Machine Learning (2021 Fall semester)

Programming Assignment: Classification of Titanic Data Set

1. **Benchmark Dataset**: 타이타닉 호에서 탑승했던 사람들의 정보를 바탕으로 생존 여부를 예측하는 문제입니다. 과제에서 제시한 머신러닝 모델들을 이용하여 각 모델의 성능을 평가해야 합니다. 데이터는 다음 주소에서 얻을 수 있습니다: https://www.kaggle.com/c/titanic

   이번 과제에서는 모델 학습 및 테스트 모두 train.csv파일을 이용합니다.

   과제를 수행하실 때 다음의 feature들은 모델 학습에 이용하지 않도록 주의해주세요.

   제외할 feature: PassengerId, Name, Ticket, Cabin

2. **Preprocessing**

   1. train data에 결측치(NULL)가 있는 feature들이 있습니다. 이 값들을 어떻게 처리할 것인지 아이디어를 제시하고 실제로 구현하세요.
   2. One-Hot encoding을 수행할 필요가 있는 feature들이 무엇이며 그 이유는 무엇인지 서술하세요. 또한 실제로 어떻게 구현했는지 보고서에 나타내세요.
   3. train.csv의 sample을 7대 3으로 학습 데이터와 테스트 데이터로 사용하세요.
   4. 그 외 진행한 전처리 과정이 있다면 서술하세요.

3. **Machine Learning Models**: scikit-learn을 이용해 후술할 세 가지 machine learning model를 구현하고 성능을 평가하세요.

3-1 K-Nearest Neighbors(KNN): K(이웃)의 개수를 [1~5]까지 변화시키면서 test data에서 결과가 어떻게 변하는지 분석하세요.

3-2 Logistic Regression: Iteration 횟수를 [0~100] 범위에서 20씩 변화시키면서 test data에서 결과가 어떻게 변하는지 분석하세요. Iteration 횟수를 100으로 고정한 후 regularization term(scikit-learn에서는 C)를 [0~5]의 범위에서 1씩 변화시키면서 test data에서 결과가 어떻게 변하는지 분석하세요.

3-3 Decision Tree: Information Gain을 통해 test data에서 결과가 어떻게 나오는지 분석하세요. 또한 적절한 tool을 이용하여 각 depth에서의 조건과 gain값을 알 수 있도록 tree를 시각화하세요.

3-3-1 Bagging with Decision Tree: Decision Tree 기반으로 Bagging 기법을 이용하여 bag의 수에 따라 test data에서 결과가 어떻게 변하는지 decision tree와 비교하여 분석하세요. Bag의 수는 [1~5]에서 1씩 변화시키세요.

4. **Evaluation Methods**: 각각 모델에 따른 성능을 Accuracy, F1-Score을 통해 나타내세요.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \quad Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}, \quad F_1 = \frac{2 \cdot Precision \cdot Recall}{Precison + Recall},$$

$$TP = true\ positive, \quad TN = true\ negative, \quad FP = false\ positive, \quad FN = false\ negative.$$

5. **Submission Form:** 제출파일은 3개 입니다. csv파일과 보고서, 그리고 python 파일을 zip 파일로 묶어 제출하시면 됩니다. 파일 이름은 학번_이름.zip형식을 반드시 지켜주세요(예시, 2020714950_홍길동.zip) csv파일과 python파일이 같은 디렉토리에 있는 상태에서 python파일을 실행했을 때 각 machine learning model에서 결과가 어떻게 나오는지 일목요연하게 표현되어야 합니다. 이는 보고서에서의 성능과 실제 실행했을 때의 성능이 비슷한 지 확인하기 위함입니다. ipynb파일로 작성하셨다면 python파일 대신 제출하셔도 무방합니다.