기계학습 과제 1 (2025.4.8.)

1. 캐글(https://www.kaggle.com)에서 설명변수의 개수가 5~8개이며 반응변수가 연속형인 데이터를 하나 선택하고 아래 정보를 제출하시오.

   1) 데이터 이름와 변수명, 데이터 크기
   2) 각 변수별 기초 통계량(평균과 분산, 범주형인 경우 범주별 비율)
   3) 반응변수와 각 변수들 간의 산점도(행렬 형태 추천)

2. 다음을 시행하시오 (모두 다 파이썬이나 R을 이용)

   1) 모든 변수를 이용하여 회귀분석을 시행하고 모형의 유의성을 유의수준 0.05에서 검정하시오.
   2) $\hat{\sigma^2}=\sum_{i=1}^{n}(\hat{y_i}-y_i)^2/(n-p-1)$를 구하시오. 모든 변수들과 절편을 포함한 회귀분석에서의 예측을 이용하며 p는 사용한 모든 변수들의 개수임.
   3) 하나의 변수나 두 개의 변수를 사용하여 각 모형에서 $C_p=RSS/\hat{\sigma^2}+2p$* 를 계산하고 결과를 산점도로 표시하시오(가로축은 사용한 변수의 개수, 세로축은 계산된 C_p*값). 여기에서 RSS는 회귀모형에서의 잔차제곱합이며 p*는 RSS에서 사용한 변수들의 개수이며, 하나의 변수를 사용하면 p개의 모형이 두 개의 변수를 사용하면 $p(p-1)$개의 모형이 고려될 수 있음(산점도에 총 $p+p(p-1)$ 개이 점이 찍혀야 함).
   4) 계산한 3)의 값들 중 가장 작은 값은 어떤 변수(들)을 사용할 때인가?

3. 반응변수가 중앙값을 초과할 때 1, 이하일 때 0으로 하여 아래를 시행하시오.

   1) 로지스틱 회귀분석을 시행해서 혼동행렬을 구하시오.
   2) LDA를 시행해서 혼동행렬을 구하시오.
   3) 2)번에서 민감도와 특이도는 어떻게 되는가. 민감도는 1을 기준으로 함.
   4) 3)번에서 예측의 규칙을 바꾸어 1에 대한 예측확률이 0.20 이상일 때 1로 예측하면 민감도와 특이도는 어떻게 변하는가?

* 위 문제들에 대해서 1. 표지(학번, 이름 포함) 1페이지, 2. 문제 및 답 서술을 포함하는 보고서 형태의 한글이나 pdf 파일을 제출. 문제를 푸는데 사용한 코드는 별도 파일로 첨부할 것.
* 표지가 없으면 부분 감점, 코드만 제출하면 제출로 인정되지 않음.