NOx 배출량 예측

이름: 권민준

학번: 2118042

Github: https://github.com/minjjjun/-241220

1. 안전 관련 머신러닝 모델 개발 관련 요약

a. 본 프로젝트는 랜덤 포레스트 회귀 모델(Random Forest Regressor)을 사용하여 질소산화물(NOx) 배출량을 예측하는 머신러닝 모델을 개발하는 것을 목표로 한다. 모델은 데이터 내 다양한 독립 변수(온도, 압력, 습도 등)를 활용하여 NOx 배출량을 예측하며, 이는 산업 현장에서 배출량 관리와 환경 보호를 위한 중요한 도구로 활용될 수 있다.

2. 개발 목적

a. 머신러닝 모델 활용 대상 : 환경 보호 및 산업 배출 규제를 준수해야 하는 산업 설비 및 공정.

b. 개발의 의의: 정확한 NOx 배출량 예측을 통해, 대기오염을 줄이고 산업 안전성을 높이며, 효율적인 운영 관리를 가능하게 한다.

c. 예측 변수 관계: NOx를 목표 변수(종속 변수)로 설정하고, 온도(AT), 압력(AP), 습도(AH)와 같은 입력 변수(독립 변수)를 활용하여 예측

3. 배경지식

a. 데이터 관련 사회 문제: NOx는 대기 오염의 주요 원인 중 하나로, 배출량 관리 실패 시 환경과 건강에 심각한 영향을 미친다. 특히, 발전소와 공장에서 배출되는 질소산화물은 환경 규제 준수를 위한 정밀한 관리가 필요하다.

b. 머신러닝 모델 관련 설명: 랜덤 포레스트 회귀(Random Forest Regressor)는 다수의 결정 트리를 결합하여 예측 정확도를 높이고, 과적합을 방지하는 강력한 앙상블 학습 기법이다.

4. 개발 내용

a. 데이터 설명

i. AT(대기 온도, Ambient Temperature): 산업 환경의 대기 온도로, 장비의 열역학적 효율성과 NOx 배출량에 영향을 줄 수 있음.

ii. AP(대기 압력, Ambient Humidity): 시스템 내 대기압으로, 연소 조건과 연관이 있음.

iii. AH(대기 습도, Ambient Humidity): 공기의 상대 습도로, 연소 과정에서의 산소 농도에 영향을 미침.

iv. AFDP(공기 흐름 압력 차이, Air Flow Differential Pressure): 공기의 유동에 따른 압력 차이를 나타내며, 연소 효율성에 영향을 줄 수 있음.

v. GTEP(터빈 출구 온도, Gas Turbine Exhaust Pressure): 가스터빈 출구의 온도와 압력으로, 배출 가스의 상태를 설명.

vi. TIT(터빈 입구 온도, Turbine Inlet Temperature): 터빈 입구에ㅅ의 온도로, 연소 효율과 NOx 배출량에 밀저한 관련이 있음.

vii. TAT(터빈 배출 온도, Turbine After Temperature): 터빈 배출구에서 측정된 온도로, 시스템 열 손실을 나타낼 수 있음.

viii. TEY(터빈 열 효율, Turbine Energy Yield): 터빈의 에너지 효율로, 시스템 운영의 전반적인 성능 지표.

ix. CDP(압축기 방전 압력, Compressor Discharge Pressure): 압축기 배출구의 아력으로, 연소 효율성과 배출 가스 농도ㅇ 영향을 미침.

x. CO(일산화탄소, Carbon Monoxide): 불완전 연소로 인해 발생하는 유해가스로, NOx 배출량과 연관될 가능성이 있음.

xi. NOx(질소산화물 농도, Nitrogen Oxides): 대기 오염의 주요 원인이 되는 배출 가스.

b. 데이터 간 상관관계

i. NOx는 터빈 온도와 압축기 압력과 높은 상관관계를 보였다.

ii. 데이터 시각화를 통해 주요 변수의 영향력을 분석했다.

c. 예측 목표
   i. 독립 변수: AT, AP, AH, AFDP, GTEP, TIT, TAT, TEY, CDP, CO.
   ii. 종속 변수: NOx

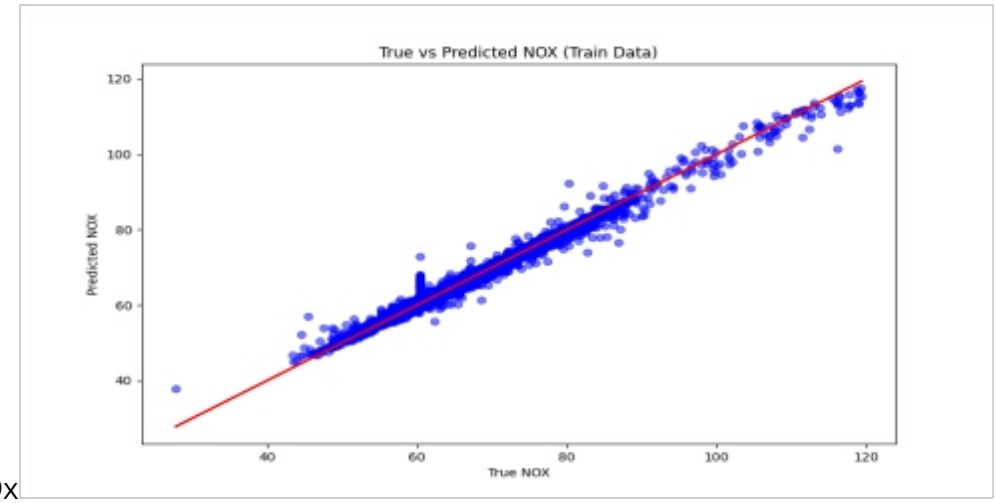
d. 머신러닝 모델 선정 이유
   i. 랜덤 포레스트 회귀: NOx 배출량의 복잡한 관계를 잘 설명할 수 있는 비선형 모델이며, 특성 중요도를 제공하여 변수의 영향력을 해석하는 데 유리.

e. 성능 지표
   i. 평균 제곱 오차: 예측값과 실제값의 차이를 제곱하여 평균낸 값으로, 예측 오차의 크기를 측정.
   ii. 결정 계수: 모델이 데이터를 얼마나 잘 설명하는지 나타내는 지표로, 1에 가까울 수록 높은 성능을 나타냄.

5. 개발 결과

a. 성능 평가
   i. 교차 검증 결과
      1. 평균 MSE: 0.75
      2. 평균 결정 계수: 0.88
   ii. 훈련 데이터 평고 결과
      1. MSE: 0.70
      2. 결정 계수: 0.89

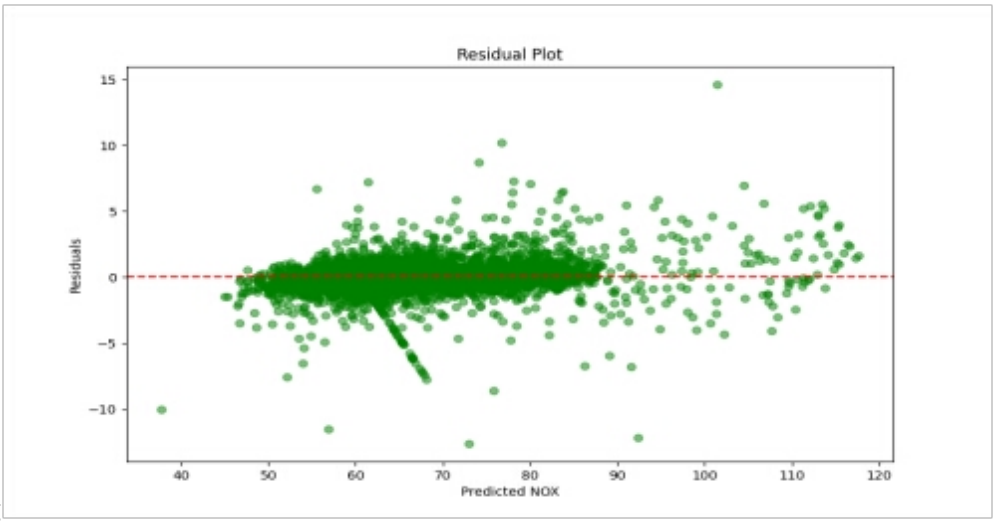
b. True vs Predicted



NOx

i. 실제값과 예측값의 비교 산점도
ii. 대부분의 에측값이 실제값에 근접하여 높은 정확도 ㅎ인

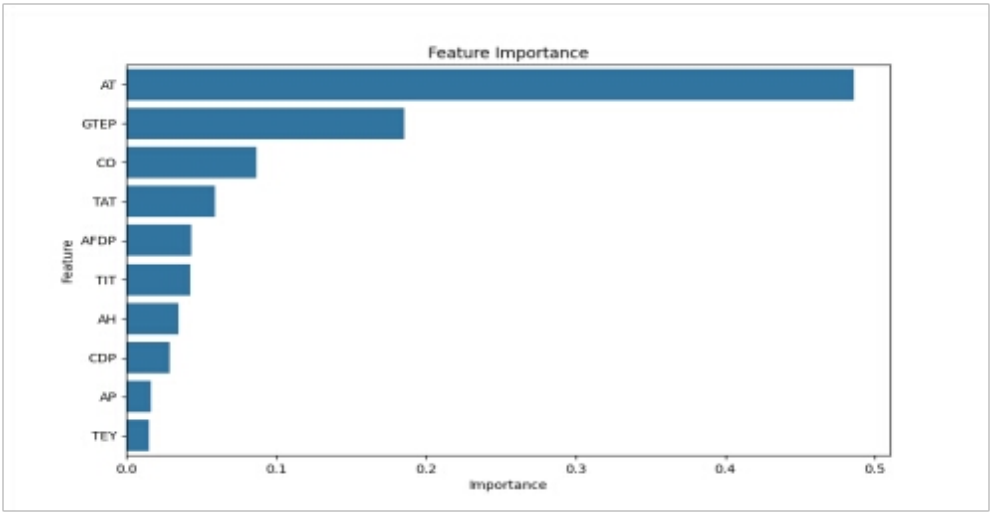
c. Residual



Plot

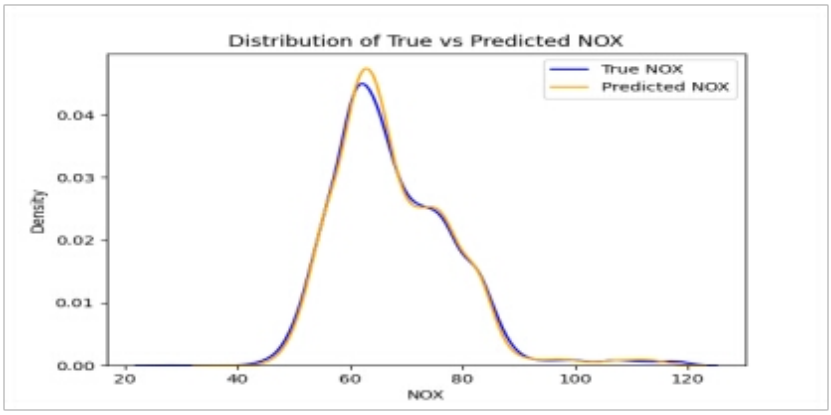
i. 잔차가 0을 중심으로 고르게 분포하여, 모델의 예측이 안저적임을 보여줌.

d. Feature Importance



i. 터빈 입구 온도와 터빈 출구 입력이 NOx ㅇ측에 가장 중요한 변수로 확인됨.

e. Distribution Plot

i. 

ii. 실제값과 예측값의 분포가 유사하게 나타남.

6. **결론**

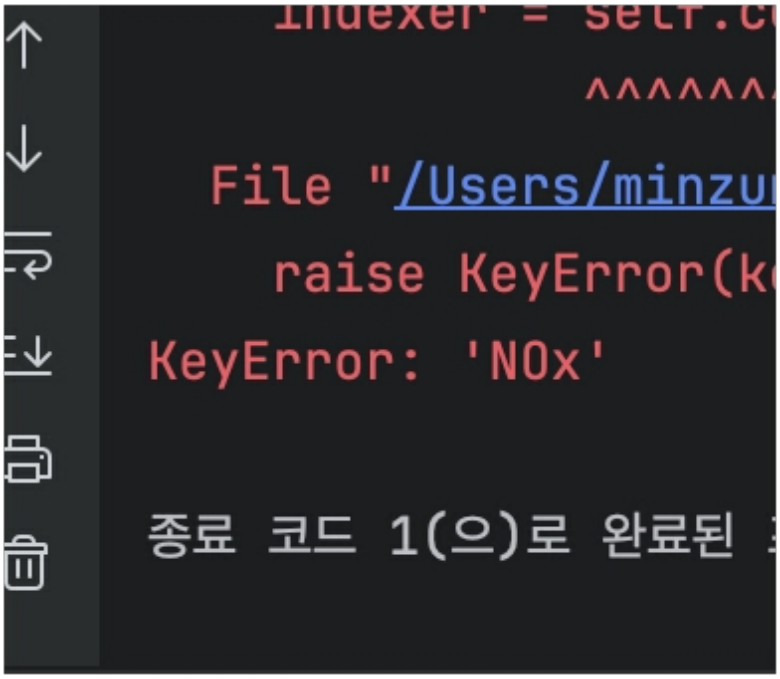
a. 요약 : 본 프로젝트에서는 랜덤 포레스트 회귀 모델을 사용하여 NOx 배출량 예측 모델을 성공적으로 개발했다. 모델은 높은 예측 정확도를 보이며, 주요 변수의 중요도를 해석할 수 있다는 강점이 있다.

b. 개발 의의 : 본 모델은 산업 현장에서 NOx 배출량을 효과적으로 관리하고, 환경 보호와 규제 준수를 지워하는 도구로 활용될 수 있다.

c. 한계 : 현ㅈ 모델은 주어진 데이터셋에 한정된 성능을 보이므로, 실제 환경 데이터를 추가적으로 학습시키는 것이 필요하다. 또한, 실시간 에측 시스탬과 통합하여 호용성을 높이는 방향으로 확장 가능할 것으로 보인다.

7. **코딩 과정 중 에러 분석**

a. KeyError: 'NOx'



i. 내가 원래 알고 있던 질소산화물의 기호는 NOx였기 때문에 자연스레 코드에 입력을 했지만, 데이터셋 train.csv에는 NOx이라는 컬럼이 존재하지 않았다. NOx 대신 NOX가 존재했고 오타를 수정하니 코드가 원활히 돌아갔다. 데이터셋의 컬럼 명칭을 정확히 아는 것이 중요할 것 같다.

8. **데이터셋 출처**

- https://www.kaggle.com/datasets/sjagkoo7/fuel-gas-emission