2021 머신러닝1 과제4

※ 첨부한 데이터 파일을 확인합니다.

train.csv/ test1.csv/ test2.csv/ gender_submission.csv

train.csv를 열어 컬럼 내용을 확인합니다. (자세한 설명은 교재 참조)

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked

1. 데이터를 로딩하세요. (csv 파일을 데이터 프레임으로 읽기)
2. Train 데이터의 각 필드 데이터분포(평균, 최소값, 최대값, 등)를 확인하세요.
3. 각 컬럼의 데이터 타입을 확인하세요.
4. 각 컬럼의 null값 유무를 확인하세요.
5. 각 컬럼별 결측치의 크기를 확인하세요.
6. 각 컬럼별 null값과 결측치의 상태를 참조해 결측치 혹은 null값의 처리 방법(제거, 할당, 혹은 추정)을 결정하고 이유를 설명하세요.
7. 6번을 실행하며 PassengerID의 생존여부를 추정하는데 사용할 컬럼과 버릴 컬럼을 결정하고 이유를 설명하세요.
8. 결측치/ null값 보정 그리고 사용 컬럼 결정 이후 사용하는 컬럼별 데이터 분포를 히스토그램으로 출력하세요.
9. 생존자와 사망자별 특징을 분석하세요. 분석 과정과 결과를 구체적으로 서술할 것. Data visualization을 적극 활용하세요.
10. 9번의 분석 결과를 토대로 test1.csv를 열어 새로운 컬럼 'Survived'를 생성, 사망자는 0 생존자는 1로 추정하세요. 만약 결측치 등의 처리를 통해 지워진 승객의 경우 사망여부를 빈칸으로 두세요.
11. Train.csv의 survived와 test1.csv의 survived를 비교하여 생존자와 사망자별 추정성공 비율을 확인하세요.
12. 추정 성공에 자신이 붙었다면 9번의 분석결과를 토대로 test2.csv를 열어 새로운 컬럼 'Survived'를 생성, 사망자는 0 생존자는 1로 추정하세요. 결과를 passengerID와 survived 컬럼만 남긴 채 '학번_submission.csv'로 저장하세요. (gender_submission.csv를 참고하세요)