서울의 대기질 예측모델

학번: 2018003

이름: 김다은

Github address: https://github.com/2018003kimdaeun/Homework.git

1. 서울의 대기질 예측모델개발의 목적

- 독립 변수(x): 이산화질소농도(ppm), 오존농도(ppm), 일산화탄소농도(ppm), 아황산가스(ppm), 미세먼지(㎍/㎥)와
- 종속 변수(y): 초미세먼지(㎍/㎥)를 사용하여 대기질 데이터를 기반으로 서울의 미래 대기질 상태를 환경 문제 예방 및 관리에 도움을 주고, 시민들에게 대기질 정보를 제공하여 건강한 환경을 조성하는 데 기여하는 데 목적이 있다.

2. 서울의 대기질 예측모델의 네이밍의 의미

- 대기질 데이터를 기반으로 서울의 대기질 상태를 예측해 사용자에게 정보를 제공한다는 것을 강조하는 의미가 있다.

3. 개발 계획

a. 데이터에 대한 요약 정리 및 시각화
   - 대기질 데이터의 기본 특성 및 통계량을 요약하여 이해한다.
   - 데이터 시각화를 통해 대기질 데이터의 분포, 추이, 상관 관계 등을 시각적으로 확인한다.

b. 데이터 전처리 계획
   - 결측치 처리를 통한 누락된 값 대체 또는 제거
   - 이상치 확인 및 처리

c. 어떠한 머신러닝 모델을 사용할 것인지 (해당 머신러닝 모델의 이론 추가)
   - DecisionTreeClassifier 를 사용한다.
   - DecisionTreeClassifier 란 분류 및 회귀작업에 사용되는 알고리즘이다.

d. 머신러닝 모델 예측 결과가 어떠할 지

- 의사결정 트리 모델은 공기 오염 특성과 초미세먼지 수준 간의 관계를 잘 파악하고 예측할 것으로 예상된다.

e. 사용할 성능 지표

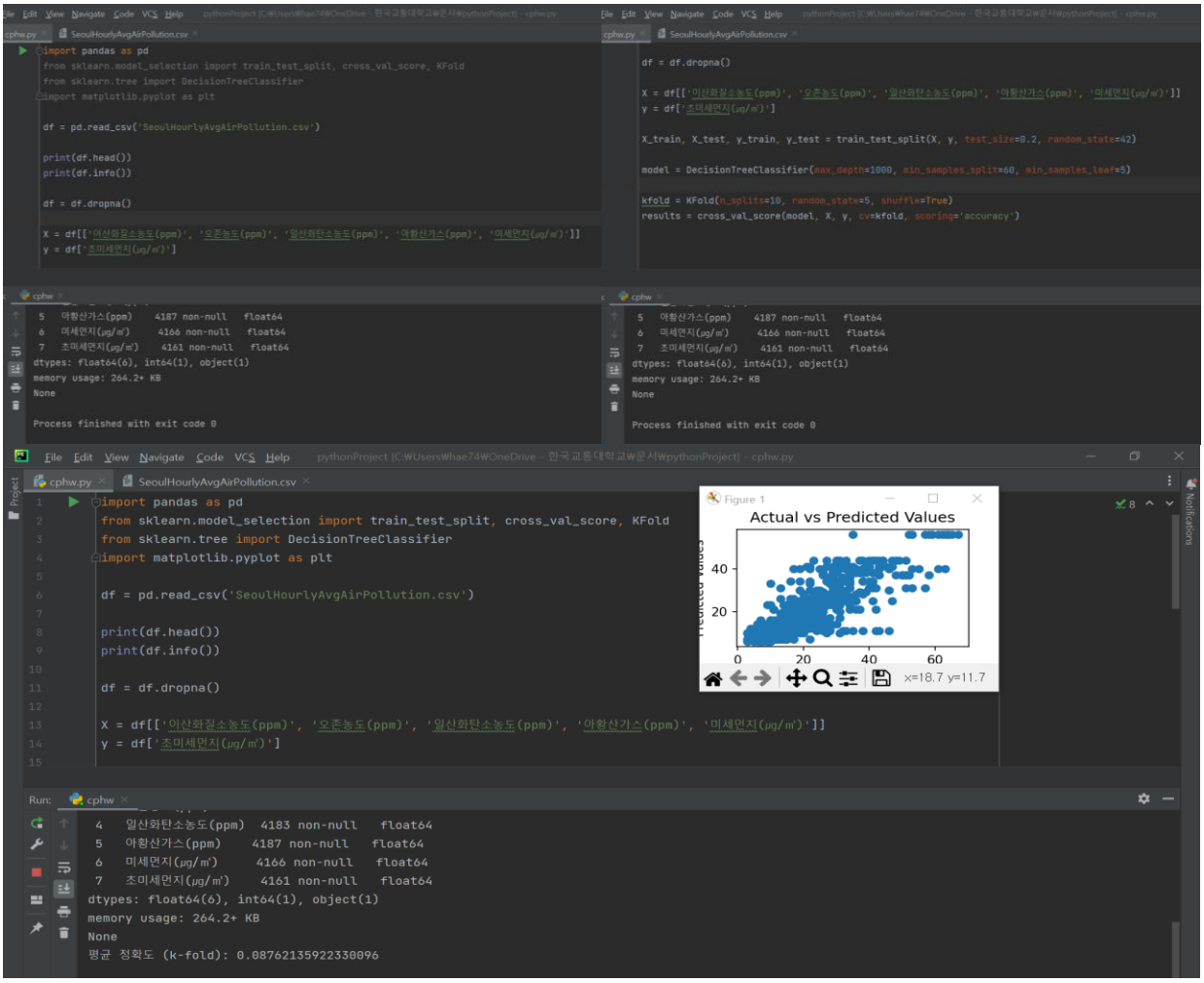
- k-fold 로 교차검증을 통해 모델의 성능을 평가 하려한다.

f. 성능 검증 방법 계획 등

- k-fold 교차검증 결과를 통해 모델의 안정성과 성능을 종합적으로 평가할 예정이다.

4. 개발 과정

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



b. 각 함수는 어떻게 동작하는 지 구체적으로 설명

- pd.read_csv('SeoulHourlyAvgAirPollution.csv'): CSV 파일을 불러와 DataFrame 으로 저장한다.
- df.dropna(): 결측치가 있는 행을 제거한다.
- train_test_split(X, y, test_size=0.2, random_state=42): 데이터를 학습용과 테스트용으로 나눈다.
- DecisionTreeClassifier(max_depth=1000, min_samples_split=60, min_samples_leaf=5): 의사결정 트리 분류기를 초기화한다.
- model.predict(X_test): 학습된 모델을 사용하여 테스트 데이터에 대한 예측을 수행한다.

c. 에러 발생 지점 및 해결 과정

- ```
  import pandas as pd
  from sklearn.model_selection import train_test_split,
  cross_val_score, KFold
  import matplotlib.pyplot as plt
  ```

  실행결과 NameError: name 'DecisionTreeClassifier' is not defined

  → from sklearn.tree import DecisionTreeClassifier 를 추가하였음

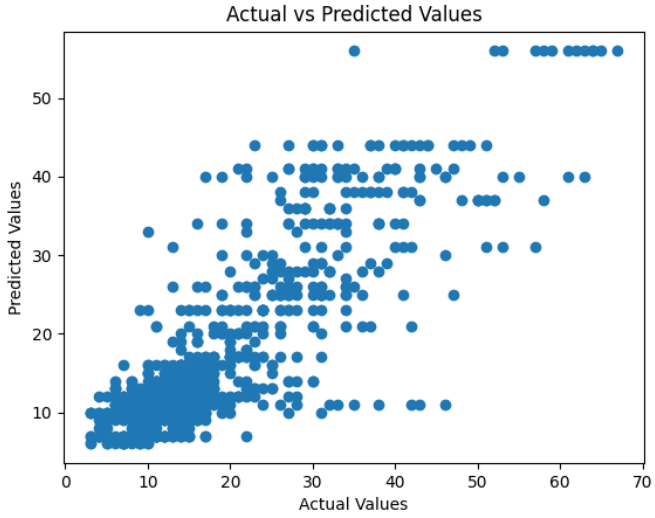
  : DecisionTreeClassifier 를 정의하지 않았기에 발생한 오류.

- DecisionTreeClassfier → DecisionTreeClassifier 오타
- n_splits-10 → n_splits=10 오타

d. 학습 모델의 성능 평가

- cross_val_score(model, X, y, cv=kfold, scoring='accuracy'): k-fold 교차검증을 통해 모델 성능을 평가한다.
- model.fit(X_train, y_train): 모델을 학습시킨다.

e. 결과 시각화
  ```



5. 개발 후기

a. 개발 후 느낀 점 설명

데이터의 이해와 전처리의 중요성을 깨달았다. 또한 상황에 따라 적절한 모델을 선택하는 것도 중요하다는 것을 깨달았다. 이러한 경험을 토대로 향후 지속적인 학습과 개선을 통해 더 나은 결과물을 얻기 위해 노력할 것이다.