# 20일차 / 수업 외 (2)

# 1. 데이콘 - 와인 품질 분류

데이콘 기본 교육 예제 2 (와인 품질 분류)

#### [화학] 와인 품질 분류

상금 : 교육 1,060명 D-18 index 구분자 quality 품질 fixed acidity 산도 volatile acidity 휘발성산 citric acid 시트르산 residual sugar 잔당 : 발효 후 와인 속에 남아있는 당분 chlorides 염화물 free sulfur dioxide 독립 이산화

https://dacon.io/competitions/open/235610/data



```
# 전체 코드
import pandas as pd
import lightgbm as lgbm
train = pd.read_csv('/content/drive/MyDrive/머신러닝/와인품질분류/data/train.csv')
test = pd.read_csv('/content/drive/MyDrive/머신러닝/와인품질분류/data/test.csv')
# 데이터 확인 시, 거의 전처리가 된 상태
# 다만, type의 red or white 때문에 에러 발생하여 replace를 통해서 치환함
train.replace('white', 0, inplace = True)
train.replace('red', 1, inplace = True)
test.replace('white', 0, inplace = True)
test.replace('red', 1, inplace = True)
# 모델링
model_LGBM = lgbm.LGBMClassifier()
model_LGBM.fit(train_x,train_y)
# 예측
LGBM_pred = model_LGBM.predict(test)
# 데이터 저장 단계
submission = pd.read_csv('/content/drive/MyDrive/머신러닝/와인품질분류/data/sample_submission.csv')
submission
```

```
submission['quality'] = LGBM_pred

# 저장
submission.to_csv('/content/drive/MyDrive/머신러닝/와인품질분류/data/submission.csv', index = False)
```

#### 앞전에 사용했던 RandomForestRegression 과의 비교

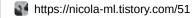
```
# RFR
model=RandomForestRegressor(n_estimators=100) # 그냥 100이라고 해도 실행됨
model.fit(train_x,train_y)
index quality
0 0 5.80
1 1 5.22
2 2 5.64
3 3 5.37
4 4 6.34
995 995 5.48
996 996 5.61
997 997 5.27
998 998 5.67
999 999 6.01
# LGBM
model_LGBM = lgbm.LGBMClassifier()
model_LGBM.fit(train_x,train_y)
index quality
0 0 6
1 1 5
2 2 5
3 3 5
4 4 6
995 995 5
996 996 6
997 997 5
998 998 6
999 999 6
# 소수점 문제가 해결되었다.
```

# 2. LGBM (Light GMB)

- GMB(Gradient Boosting Machine) 이란?
  - 。 틀린부분에 가중치를 더하면서 진행하는 알고리즘
- GMB와 LGBM의 차이
  - 기존 GBM과 다른점은 GBM은 균형 트리분할(Level Wise) 방식
  - LGBM은 **리프중심 트리분할(Leaf Wise)** 방식
  - 균형 트리분할: 최대한 균형 잡힌 트리를 유지하며 분할하여 트리의 깊이를 최소화하여오
     버피팅에 강한구조이지만 균형을 맞추기 위한 시간이 필요
  - 리프중심 트리분할 : 최대 손실 값을 가지는 리프노드를 지속적으로 분할하면서트리가 깊 어지고 비대칭적으로 생성하며 이로써 예측 오류 손실을 최소화.
- 과적합 우려가 있어. 데이터가 큰 경우에 적합함.

#### Light GBM(LGBM)의 개요와 파라미터 정의에 대해

Light GBM은 Kaggle 데이터 분석 경진대회에서 우승한 많은 Tree기반 머신 러닝 알고리즘에서 XGBoost와 함께 사용되어진것이 알려지며 더욱 유명해 지게 되었습니다. GMB(Gradient Boosting Machine) 이란? 틀린부분에 가





### 3. LGBM, GBM, XGBoost

# [머신러닝] LGBM, XGBoost, GBM LGBM(Light GBM) 데이터셋 작으면 과적합하기 쉽다. 문서상 10,000개 데이터 이상인 데이터셋에 적합 하지만 많은 데이터셋에서는 XGBoost보다 빠른 학습속도, 적은 메모리사용량 카테고리형 데이터에 대해서 원핫.. ✔ https://jaeyung1001.tistory.com/200