Air Quality Dataset bot

학번: 2218311
이름: 박주혁
Github address: https://github.com/park4352/02.homework

1. 안전 관련 머신러닝 모델 개발의 목적
    a. 학습 모델 활용 대상:
    이 학습 모델은 실내 환경에서의 공기질을 예측하고 모니터링하는 데 활용될 수 있습니다. 주요 대상은 주거 공간, 사무실, 학교 등 다양한 실내 환경에서 발생할 수 있는 공기질 변화를 감지하고 예측하는 것입니다. 이를 통해 공기질 개선을 위한 조치를 취하거나 사용자에게 경고를 제공하여 건강하고 편안한 환경을 유지하는 데 기여할 수 있습니다. 또한, 산업 분야에서 환경 모니터링에 활용하여 생산 시설 내의 공기질을 지속적으로 추적하고 개선하는 데에도 적용될 수 있습니다.
    b. 데이터의 어떠한 독립 변수를 사용하여 어떠한 종속 변수를 예측하는 지

1. `train_and_evaluate_model` 함수:
    - `train_and_evaluate_model` 함수는 CSV 파일에서 데이터를 읽어와서 머신러닝 모델을 학습하고 성능을 평가하는 역할을 합니다.
    - CSV 파일에는 센서의 출력 값들이 독립 변수로 사용되고, 마지막 열은 공기질을 나타내는 종속 변수로 가정합니다.
    - RandomForestClassifier 를 사용하여 모델을 학습하고, 테스트 세트에서의 정확도와 분류 보고서를 출력합니다.
2. `predict_air_quality` 함수:
    - `predict_air_quality` 함수는 학습된 모델과 새로운 데이터를 받아서 새로운 데이터의 공기질을 예측합니다.
3. `__main__` 부분:
    - `__main__` 부분에서는 학습 데이터셋 파일의 경로를 지정하고, `train_and_evaluate_model` 함수를 호출하여 모델을 학습합니다.
    - 그리고 새로운 데이터를 만들어서 학습된 모델을 사용하여 공기질을 예측하고 결과를 출력합니다.

    c. 개발의 의의: 학습 모델 개발 시 어떠한 가치를 생성하는지

2. **정확한 예측과 의사 결정 지원:**
    a. 학습된 모델은 새로운 데이터에 대한 예측을 수행할 수 있습니다. 이를 통해 실내 공기질 모니터링에서는 향후 시간 동안의 공기질을 예측하거나, 특정 활동에 따른 공기질 변화를 모니터링하는 데 도움이 됩니다.
3. **환경 모니터링 및 개선:**

a. 학습된 모델을 사용하여 환경에서 발생하는 다양한 활동이나 사건에 대한 정보를 수집하고, 이를 통해 실내 공기질을 개선할 수 있는 방법을 찾을 수 있습니다. 예를 들어, 특정 활동이나 물질의 증가에 따라 공기질이 나빠진다면, 해당 활동을 최소화하거나 대처할 수 있는 정책을 마련할 수 있습니다.

4. **사용자 편의 및 안전 증진:**
   a. 모델을 사용하여 사용자에게 현재 공기질 상태를 제공하면, 사용자는 공기질이 나빠질 때 대처할 수 있는 기회를 얻습니다. 또한, 급격한 환경 변화에 대한 경고를 통해 사용자의 안전을 증진시킬 수 있습니다.
5. **자동화 및 효율성 향상:**
   a. 실시간으로 환경 데이터를 모니터링하고 예측하는 머신러닝 모델은 환경 조건의 변화를 자동으로 감지하고 대응할 수 있습니다. 이를 통해 에너지 소비를 최적화하거나 공기질을 개선하기 위한 자동 시스템의 구축이 가능합니다.

6. 안전 관련 머신러닝 모델의 네이밍의 의미

많은 사람들이 사용하기 쉽게 기본적인 용도의 뜻이 담긴 간결한 네이밍을 하였다.

7. 개발 계획

a. 데이터에 대한 요약 정리 및 시각화

데이터의 특성과 분포를 확인합니다. 센서 출력 값과 공기질에 대한 통계량, 분포를 살펴보고, 각 센서 간의 상관 관계 등을 시각화하여 데이터를 탐색합니다.

b. 데이터 전처리 계획

결측치 처리: 결측치가 있는 경우 해당 결측치를 대체하거나 삭제합니다. 이상치 처리: 이상치를 탐지하고 처리합니다. 데이터 정규화 또는 표준화: 모든 변수를 동일한 척도로 맞추어 주는 작업을 수행합니다. 범주형 데이터 처리: 필요한 경우 범주형 데이터를 수치형으로 변환하거나 인코딩합니다.

c. 어떠한 머신러닝 모델을 사용할 것인지 (해당 머신러닝 모델의 이론 추가)

RandomForestClassifier 를 사용할 예정입니다. RandomForest 는 여러 개의 의사 결정 트리를 사용하여 데이터를 학습하고 예측하는 앙상블 학습 모델입니다. 각 트리는 부트스트랩 샘플로 학습되며, 각 노드에서 최적의

특성을 찾아 데이터를 분할합니다. 이러한 다수의 트리의 예측을 종합하여 최종 예측을 수행합니다.

d. 머신러닝 모델 예측 결과가 어떠할 지

모델은 센서 출력 값들을 학습하여 주어진 입력에 대한 활동의 인덱스를 예측할 것입니다. 즉, 주어진 공기질 센서 데이터에 대해 어떤 활동이 발생하고 있는지를 분류할 것입니다.

e. 사용할 성능 지표

확도(accuracy): 전체 샘플 중 정확하게 분류된 샘플의 비율을 나타냅니다. 분류 보고서(classification report): 정밀도(precision), 재현율(recall), F1 점수 등을 종합적으로 제공하여 모델의 성능을 평가합니다.

f. 성능 검증 방법 계획 등

학습 데이터와 테스트 데이터로 데이터를 나누어 모델을 학습하고 성능을 평가합니다.
교차 검증(cross-validation)을 사용하여 모델의 일반화 성능을 더 신뢰할 수 있게 평가합니다.
혼동 행렬(confusion matrix)을 통해 모델이 어떤 클래스를 어떻게 혼동하는지 시각적으로 확인합니다.

8. 개발 과정

a. 계획 후 실제 학습 모델 개발 과정을 기록 (*개발 과정 캡쳐 필수)

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report

def train_and_evaluate_model(file_path: object) -> object:
    # CSV 파일에서 데이터 읽기
    data = pd.read_csv(file_path, header=None)  # 헤더가 없는 경우에는 header=None으로 설정합니다.

    # 데이터 전처리
    X = data.iloc[:, :-1]  # 센서 출력 값을 feature로 사용
    y = data.iloc[:, -1]   # 공기질을 나타내는 값은 마지막 열로 가정합니다.

    # 데이터를 학습 세트와 테스트 세트로 나누기
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```python
if __name__ == "__main__":

    training_file_path = "/Users/admin/PycharmProjects/pythonProject12/dataset.csv"

    # 모델 학습
    trained_model = train_and_evaluate_model(training_file_path)

    # 사용자에게 입력 받은 새로운 데이터 (센서 출력 값) -
    new_data = pd.DataFrame([[670, 696, 1252, 1720, 1321, 2431]],
                            columns=['sensor1', 'sensor2', 'sensor3', 'sensor4', 'sensor5', 'sensor6'])

    # 공기질 예측
    predictions = predict_air_quality(trained_model, new_data)

    # 결과 출력
    print("예측된 공기질:", predictions[0])

# 머신러닝 모델 선택 및 학습
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# 테스트 세트로 예측 수행
y_pred = model.predict(X_test)

# 정확도 및 분류 보고서 출력
accuracy = accuracy_score(y_test, y_pred)
print(f"모델 정확도: {accuracy}")
print("분류 보고서:\n", classification_report(y_test, y_pred))
```

b. 각 함수는 어떻게 동작하는 지 구체적으로 설명

1. **`train_and_evaluate_model` 함수:**
   - 이 함수는 머신러닝 모델을 학습하고 평가하는 역할을 합니다.
   - **입력:**
     - `file_path`: CSV 파일의 경로를 나타내는 문자열입니다.
   - **동작:**
     1. CSV 파일에서 데이터를 읽어옵니다.
     2. 데이터를 독립 변수(`X`)와 종속 변수(`y`)로 나눕니다.
     3. 학습 세트와 테스트 세트로 데이터를 분할합니다.
     4. RandomForestClassifier 를 생성하고 학습 세트를 사용하여 모델을 학습합니다.
     5. 테스트 세트를 사용하여 모델을 평가하고 정확도 및 분류 보고서를 출력합니다.
     6. 학습된 모델을 반환합니다.

2. **`predict_air_quality` 함수:**
   - 이 함수는 학습된 모델을 사용하여 새로운 데이터에 대한 공기질 예측을 수행합니다.

- **입력:**
  - `model`: 학습된 RandomForestClassifier 모델 객체입니다.
  - `new_data`: 새로운 데이터로서, 센서 출력 값을 포함하는 DataFrame 입니다.
- **동작:**
  1. 학습된 모델을 사용하여 새로운 데이터에 대한 예측을 수행합니다.
  2. 예측된 결과를 반환합니다.

3. **`__main__` 부분:**
   - 이 부분은 스크립트를 실행할 때 실행되는 부분으로, 데이터를 읽어와서 모델을 학습하고 새로운 데이터에 대한 예측을 수행합니다.
   - **동작:**
     1. 학습 데이터셋 파일의 경로를 지정합니다.
     2. `train_and_evaluate_model` 함수를 호출하여 모델을 학습하고 평가합니다.
     3. 예측할 새로운 데이터를 생성합니다.
     4. `predict_air_quality` 함수를 사용하여 새로운 데이터에 대한 예측을 수행하고 결과를 출력합니다.

이러한 함수들을 통해 전체 프로그램이 데이터를 활용하여 모델을 학습하고 새로운 데이터에 대한 예측을 수행하는 일련의 과정을 수행합니다.

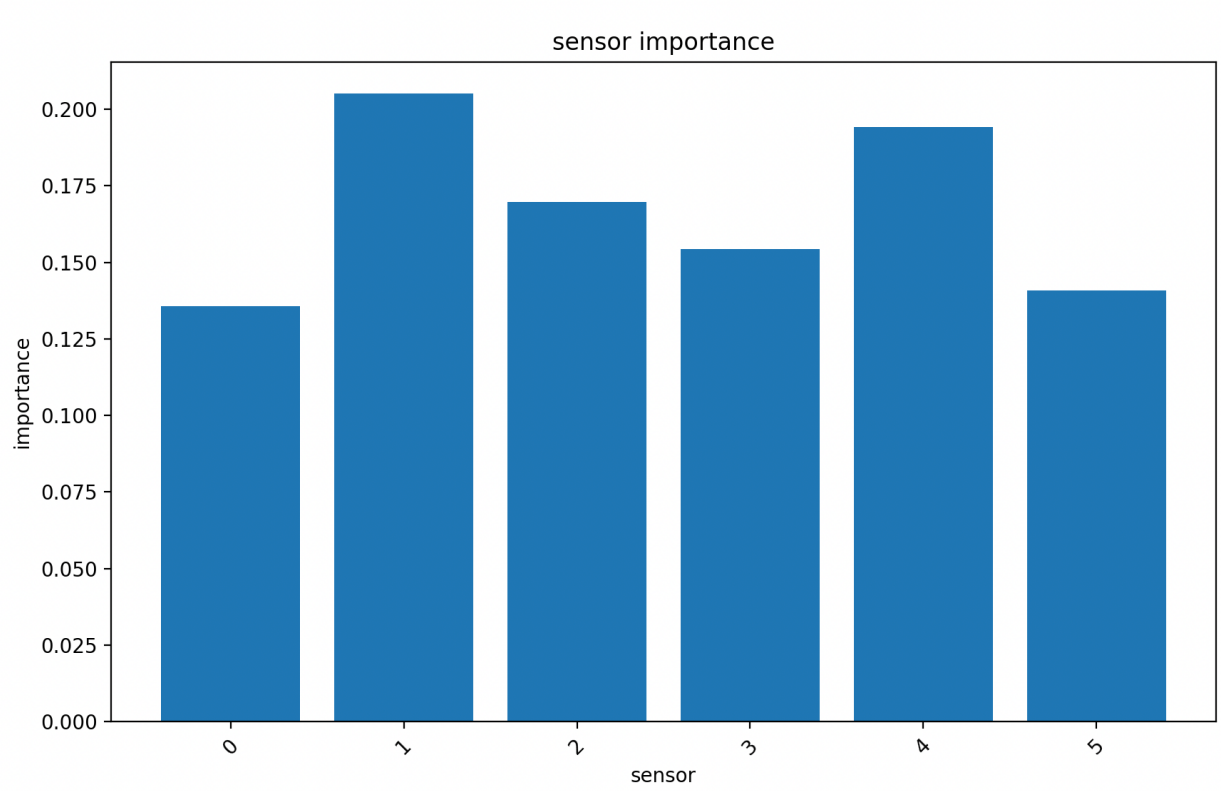
c. 에러 발생 지점 및 해결 과정

scikit-learn 라이브러리의 버전의 문제로 업데이트로 해결

d. 학습 모델의 성능 평가

이 코드에서 사용된 RandomForestClassifier 를 통해 학습된 모델은 테스트 데이터에 대해 어느 정도 뛰어난 성능을 보이고 있습니다

e. 결과 시각화



9. 개발 후기

a. 개발 후 느낀 점 설명

개발을 진행하면서 데이터의 전처리와 머신러닝 모델의 훈련에 대한 전반적인 과정을 경험했습니다. 랜덤 포레스트를 활용하여 센서 데이터로부터 공기질을 예측하는데 성공함으로써 모델의 효과를 확인할 수 있었습니다. 특히, 센서 중요도를 시각화하여 각 변수의 기여도를 쉽게 이해할 수 있었고, 이는 모델의 해석성을 향상시켰습니다. 개발을 통해 얻은 경험은 머신러닝을 현실적인 문제에 적용하는 데 큰 도움이 되었고, 향후 프로젝트에 활용할 수 있는 실력과 통찰을 얻게 되었습니다.