데이터마이닝 과제4 – 2022920024 박동찬

1. 실험

Titanic dataset을 Decision Tree Classifier를 통해 분류했다. Categorical features(sex, class, deck, embark_town, alone)에 대해 Label Encoding과 One-Hot Encoding을 각각 적용해보고 GridSearchCV를 통해 다양한 hyper-parameter를 탐색했다. 아래와 같은 hyper-parameter 조합을 탐색했다.

```
param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20],
    'min_samples_split': [2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20],
    'min_samples_leaf': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
}
```

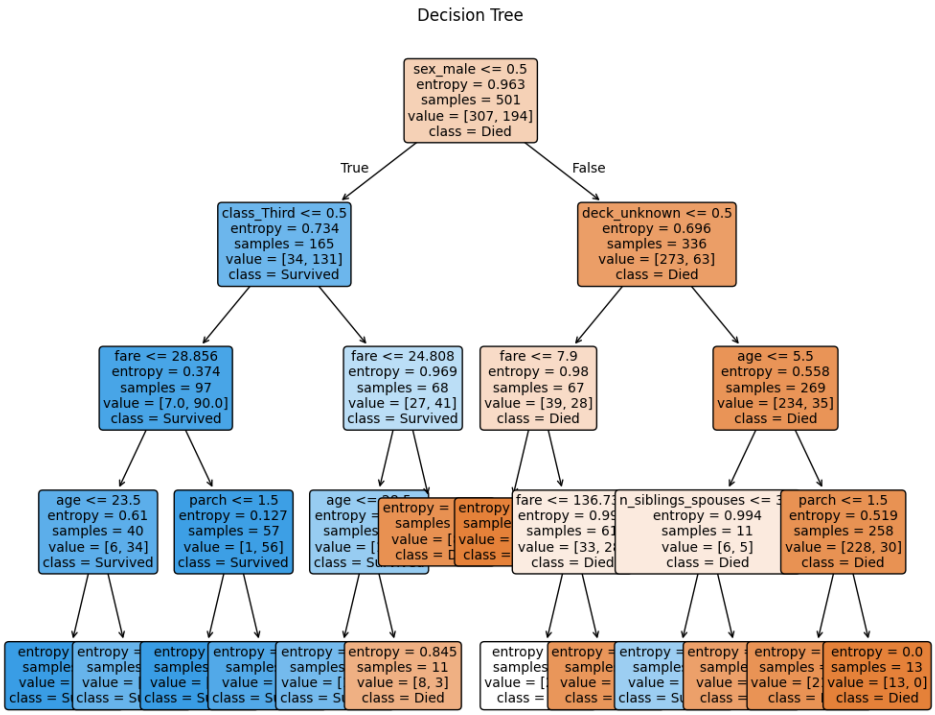
또한 5-fold CV를 통해 가장 높은 정확도를 보인 조합을 선택하였다.

2. 실험 결과

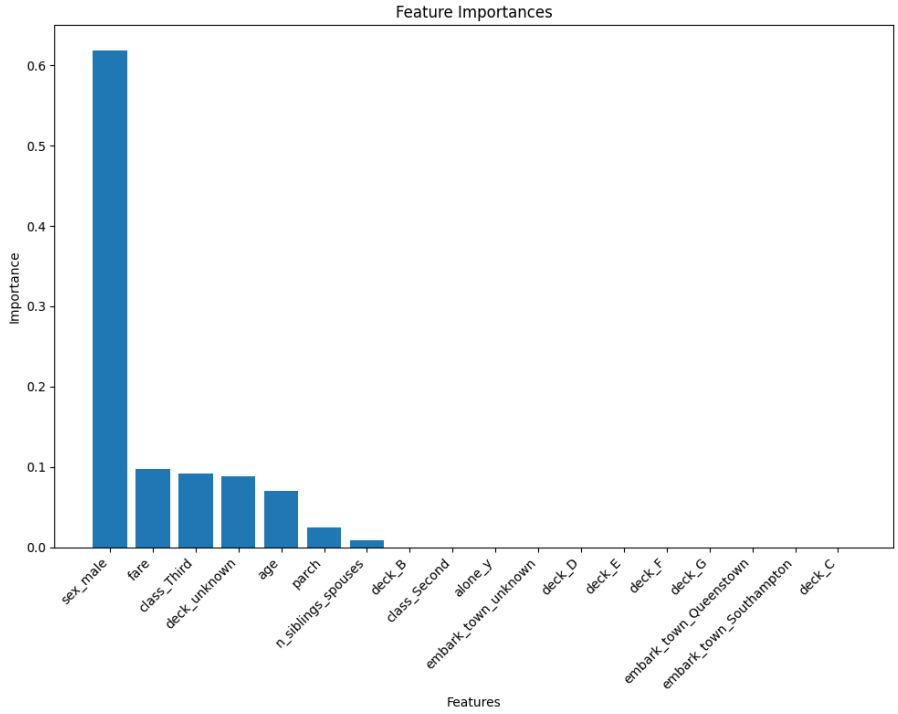
A. Best Accuracy

```
----- Encoding Method: label -----
Fitting 5 folds for each of 2860 candidates, totalling 14300 fits
Best params: {'criterion': 'entropy', 'max_depth': 4, 'min_samples_leaf': 4, 'min_samples_split': 2}
CV Best Accuracy: 0.8264
Test Accuracy: 0.8095
----- Encoding Method: onehot -----
Fitting 5 folds for each of 2860 candidates, totalling 14300 fits
Best params: {'criterion': 'entropy', 'max_depth': 4, 'min_samples_leaf': 5, 'min_samples_split': 2}
CV Best Accuracy: 0.8244
Test Accuracy: 0.8413
========================================
Best performing encoding method: onehot
Test Accuracy: 0.8413
```

B. Plot tree image



C. Feature importances image



3. Discussion

A. Encoding Methods

Label Encoding 방식에서는 {'criterion': 'entropy', 'max_depth': 4, 'min_samples_leaf': 4, 'min_samples_split': 2} 조합에서 CV Best Accuracy: 0.8264, Test Accuracy: 0.8095가 도출되었고 One-Hot Encoding 방식에서는 {'criterion': 'entropy', 'max_depth': 4, 'min_samples_leaf': 5, 'min_samples_split': 2} 조합에서 CV Best Accuracy: 0.8244, Test Accuracy: 0.8413가 도출되었다. Train set에 대한 Accuracy는 Label Encoding 방식이 조금 높았지만 test set에서 One-Hot Encoding 방식이 더 높은 Accuracy를 보였다. Label Encoding 방식은 Categorical features에 순서가 없음에도 불구하고 정수 값이 부여되면서 암묵적인 순서 정보가 들어갈 수 있고 이로 인한 오류가 나타날 수 있다. One-Hot Encoding은 순서 관계 없이 모델에게 정확한 정보를 학습시킬 수 있어 더 좋은 일반화 성능을 보인 것이라고 할 수 있다.

B. Criterion Methods

Criterion은 모두 entropy가 선택되었는데 Entropy가 Gini impurity에 비해 극단적인 확률 분포에 더 민감하게 반응하고 실제로 노드가 얼마나 순수해지는지 더 세밀하게 평가할 수 있다. Titanic dataset에는 sex, fare, class와 같은 생존 여부에 큰 영향을 미치는 features가 존재하므로 분할할 때 얻을 수 있는 정보 이득을 Entropy가 더 명확하게 측정할 수 있었을 것이다.

C. Max depth

트리의 깊이가 너무 얕으면 underfitting이 발생하고 너무 깊어지면 overfitting이 발생할 수 있으므로 적절한 깊이 제한이 모델의 일반화 성능을 향상시킨다고 볼 수 있다.

D. Min samples split

탐색한 값들 중 가장 작은 값인 2일 때 가장 좋은 성능을 보였다. 이는 내부 노드를 분할하기 위한 최소 샘플 수인데 작은 값을 사용할수록 더 세밀하게 분할할 수 있어 dataset의 미세한 차이도 학습할 수 있다. 너무 작은 값을 사용하면 overfitting 될 수 있지만 Titanic dataset에서는 최적의 성능을 보였다.

E. Min samples leaf

탐색한 값들 중에서 4(label encoding), 5(one-hot encoding)일 때 가장 좋은 성능을 보였다. 분할 후 leaf node에 존재해야 하는 최소 샘플 수로써 너무 작은 값을 사용하면 leaf node에 작은 샘플만 포함해서 noise에 민감해지고 너무 큰 값을 사용하면 세밀한 부분을 학습하지 못하여 모델의 표현력이 감소한다. 따라서 비교적 중간 값인 4, 5일 때 좋은 성능을 보였으며 One-Hot Encoding 하면 feature의 차원이 category 수만큼 늘어나므로 해당하는 데이터가 sparse해져 leaf node에 더 많은 샘플이 존재해야 안정적인 예측을 할 수 있으므로 약간 더 높게 선택된 것으로 볼 수 있다.

F. Decision Tree 구조 분석

Root node에서 sex_male <= 0.5로 분할이 이루어진다. 이는 여성(sex_male=0)과 남성(sex_male=1)의 구분을 의미한다. 성별에 따라 생존 여부가 크게 바뀐다는 것을 알 수 있다. 하위 분할된 왼쪽 노드(여성)는 class_Third <= 0.5로 분기되는데 3등실보다 1, 2등실일 때 생존하기 유리하다는 것을 알 수 있다. 실제로 타이타닉에서 객실 등급에 따라 탈출 기회가 달라졌다. 이보다 더 하위 노드에서는 fare와 age가 많이 등장하며 요금이 높은 사람일수록 높은 등급의 객실을 사용했을 가능성이 커 생존에 유리하다는 것을 볼 수 있다. 하위 분할된 오른쪽 노드(남성)는 deck_unknown <= 0.5로 이루어져 있는데 이는 남성 집단에서 갑판 정보의 유무와 생존 여부가 어느정도 관련이 있음을 알 수 있다. 전체적으로 남성은 생존율이 낮지만 그 중에서 상위 객실에 배정된 사람들은 갑판 정보가 기록되어 간접적으로 생존율이 올라갔다고 추측해볼 수 있다. 이보다 더 하위 노드에서는 fare, age, n_siblings_spouses 등 다양한 feature가 등장한다.

G. Feature Importances 분석

그래프를 보았을 때 여성 여부, 요금, 3등실 여부, 갑판 정보 여부, 나이가

생존 여부에 결정적으로 작용한다는 것을 알 수 있다. 나머지 features는 상대적으로 중요도가 낮았다. 따라서 여성이고 상위 객실이며(높은 요금) 나이가 어리고 갑판 정보가 존재할수록 생존율이 높아진다는 사실을 알 수 있다.