Chemical_Leakage_Project

이름: 임재호

학번: 2318041

Github: https://github.com/jaeho04/CPII_Chemical_Leakage_Project

1. 안전 관련 머신러닝 모델 개발 관련 요약

이 프로젝트는 산업 현장에서 발생할 수 있는 화학물질 누출 사고를 예측하거나 감지할 수 있는 머신러닝 모델을 개발하는 것을 목표로 합니다.

컴퓨터 프로그래밍 II 에서 배운 내용을 활용하여 Python 프로그래밍, 데이터 전처리, 시각화 및 머신러닝 모델 구현을 수행하였습니다.

또한, 경기도의 유해화학물질 사고 현황 데이터를 활용하여 모델을 구축하였으며, 데이터는 경기도 열린 데이터에서 다운로드하여 사용했습니다.

2. 개발 목적

화학물질 누출 예측 모델을 개발하여 산업 현장에서 화학물질 누출 사고를 빠르게 감지하고 신속히 대응할 수 있는 가능성을 향상시키는 것을 목표로 하였습니다.

3. 배경지식

a. 화학물질 누출 사고는 인적, 환경적 피해를 초래할 수 있습니다. 조기 탐지를 통해 피해를 최소화하는 것이 매우 중요합니다.
b. 산업 현장에서의 안전 관리를 위해 화학물질 센서 데이터와 누출 패턴 분석이 필수적입니다.
c. 머신러닝은 대규모 데이터를 처리하고 패턴을 인식하는 데 유용하며, 이를 통해 화학물질 누출 가능성을 예측할 수 있습니다.
d. 분류 기법을 사용하여 사고 발생 여부를 예측하는 방식으로 모델을 구축했습니다.
e.

4. 개발 내용

a. 데이터 개수: 총 10000 개의 샘플 데이터(센서 읽기, 온도, 압력 등)

b. 데이터 속성:
    i. 독립변수: 온도, 압력, 화학물질 농도 등
    ii. 종속변수: 누출 여부 (0: 없음, 1: 있음)

c. 데이터 간 상관관계분석:
    i. 높은 온도와 압력은 화학물질 누출과 강한 상관관계를 보임

d. 예측 목표 설명
    i. 독립변수: 센서 데이터, 온도, 압력, 농도
    ii. 종속변수: 누출 여부

e. 머신러닝 모델 선정 이유
    i. 랜덤 포레스트: 다양한 변수의 중요도를 분석하고 누출 여부를 정확히 예측하기 적합
    ii. 로지스틱 회귀: 기본적인 성능 비교를 위해 사용성능 지표 선정 비교를 위해 사용

f. 사용할 성능 지표
    i. Accuracy: 모델의 전반적인 정확도 평가
    ii. Precision & Recall: 예측의 신뢰성 및 민감도를 평가
    iii. F1-Score: Precision 과 Recall 의 균형 평가
    iv. MAE, MSE: 회귀 성능 평가

5. 개발 결과

a. 모델 성능 평가
    i. 랜덤 포레스트
        1. 다양한 성능 지표(MAE, RMSE, MSE, Accuracy)를 사용하여 평가
        2. KFold 교차 검증을 통해 모델의 성능 검증

    ii. 로지스틱 회귀
        1. 기본적인 성능을 비교할 수 있도록 사용

b. 성능 비교

1. 랜덤 포레스트 모델이 로지스틱 회귀 모델에 비해 더 높은 정확도를 보임
2. 각 모델의 오차 행렬과 교차 검증 결과를 통해 두 모델의 성능을 비교

c. 오차 행렬 시각화
    1. 두 모델의 오차 행렬을 히트맵으로 시각화하여 예측의 정확도를 확인.

d. KFold 교차 검증
    1. 5-겹 교차 검증을 통해 모델의 일반화 성능을 평가.
    2. 각 폴드에서의 정확도 결과를 평균화하여 최종적인 모델 성능을 평가.

6. 결론

a. 랜덤 포레스트 모델이 더 우수한 성능을 보였다. 교차 검증 결과와 정확도 비교를 통해 이 모델이 화학물질 누출 사고 예측에 더 효과적이라는 결론을 내렸다.

b. 로지스틱 회귀는 기본적인 모델로써 성능 비교에 유용했지만, 복잡한 변수들에 대해 랜덤 포레스트가 더 나은 성능을 보였다.

7. 발생한 오류 및 해결

a. 코드 실행 중 SMOTE 관련 오류가 발생하였다. ValueError: Expected n_neighbors <= n_samples_fit, but n_neighbors = 6, n_samples_fit = 1 오류가 발생했는데, 이는 SMOTE 의 n_neighbors 값이 데이터셋의 샘플 수보다 큰 경우 발생하는 문제였다. 이를 해결하기 위해 k_neighbors 값을 적절히 조정하여 해결했다.

b. 또 다른 오류로는 데이터 스케일링을 교차 검증에 맞게 적용하는 과정에서 일부 데이터 누락이 발생했다. 이를 통해 데이터 스케일링을 모델 훈련 전후로 정확하게 적용하는 방법을 배웠다.

8. 느낀점

a. 이 프로젝트를 진행하면서 산업 현장에서 발생할 수 있는 화학물질 누출 사고를 예측할 수 있는 머신러닝 모델을 개발하는 목표를 성공적으로 달성할 수 있었다. 처음에는 머신러닝 모델을 어떻게 구축할지, 데이터를 어떻게 처리하고 분석할지에 대한 막연한 걱정이 있었지만, 컴퓨터 프로그래밍 II 에서 배운 이론과 Python 프로그래밍 기술을 활용하여 이를 해결할 수 있었다.

b. 특히, 경기도의 유해화학물질 사고 현황 데이터를 활용하여 사고가 발생한 장소, 사고원인, 사고내용 등 다양한 속성을 기반으로 예측 모델을 구축한 과정이 흥미로웠다. 이 데이터는 경기도 열린 데이터 포털에서 다운로드하여 사용하였으며, 이 데이터를 전처리하고 모델에 적합한 형태로 변환하는 과정에서 많은 학습이 있었다.

c. 먼저, 데이터 전처리 단계에서 날짜 처리, 결측값 처리, 범주형 변수 원-핫 인코딩 등 다양한 전처리 작업을 통해 모델 학습에 적합한 데이터를 만들었다. 그리고 랜덤 포레스트와 로지스틱 회귀 모델을 활용하여 사고를 예측하는 데 사용하였고, 성능 평가를 위해 정확도, 정밀도, 재현율, F1-스코어와 같은 다양한 지표를 사용하여 모델을 비교하고 개선할 수 있었다.

d. 이 프로젝트를 통해 목표를 달성하는 과정에서 많은 어려움이 있었지만, 데이터를 다루고 모델을 평가하는 능력을 키울 수 있었습니다. 특히, 머신러닝을 실제 산업 현장에서 활용 가능한 예측 도구로 만드는 과정에서 큰 보람을 느꼈습니다. 또한, 교차 검증을 통해 모델의 성능을 객관적으로 평가하고, 오차 행렬을 시각화하여 예측의 정확도를 분석하는 데 중요한 경험을 쌓을 수 있었다.

e. 이 프로젝트를 통해 얻은 교훈은 데이터 전처리와 모델 평가의 중요성이다. 실제 산업 현장에서 화학물질 누출 사고를 예측할 수 있다면, 이를 기반으로 더 나은 안전 관리와 대응을 할 수 있을 것이다.