Water Quality

수질 데이터 셋을 활용한 사람이 마실 수 있는 물 여부 예측

학번 : 2018062

이름 : 김태현

Github address :

1. 안전 관련 머신러닝 모델 개발의 목적

물의 안전성에 관련된 데이터셋을 사용하여 RandomForestClassifier를 학습하고, 해당 모델을 사용하여 물의 품질이 가공 가능한지 여부를 예측하는 것입니다. "Potability" 열은 물이 가공 가능한지 여부를 나타내며, 모델은 다양한 물의 특성을 기반으로 이진 분류를 수행합니다. 이는 물의 안전성을 판단하고 안전하지 않은 물을 식별하는 데 도움이 될 수 있습니다.

2. 안전 관련 머신러닝 모델의 네이밍의 의미

a. WaterSafetyPredictor:

"Water Safety"는 모델이 다루는 주제를 나타내며, "Predictor"는 해당 모델이 예측을 수행하는 역할을 나타냅니다.

b. PotabilityPredictor:

"Potability"는 모델이 예측하는 대상인 마실 수 있는 물의 품질을 나타냅니다.

"Predictor"는 마찬가지로 모델의 예측 기능을 나타냅니다.

3. 개발 계획 및 과정

a. 데이터 불러오기 : 'water_potability.csv' 파일을 불러와서 데이터프레임으로 저장

```
filename = './data/water_potability.csv'
data = pd.read_csv(filename)
```

b. 데이터 전처리 :

Unnamed: 0' 열을 삭제하고, 결측치가 있는 행을 삭제하는 간단한 전처리를 수행

```
data = data.drop(['Unnamed: 0'], axis=1)
data = data.dropna()
```

c. 데이터 시각화 :

히스토그램을 사용하여 데이터의 특성 분포를 시각화합니다. 각 특성에 대한 히스토그램이 그려지고, 이를 통해 데이터의 분포를 확인

```
data.hist(figsize=(12, 10))
plt.show()
```

d. 데이터 분할 : 데이터를 특성(X)과 레이블(y)로 분할하고, 학습 데이터와 테스트 데이터로 나눔

```
X = data.drop("Potability", axis=1)
y = data["Potability"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

e. 머신러닝 모델 학습 : RandomForestClassifier를 사용하여 머신러닝 모델을 학습

```
model = RandomForestClassifier()
model.fit(X_train, y_train)
```

f. 예측 : 학습된 모델을 사용하여 테스트 데이터에 대한 예측을 수행

```
predictions = model.predict(X_test)
```

g. 성능평가 : 정확도와 분류 보고서를 계산하여 모델의 성능을 평가

```
accuracy = accuracy_score(y_test, predictions)
report = classification_report(y_test, predictions)
```

h. 결과 출력 : 정확도와 분류 보고서를 출력하여 모델의 성능을 확인

```
print(f"Accuracy: {accuracy}")
print("Classification Report:\n", report)
```

4. 개발한 학습 모델의 효과

높은 정확도와 균형 잡힌 정밀도 및 재현율, 그리고 특성 분포를 고려한 히스토그램 분석을 통해 모델이 물의 안전성을 예측하는 데 효과적으로 기여하는 것을 확인할 수 있을것입니다.