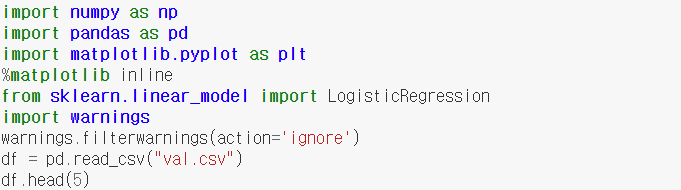
**로지스틱 회귀 분석을 이용한 질병 발생 예측**

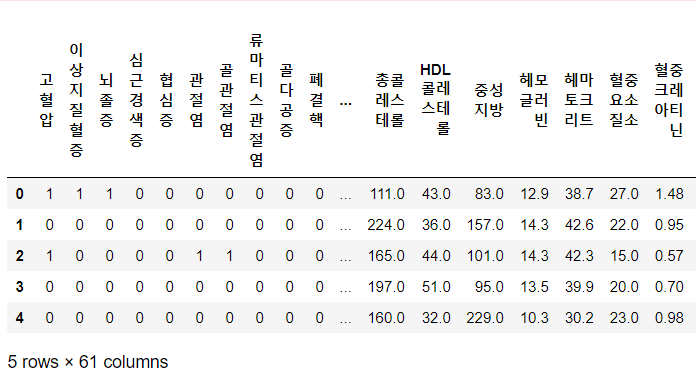
\*\*로지스틱 회귀 - 선형 회귀 방식을 분류에 적용한 알고리즘으로,

주로 이진 분류에 사용됨.

1. **이전 환자 데이터를 이용한 질병 발생 예측**
2. **필요한 모듈을 불러온 후 분석에 필요한 데이터를 불러온다.**

**(분석할 때 용이하기 위해 속성 명을 한글로 사용했다.)**





* 고혈압 ~ C형 간염까지의 변수들은 환자가 질병에 걸렸는지 걸리지 않았는지를 나타낸다. 질병에 걸리면 1, 걸리지 않으면 0이 입력되어 있다.
* 신장 ~ 혈소판까지의 변수들은 환자 개인 특성에 대한 데이터이다.

1. **결측 치 제거**





* 데이터 분석하기에 앞서 결측 치를 찾아내서 제거하였다.

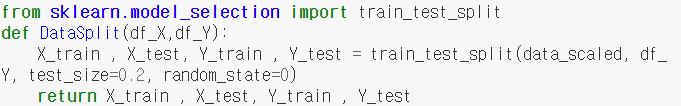
1. **필요한 작업들의 함수화**

* 로지스틱 회귀 분석에 필요한 작업들을 함수화 한다.

1. 데이터 전처리 함수 - 로지스틱 회귀분석에는 데이터 전처리가 필요하여 DataScaler이라는 전처리 함수를 생성하여 평균이 0, 분산이 1인 데이터 분포로 변환한다.

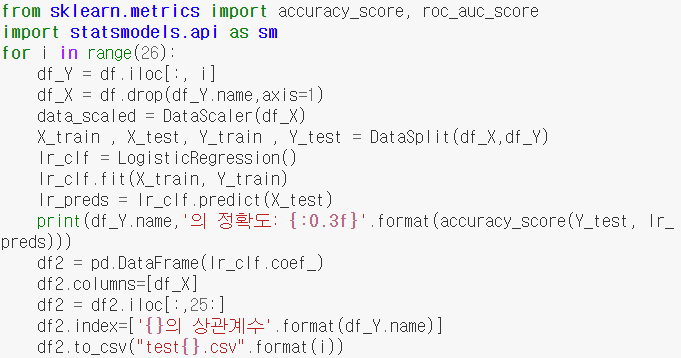


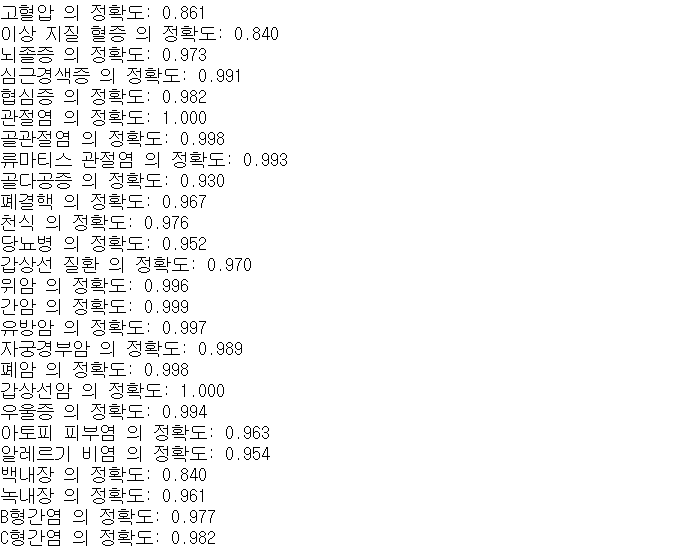
1. 데이터 스플릿 함수 – 학습 데이터와 테스트 데이터를 나누는 함수, 학습 데이터를 80%, 테스트 데이터를 20%로 설정하였다.



1. **로지스틱 회귀 분석**

* 각종 질병 발생 여부를 종속 변수로, 나머지 변수들은 독립변수로 설정한 후, 각종 질병에 대한 발생 가능성을 분석한다.
* 각각의 질병 발생 예측 분석의 정확도를 확인한다.
* 그 후, 환자 특성에 대한 변수들의 상관계수를 따로 추출하여 csv파일로 저장한다.

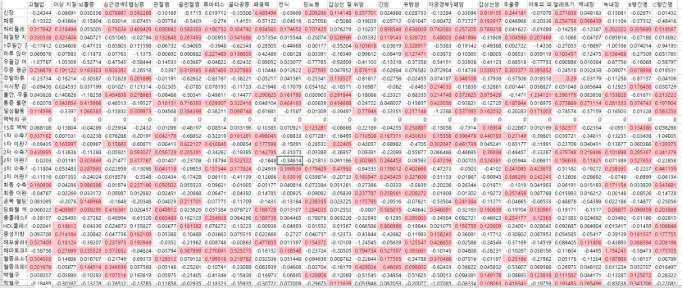




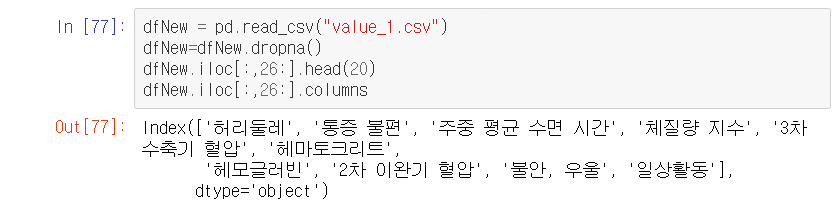
1. **상관계수 정리**

* 각 상관계수들의 상위 30%만 색깔을 다르게 하였다.





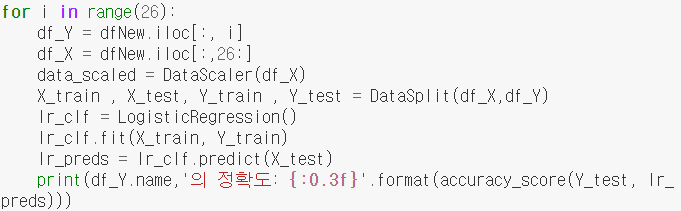
* 그 중 가장 많이 중복된 변수 중 상위 10개만 선택하여 새로운 환자 개인 특성에 대한 데이터를 만들었다.
* 선택한 변수들은 아래와 같다.

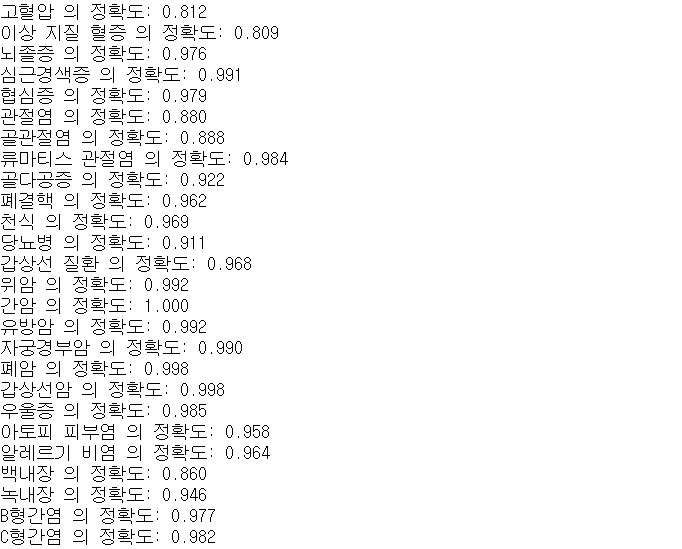




1. **정리된 상관계수를 이용한 질병 발생 예측**

* 선택한 변수들을 독립변수로 설정하여 다시 한번 질병 발생 가능성에 대한 로지스틱 회귀 분석을 진행하였다.





* 대부분 90% 이상의 정확도를 가지는 예측 모델이 생성되었다.

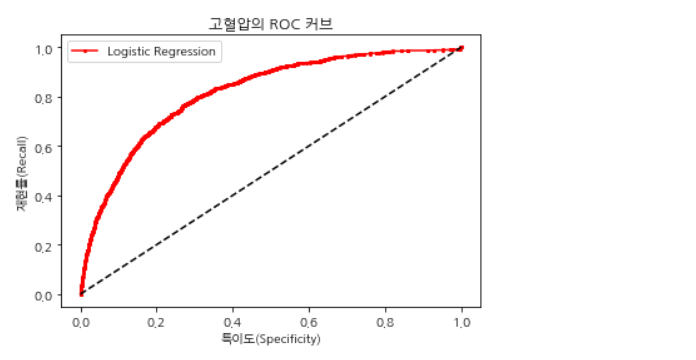
1. **모델의 신뢰도 평가**

* **모든 모델의 신뢰도를 알아보는 것은 많은 시간이 걸리므로 예시로 ‘고혈압’과 ‘위암’ 모델의 신뢰도를 알아보았다.**

1. **고혈압 모델의 정확도와 신뢰도**



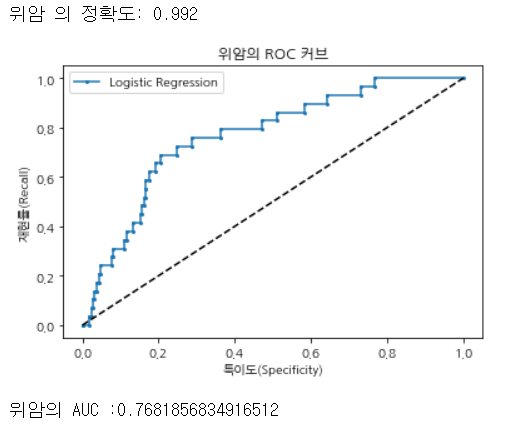






1. **위암의 모델의 정확도와 신뢰도**

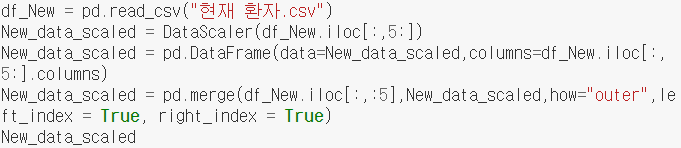


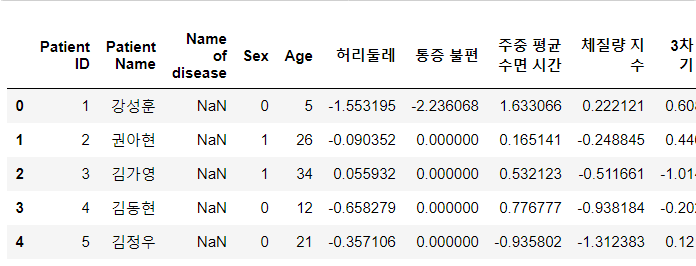


* **모든 모델이 70%이상의 신뢰도를 가지고 있다.**

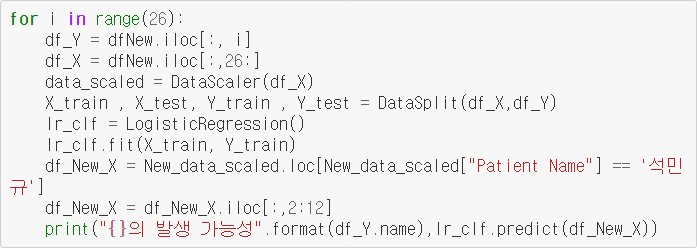
1. **현재 환자 데이터를 이용한 질병 발생 예측**

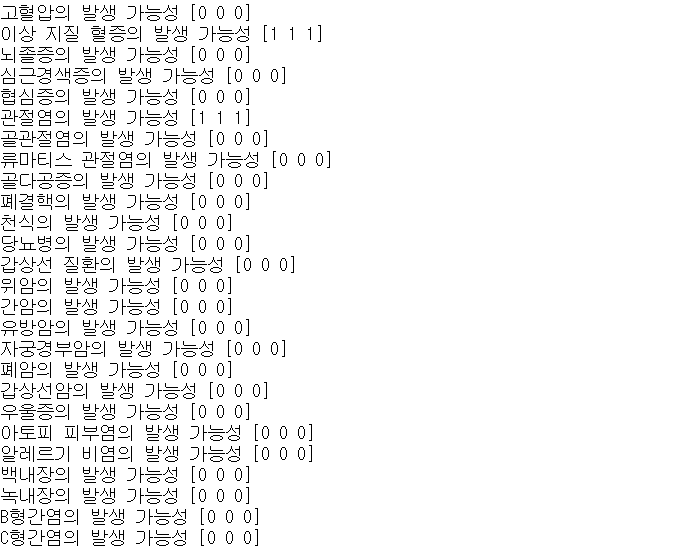
* 현재 환자 데이터를 위 예측 모델에 적용하였다.
* 예측을 위하여 환자 개인 특성에 대한 값 들은 전처리 하였다.



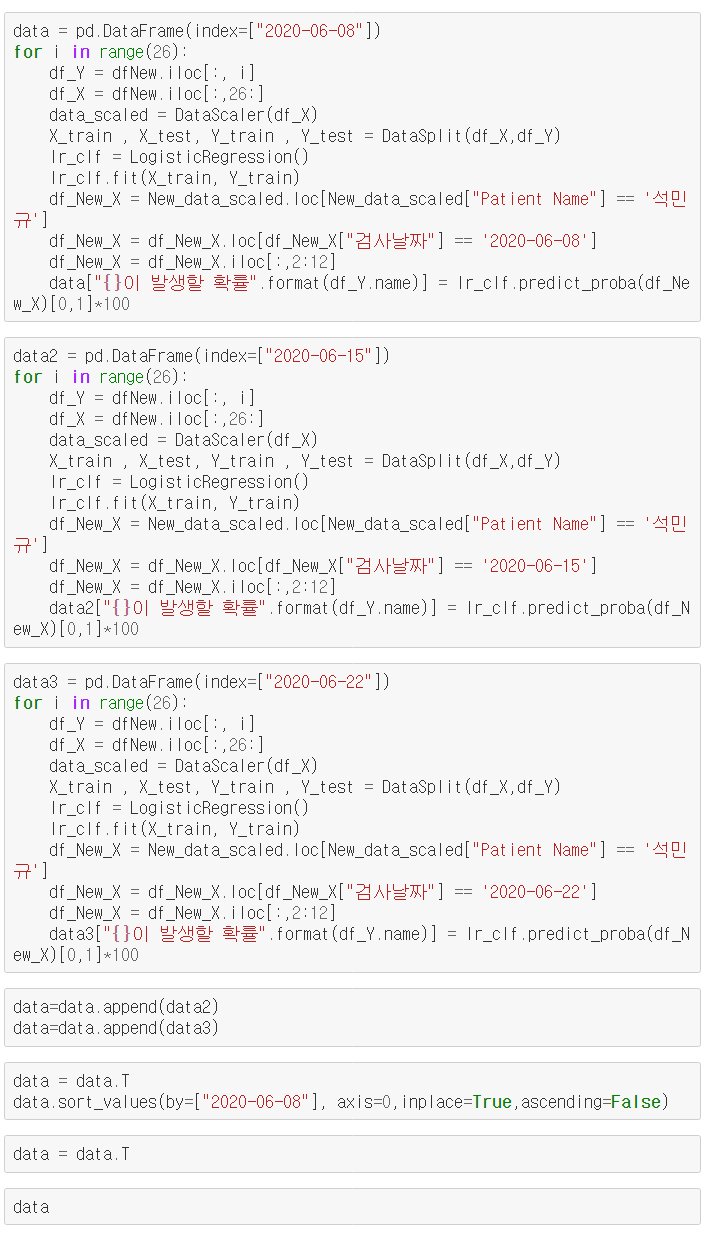


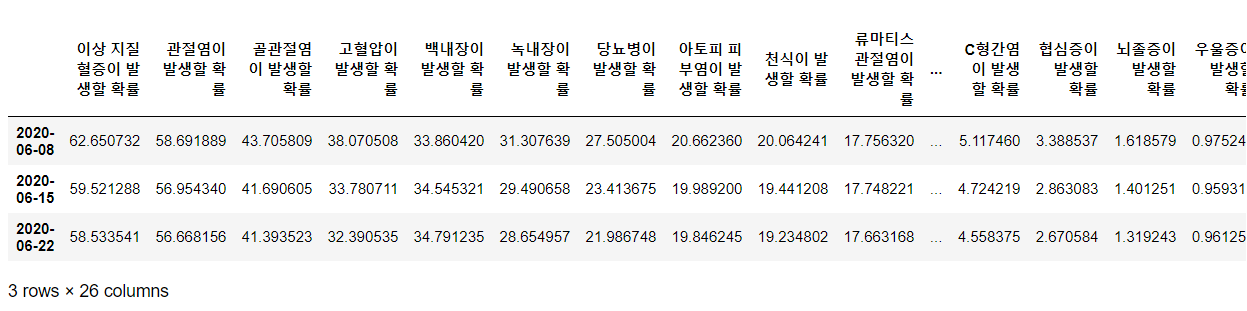
* 예시로 ‘석민규’ 환자의 질병 발생 가능성을 예측하였다.



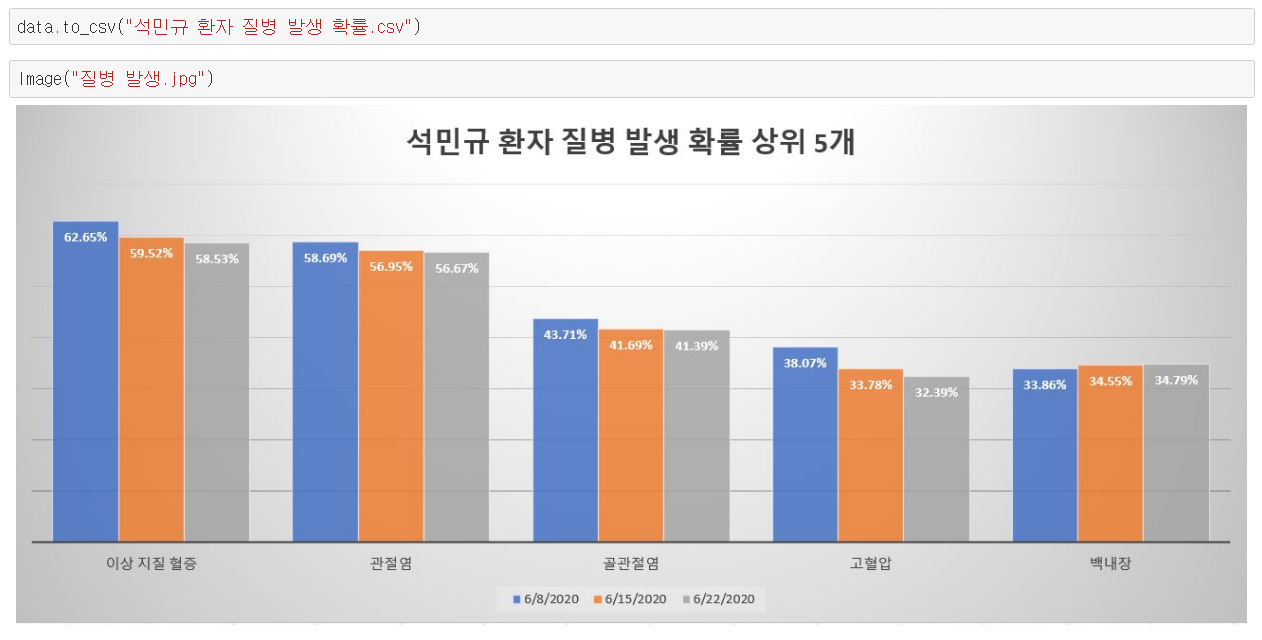


* 숫자가 0인 경우 질병이 발생하지 않을 확률이 더 큰 것이고, 숫자가 1인 경우 질병이 발생할 확률이 더 큰 것이다.
* 각종 질병의 발생확률을 표로 만들고, 확률이 큰 순서에서 낮은 순서로 내림차순 정렬하였다.





* 확률 상위 5개의 변화에 대해 알아보았다.



* 대부분의 발생 확률이 일주일 전보다 감소하였다.
* 따라서, ‘석민규’ 환자의 상태는 호전되고 있다고 할 수 있다.