기말과제

[타이타닉 탑승승객 정보를 이용하여 생존자 수 예측하기]

◆ 과제: 아래의 조건을 만족하는 타이타닉 승객의 생존 추정 모델 만들기

1. 데이터의 결측치는 0 으로 채우기
2. Train data 중 datatype 이 string 인 칼럼들 제거
3. Test data 중 성별 칼럼을 제외한 datatype 이 string 인 칼럼들 제거
4. 학습에는 Pclass, Age, SibSp, Parch, Fare 칼럼만 사용
5. 테스트 데이터 중 여성 승객의 데이터만 사용하여 테스트

◆ 데이터 정보

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen ...	male	22.0	1	0	A/5 21171	7.25	null	S
2	1	1	Cumings, Mrs. Joh...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. ...	female	26.0	0	0	STON/O2. 3101282	7.925	null	S
4	1	1	Futrelle, Mrs. Ja...	female	35.0	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. Willia...	male	35.0	0	0	373450	8.05	null	S

PassengerId: 탑승객 ID

Survived: 생존여부(생존=1.죽음=0)

Pclass: 승객 등급

Name: 이름

Sex: 성별

Age: 나이

SibSp: 함께 탑승한 형제, 배우자 수

Parch: 함께 탑승한 부모, 자녀 수

Ticket: 티켓번호

Fare: 요금

Cabin: 선실번호

Embarked: 탑승장소

◆ 데이터 파일

1. 학습에 사용할 데이터: train.csv
2. 학습된 모델을 테스트할 데이터: test.csv

◆ 제출 파일

1. 스크린샷 제출(코드와 출력 결과가 담기게 스크린샷을 출력하여 제출)

1) train데이터와 test데이터의 개수

2) 학습에 사용된 train데이터 테이블(모든 칼럼 5개까지 표시)

3) 테스트에 사용된 test데이터 테이블(모든 칼럼 5개까지 표시)

4) Lab8에서 다룬 Logistic regression 모델을 사용한 예측 모델의 정확도 결과

5) (추가 가산점) Logistic regression 이외의 다른 기계학습 기법(램덤포레스트, XGBoost 혹은 lightGBM 등등)을 사용하여 예측 모델을 만들고 Logistic regression 모델과 정확도를 비교 평가한 결과

2. 주피터 노트북 소소 코드 파일 제출