**공기질 예측**

**학번: 2018054**

**이름: 공왕택**

**Github address: https://github.com/2018054kongwangtaek/HW\_2**

1. **안전 관련 머신러닝 모델 개발의 목적**

목적은 시간 경과에 따른 가스 농도의 변화를 모니터링하여 실내 환경에서 일어나는 활동 유형을 평가하는 것에 있습니다. 이 모델은 정상 활동, 식사 준비, 연기의 존재, 청소와 같은 여러 상황을 식별하고, 이러한 활동들로부터 발생할 수 있는 안전 문제를 감지합니다. 센서를 통해 수집된 데이터를 활용하여 머신러닝 알고리즘을 훈련시켜 모델을 개발하고, 이를 통해 활동 패턴을 식별함으로써 안전 상태를 모니터링하고 사용자에게 필요한 경고 및 안전 조치를 제공하는 것이 최종 목표입니다.

독립 변수(X): MQ1, MQ2, MQ3, MQ4, MQ5, MQ6= 아날로그 CO2 가스 센서 (가스 센서 출력 값)

종속 변수(y): CO2 농도

1. **안전 관련 머신러닝 모델의 네이밍의 의미**

모델의 이름 "RandomForestClassifier"는 마치 숲 안에서 나무들이 모여 안정성을 더하는 개념입니다. "랜덤"은 각 나무가 다양한 특징들을 학습하도록 하는데, 이는 모델이 다양한 상황에서 유연하게 대응할 수 있게 만들어줍니다. "Classifier"는 모델이 데이터를 분류하고 예측하는 역할을 말하며 그래서 RandomForestClassifier는 여러 결정 트리를 모아 안정적이면서도 정확한 예측을 해내는 머신러닝 모델을 지칭합니다.

1. **개발 계획**

데이터에 대한 요약 정리 및 시각화: 데이터를 살펴보고 각 가스 센서의 특성을 이해하기 위해 시각화를 활용할 것입니다. 이를 통해 데이터의 분포와 이상치 등을 확인하며, 어떤 센서가 중요한 역할을 하는지 파악합니다.

데이터 전처리 계획 : 데이터를 학습과 테스트로 나누고, 결측치를 처리하며, 각 센서의 값들을 일정한 범위로 스케일링하여 모델이 더 잘 학습할 수 있도록 전처리를 진행했습니다.

어떠한 머신러닝 모델을 사용할 것인지: RandomForestClassifier를 사용할 것이며, 각 트리가 다양한 측면에서 학습함으로써 안정성을 확보하고, 모델이 과적합되지 않도록 설정합니다.

RandomForestClassifier는 여러 의사결정 트리를 조합해 만든 머신러닝 모델입니다. 각 트리는 데이터 일부를 학습하고, 무작위로 일부 특성을 선택해 다양성을 확보합니다. 이렇게 함으로써 모델은 안정적이고 정확하게 분류 작업을 수행할 수 있습니다.

머신러닝 모델 예측 결과가 어떠할 지 : 데이터 베이스를 기반으로 모델은 센서의 출력을 각 상황별로 식별하여 사용자에게 경고 혹은 위험 상황을 알릴 수 있습니다.

사용할 성능 지표 : 모델의 정확도를 주요 성능 지표로 사용합니다.

성능 검증 방법 계획 등 : SHAP을 활용하여 모델의 결정 과정을 해석하고, 학습 데이터와 테스트 데이터를 사용하여 모델의 일반화 성능을 검증합니다. 이를 통해 모델의 안정성과 신뢰성을 평가합니다.

SHAP (Shapley Additive exPlanations) : 머신러닝 모델의 예측을 설명하는 도구로, 각 특성이 예측에 미치는 영향을 알려줍니다.

1. **개발 과정** 
   1. 계획 후 실제 학습 모델 개발 과정을 기록

실제 모델 개발에 앞서 데이터를 불러오고,

기본적인 통계적 분석을 통해 데이터의 특성을 파악합니다. preprocess\_inputs 함수를 통해 데이터를 학습용과 테스트용으로 나누고, 센서 데이터와 CO2 농도를 독립 및 종속 변수로 나눕니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

이후 RandomForestClassifier를 초기화하고 학습 데이터로 모델을 훈련시킵니다.

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

* 1. 각 함수는 어떻게 동작하는 지 구체적으로 설명

Def preprocess\_inputs(df) : 이 함수는 데이터를 모델에 입력으로 넣을 수 있도록 처리하는 역할을 합니다. 즉 데이터를 학습과 테스트로 나누고, 모델이 학습할 때 사용할 독립 변수와 종속 변수를 정리하여 반환하는 역할을 합니다.

X = df.drop('CO2', axis=1): 'CO2' 열을 제외한 나머지 열들을 독립 변수로 설정합니다. 독립 변수는 모델이 학습할 때 사용되는 입력 특성들로 구성됩니다.

y = df['CO2']: 데이터프레임에서 'CO2' 열을 종속 변수로 지정합니다. 종속 변수는 모델이 예측하려는 대상이 되는 변수입니다.

return train\_test\_split(X, y, train\_size=0.7, shuffle=True, random\_state=1): 마지막으로, 독립 변수(X)와 종속 변수(y)를 train\_test\_split 함수를 사용하여 학습 데이터와 테스트 데이터로 나눕니다. 이 함수는 데이터를 주어진 비율에 따라 나누고, 데이터를 무작위로 섞어줍니다. 반환값으로는 학습용 데이터(X\_train, y\_train)와 테스트용 데이터(X\_test, y\_test)가 제공됩니다.

전처리 함수: preprocess\_inputs 함수는 데이터프레임을 복사하고, 종속 변수와 독립 변수를 나누어 학습용과 테스트용 데이터로 분할합니다.

모델 학습: RandomForestClassifier 모델을 초기화하고, fit 메서드를 사용하여 학습 데이터에 모델을 훈련시킵니다.

성능 평가: model.score 메서드를 사용하여 테스트 데이터에서 모델의 정확도를 계산합니다.

* 1. 에러 발생 지점 및 해결 과정



X\_test의 컬럼 순서가 모델이 훈련된 데이터의 컬럼 순서와 일치하지 않는 경우 오류가 발생했습니다. X\_test의 컬럼 순서를 모델을 훈련할 때의 순서와 동일하게 맞춰야 합니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

* 1. 학습 모델의 성능 평가

모델의 성능은 가스감지기 정확도 측정입니다. 테스트 데이터를 사용하여 모델이 새로운 데이터에서 얼마나 잘 수행되는지 평가했습니다.

* 1. 결과 시각화 텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

     자동 생성된 설명

SHAP (SHapley Additive exPlanations) 라이브러리를 활용하여 모델의 해석가능성을 높이기 위해 시각화를 수행했습니다. 이를 통해 각 센서의 기여도를 이해하고 모델의 예측 과정을 시각적으로 확인할 수 있습니다.

1 : 정상 상황 - 활동 : 깨끗한 공기, 잠을 자거나 공부하거나 쉬는 사람

2 : 식사 준비 – 활동 : 고기 또는 파스타 요리, 튀긴 야채. 방에 한두 사람, 강제 공기 순환

3 : 연기의 존재 – 활동 : 창문과 문이 닫힌 방에서 짧은 시간 동안 종이와 나무를 태우는 행위

4 : 청소 - 활동 : 암모니아 또는 알코올이 함유 된 스프레이 및 액체 세제 사용. 강제 공기 순환을 활성화하거나 비활성화할 수 있습니다.

1. **개발 후기**
   1. 개발 후 느낀 점 설명

실제로 빅데이터를 전처리하고 머신러닝 모델을 도입하여 원하는 정보를 추출한 것이 처음이라 코드를 작성하는데 많은 어려움이 겪었습니다. 하지만 수업시간에 배운 내용을 활용하여 원하는 데이터를 가공하고 시각적 정보로 변환해 볼 수 있는 뜻깊은 시간이 였습니다.

해당 과제를 통하여 러신머닝에 대한 이해도를 높일 수 있었으면 다음에는 전처리하려는 데이터까지 직접 수집하여 원하는 러신머닝 모델을 만들어 보고 싶습니다.