과목: 기계학습
과제명: 분류 예측 모델 구현 및 분석
학번 / 이름: 202021065 / 최세진
제출일: 2025.04.26

1. 프로젝트 개요

-문제 정의 :
이 프로젝트에서는 버섯 데이터셋(Mushroom Dataset)을 활용하여, 주어진 버섯 샘플이 식용 가능한지 또는 독성이 있는지를 분류하는 이진 분류 모델을 개발하였다. 주어진 특성(feature)들을 기반으로 독성 여부를 정확히 예측하는 것이 목표다.

-데이터셋 설명
1. 출처: Kaggle Mushroom Dataset
2. 데이터셋 크기: 8124 개 샘플 × 22 개 변수
3. 종속변수(Target): Mushroom_quality (e: edible, p: poisonous)
4. 주요 독립변수(Features): cap_shape, odor, gill_size, stalk_root 등

2. 데이터 전처리 및 탐색적 분석 (EDA)

-결측치 처리 :
stalk_root 변수에 '?' 값으로 존재하는 결측치를 확인하고, 해당 레코드를 제거하였다.

-범주형 변수 처리 :
모든 특성이 범주형(categorical) 데이터로 구성되어 있어, pd.get_dummies 를 이용해 One-Hot Encoding 을 수행하였다.

-스케일링 :
데이터가 모두 범주형 특성으로 구성되어 있기 때문에 별도의 스케일링은 적용하지 않았다.

-EDA 시각화 :
1. 클래스 분포(식용 vs 독성)를 시각화하여 데이터 불균형 여부를 확인하였다.
2. 주요 변수(예: odor)와 타겟 간 관계를 시각화하여 변수의 중요성을 탐색하였다.

3. 모델 구축 및 학습

-사용한 알고리즘 :
RandomForestClassifier (기본 설정)
RandomForestClassifier (GridSearchCV 를 통한 하이퍼파라미터 튜닝 버전)

-데이터 분할 방식 :
train_test_split 을 사용하여 데이터를 학습용(80%)과 테스트용(20%)으로 분할하였다.

-파이프라인 사용 여부 :
본 프로젝트에서는 별도의 파이프라인은 사용하지 않았으며, 전처리와 모델링을 단계별로 수행하였다.

-학습 코드 요약
1. 인코딩된 특성을 기반으로 RandomForest 모델을 학습
2. 기본 모델과 튜닝된 모델 모두 테스트 데이터에 대해 성능 평가 진행

4. 성능 평가

-사용한 지표
1.Accuracy
2. Precision
3. Recall
4. F1-score
5. ROC-AUC

-예측 결과 시각화
1. Confusion Matrix 를 통해 분류 정확도를 시각적으로 분석하였다.
2. ROC Curve 를 그려 모델의 분류 임계값에 따른 민감도와 특이도 변화를 시각화하였다.

-해석
1. 기본 RandomForest 모델은 100%의 Accuracy 와 F1-score 를 달성하였다.
2. 튜닝된 모델 역시 완벽한 성능을 유지했으며, AUC 점수 또한 1.0 을 기록했다.
3. 클래스 간 성능 차이는 존재하지 않았고, 모든 클래스에서 균등한 성능을 보였다.

5. 하이퍼파라미터 튜닝

-튜닝 방법
1. GridSearchCV 를 활용하여 RandomForestClassifier 의 최적 하이퍼파라미터를 탐색하였다.
2. 튜닝한 하이퍼파라미터
    n_estimators: 100 또는 200
    max_depth: 5, 10, None
    min_samples_split: 2, 5, 10

-튜닝 결과 분석
1. 최적 파라미터 조합은 n_estimators=100, max_depth=10, min_samples_split=2 로 확인되었으며,
2. 기본 모델 대비 변화는 없었지만, 하이퍼파라미터 최적화 과정의 유효성을 검증할 수 있었다.

6. 결론 및 고찰

-최종 모델 성능 종합 평가 :
최종 RandomForest 모델은 매우 높은 정확도와 안정적인 분류 성능을 달성하였다.

-데이터 또는 모델의 한계 :
데이터가 매우 깨끗하여 추가적인 전처리 작업이 필요 없었지만, 실생활 데이터에서는 오염된 샘플이나 레이블 오류가 발생할 수 있다는 점을 고려해야 한다.

-실생활 응용 가능성 :
버섯 데이터와 유사한 생물학적 샘플 분류 문제에 본 모델을 적용할 수 있으며, 예를 들어 자동화된 식용 가능성 판별 시스템 구축에 활용될 수 있다.

-다음 단계에서 고려할 점 :
1. 다양한 모델(예: XGBoost, LightGBM) 추가 실험
2. Feature Importance 를 분석하여 모델 해석성 강화
3. 클래스 불균형 데이터셋에 대비한 추가 실험 진행

7. 참고자료

https://www.kaggle.com/datasets/rinichristy/uci-mushroom-dataset
Scikit-learn 공식 문서 (version: 1.3.0)

8. 부록

중요 코드 스니펫

```
# 데이터 분할 및 모델 학습
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
rf = RandomForestClassifier()
rf.fit(X_train, y_train)


# 하이퍼파라미터 튜닝
param_grid = {'n_estimators': [100, 200], 'max_depth': [5, 10, None], 'min_samples_split'
: [2, 5, 10]}
grid_search = GridSearchCV(rf, param_grid, cv=5)
grid_search.fit(X_train, y_train)
```