타이타닉 학습하기

타이타닉 머신러닝 과정

1. **데이터 수집 및 전처리**

데이터는 어차피 있으니까 전처리 과정이 중요하다 이말이야!

전처리 과정에서 결측치를 채워줘야 하는데

<class 'pandas.core.frame.DataFrame'>

Int64Index: 891 entries, 1 to 891

Data columns (total 11 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Survived 891 non-null int64

1 Pclass 891 non-null int64

2 Name 891 non-null object

3 Sex 891 non-null object

4 Age 714 non-null float64

5 SibSp 891 non-null int64

6 Parch 891 non-null int64

7 Ticket 891 non-null object

8 Fare 891 non-null float64

9 Cabin 204 non-null object

10 Embarked 889 non-null object

dtypes: float64(2), int64(4), object(5)

memory usage: 83.5+ KB

위와 같이 나이, 객실, 탑승지 이 3개에 결측치가 존재한다.

결측치를 대충 넣으면 나중에 정확도가 떨어지기 때문에, 정확도가 생명이다.

나이의 경우 이름에 있는 호칭을 통해 유추

객실번호의 경우 티켓의 가격이나 객실등급으로 유추해보고

탑승지의경우 조사를 해봐야겠다…

1. 모델 학습 및 평가
2. 모델 적용