안녕하십니까 허의남 교수님의 지도하에 있는10학번 박재호라고 합니다.

저의 주제는 머신러닝을 활용한 경기도권 대기환경 빅데이터 분석 및 예측입니다.

<1> 발표는 총 4개의 단락으로 이루어져 있습니다. 왜 이 주제를 선정하게 되었고 어떻게 하여 분석을 하고 결론을 도출하는지 흐름대로 소개하도록 하겠습니다.

<2> 대기오염, 제가 신입생.. 시절일때만 해도 대기오염은 그렇게 큰 이슈가 아니었습니다. 하지만 최근들어 보시는 바와 같이 우리는 지금 심각한 대기오염문제를 직면하고있습니다. 그 중 가장 골치거리는 바로 미세먼지입니다. 우리는 이제 미세먼지를 일기예보처럼 보도받는 시대에 살고있습니다.

<3> 국내의 주요 대기오염물질은 6가지로 나뉩니다. 우측 3개의 성분과 오존은 산업혁명 시대부터 발생했었던 대기오염물질이지만 좌측 상단 2개는 최근에 지정된 대기오염물질 바로 미세먼지입니다. 미세먼지란 보시는바와 같이 매우 작은 입자이며 여러가지 해로운 공해물질이 섞여있다고 합니다.

<4> 국내의 미세먼지 정책 및 예보 인프라는 아직 확고히 수립되지 않았으며 현 정부는 대선 공약으로 내세울 정도로 매우 중요한 사항이 되었습니다.

-------------------다음은 2번째 단락인 어떻게 입니다. -------------------

<5> 본 프로젝트는 대기오염물질과 기상데이터를 정제한 후

이를 토대로 분석하여 시각화 및 결론을 도출한 후

k-means 와 ARIMA 모형 시계열 분석을 실시합니다.

마지막으로 분석 모델을 평가 및 향후 동향을 예측합니다.

<6> 데이터 전처리 과정입니다. 기반데이터는 2014년부터 2017년까지의 대기오염물질 데이터와 종관기상관측 데이터입니다.

<7> 대기오염물질의 경우 결측치 존재 및 측정일시 변수가 표준형태를 따르지 않았으며 하나의 지역에 많은 측정소들이 존재하기도 했습니다.

<8> 따라서 첫째로, 측정일시 변수를 표준형태의 Date변수로 바꾸었고 둘째로, 결측치를 제거하였으며 마지막으로 측정소들을 시 별 평균값으로 대체하여 통합하였습니다.

<9> 전처리의 마지막 단계로 대기오염물질 데이터와 기상데이터를 R의 merge함수를 활용하여 통합하였습니다. 통합데이터는 대기오염 현황 분석에 사용되며 대기오염물질 데이터는 머신러닝 분석에 사용됩니다.

-------------------다음은 3번째 단락인 분석 입니다. -------------------

<10> 분석은 시각화 분석, 공간 분석, 모델 분석으로 이루어져 있습니다.

우선 시각화 분석입니다. 전처리작업을 실시한 대기오염물질 중 미세먼지의 농도를 채색달력으로 시각화한 자료입니다. 빨갈수록 농도가 높은 것이며 경기도의 미세먼지는 봄과 겨울에 집중되어 있음을 알 수 있습니다.

<11> 대기오염물질과 기상자료를 통합한 자료를 기반으로 Wind Rose 기법으로 시각화 하였습니다. 경기도의 일부지역은 북서풍이 불 때 미세먼지의 농도가 높았지만 경기도 전체 평균자료의 시각화 결과는 남동풍이 불 때 크게 상승하였습니다.

<12> 대기 오염물질 요소의 패턴을 분석한 그래프입니다. 미세먼지의 경우 겨울과 봄에 상승하지만 여름에는 낮아지는 경향을 보입니다. 그 외 대기오염물질의 그래프는 보시는 바와 같습니다. 오존을 제외한 나머지 3개의 오염물질은 비슷한 패턴을 보입니다. 오존의 경우 NO2 수치와 연관이 있는데 NO2는 주로 자동차의 매연으로 경기도의 요일별 그래프를 보시면 차운행이 적은 주말에는 O3가 증가하는 것을 알 수 있습니다. 본 프로젝트는 30개의 지역별로 NO2와 O3의 수치를 측정하였으나 발표에는 넣지 않았습니다. 하지만 이 데이터는 도시 별 ‘자동차 없는 요일’ 제도 등 여러 행정업무에 유용할 것으로 보입니다.

<13> 다음은 공간 분석을 통한 새로운 대기오염 관측소의 최적 설치 위치 선정입니다. R을 이용하여 경기도 내의 관측소 위치를 표시하였고 관측범위인 4km를 지도에 표기하였습니다. 환경부의 인구 10만명 이상의 도시에 측정소를 설치해야 한다 라는 지침에 따라 아래 나열된 지역은 추가적인 대기오염 관측소가 필요한 것으로 보입니다.

<14> 마지막 모델 분석으로 k-means 분석과 ARIMA 모델분석으로 구성되어 있습니다.

K-means 분석을 위하여 Decision Tree를 통해 대기오염물질 유의 변수를 파악하였습니다.

Decision Tree는 대기오염물질이 국가에서 정한 한도를 넘을 경우 종속변수인 Level 변수를 counting하였습니다. 그 결과 경기도의 대기오염에는 NO2와 PM10이 유의한 변수로 선정되었습니다.

<15> 경기도의 30개 도시를 군집분석한 결과 3개의 군집이 확인되었습니다.

<16> 대기오염 유의 변수와 3개의 군집을 바탕으로 k-means 분석을 실시하였습니다.

30개의 도시는 특정한 군집을 이루며 분포하는 것을 확인하였습니다. 또한 30개의 도시를 연도 별로 변수를 두어 분석한 결과 다음과 같은 분포를 확인하였고 이를 통하여 해마다 군집이 바뀌는 도시를 확인 할 수 있었습니다.

<17> 다음은 ARIMA 모델 분석입니다. ARIMA 모델은 정상시계열 데이터를 필요로 하지만 본데이터는 요소별 분해 결과 경향성과 계설성이 보였습니다.

<18> 정상 시계열을 위하여 계절성을 제거 및 1차 차분 후 시계열 데이터가 정상성이 라는 대립가설을 채택하였습니다.

<19> 여러 모형을 추정해본 결과 다음과 같은 ARIMA 모형이 가장 적합하다고 판단했고 잔차 검정해본 결과 ~~ 였었습니다.

마지막으로 해당 모델을 통하여 향후 미래 미세먼지의 동향을 예측한 결과입니다.

-------------------프로젝트의 결론입니다. -------------------