< 과제 #4 : Classification >

☞ **HW4.csv 파일은 은행 마케팅 데이터로서, 은행 고객의 정보(age, job, marital, education, default, balance, housing, load)와 캠페인(정기예금 유치 마케팅)과 관련된 정보(contact, day, month, duration, campaign, pdays, previous, poutcome), 그리고 마지막으로 정기예금 유치결과(y)를 저장하고 있다. y를 레이블로 하는 분류 문제에 대하여 아래의 순서로 분석을 실행하시오.**

1) 파일을 읽어 데이터프레임을 생성한 후, 분석에 적절하지 않은 day와 month 칼럼은 삭제하시오. 수치형 특성의 이름과 범주형 특성의 이름의 리스트를 출력하시오.
2) 레이블의 범주별 비율을 구하시오.
3) 수치형 특성의 요약통계(평균, 표준편차 등)와 범주형 특성의 막대 그래프를 구하시오.
4) 특성행렬을 만든 후 범주형 특성은 원-핫-인코딩한 데이터프레임(4521rows×38columns)을 구하시오. 그리고 레이블에 대하여 'yes'는 0으로 'no'는 1로 인코딩하시오.
5) 데이터를 훈련용과 테스트용으로 8:2로 분할한 후, 표준화하시오.
6) cross_val_score 함수(cv=5)를 사용하여 로지스틱 회귀의 초모수 C의 값을 [0.01, 0.1, 1, 10, 100]로 바꾸어 가며 최적값을 구한 후, 이 값을 사용한 테스트 스코어를 구하시오. 테스트 데이터에 대한 정오분류표를 작성하고 정기예금에 가입하지 않은 고객의 정밀도와 재현율을 구한 후 그 의미를 기술하시오.
7) GridSearchCV(cv=5)를 사용하여 결정트리의 최적 최대 깊이(1~10)를 구하고, 이 값을 사용한 테스트 스코어를 구하시오.
8) StandardScaler, PCA(n_componetns=5), SVC(kernel='rbf')를 순서대로 연결한 파이프라인을 생성한 후, C와 gamma의 값들을 [0.01, 0.1, 1, 10, 100]으로 바꾸면서 GridSearchCV(cv=5)를 사용하여 최적 C와 gamma를 구하시오. 이 모형의 테스트 스코어는 얼마인가?