질소화합물 배출량 예측

발표자: 하현수

팀원: 하현수(조장), 박민상, 이종민, 임채원

Github: https://github.com/1000mmoney/1TEAM.NOX1

1. 요약

a. 질소화합물(NOx)의 배출량 예측을 위해 다양한 머신 러닝 모델을 활용하여 데이터 분석하고 평가한 내용을 다루고 있습니다. 예측의 주요 목표는 산업현장 및 건설현장에서 질소화합물의 배출량을 정확하게 예측하여 공기질 개선과 인체 건강 보호에 기여하는 것입니다.

2. 배출량 예측 목적

a. 산업현장 및 건설현장의 질소산화물 발생 원리
   i. 질소화합물 (NOx)는 공기 중으로 배출되었을 때 햇빛과 광화학 반응을 통해 미세먼지와 오존 등을 생성하여 대기오염의 원인이 되며 현장 작업자의 호흡계 질환, 심혈관계 질환을 유발함.

b. 질소화합물이 인체에 미치는 영향
   i. 호흡기 문제
      1. 질소화합물(NOx)은 호흡기를 자극하여 기침, 가래, 호흡곤란 등의 증상을 유발할 수 있습니다. 특히 천식 환자는 NOx 에 노출될 경우 상태가 악화될 수 있다. 장기적인 노출은 기관지염과 같은 만성 호흡기 질환을 유발하거나 악화시킬 수 있으며, 일부 연구에서는 폐암 위험 증가에 연관되어 있다고 제안되고 있다.
   ii. 면역체계 영향
      1. 일부 연구에서는 질소화합물(NOx)이 면역체계에 영향을 미치는 것으로 나타난다. 예를 들어, 이산화질소(NO2)는 폐 내

면역 반응을 변경하여 감염에 대한 저항력을 감소시켜 바이러스와 박테리아로 인한 질병에 취약하게 만들 수 있다.

iii. 심혈관 질환

1. 최근의 연구에서는 질소화합물(NOx)와 심혈관 질환 사이의 관련성도 지적되고 있다. 공기 오염과 심장질환 및 뇌졸중 사이의 관련성이 확인되면서 NOx 도 그 원인 중 하나로 보고 있다.

c. 모델 선택의 의의

i. 물리적 모델을 기반으로 한 질소화합물(NOx) 배출량은 연료 및 공기의 유체거동과 화학반응을 동시에 고려한 모델이 필요하기 때문에 실시간 질소화합물(NOx) 배출량 모니터링의 한계를 지니기 때문에 파이썬을 활용하여 데이터 값을 통해 인공지능을 머신 러닝으로 학습시켜 질소화합물(NOx) 배출량을 실시간으로 산출하여 모니터링할 수 있는 모델을 선택.

3. 배경지식

a. 데이터 (변수, 이름, 단위)

i. 주변 온도 (AT, Ambient Temperature) ℃

1. 측정된 지역의 대기 온도를 나타내며, 가스터빈 성능에 영향을 미치며, 온도가 높을수록 질소화합물의 발생량은 감소한다.

ii. 주변 압력 (AP, Ambient Pressure) mbar

1. 측정된 지역의 대기 압력을 나타내며, 대기 압력이 높을수록 질소화합물의 발생량은 증가한다.

iii. 주변 습도 (AH, Ambient Humidity) **%**

1. 공기 중의 수분 함량을 나타내며, 높은 습도는 공기의 밀도를 낮추어 주변습도가 높을수록 질소화합물의 발생량은 증가한다.

iv. 공기 필터 차압 (AFDP, Air Filter Differential Pressure) mbar

1. 공기 필터를 통과하는 공기의 압력 차이를 나타내며, 공기 필터 차압이 증가하면 질소화합물의 발생량은 감소한다.

v. 가스터빈 배기 압력 (GTEP, Gas Turbine Exhaust Pressure) mbar
    1. 가스터빈에서 배출되는 가스의 압력을 나타내며, 가스터빈 배기 압력이 증가하면 질소화합물의 발생량은 감소한다.
vi. 터빈 입구 온도 (TIT, Turbine Inlet Temperature) °C
    1. 가스터빈 터빈에 들어오는 가스의 온도를 나타내며, 터빈 입구 온도가 증가하면 질소화합물의 발생량은 감소한다.
vii. 터빈 통과 후 온도 (TAT, Turbine After Temperature) °C
    1. 터빈을 통과한 후의 가스 온도를 나타내며, 터빈 통과 후 온도가 증가하면 질소화합물의 발생량은 감소한다.
viii. 압축기 배출 압력 (CDP, Compressor Discharge Pressure) mbar
    1. 압축기에서 압축된 공기의 배출 압력을 나타내며, 압축기 배출 압력이 증가하면 질소화합물의 발생량은 감소한다.
ix. 터빈 에너지 생산량 (TEY, Turbine Energy Yield) MWH
    1. 터빈이 생성하는 에너지의 양을 나타내며, 터빈 에너지 생산량이 증가하면 질소화합물의 발생량은 감소한다.
x. 일산화탄소 (CO, Carbon Monoxide) mg/m3
    1. 연료 연소 과정에서 생성되는 일산화탄소의 농도를 나타내며, 일산화탄소량이 증가하면 질소화합물의 발생량은 증가한다.
xi. 질소화합물(NOx) mg/m3
    1. 질소와 산소가 결합하여 만들어진 화합물을 나타내며 주로 공장, 자동차, 발전소 등에서 연료를 태울 때 발생하는데, 대표적인 종류로는 이산화질소(NO2)와 일산화질소(NO)가 있습니다.

b. 머신 러닝 모델

i. 선형 회귀 (Linear Regression)
    1. 선형 회귀는 가장 기본적인 회귀 분석 방법입니다. 데이터의 독립변수(X)와 종속변수(Y) 간의 관계를 직선으로 표현하려고 합니다. 예를 들어, 키와 몸무게의 관계를 직선으로 표현하는 것입니다.
ii. 라쏘 회귀 (Lasso)
    1. 라쏘 회귀는 선형 회귀와 비슷하지만, 모델이 너무 복잡해지는 것을 방지하기 위해 일부 변수의 계수를 0 으로

만듭니다. 즉, 덜 중요한 변수는 무시하고 중요한 변수만 사용하는 회귀 방법입니다.

iii. 리지 회귀 (Ridge)
    1. 리지 회귀는 선형 회귀와 유사하지만, 회귀 계수가 너무 크지 않도록 제약을 가하는 방식입니다. 즉, 너무 큰 계수가 나오지 않도록 제어하여 모델이 과적합(overfitting)되지 않도록 합니다.

iv. K-최근접 이웃 회귀 (K-Nearest Neighbors Regressor)
    1. K-최근접 이웃(KNN) 회귀는 특정 데이터 포인트를 예측할 때, 그 데이터와 가장 가까운 K 개의 이웃 데이터를 참고하여 평균값을 구하는 방식입니다.

v. 결정 트리 회귀 (Decision Tree Regressor)
    1. 결정 트리 회귀는 데이터를 여러 조건에 따라 분할하여 트리 구조로 예측을 수행하는 방식입니다. 각 노드에서 데이터가 특정 조건에 따라 분기되며, 마지막 노드(리프 노드)에서 예측 값을 제공합니다.

vi. 엑스트라 트리 회귀 (Extra Trees Regressor)
    1. 엑스트라 트리 회귀는 여러 결정 트리를 만들어 그 결과를 평균 내어 예측을 수행하는 방식입니다. 랜덤 포레스트 라는 모델과 유사하지만, 더 많은 무작위성을 부여하여 다양성을 높입니다.

4. 개발 내용

a. 데이터 설명

i. 데이터 개수 : 36,733 개의 데이터

ii. 데이터 속성 : 11 개의 속성(이름, 변수, 단위)

1. 주변 온도 (AT) °C
2. 주변 압력 (AP) mbar
3. 주변 습도 (AH) %
4. 공기 필터 압력 차이 (AFDP) mbar
5. 가스터빈 배기 압력 (GTEP) mbar
6. 터빈 입구 온도 (TIT) °C
7. 터빈 배출 온도 (TAT) °C
8. 압축기 배출 압력 (CDP) mbar
9. 터빈 에너지 출력 (TEY) MWH
10. 일산화탄소 (CO) mg/m3
11. 질소 산화물 (NOx) mg/m3

b. 데이터 불러오기

```
data = pd.read_csv("./data/5.gt_full.csv")
```

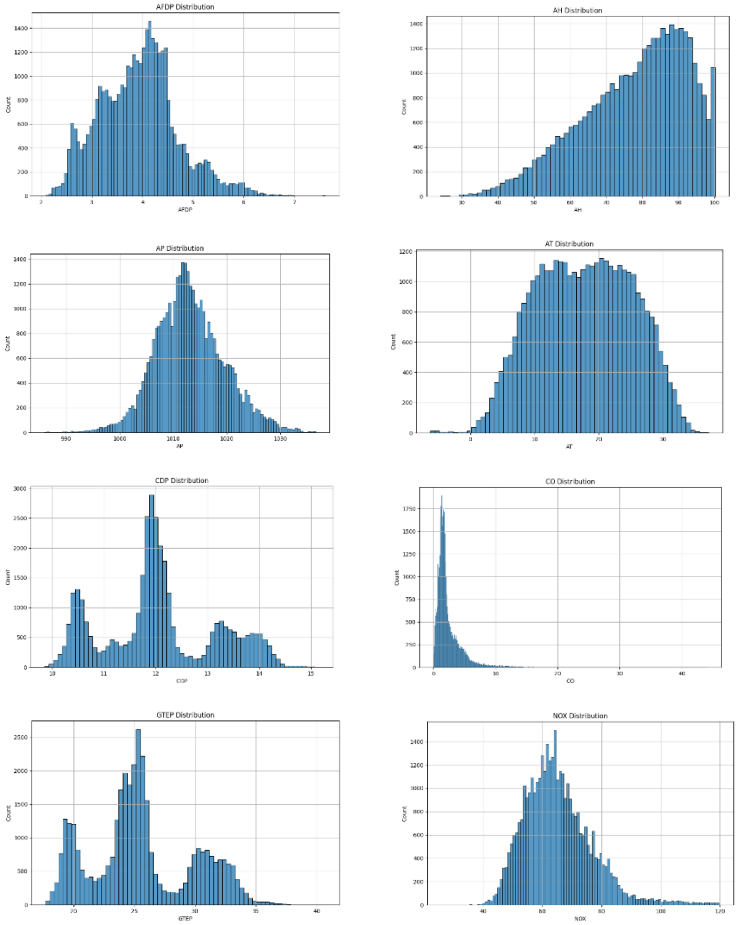
c. 데이터확인

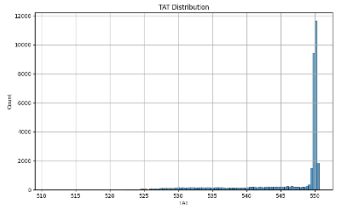
```
Gas Turbine Dataset:

   Unnamed: 0      AT      AP      AH  ...     TEY     CDP      CO     NOX
0           1  4.5878  1018.7  83.675  ...  134.67  11.898  0.32663  81.952
1           2  4.2932  1018.3  84.235  ...  134.67  11.892  0.44784  82.377
2           3  3.9045  1018.4  84.858  ...  135.10  12.042  0.45144  83.776
3           4  3.7436  1018.3  85.434  ...  135.03  11.990  0.23107  82.505
4           5  3.7516  1017.8  85.182  ...  134.67  11.910  0.26747  82.028
```

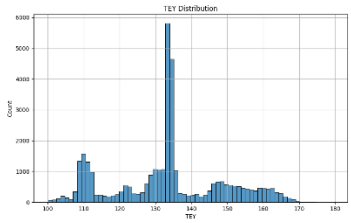
d. 불필요한 데이터 제거

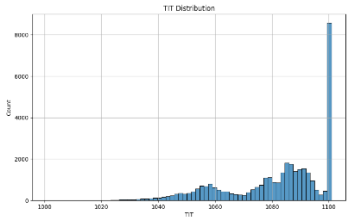
```
Dropping unnecessary columns:
      AT      AP      AH    AFDP  ...     TEY     CDP      CO     NOX
0  4.5878  1018.7  83.675  3.5758  ...  134.67  11.898  0.32663  81.952
1  4.2932  1018.3  84.235  3.5709  ...  134.67  11.892  0.44784  82.377
2  3.9045  1018.4  84.858  3.5828  ...  135.10  12.042  0.45144  83.776
3  3.7436  1018.3  85.434  3.5808  ...  135.03  11.990  0.23107  82.505
4  3.7516  1017.8  85.182  3.5781  ...  134.67  11.910  0.26747  82.028
```

e. 데이터 속성 그래프









f. 독립변수와 종속변수 나누기

i. 독립변수

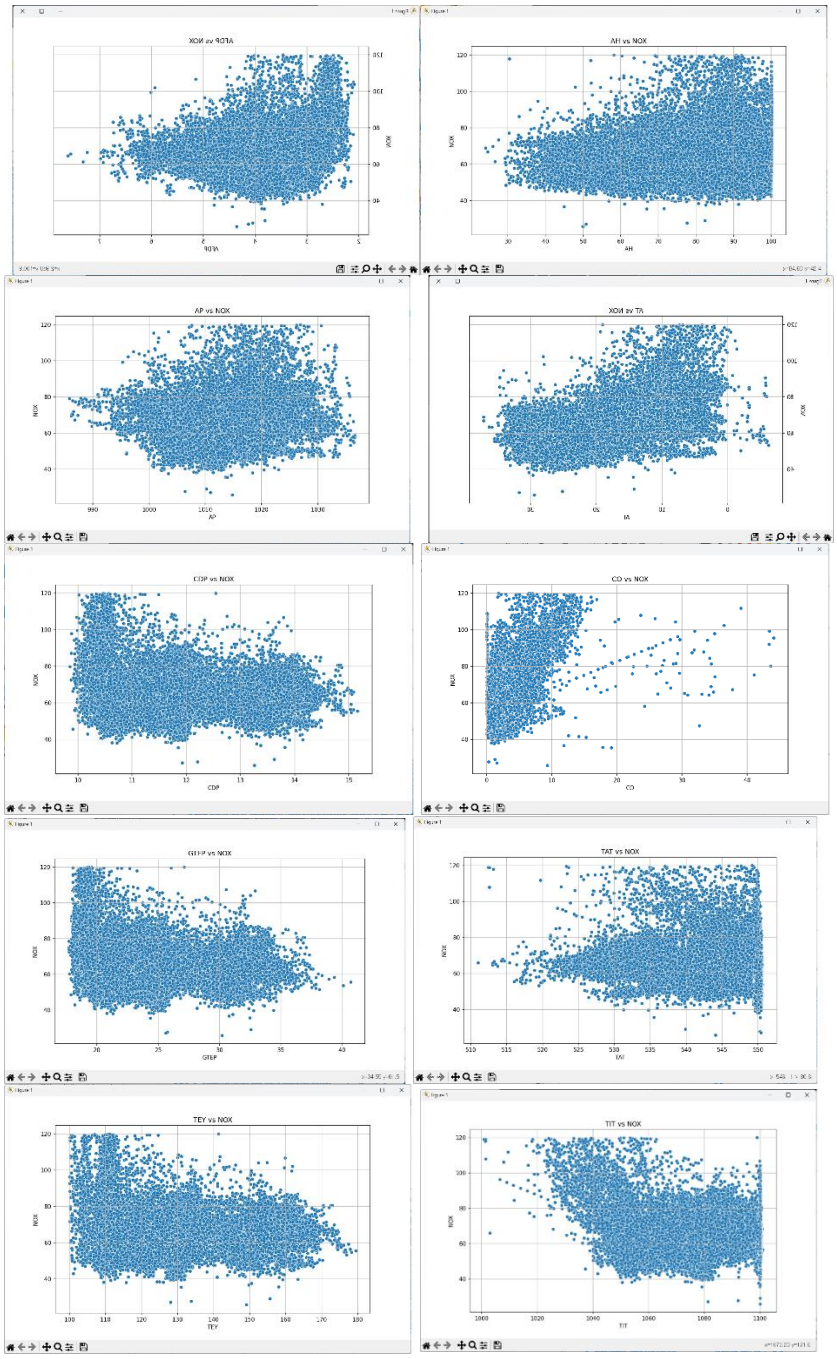
1. 주변 온도 (AT) °C
2. 주변 압력 (AP) mbar
3. 주변 습도 (AH) %
4. 공기 필터 압력 차이 (AFDP) mbar
5. 가스터빈 배기 압력 (GTEP) mbar
6. 터빈 입구 온도 (TIT) °C
7. 터빈 배출 온도 (TAT) °C
8. 압축기 배출 압력 (CDP) mbar
9. 터빈 에너지 출력 (TEY) MWH
10. 일산화탄소 (CO) mg/m3

ii. 종속변수

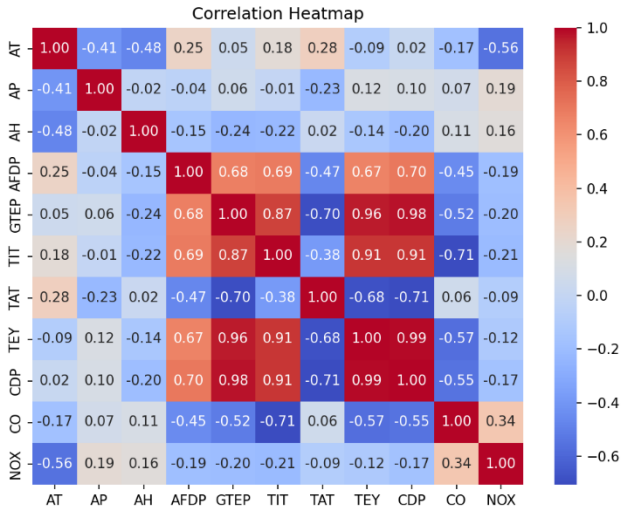
1. 질소 산화물 (NOx) mg/m3

```
Input Variables:
       AT      AP      AH    AFDP  ...     TAT     TEY     CDP       CO
0  4.5878  1018.7  83.675  3.5758  ...  549.83  134.67  11.898  0.32663
1  4.2932  1018.3  84.235  3.5709  ...  550.05  134.67  11.892  0.44784
2  3.9045  1018.4  84.858  3.5828  ...  550.19  135.10  12.042  0.45144
3  3.7436  1018.3  85.434  3.5808  ...  550.17  135.03  11.990  0.23107
4  3.7516  1017.8  85.182  3.5781  ...  550.00  134.67  11.910  0.26747

[5 rows x 10 columns]

Output variable:
0    81.952
1    82.377
2    83.776
3    82.505
4    82.028
Name: NOX, dtype: float64
```

g. 데이터 속성과 질소화합물 그래프



h. 데이터 속성별 상관관계 그래프



i. 데이터 스케일러

```
Standard scaling the input training variables:
         AT        AP        AH  ...       TEY       CDP       CO
0  1.056745 -0.656642  0.462865  ... -0.041464  0.087934 -0.286281
1  0.092683 -0.099450 -0.004915  ...  1.461932  1.462843 -0.848781
2  0.387809 -1.678163 -0.980067  ... -0.280494 -0.268660 -0.067691
3  1.146035  0.024371 -1.710895  ...  0.990919  1.215616 -0.638270
4 -0.546847 -0.842373 -0.471728  ... -0.177320 -0.279689  0.305257

[5 rows x 10 columns]

Standard scaling the input testing variables:
         AT        AP        AH  ...       TEY       CDP       CO
0 -1.691422  1.510219  0.938453  ...  1.499741  1.480305 -0.674261
1 -0.180154 -2.896248 -1.132078  ... -0.251016 -0.348618  0.563219
2 -1.063963  0.411310  1.264586  ... -1.557035 -1.418400  4.540978
3 -1.025924  0.395833  0.669808  ... -1.225083 -1.313628  0.781287
4 -1.130104  1.061369  0.356941  ...  0.186673 -0.046248  0.089232
```

j. 데이터 모델 평가
    i. 선형 회귀 (LinearRegression)
    ii. 라쏘 회귀 (Lasso)
    iii. 리지 회귀 (Ridge)
    iv. K-최근접 이웃 회귀 (KNeighborsRegressor)
    v. 결정 트리 회귀 (DecisionTreeRegressor)
    vi. 엑스트라 트리 회귀 (ExtraTreesRegressor)

```
Model Training:

Model used: LinearRegression().
Accuracy Acquired: 0.5667.

Model used: Lasso().
Accuracy Acquired: 0.357.

Model used: Ridge().
Accuracy Acquired: 0.5667.

Model used: KNeighborsRegressor().
Accuracy Acquired: 0.8719.

Model used: DecisionTreeRegressor().
Accuracy Acquired: 0.7555.

Model used: ExtraTreesRegressor().
Accuracy Acquired: 0.9012.

Best R2 Score Recorded: 0.9012.
```

k. 사용할 성능 지표
    i. MSE (Mean Squared Error)
        1. 예측값과 실제값의 차이를 제곱해 평균낸 값으로, 값이 작을수록 예측이 실제와 가깝다는 것을 의미합니다.
    ii. MAE(Mean Absolute Error)
        1. 예측값과 실제값의 차이를 절대값으로 변환해 평균값입니다. MSE 와 유사하게, 값이 작을수록 모델의 예측이 실제값에 가깝다는 것을 의미합니다.

l. 모델 2 개 비교 분석 및 모델 선정

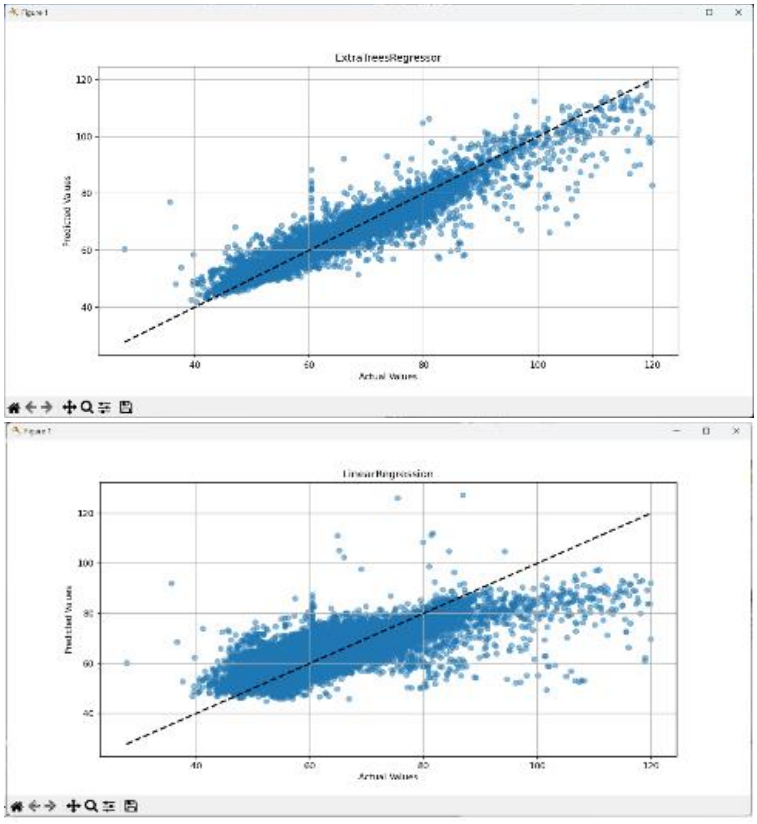
```
Model1 ExtraTreesRegressor:
R2 Score: 0.9007.
Mean Squared Error: 13.509830773542323.
Mean Absolute Error: 2.2664445444646093.
Model2 LinearRegression:
R2 Score: 0.5667.
Mean Squared Error: 58.97808375073865.
Mean Absolute Error: 5.381889286829399.
```

i. 엑스트라 트리 회귀 모델과 선형 회귀 모델을 비교 분석

ii. R2 Score

1. 모델이 실제 데이터를 얼마나 잘 설명하는지를 나타내는 지표입니다. 값이 1 에 가까울수록 모델이 데이터를 잘 설명한다는 의미입니다.

iii. 두 모델의 실제값과 예측값 그래프



iv. MSE, MAE 결과 값이 작을수록 예측값과 실제값이 가깝다는 것을 알 수 있습니다. 그 결과 가장 정확도가 높은 엑스트라 트리 회귀 (ExtraTreesRegressor)모델로 선정하였습니다.

5. 예측 결과

a. 엑스트라 트리 회귀 (ExtraTreesRegressor)모델로 데이터 분석

```
Best Model Performance: ExtraTreesRegressor().
Model Saved.
Model to be extracted from ./data/Extra_Trees_Regressor_Model.joblib.
Extracted Model:
ExtraTreesRegressor()
Enter the input details:
New data for prediction:
[[   7.7533 1011.8      73.067     2.6621   20.886  1039.7     545.41
   109.27     10.39      8.0651]]
Scaled new data:
[[-1.33698826 -0.19231511 -0.33256825 -1.63541476 -1.11551049 -2.38178007
  -0.10777482 -1.55318979 -1.53512021  2.47463614]]
Predicted Nitrous Oxides Emission Amount: 82.26
```

6. 결론

a. 가스터빈 배출가스 예측을 위한 다양한 머신 러닝 모델을 개발하고 평가한다. 최종적으로 가장 높은 성능을 보인 모델을 선정하였으며, 이 모델은 가스터빈의 안전성과 효율성을 높이는 데 기여할 수 있다.

b. 여러 모델을 비교함으로써 배출가스의 예측 정확도를 높이고, 위해노출 농도에 따른 인체에 미치는 영향을 예측할 수 있다. 이를 보완하기 위해 추가적인 데이터 수집과 모델 개선이 필요할 수 있다.

c. 모델의 한계로는 데이터의 품질과 양, 그리고 특정 환경 변수에 대한 반응 예측의 어려움이 있을 수 있다.

d. 모델 선정 및 예측 결과 등을 바탕으로 한 활용가치로는 이 모델은 미세먼지 등의 비산물질 발생 사업장에서 데이터를 추출하여 배출가스의 예측 정확도를 상승시켜 작업자의 인체에 미칠 수 있는 영향을 예측하고 이에 따른 보호조치 및 환경조성 등을 기대 할 수 있기 때문에 해당 작업에 활용 될 수 있다.

e. 작업장 환경에 대한 실시간 데이터를 통해 질소산화물의 배출량을 정확하게 예측하여 근로자의 건강을 보호하고, 쾌적한 작업환경을 조성하여 산업재해 예방의 기여할 수 있을 것으로 예상할 수 있다.

f. 질소화합물 뿐 아니라 건설분야에서 소요되는 에너지와 이산화탄소 등 다른 오염물질의 배출량 모니터링에 활용 가능할 것으로 보인다.