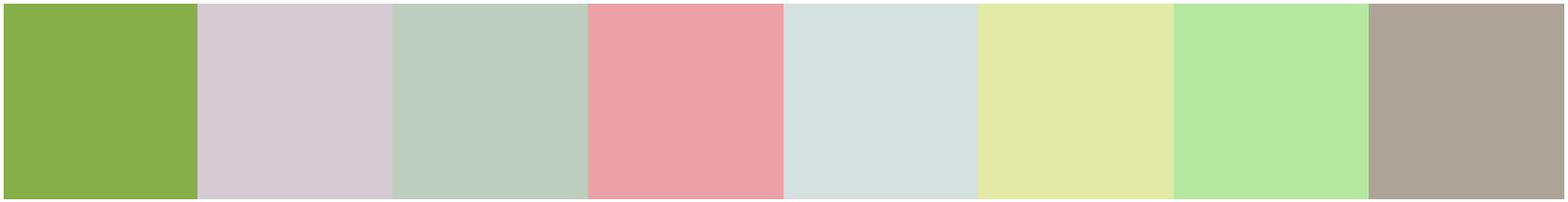
머신러닝 기법을 활용한

미세먼지 영향 변수 평가 연구 프로포절

- 의사결정나무를 중심으로 -



김 진 형

2017.04.13

연구의 필요성 및 목적

미세먼지의 영향

- WHO 산하 국제암연구소(IARC)에서 미세먼지를 사람에게 발암이 확인된 1군 발암물질로 지정 (1군 발암물질에는 석면, 벤젠, 카드뮴 등이 있음)
- 미세먼지는 호흡기질환 및 천식, 심혈관질환을 일으키며, 조기사망에 영향을 미침. 우리나라에서 대기오염으로 조기사망하는 인구가 10,000여명으로 추정됨
- 농작물, 토양, 수생생물에 피해를 줄 수 있으며, 기공을 막고 광합성을 저해함
- 교통·운항 장애 및 실외 기계·건물 설비 고장 등 대기오염으로 인한 사회적 비용이 연간 10조원에 달하는 것으로 추정됨

국제암연구소(IARC)에 따른 발암물질 분류

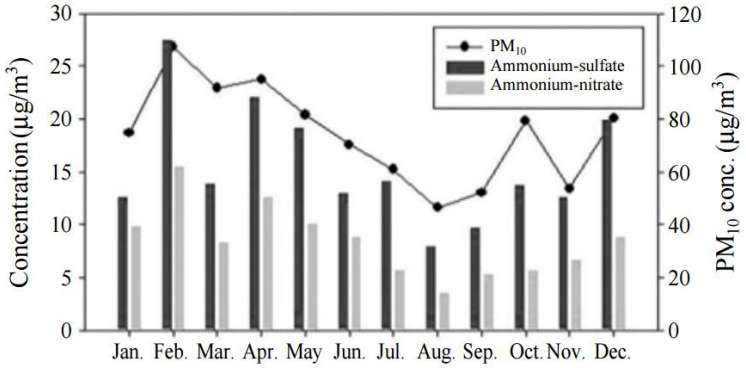
구분	주요 내용	예시
1군(Group 1)	인간에서 발암성이 있는 것으로 확인된 물질	석면, 벤젠, 미세먼지
2A군(Group 2A)	인간에서 발암성이 있을 가능성이 높은 물질	DDT, 무기납화합물
2B군(Group 2B)	인간에서 발암성이 있을 가능성이 있는 물질	가솔린, 코발트
3군(Group 3)	발암성이 불확실하여 인간에서 발암성이 있는지 분류하는 것이 가능하지 않은 물질	페놀, 톨루엔
4군(Group 4)	인간에서 발암성이 없을 가능성이 높은 물질	카프로락담

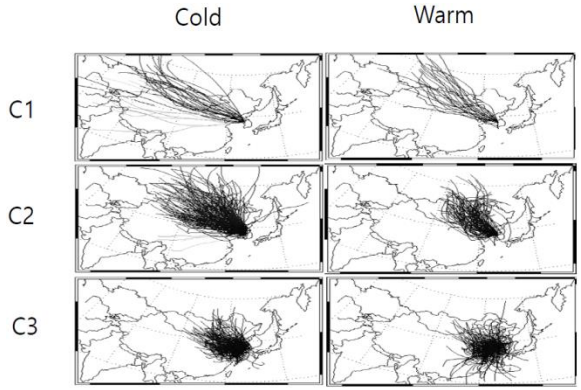
미세먼지로 인한 가시거리 저하

연구의 필요성 및 목적

미세먼지 연구

- 미세먼지의 화학적 성분에 대한 분석을 통해 미세먼지의 특성 파악(원인 및 경로)
- 기상·기후 요인과 외부요인(황사, 오염물질 장거리 이동 등)이 미세먼지 농도에 미치는 영향을 파악
- 미세먼지 농도에 대한 시계열·공간 분석
- 미세먼지의 건강에 미치는 영향에 대한 연구





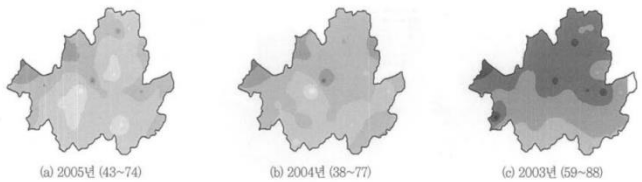
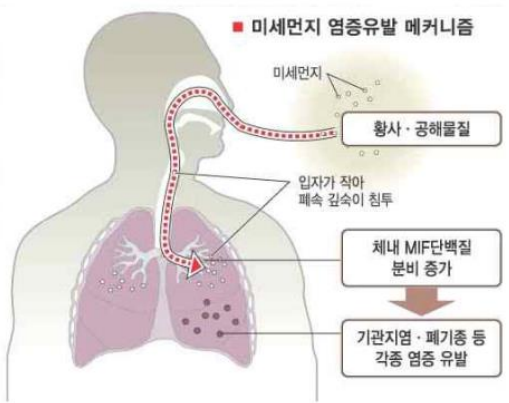


그림 5. RBF 공간보간기법을 이용한 서울시 PM10 연평균농도(μg/m³)의 공간적 분포 (최소값~최대값)

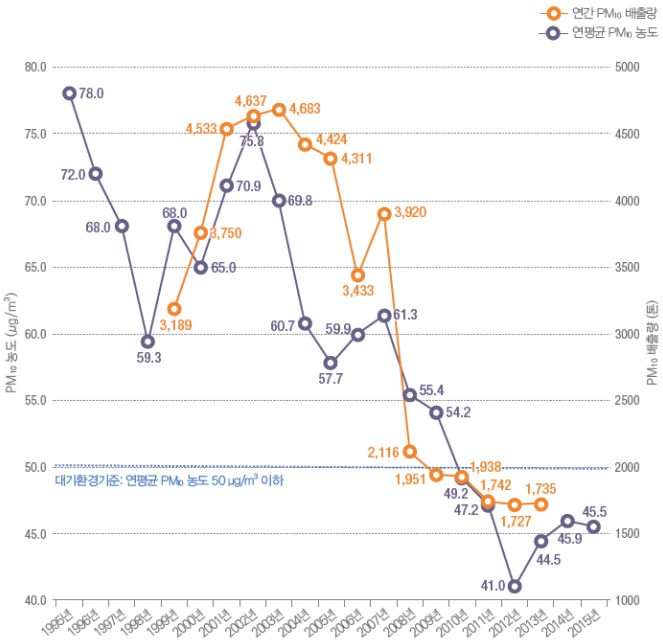
(범례: 38 - 45 | 46 - 50 | 51 - 55 | 56 - 60 | 61 - 65 | 66 - 70 | 71 - 75 | 76 - 80 | 81 - 85 | 86 - 96)

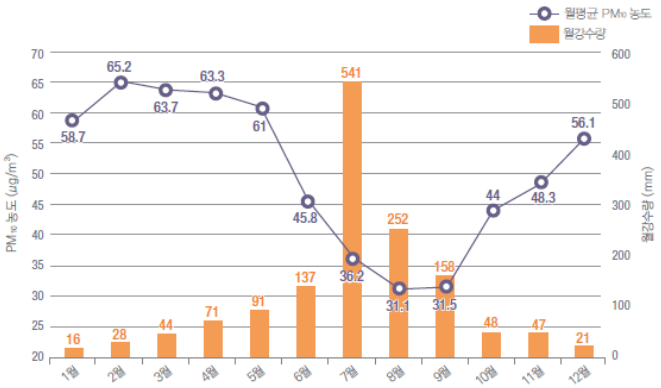


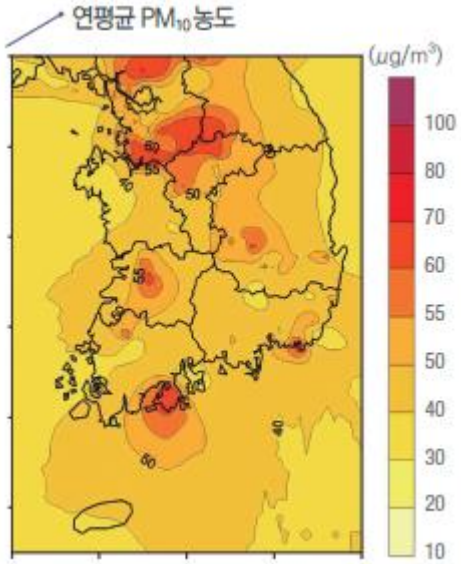
연구의 필요성 및 목적

우리나라 미세먼지의 특성

- 2000년대 들어 연평균 PM10 농도가 지속적으로 감소하였으나 여전히 다른 선진국에 비해 높은 수준이며 일평균 대기환경기준을 넘는 고농도 사례도 빈발하고 있는 실정
- 국내 약 300개 측정소 별 미세먼지 농도와 변동성에 차이가 존재
- 계절별(월별) 농도에 큰 차이가 존재







연구의 필요성 및 목적

우리나라 미세먼지 정책의 필요성

- 미세먼지가 미치는 영향과 국내 미세먼지 현황을 고려하였을 때, 국내 미세먼지 문제는 심각하며, 법, 정책, 제도적 측면의 개선이 필요함
- 미세먼지를 예측하는 연구와 저감하기 위한 정책을 진행하기 위해서는 미세먼지와 관련이 있는 요인들에 대한 해석이 필요함
- 각 요인들의 정량적 분석에 기반한 정책이 필요함

선행연구 현황 및 선행연구와 본 연구의 차별성

구 분		연구목적	연구방법	주요 연구내용
주요 선행연구	1	**과제명**: 통계-역학 융합방식의 미세먼지 예보기법 개발 연구(I) **연구자**: 허창회 외(2014) **연구목적**: 고농도 미세먼지 예보 의사결정 가이드라인 구축	- 군집분석 - 유클리디안 유사도 - 코사인 유사도 - 상관분석 - 역궤적 분석	- 과거 10년 이상 고농도 발생 기간 중 국내외 자료 수집 - 미세먼지 고농도 사례의 군집분석 및 유형별 상세특성 정보 정리(기상요소, 오염물질) - 웹기반 미세먼지 예보 확정 지원 모듈 개발
	2	**과제명**: 서울시 고농도 미세먼지 오염현상의 원인분석 및 지역별 맞춤형 관리대책 **연구자**: 김운수 외(2011) **연구목적**: 대기질을 개선하고, 환경경쟁력을 높이기 위한 미세먼지 관리전략 마련	- 문헌 연구 및 자료조사 - 통계분석(상관분석, 회귀분석	- 지역별 미세먼지 오염현상 진단 - 지역별 고농도 미세먼지 원인분석 - 지역별 미세먼지 배출량 DB 구축 - 해외도시의 미세먼지 배출원 분류체계 및 관리대책 사례 분석 - 서울시 지역별 맞춤형 미세먼지 관리전략
	3	**과제명**: PM2.5 배출특성 및 기여도 추정 연구 **연구자**: 국립환경과학원(2009) **연구목적**: PM2.5 대기환경기준 설정을 위한 국내 미세먼지의 배출량 산출 및 기여도 추정	- 화학수송모델링 - 배출량을 활용한 기여율 추정	- PM2.5 배출특성 조사와 배출목록 개선 - PM2.5 배출량 산정 모듈의 검토와 개선 - 동북아 규모 격자별 배출량 검토 및 모델링 - 2차 생성 PM2.5 특성 분석 및 배출원 별 기여율 추정
본 연구		- 기존의 연구는 활용한 데이터의 측면에서 기상 혹은 배출량 데이터만을 이용하는 데 한계를 가짐 - 방법론 측면에서 상관분석, 회귀분석을 활용함으로써 이용한 변수에 대한 분석에 한계가 있음 - 본 연구에서는 발생요인, 기상기후요인, 사회경제적 요인, 외부요인 변수를 활용함으로써 설명력을 높임 - 변수선택법을 활용하여 변수에 대한 정량적 분석 및 정책적 제언 가능	- 데이터 수집 - 머신러닝 알고리즘 적용	- 선행연구 및 문헌 분석을 통한 변수선정 - 데이터 수집 및 데이터 전처리 - 머신러닝 알고리즘 적용을 통한 변수 중요도 측정 - 알고리즘 변형을 통한 정확도 비교 - 결과해석을 통한 정책적 제언

주요 연구내용 및 연구추진방법

데이터 수집 및 클리닝

- 미세먼지 농도에 영향을 미치는 요인에 대한 데이터 수집
- 비정상·결측값 처리 등 데이터 클리닝 작업
- 시공간 해상도 통일, 데이터 조인 등 알고리즘 적용을 위한 변수화

머신러닝 알고리즘 적용

- 앞서 구축한 데이터에 머신러닝 알고리즘 적용
- 알고리즘 변형을 통한 정확도 또는 안정성 향상

결과 해석 및 정책적 제언

- 분석 결과를 통해 미세먼지의 농도에 대한 설명력이 높은 변수를 제안
- 해당 변수를 바탕으로 정책적 시사점 제안

주요 연구내용 및 연구추진방법

데이터 수집 및 클리닝

- **미세먼지 관련 문헌 및 이론 분석을 통한 변수 후보군 선정**

변수 분류	변수	
발생원인 변수	직접 발생 원인	NOx, SOx 등
	간접 발생 원인	VOCs, O3, NH3, 빛에너지 등
기후기상요인 변수	기온, 기압, 강수량, 바람 등	
사회경제적 변수	가계	인구, 인구밀도, 가계 소득, 난방비 등
	기업	대기오염 배출 사업장 및 배출량 정보 등
외부요인 변수	중국 동북부 미세먼지 농도, 황사발생일수 등	

- **데이터 수집 (각종 오픈 데이터)**

데이터 출처	URL	변수
기상청 국가기후데이터센터	http://sts.kma.go.kr/jsp/home/contents/main/main.do	기후 기상요인 변수
기상자료개방포털	https://data.kma.go.kr/cmmn/main.do	
에어코리아	http://www.airkorea.or.kr/index	발생원인 변수
국가통계포털	http://kosis.kr/	사회경제적 변수
환경공간정보서비스	https://egis.me.go.kr/main.do	
World Air Quality Index Sitemap	http://aqicn.org/map/china/kr/	외부요인 변수

주요 연구내용 및 연구추진방법

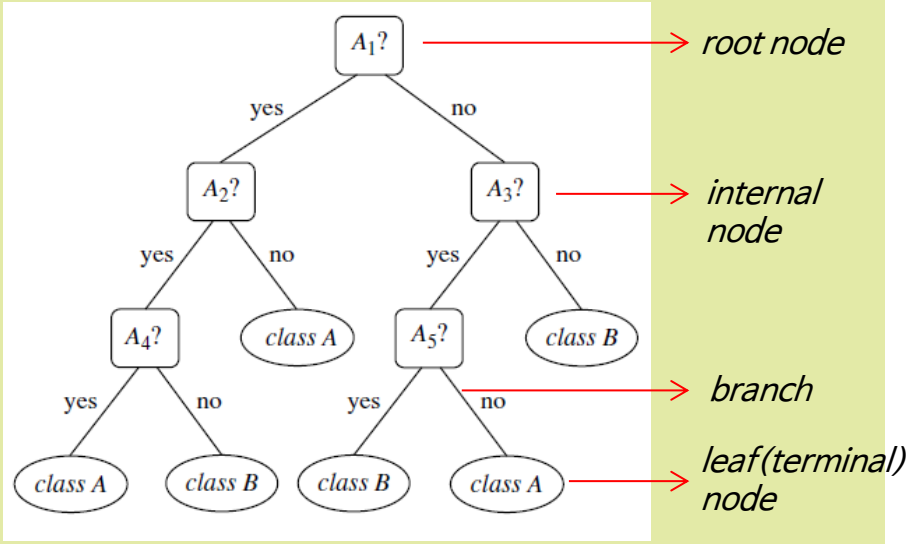
데이터 수집 및 클리닝

- 데이터 클리닝 및 입력변수로의 가공
  - 머신러닝 기법을 적용할 수 있는 형태로 데이터 생성
  - 데이터의 특성을 반영한 결측값 처리 (최빈값, 중앙값, 평균값 등)
  - 이상치 제거
  - 확보한 데이터의 시·공간 해상도를 통일
  - 데이터 조인을 통한 데이터 통합 (integration)

주요 연구내용 및 연구추진방법

머신러닝 알고리즘 적용

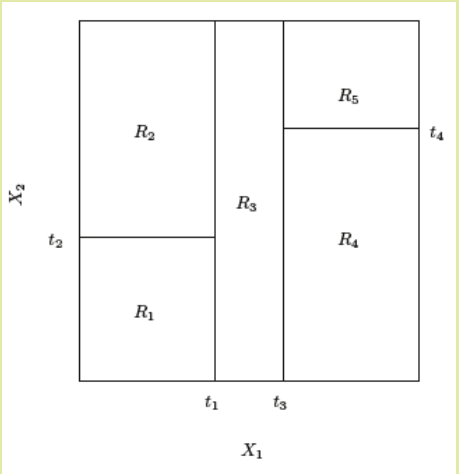
- 의사결정나무(Decision tree)
  - 결과를 설명하기 쉬움
  - 사람의 의사결정과 유사한 측면이 강함
  - 시각적으로 결과의 표현이 가능함
  - 비전문가도 쉽게 해석할 수 있음
  - dummy variable 없이도 범주변수를 쉽게 관리할 수 있음
  - 그러나 예측 정확도는 떨어지는 편임



- 의사결정나무의 유도(induction) - regression tree
  - recursive binary splitting (top-down, greedy)
  - goal: RSS 최소화
  - 모든 설명변수 X_j와 s(splitting point) 에 대해 RSS를 최소화하는 X_j와 s를 선택

$$\sum_{j=1}^{J}\sum_{i\in R_j}(y_i-\hat{y}_{R_j})^2 \qquad \sum_{i:\,x_i\in R_1(j,s)}(y_i-\hat{y}_{R_1})^2+\sum_{i:\,x_i\in R_2(j,s)}(y_i-\hat{y}_{R_2})^2$$

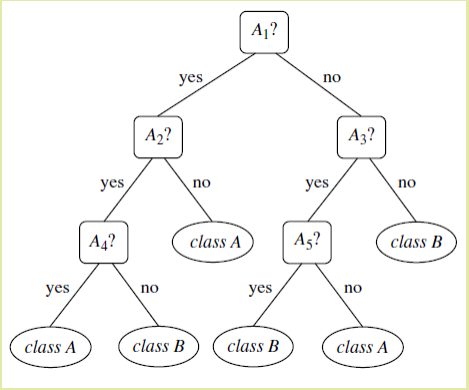
  - leaf node에 들어오는 case의 개수가 특정 수준 이하로 될 때까지 앞선 과정을 반복하여 내려감

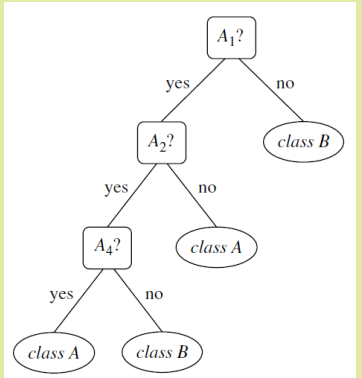


머신러닝 알고리즘 적용

- 가지치기(tree pruning)
  - variance를 줄일 수 있음
  - 해석을 쉽게 할 수 있음
  - overfitting 문제 해결(noise, outlier 〉 anomaly)
- postpruning
  - prepruning vs. postpruning
  - “full-grown” tree로부터 sub-tree 제거
  - Cost complexity pruning

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|$$

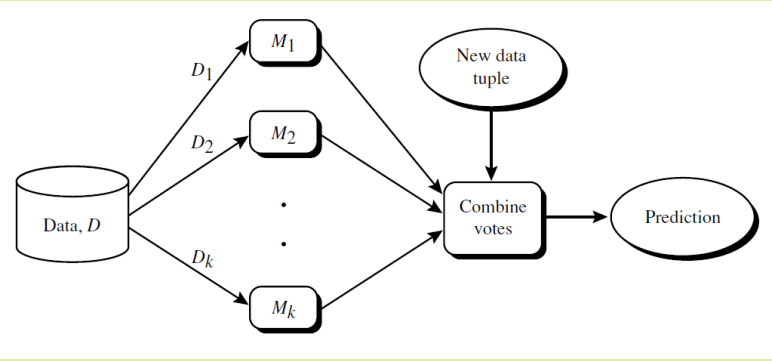




주요 연구내용 및 연구추진방법

머신러닝 알고리즘 적용

- 정확성을 향상시키기 위한 방법
  - ensemble method: 분류기(classifier)들의 조합
  - 분류기 간 상관성이 작음(little correlation)
  - 각 분류기를 다른 CPU에 배치하여 병렬처리 가능
  - 정확성이 높아짐



- Bagging
  - bootstrap aggregation
  - bootstrap method: sampling the training tuples uniformly with replacement
  - variance 낮출 수 있음
  - regression tree에 적용하는 경우, 가지치기 하지 않은 tree를 모두 이용함으로써 variance를 줄임

- Boosting
  - combination of weighted classifiers
  - regression tree의 경우 나무가 순차적으로 생성됨
  - 나무가 생성될 때 이전에 생성된 나무의 결과를 반영함

- Random Forest
  - 분류기가 의사결정나무로만이루어진 Bagging 기법
  - major predictor -〉 high correlation
  - 이를 해결하기 위해 split이후, subset에서 변수를 선택하게 함

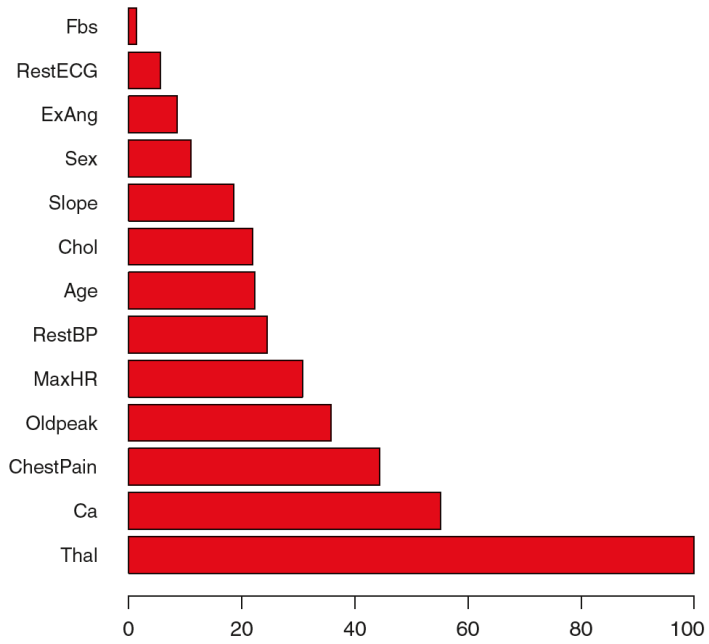
결과 해석 및 정책적 제언

- 변수의 중요도 측정 (variable importance measure)

– 정확성을 높이기 위해 bagging, random forest 기법을 적용하게 되면 결과를 해석하기 쉽다는 의사결정나무의 장점이 희석됨

– root node에 가까운 나무의 상단에 변수가 위치하는 경우 또는 나무에서 가장 많이 나타나는 변수가 중요변수라고 직관적으로 추정할 수 있음

– 정량적으로는 전체 트리에서 각 설명변수로 인해 줄어든 RSS 감소량의 평균을 통해 변수의 중요도를 측정할 수 있으며, 큰 값을 가진 변수를 중요한 변수라고 할 수 있음



- 변수 활용

– 중요하다고 판단된 변수를 다른 머신러닝 기법 (Linear model, 신경망 등)에 활용

– 변수 중요도를 기반으로 정책 우선순위 제안

- Counterfactual experiment

기대효과

데이터 수집 및 전처리, 통합을 통해 연구 기반 마련

- 미세먼지와 관련된 다양한 측면의 데이터를 수집함으로써 향후 미세먼지 연구 기반 마련
- 전처리 및 통합을 통해 향후 알고리즘을 쉽게 적용할 수 있음

환경분야 머신러닝 알고리즘 연구 확대

- 환경분야 데이터에 머신러닝 알고리즘을 적용하는 연구 촉진
- 머신러닝 알고리즘 비교를 통해 향후 연구에 기여

미세먼지에 대한 이해도 상승 및 정책 기여

- 미세먼지에 영향을 미치는 요인들을 정량적으로 평가함으로써 미세먼지에 대한 이해도를 높일 수 있음
- 정량적인 평가는 정책 수립측면에서 우선순위 선정에 기여함
- 향후 정책의 효과 평가 측면에 기여 가능할 것으로 예상

미세먼지 저감에 기여

- 미세먼지 저감정책을 통해 미세먼지 저감에 기여함으로써 환경 및 국민 건강 개선에 기여할 수 있음

연구추진계획

구분		4	5	6	7	8	9	10
데이터 수집 및 클리닝	문헌 연구							
	데이터 수집							
	데이터 전처리							
머신러닝 알고리즘 적용	기존 툴 조사							
	알고리즘 적용							
	알고리즘 변형							
결과 해석 및 정책적 제언	결과해석 및 정책 제언							
	결과 정리							

출처

- Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2011.
- Gareth James and others, *An Introduction to Statistical Learning with Applications in R*, New York: Springer Science+Business Media, 2013.
- 조홍래, “공간보간기법에 의한 서울시 미세먼지(PM10)의 분포 분석”, 『환경영향평가』, 2009, 18(1), pp.31-39.
- 한국환경정책평가연구원, 『KEI포커스』, 2016, 4(3), 세종: 한국환경정책평가연구원.
- 환경부, 『바로 알면 보인다. 미세먼지, 도대체 뭘까?』, 세종: 환경부, 2016.
- 국립환경과학원, 『통계-역학 융합방식의 미세먼지 예보기법 개발 연구 (I)』, 2014.
- 윤원서, “미세먼지에 따른 우리나라 사회적 비용 연간 10조원… 2060년에는 20조원”, 『그린포스트코리아』, 2017. 3. 23

Thank you!