

# 5주차 과제

```
# Week_5 Quiz

# 문제 바로 아래에 답안을 작성하여, <<r script 형태>>로 저장 후 제출해주시면 됩니다. (.R 파일업로드 / Rproj파일은 업로드하지 않으셔도 됩니다.)
# 답안인 코드만 작성하시면 됩니다. 코드 실행 결과 포함 X
# [R Script Encoding Error] : File > Reopen with Encoding > 'UTF-8' > OK

# 아래 4줄의 코드를 실행 후 문제를 풀어주세요.
# titanic는 수업 > [4주차 과제 외부데이터]에 있습니다.
# 출처 : https://www.kaggle.com/c/titanic/data?select=train.csv
# 데이터 설명 : 타이타닉호의 생존자 분류

library(caret)
library(dplyr)
train <- read.csv("경로/titanic_train.csv") # 경로는 본인의 컴퓨터 환경에 맞게 작성하세요.
ctrl <- trainControl(method = "repeatedcv", number = 5, repeats = 2)

# (5점) 1. train와 데이터의 타겟변수 Survived를 범주형 데이터로 변경해주세요.
# hint. as.factor함수 이용

# (5점) 2. Pclass는 Ticket 등급을 나타내는 변수입니다. 변수의 고유한 값을 확인할 수 있도록 코드를 작성해주세요.
# hint. unique함수 이용

# (5점) 3. Pclass, Age, Fare 변수들의 히스토그램을 한 눈에 볼 수 있도록 시각화 해주세요 (1x3)
# hint. par, hist함수 이용

# (5점) 4. train 전체 데이터에 대해 변수산점도를 그려주세요.
# hint. plot함수 이용

# (5점) 5. Age, Fare 변수들을 표준화해주세요.

# (5점) 6. train에서 na가 포함된 row는 모두 제거해주세요
# hint. na.omit함수 이용

## 아래부터는 6번에서 생성한 데이터를 활용하여 머신러닝 모델을 만들고 해석하는 부분입니다.
# (5점) 7. k를 1~5까지 정의하여 kNN 모델을 생성해주세요. (preProcess 불필요, metric = "Accuracy")

# (5점) 8. 7번 모델에 대한 실행 결과를 해석해주세요.
# k가 ( )일 때 Accuracy가 ( )%로 가장 높습니다.

# (5점) 9. Boosted Logistic Regression 모델을 생성해주세요. (preProcess 불필요, metric = "Accuracy")

# (5점) 10. 9번 모델에 대한 실행 결과를 해석해주세요.
# nIter가 ( )일 때 Accuracy가 ( )%로 가장 높습니다.

# (5점) 11. Naive Bayes 모델을 생성해주세요. (preProcess 불필요, metric = "Accuracy")
```

```
# (5점) 12. 11번 모델에 대한 실행 결과를 해석해주세요.  
# useKernel이 (    )일 때 Accuracy가 (    )%로 가장 높습니다.
```

```
# (5점) 13. 연속형 숫자 피쳐만 선택하여 주성분 분석을 진행해주세요  
# hint. 연속형 숫자 피쳐 = 데이터 타입이 int와 num인 변수  
# hint. prcomp 함수 활용
```

```
# (5점) 14. 13번 모델의 summary 결과를 해석해주세요.  
# 축이 4개 일 때 전체 변동성의 (    )를 설명합니다.
```