

**1. 서론**

프로젝트에서는 암 환자의 사망확률을 로지스틱 회귀분석을 통해 나타낼 것인데 이때 활용한 자료는 폴란드 브로츠와프 의과대학에서 2013년에 공개한 폐암환자의 수술 전 진단 데이터와 수술 후 생존 결과를 기록한 실제 의료 기록을 사용하였고. 이를 머신 러닝으로 교육시켜 암 환자의 상황에 따른 생존 및 사망확률을 알아보았습니다.

**2. 실행 방법**

동봉된 폴더의 machine\_run 코드 파일과 ThoraricSurgery 엑셀 파일을 다운받고 실행하는데 실행기에 따라서는 numpy, pandas, matplotlib , sklearn 라이브러리를 다운받아야 하고 코랩 의 경우는 드라이브 접근 허용 후 엑셀파일을 import 해야 합니다.

**3. 코드 설명 \*폴더 구성 포함**

폴더에는 machine\_run 코드 파일과 ThoraricSurgery 엑셀 파일 지금의 word파일로 구성되어 있을 것입니다

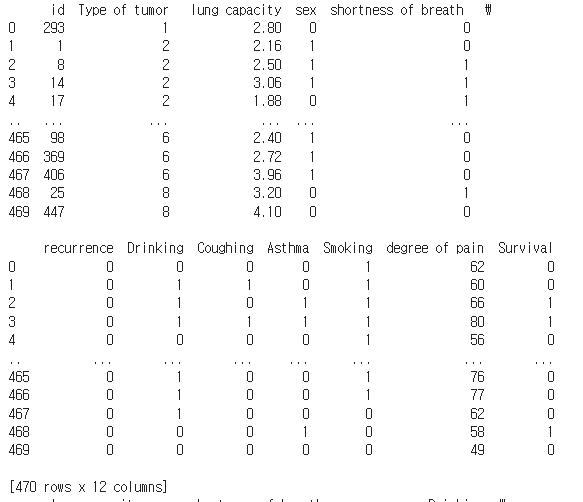
붉은색은 사이킷런의 함수이다.

|  |  |
| --- | --- |
| sklearn | 머신러닝의 대표적인 라이브러리 |
| read\_csv() | 표 형태의 파일을 읽는다. |
| head() | 앞에서 5행를 나타낸 |
| drop() | 특정 열을 삭제한 |
| train\_test\_split() | 훈련시킬 데이터(학습 데이터셋)와 그렇지 않은 데이터(테스트 데이터셋)로 나눠준다. . |
| StandardScaler() | 평균을 제거하고 데이터를 단위 분산으로 조정한다. |
| fit() | 모델을 학습시키는 함수입니다. |
| predict() +\_proba | 결과를 예측해준다. +\_proba가 붙으면 확률로 나타내준다. |
| arange() | 수열을 생성해준다 (시작점(생략 시 0), 끝점(미포함), step size(생략 시 1)) |

**4. 실행 결과**

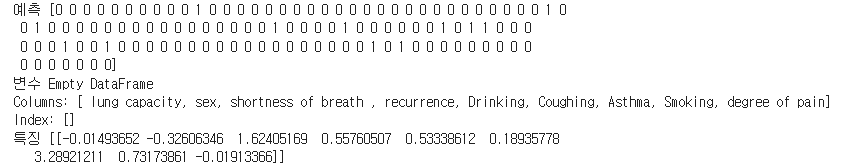
14번줄

읽어들인 데이터의 모습은 [470x12]의 크기로 모습은 아래와 같다.



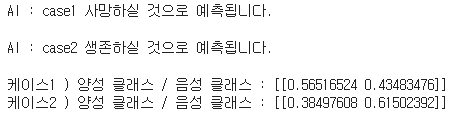
37~49번줄

로지스틱 회귀 인스턴스를 생성 후 모델을 훈련시킨 후 결과를 예측해보고 그 후에는 변수의 종류와 가중치를 알아봤다. 예측에서 0은 사망한 사람,1은 생존한 사람을 나타냅니다.



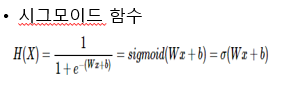
52~68번줄

임의로 폐암에 걸렸을때 음주,흡연등의 변수를 설정하고 생존여부를 확인하고 생존 확률도 알아본다.



70~77번줄

그래프 그려보든데 로지스틱 회귀에서 주로 사용되는 시그모이드 함수를 사용하였습니다.



그래프를 보면 x축의 값이 0일 경우 0.5의 값이 도출 되는데, 이를 토대로 x가 0보다 크면 y는 0.5를 기준으로 긍정의 결과가 나온다는 것을 알 수 있습니다. 그렇기 때문에 위의 생존 여부도 양성 클래스가0.5이상이면 생존, 이하면 사망으로 나타낸 것 입니다.

