

서울의 대기오염도를 통한 미세먼지 예측

-Prediction of fine dust based on air pollution in Seoul-

이름: 이민영

학번: 2118309

Github: <https://github.com/leeminyoung7/Prediction-of-fine-dust-based-on-air-pollution-in-Seoul.git>

1. 안전 관련 머신러닝 모델 개발 관련 요약

a. 프로젝트에 관한 전체 내용을 요약

서울의 시간별 평균 대기 오염 데이터 셋을 활용하여서 미세먼지(PM) 농도를 예측하는 머신러닝 모델을 개발하였으며 선형 회귀 모델을 활용하였다.

2. 개발 목적

a. 머신러닝 모델 활용 대상:

이 모델은 대기 오염 관리 기관, 환경 연구자 및 정책 관리자, 환경 정보 서비스 제공자들에게 대기 질 모니터링 및 개선 전략 수립에 많은 기여를 할 수 있을 것이라고 기대한다.

b. 개발의 의의:

이 서울 미세먼지 예측 모델을 통해 건강 위험 요소를 사전에 관리하고 대기 오염 문제를 보다 효과적으로 해결할 수 있는 기반을 마련하고 실시간 데이터 분석을 통해 대중의 건강을 보호하며 정책적으로 국민의 안전에 대해서 대응 시스템을 구축할 수 있다.

c. 데이터의 어떠한 독립 변수를 사용하여 어떠한 종속 변수를 예측하는지

독립 변수는 대기 오염에 영향을 미치는 다양한 환경적 요인(NO₂, CO, O₃, SO₂) 이라고 할 수 있으며 종속 변수는 미세먼지 농도(PM₁₀, $\mu\text{g}/\text{m}^3$)이다.

3. 배경지식

a. 데이터 관련 사회 문제 설명

대기 오염은 건강에 심각한 영향을 미치며, 특히나 미세먼지는 호흡기 질환 및 심혈관 질환과 관련이 있다. 특히나 서울과 같은 대도시에서는 대기질 관리를 위한 데이터관리가 필요하다.

b. 머신러닝 모델 관련 설명 등

선형 회귀는 예측 결과 해석이 용이하고, 분석 패턴을 학습하고 예측하는데 용이한 기술이다. 이로 인해 환경 데이터 분석에 적합하다고 생각하며 복잡한 환경적 요인 간의 관계를 이해하고 예측할 수 있다.

4. 개발 내용

a. 데이터에 대한 구체적 설명 및 시각화

i. 데이터 개수, 데이터 속성 등

데이터셋은 서울의 시간별 평균 대기 오염 데이터를 포함하고 4226 개의 샘플과 여러 환경적 속성을 포함하고 있다. (미세먼지, NO₂, CO, O₃, SO₂)

ii. 데이터 간 상관관계 설명 등

데이터 간 상관관계를 분석하여 독립 변수와 종속 변수 간의 관계를 파악하고, 특정 변수들이 미세먼지 농도에 미치는 영향을 보여준다.
CO(일산화탄소)와 미세먼지 농도가 상관관계를 가질 가능성이 크다고 본다.

b. 데이터에 대한 설명 이후, 어떤 것을 예측하고자 하는지 구체적으로 설명

i. 독립변수, 종속변수 설정

독립변수: 환경적 요인(NO₂, CO, O₃, SO₂)

종속변수: 미세먼지 농도(PM₁₀)

이러한 독립변수를 통해서 종속변수의 수치가 얼마나 늘어나는지에 대한 상관관계를 파악하여 미래의 종속변수인 미세먼지 농도(PM₁₀)를 예측하고자 한다.

c. 머신러닝 모델 선정 이유

i. 설명한 데이터를 기반으로 머신러닝 모델 선정 이유 설명

선형 회귀 모델을 선택한 이유는 독립변수가 상대적으로 양이 적기 때문에 선형 회귀 모델이 적합하다고 생각했다. 왜냐하면 상대적으로 단순하면서도 해석이 용이하고 결과를 쉽게 이해할 수 있기 때문이다.

ii. 성능 비교를 위한 머신러닝 모델 선정 이유

SVR 은 비선형 데이터 처리에 강점이 있는 회귀 모델이며 뛰어난 성능을 발휘하며, 작은 데이터셋에도 적합하다.

랜덤 포레스트는 비선형 관계를 잘 처리하며, 변수 간 복잡한 관계를 모델링할 수 있기에 SVR 과 랜덤 포레스트는 선형 회귀와 함께 비교하기 적합하다고 생각했다.

d. 사용할 성능 지표

i. 머신러닝 모델의 성능을 평가하기 위해 사용하는 성능 지표에 관한 설명 등

모델 성능 평가를 위해 평균 제곱 오차(MSE), 결정 계수(R^2), 평균 절대 오차(MAE)를 사용했으며, 이 지표들은 모델의 예측 정확성을 다각도로 평가하는 데 유용하다.

ii. 성능 지표 선정 이유 등

회귀 모델의 일반적인 평가 기준이 MSE, RMSE 를 사용한다고 하였고 R^2 는 모델의 적합도를 측정하는 데 유용하다고 하였기 때문이다.

5. 개발 결과

a. 성능 지표에 따른 머신러닝 모델 성능 평가

i. 수치 자료 및 시각화 자료를 사용

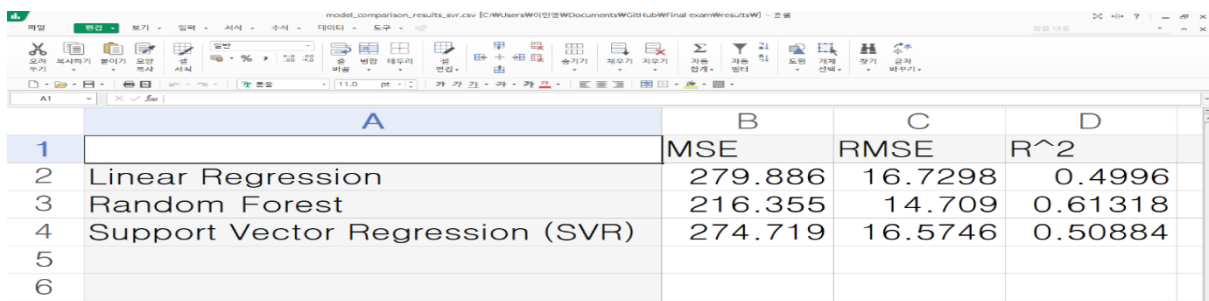
1. MAE, RMSE, MSE, Accuracy, 오차행렬 등

MAE : 8.163033

RMSE : 16.7298

MSE : 279.8862

R^2 : 0.4996



	A	B	C	D
1		MSE	RMSE	R^2
2	Linear Regression	279.886	16.7298	0.4996
3	Random Forest	216.355	14.709	0.61318
4	Support Vector Regression (SVR)	274.719	16.5746	0.50884
5				
6				

2. KFold 결과

KFold MSE: 211.582

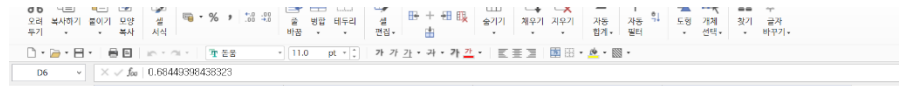
KFold RMSE: 14.3149

KFold R^2 : 0.5916

ii. 다른 머신러닝 모델과 성능 비교

MSE : 예측값과 실제값의 차이를 제곱하여 평균화한 값, 작을수록 좋다.

RMSE : MSE 의 제곱근으로, 예측 오차를 실제 단위로 측정



	A	B	C	D
1	Fold	MSE	RMSE	R ²
2	1	279.886	16.7298	0.4996
3	2	210.385	14.5047	0.59684
4	3	127.694	11.3002	0.67229
5	4	307.778	17.5436	0.50462
6	5	132.167	11.4964	0.68449
7				

R²(결정 계수) : 모델이 데이터를 얼마나 잘 설명하는지 나타내며 1에 가까울수록 좋다.

랜덤 포레스트는 가장 낮은 MSE와 높은 R²을 기록하며, 미세먼지 농도 예측에서 가장 우수한 성능을 보였다.

SVR은 랜덤 포레스트보다 약간 낮은 R²을 기록했으나, 여전히 높은 예측 성능을 나타냈다.

선형 회귀는 상대적으로 높은 MSE를 기록하며 위 두 모델에 비해 성능이 다소 떨어진다. 하지만 실제로는 R² 수치가 랜덤 포레스트는 0.61, SVR은 0.5, 선형 회귀 모델은 0.49로 많은 차이를 보이는 것은 아니기에 실용적이며 해석의 용이한 선형 회귀 모델을 사용하게 되었다.

b. 머신러닝 모델의 성능 결과에 대한 해석

랜덤 포레스트와 SVR는 비선형 관계를 잘 반영하여 조금 더 나은 성능을 보이지만 날씨와 매일 변하는 미세먼지 농도와 같은 대량의 데이터셋을 가지고 있는 것은 간단하고 해석의 용이성과 간단한 구현으로 실행되는 모델이 실용적이다. 선형 회귀 모델은 변수 간의 관계를 직관적으로 파악할 수 있고 계산 비용이 낮고 학습 속도가 빠르며, 대규모 데이터에 적합하다고 볼 수 있다.

6. 결론

a. 머신러닝 모델 개발에 관한 간략한 요약 및 결과 설명

서울의 대기오염 데이터를 바탕으로 미세먼지 농도(PM10)를 예측하는 머신러닝

모델을 개발하였고 그에 대한 타 모델과의 성능을 분석해보았다.

또한 머신러닝 비교를 위해 비교 지표를 시각화한 것을 결과 폴더에 저장되는 것을 볼 수 있다.

b. 개발 의의 등

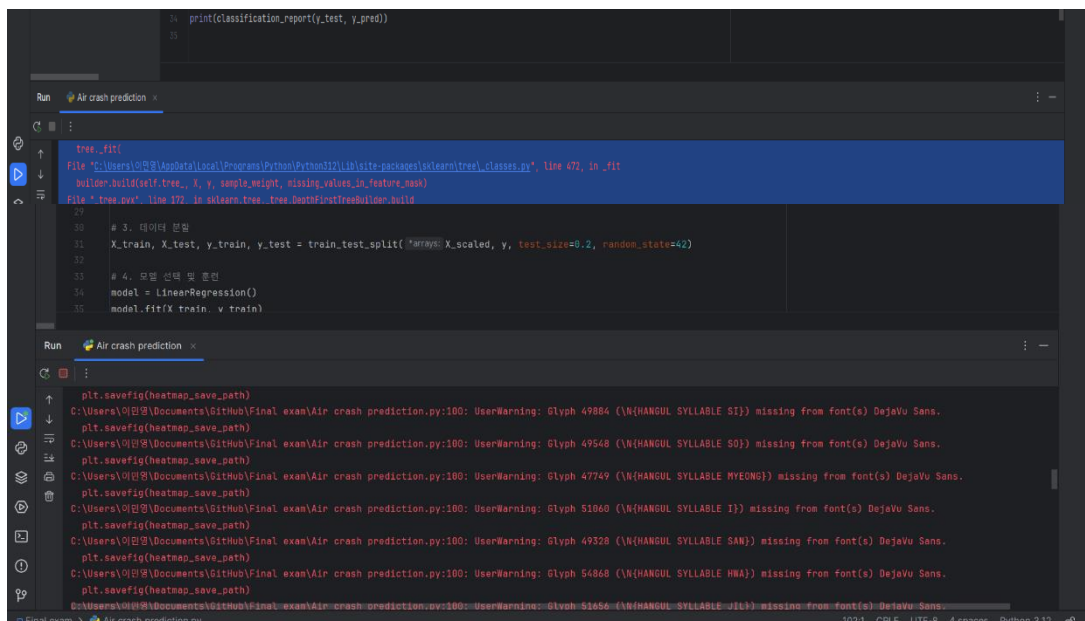
미세먼지 농도 예측으로 모든 국민이 조금이나마 건강하고 안전한 삶을 살아갈 수 있는 계기가 되기를 바라며 실시간 데이터 분석을 통해 정책 결정 과정에 도움을 주길 바란다. 또한 선형 회귀모델이 타 머신러닝 모델에 비해 조금 성능이 떨어지더라도 방대한 데이터가 있다면 그때 선형 회귀 모델이 빛을 바란다고 생각하기에 선형 회귀 모델을 선정하였다.

c. 머신러닝 모델의 한계

머신러닝 모델은 특정 환경적 요인에 의존하고 모든 변수를 고려하지 않기에 한계가 있을 수 있으며 데이터의 품질과 양이 모델 성능에 큰 영향을 미칠 수 있다.

d. 계획했던 것과 다른 점

계획은 가장 대중적인 NumPy, Pandas, Matplotlib 만을 사용하면 쉽게 예측하는 문제를 쉽게 해결해 나아갈 줄 알았다. 하지만 데이터 양과 어떤 주제를 선정함에 따라서 어려움이 있었으며 실용성이 편한 머신러닝을 찾는 부분에서 많은 시간이 소요되었으며 여러 번의 주제 변경이 있었고 중간에는 비행기 사고 예측에 대해서 머신러닝을 만들려고 했지만 데이터 셋의 부족함에 있어서 변경하게 되었다. 또한 데이터셋만 잘 활용하는 코드만 사용하면 될 줄 알았지만 메모리 부족, 폰트



```
print(classification_report(y_test, y_pred))
```

```
tree_.fit(  
    File ~\Users\이민준\AppData\Local\Programs\Python\Python37\Lib\site-packages\sklearn\tree\_classes.py, line 472, in _fit  
    builder.build(self, X, y, sample_weight, missing_values, is_feature_node)  
    File ~\tree.pyx, line 172, in sklearn.tree._tree.TreeBuilder.build
```

```
# 3. 데이터 분할  
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)  
  
# 4. 모델 선택 및 훈련  
model = LinearRegression()  
model.fit(X_train, y_train)
```

```
plt.savefig(heatmap_save_path)  
plt.savefig(heatmap_save_path)  
plt.savefig(heatmap_save_path)  
plt.savefig(heatmap_save_path)  
plt.savefig(heatmap_save_path)  
plt.savefig(heatmap_save_path)  
plt.savefig(heatmap_save_path)  
plt.savefig(heatmap_save_path)  
plt.savefig(heatmap_save_path)
```

```
C:\Users\이민준\Documents\GitHub\Final_exam\Air crash prediction.py:100: UserWarning: Glyph 49884 (\N(HANGUL SYLLABLE SI)) missing from font(s) DejaVu Sans.  
C:\Users\이민준\Documents\GitHub\Final_exam\Air crash prediction.py:100: UserWarning: Glyph 49548 (\N(HANGUL SYLLABLE SO)) missing from font(s) DejaVu Sans.  
C:\Users\이민준\Documents\GitHub\Final_exam\Air crash prediction.py:100: UserWarning: Glyph 47749 (\N(HANGUL SYLLABLE MYEONG)) missing from font(s) DejaVu Sans.  
C:\Users\이민준\Documents\GitHub\Final_exam\Air crash prediction.py:100: UserWarning: Glyph 51860 (\N(HANGUL SYLLABLE I)) missing from font(s) DejaVu Sans.  
C:\Users\이민준\Documents\GitHub\Final_exam\Air crash prediction.py:100: UserWarning: Glyph 49328 (\N(HANGUL SYLLABLE SAN)) missing from font(s) DejaVu Sans.  
C:\Users\이민준\Documents\GitHub\Final_exam\Air crash prediction.py:100: UserWarning: Glyph 54868 (\N(HANGUL SYLLABLE HMA)) missing from font(s) DejaVu Sans.  
C:\Users\이민준\Documents\GitHub\Final_exam\Air crash prediction.py:100: UserWarning: Glyph 54656 (\N(HANGUL SYLLABLE JIL)) missing from font(s) DejaVu Sans.
```

깨짐 등 결과물에도 많은 신경을 써야하는 일이 있었다.

e. 느낀점

머신러닝 모델을 대중적인 것을 사용하면 가장 좋을 줄 알았는데 무슨 주제로 어떤 데이터를 사용하는가에 따라서 결과 도출의 질이 달라지는 것에 편하기만 할 줄 알았던 머신러닝에 벅을 느꼈지만 하지만 이번 과제를 해결하면서 사용자의 대상과 개발자인 내가 실용성인지 정확성인지 이러한 방향에 따라서 머신러닝 선정에 대한 중요함을 느끼게 되었다. 또한 결과물에 대한 오류 해결에 대한 어려움도 있었지만 잘 해결되어서 성취감을 얻고 가는 것 같다.