성에 따른 미세먼지 정체/확산 가능성을 시사
- **풍속**: 낮음 – 모델 내에서 기여도 미미, 다른 변수에 비해 분기 기준으로 자주 선택되지 않음
- **해안지역 여부**: 매우 낮음 – 이진 변수이며, 결정트리에서는 정보 이득이 낮아 분기 조건으로 활용 빈도 거의 없음
- 주요 예측 변수는 평균기온과 습도로, 기상 요인의 영향이 지리적 요인보다 훨씬 강하게 나타났음



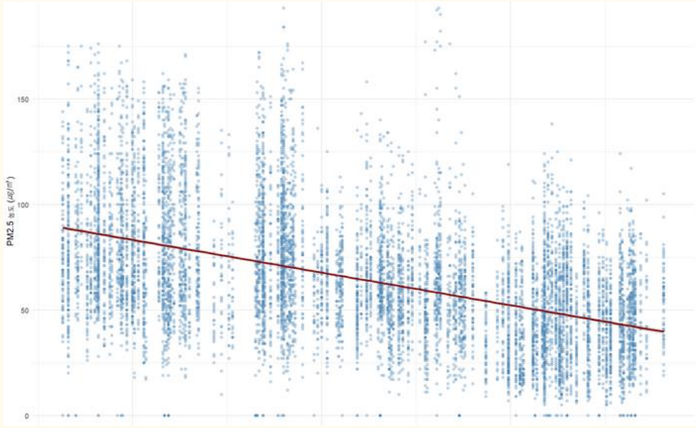
08 주요 변수와 PM2.5 간 상관관계 분석 (산점도 기반)

평균기온과 PM2.5의 상관관계

평균기온과 PM2.5는 음의 상관관계

기온이 낮을수록 미세먼지 농도가 높아지는 경향이 있으며, 이는 겨울철 PM2.5 고농도 현상과도 일치

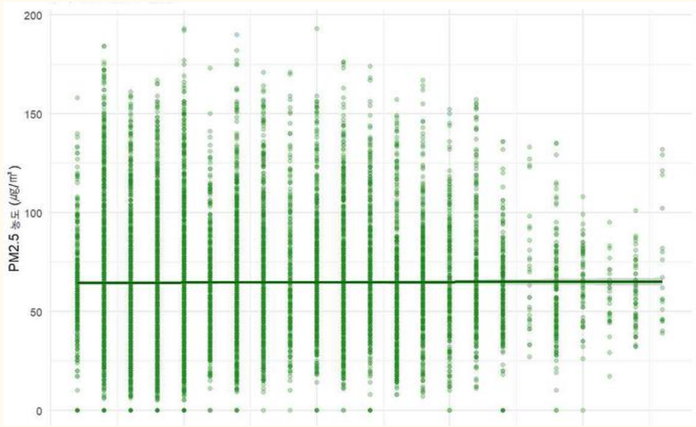
평균기온은 PM2.5 예측에 있어 가장 영향력 높은 변수 중 하나로 판단



풍속과 PM2.5의 상관관계

풍속과 PM2.5 사이에는 뚜렷한 상관관계가 나타나지 않음.

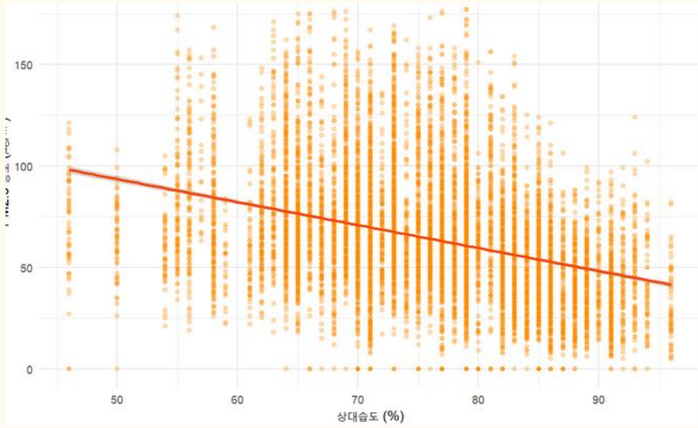
풍속이 강해진다고 해서 PM2.5 농도가 일정하게 낮아지거나 높아지는 경향은 보이지 않으며, 예측 변수로서의 영향력은 미약한 것으로 해석



습도와 PM2.5의 상관관계

상대습도와 PM2.5 농도는 음의 상관관계에 있음.

습도가 높을수록 미세먼지 농도가 낮아지는 경향이 있으며, 이는 “대기 중 수분이 미세먼지를 흡착·제거하는 역할을 한다”는 기존 환경보건 연구 결과와도 일치



09 결론 및 한계점

1.연구 결론

- 평균기온과 습도는 PM2.5 예측에 가장 핵심적인 변수로 나타남
- 고도는 중간 수준의 영향력, 해안지역 여부는 영향 미미
- 선형회귀·결정트리는 안정적인 예측 성능, KNN은 데이터 구조에 따라 성능 변동이 커 신뢰도 낮음

2.시사점 및 향후 연구방향

- 기상 변수만으로도 일정 수준 이상의 예측력 확보 가능
- 지리 정보는 선택적·보조적 변수로 활용하는 것이 더 효율적
- 주요 기상 변수 기반의 정교한 모델 설계 및 경량화 필요
- 계절성·지역성 반영한 지역 맞춤형 미세먼지 예보 시스템 개발 필요

3.한계 및 향후 보완 방향

- 연속형 변수의 나이브 베이즈 적용 시, 단순 구간화보다 PM2.5의 분포 기반 분위수 기준 분할이 효과적
- 다만, 본 데이터는 일별 시계열 특성을 가지므로 관측치 간 독립성 가정이 충족되지 않아 분석 결과에 제약이 있음 → 모델 성능보다 데이터 구조 이해가 더욱 중요

THANK YOU!