자작 모의 문제

문제준비 : 신언석 (2021. 8. 7.)

주제 : 정형데이터 마이닝 분류

요구지식

1. 데이터 핸들링 관련 지식

- knn을 이용한 결측치 처리 :

방법 1 : DMwR 패키지 사용하기

방법 2 : Caret 패키지 사용하기 knnImputation

- EDA를 위한 데이터 가시화

- 데이터 Scaling, Centering

- 데이터 분할 : 80:20 (Train : Test)

- 10-Fold Cross Validation 을 적용하여 오버피팅 줄이기

2. 4가지 이상 분류 알고리즘으로 펭귄 종 분류하기

- 분류된 결과에 대한 평가.

- 평가결과 가장 우수한 알고리즘으로 Test 데이터 예측하기

- 예측결과 제시

- 성능을 향상시키기 위한 아이디어 방법 제시.

문제 :

palmer penguin 데이터를 이용하여 다음을 해결하시오.

1. 결측치 존재 여부를 파악하고
   * numeric 변수일 경우 knn 방법을 이용하여 결측치를 처리하고
   * char 또는 factor형 변수인 경우 최빈값으로 대체하시오.
2. EDA를 위한 데이터 가시화
3. 4가지 이상 모델로 펭귄의 종을 분류하라
4. 분류모델의 성능을 비교평가하고
5. 최상의 모델을 선정하여 검증용 데이터로 예측을 수행하라.