< 머신러닝 20220413 >

# 외부강사

이번주 금요일

# 과제 3

- (1) 선형회귀

pisa2009train

pisa2009test

두 데이터로 선형회귀

global 학생성적에 관한 데이터

\*대부분의 속성은 binary

schoolSize는 continous 변수

readingSize가 Dependent 변수(DV)

-> 모든 학생들에 대한 DV를 예측하는 것이 목적이다.

A 선형회귀

A1. 기술통계를 수행

ex. 총 학생수, 평균, 변수들 간의 상관관계(상관관계는 필수) , + 상관관계 matrix 혹은 heatmap 등 다양한 시각화 수행

A2. MISSING VALUES를 모두 제거(DROP) 혹은 다른 방법을 이용하여 MISSING VALUES 전처리 -> 교수님은 DROP하였음

(+ GROUP MEAN 등 )

A3. 카테고리컬 변수들을 더미Variables로 변환

ex)

color변수 => greenColor 더미 Binary(0/1) Variables

( afterN= beforeN-1 )...etc

A4. Linear Regression Model 설계(training set 사용) readingScore를 예측하기 위한

A5. 어떤 독립변수들이 DV를 설명하기 위한 KEY 설명변수인지 선정하고 이에 대해 설명을 출력하라

A6. 선형회귀 모델 성능 평가지표 MAE,MSE, RMSE를 보이라

B 로지스틱 회귀

준비 ) Kickstarter data set 다운로드 : 크라우드 펀딩 프로젝트들이 성공적이었는지 아닌지에 관한 예측 :

https://www.kaggle.com/code/jaylew/logistic-regression-2016-kickstarter-projects/data?select=ks-projects-201612.csv

- DV is Binary

-> binary prediction = Logistic Regression

B1. Missing variable 전처리 ( + column names의 끝에 있는 공백도 제거 )

B2. 데이터에 관한 시각화 ( 히스토그램, 바차트 등 )

- seaborn 과 matplotlib을 사용하여

- continuous variables의 분포와 내 관심에 따라 cateogrical variables의 compositions를 시각화

B3. 카테고리컬 데이터의 더미변수를 생성

B4. SUCCESSFUL / FAILED 두가지의 BINARY 결과값만 가지는 DataFrame을 처리->생성

B5. DATA를 TRAIN/TEST 셋으로 분할함으로서 REGRESSION MODEL을 생성

B6. Logistic Regression에 대한 모델 성능 평가지표인 Accuracy / precision Metrics를 생성 및 출력

B7. Accuracy와 precision이 향상된 예측들을 위한 개선을 어떻게 수행해야 할지 내 의견과 함께 제안하기

B8.