머신러닝을 활용한 대기환경

빅데이터 분석 및 예측

2010104050

박재호

요약

본 프로젝트는 인체에 유해한 성분의 물질이 다수 포함된 인위적인 대기오염물질 데이터를 분석하여 서울, 경기도 지역의 대기오염물질 데이터를 지역별, 연도별로 분석하여 새로운 관측소 최적 설치 장소와 미세먼지의 이동 흐름 및 미래의 미세먼지 추이를 예측하고자 한다.

1. 서론

1.1 연구배경

2016년 아시아 환경오염의 화두는 미세먼지에 의한 대기오염이었다. 미세먼지는 인체에 유해한 성분의 물질이 다수 포함된 인위적인 대기오염물질이다. 많은 양의 미세먼지에 지속적으로 노출될 경우 인체에 치명적인 질병을 유발할 수 있을 뿐만 아니라 자연 생태계에도 큰 영향을 준다. 이러한 대기오염의 피해를 최소화 하기 위해서는 관련 예방책이 필수적이다. 본 프로젝트는 대기오염의 피해를 최소화 하기 위한 예방책에 도움을 주고자 인체와 생태계에 해로운 대기오염물질 관련 데이터를 지역별, 연도별로 분석하여 최적의 새로운 관측소 설치 장소와 미세먼지의 이동 흐름 및 미래의 미세먼지 추이를 예측하고자 한다.

1.2 연구목표

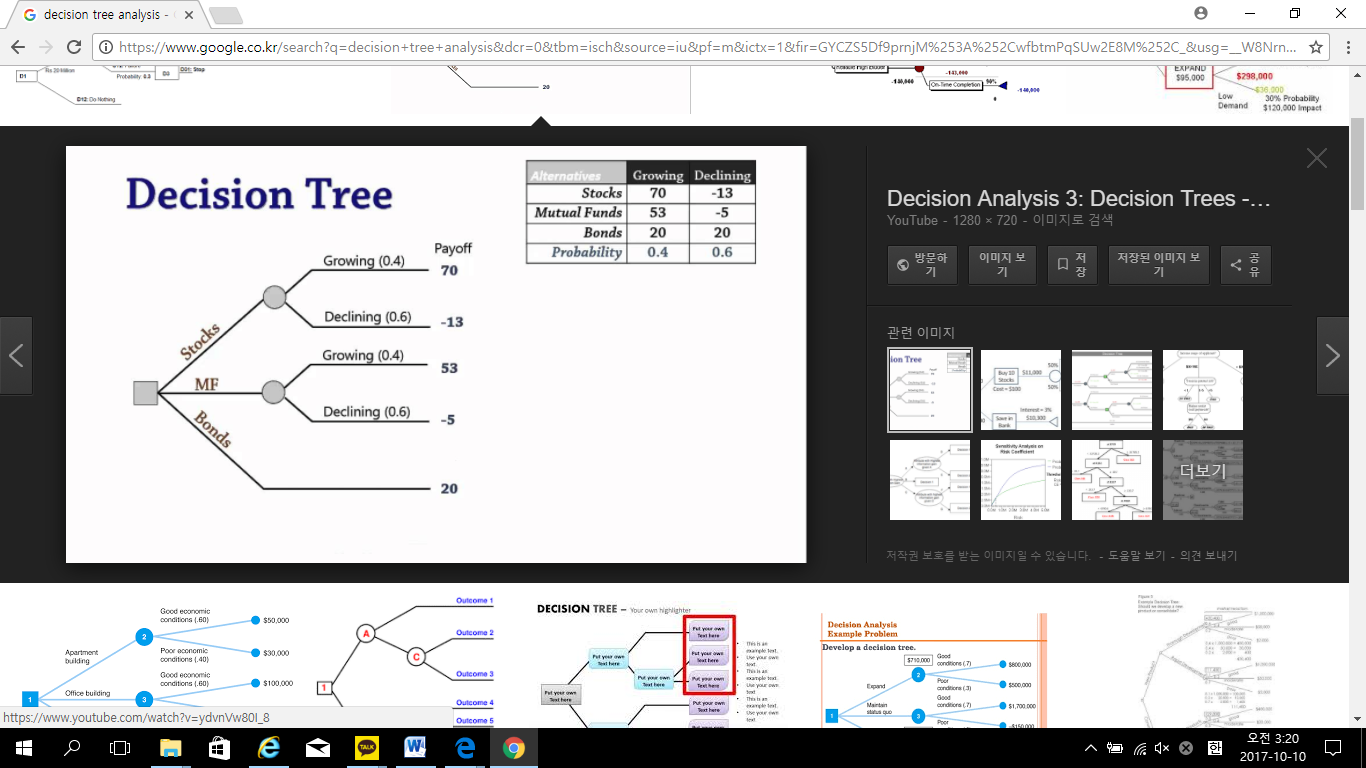
본 프로젝트는 경기도와 서울지역으로 범위를 한정하고 다음 3가지를 목표로 한다.

1. 각 지역의 인구밀도와 대기오염물질의 지역특성, 공간분포 상 취약지를 고려하여 추가 대기환경 측정소 위치를 선정한다. 각 데이터는 정부3.0, 한국환경공단, 기상청 기상자료를 이용하여 수집한다.
2. 지역별, 년도 별 대기 오염물질 변화 추이를 분석하여 미세먼지에 영향을 주는 유의한 변수를 파악한 후 비슷한 특징을 가지는 지역을 군집으로 파악하여 미세먼지의 흐름을 예측한다.
3. 다양한 머신 러닝 방법을 이용하여 가장 성능이 좋은 모델을 선출 후 미래 미세먼지 추이를 예측한다.

2. 관련연구

2.1 데이터 마이닝- Dicision tree

어떤 항목에 대한 관측값과 목표값을 연결시켜주는 예측 모델이다. 본 연구는 Dicision tree 알고리즘을 이용하여 공기 중의 성분을 변수로 두고 실질적으로 미세먼지에 영향을 끼치는 유의한 변수를 도출할 예정이다.



3. 데이터 전처리

본 프로젝트는 머신러닝 기반 분석을 위한 대기오염물질 데이터와 기술통계학 분석에 사용할 대기오염물질과 기상관측 데이터의 통합데이터가 필요하다.

다음은 대기 오염물질 데이터와 종관 기상관측 데이터의 원 자료 형태이다.

4. 데이터 분석

4.1 대기 오염물질의 패턴 분석

4.2 공간 분석

다음은AirKorea의 경기도 내의 81개 대기오염 관측소 좌표 데이터를 R을 이용하여 구글 지도 상에 표시한 그림이다.

다음은2015년 경기도의 인구를 밀도에 따라 시각화한 후 현재 대기오염 관측소의 위치를 표시한 그림이다.

다음은 환경부에서 발행한 대기오염 측정망 설치운영 지침에 따라 대기 오염관측소의 관측 범위인 반경 4km로 관측 범위를 표기하였다.

4.3 머신 러닝

4.3.1 대기오염물질 유의변수 파악 및 주성분 분석

4.3.2 대기오염물질에 따른 도시 별 군집 분석

4.3.3 머신 러닝을 활용한 미세먼지의 추이 예측

5. 결론(기대효과)

본 프로젝트는 다음과 같은 3가지의 기대효과가 있다. 첫째로, 인구밀도와 대기오염물질의 지역특성, 공간분포상의 취약지를 고려하여 추가 대기환경 측정소 설치지역을 파악하였다. 이 데이터는 정부의 미세먼지와 같은 대기오염의 예보, 경보제 운영인프라 확충에 쓰일 수 있으며 추가 측정망 설치에 활용할 수 있다. 둘째로, 지역 별, 년도 별 대기오염물질 변화 추이를 분석하여 미세먼지에 영향을 주는 유의한 변수를 파악 및 비슷한 특징을 가지는 지역을 군집으로 파악하여 미세먼지의 흐름을 예측하였다. 이 데이터는 정부가 대기오염 관련 대책을 수립할 때 지역 군집을 바탕으로 효율적인 연계 대책을 수립할 수 있을 것이다. 마지막으로, 다양한 머신 러닝 방법을 이용하여 가장 성능이 좋은 모델을 활용 및 미래 시점 미세먼지의 추이를 예측하였다. 이 데이터는 한국형 미세먼지 예보모델의 개발 및 예보 권역 별 상세모델 구축의 자료로 활용될 수 있다.

6. 참고자료

예측방법론(저자 : 이긍희 외 2인 / 출판사 : 한국방송통신대학교 출판문화원)

시계열 분석과 예측(저자 : 이우리 / 출판원 탐진)

R을 이용한 데이터 처리&분석 실무(저자 : 서민구 / 출판사 : 한빛미디어)

R을 이용한 데이터 시각화(저자 : 전희원 / 출판사 : 자유아카데미)

R을 이용한 데이터 마이닝(저자 : 박창이 외 4인 / 출판사 : 교우사)

SAS 다변량 자료분석 입문(저자 : 강현철 외 3인 / 출판사 : 자유아카데미)

7. 상세일정

11.14일 ~ 17일 데이터 전처리(기존의 초기 논문을 파기하고 내용 및 방향성을 수정하였습니다.)

11.17일 ~ 27일 데이터 분석

11.27일 ~ 30일 데이터 시각화 및 최종 보고서 작성