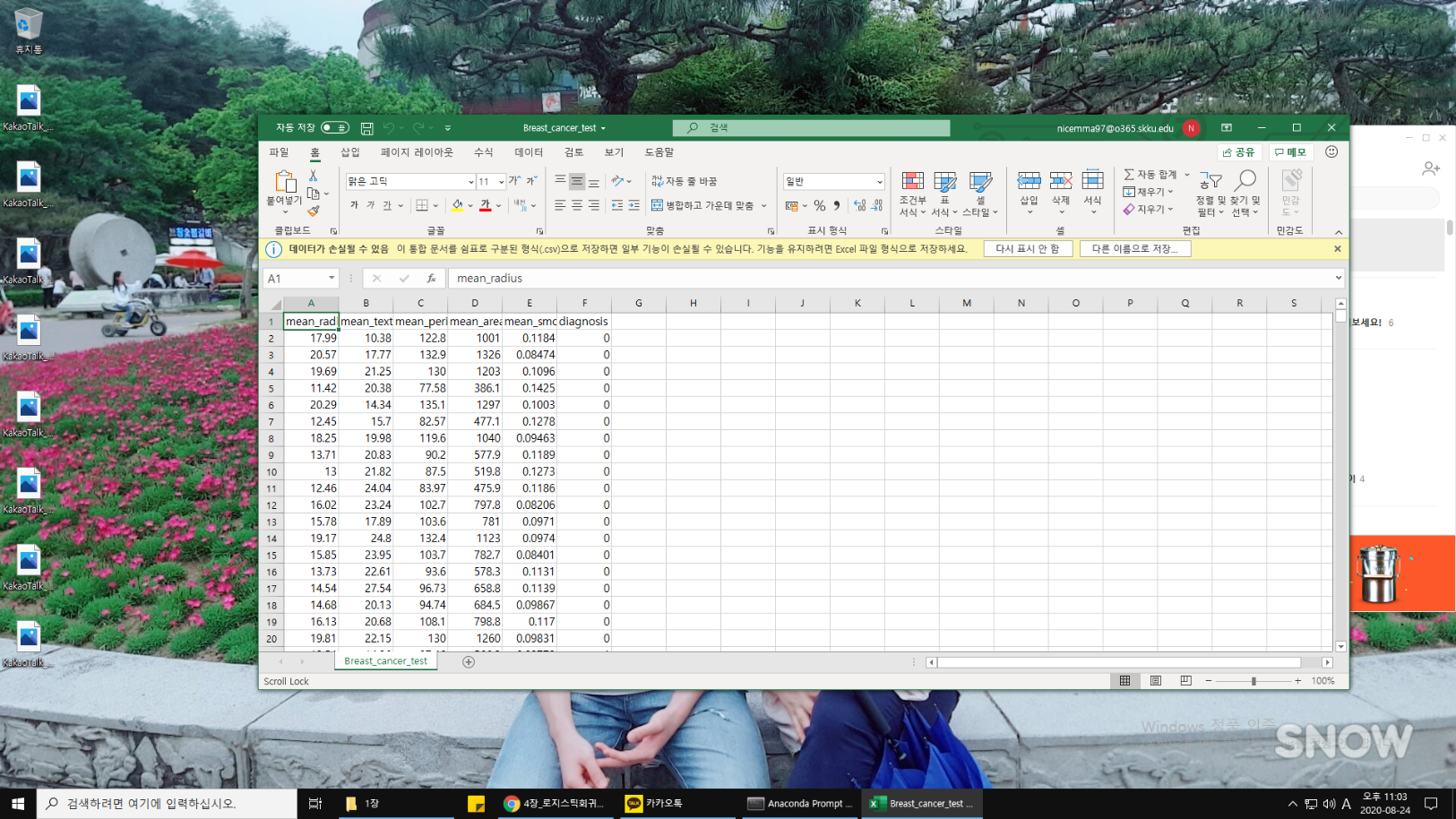
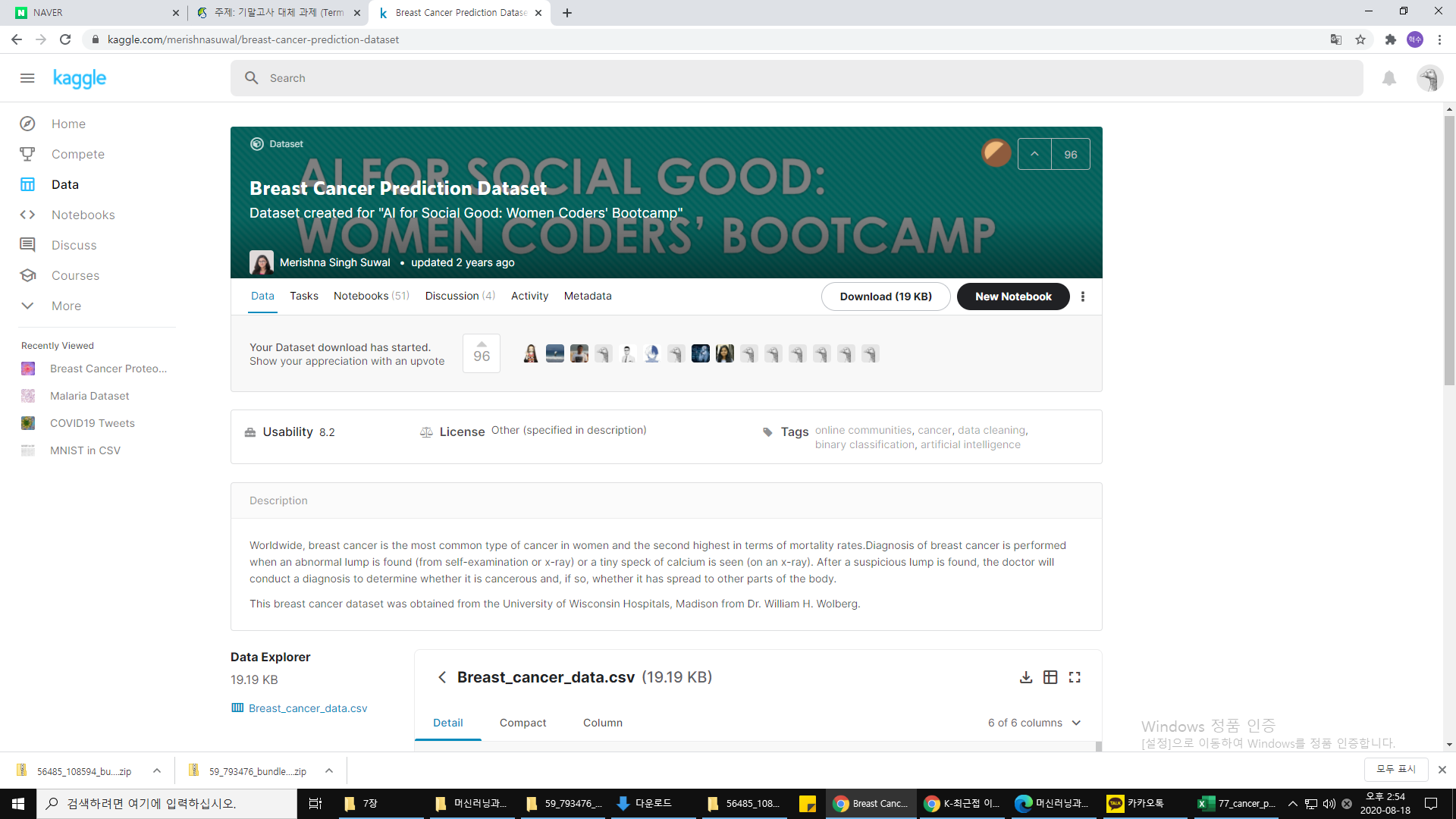
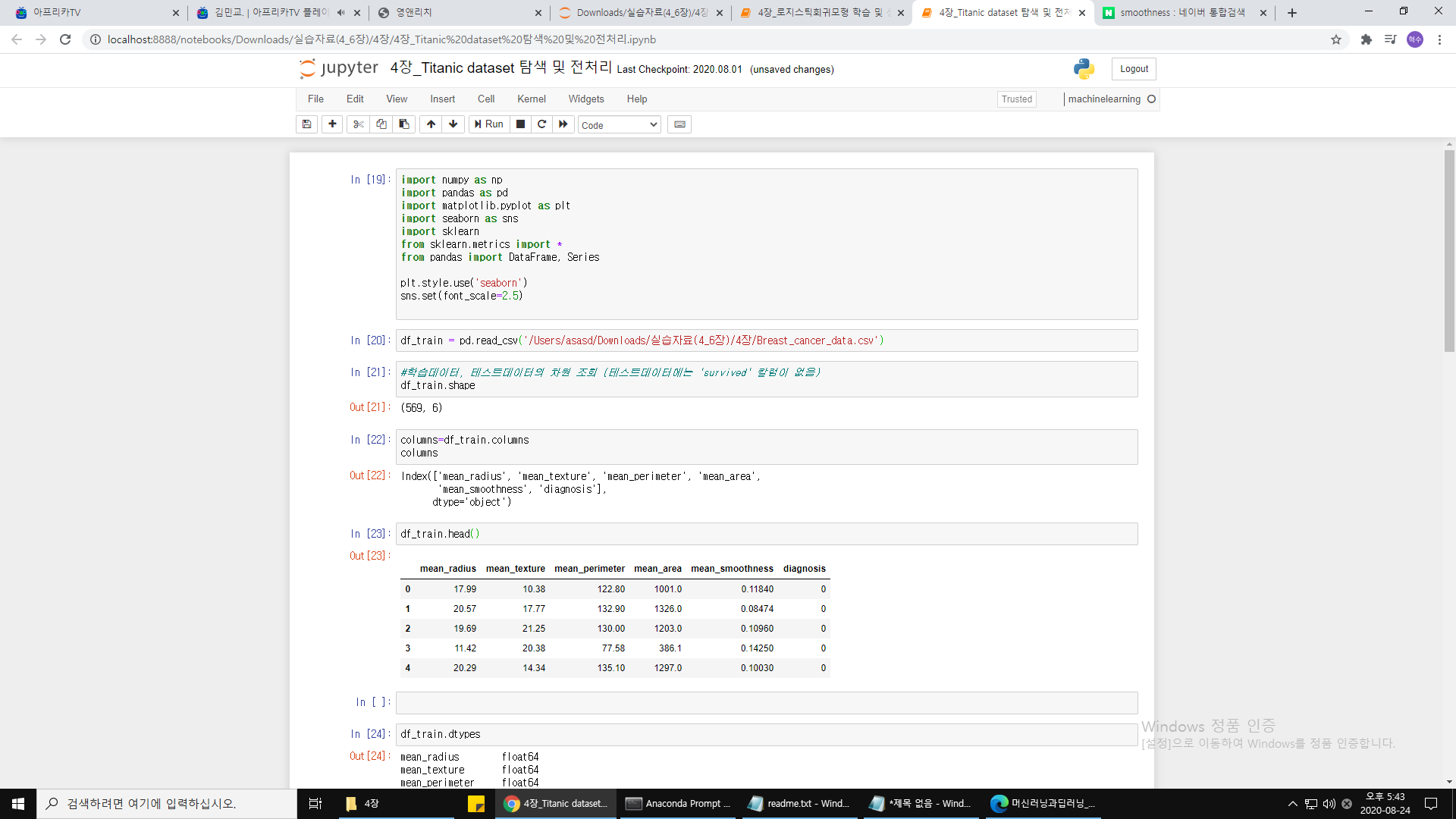
2016312761 여혁수 기말 대체 과제

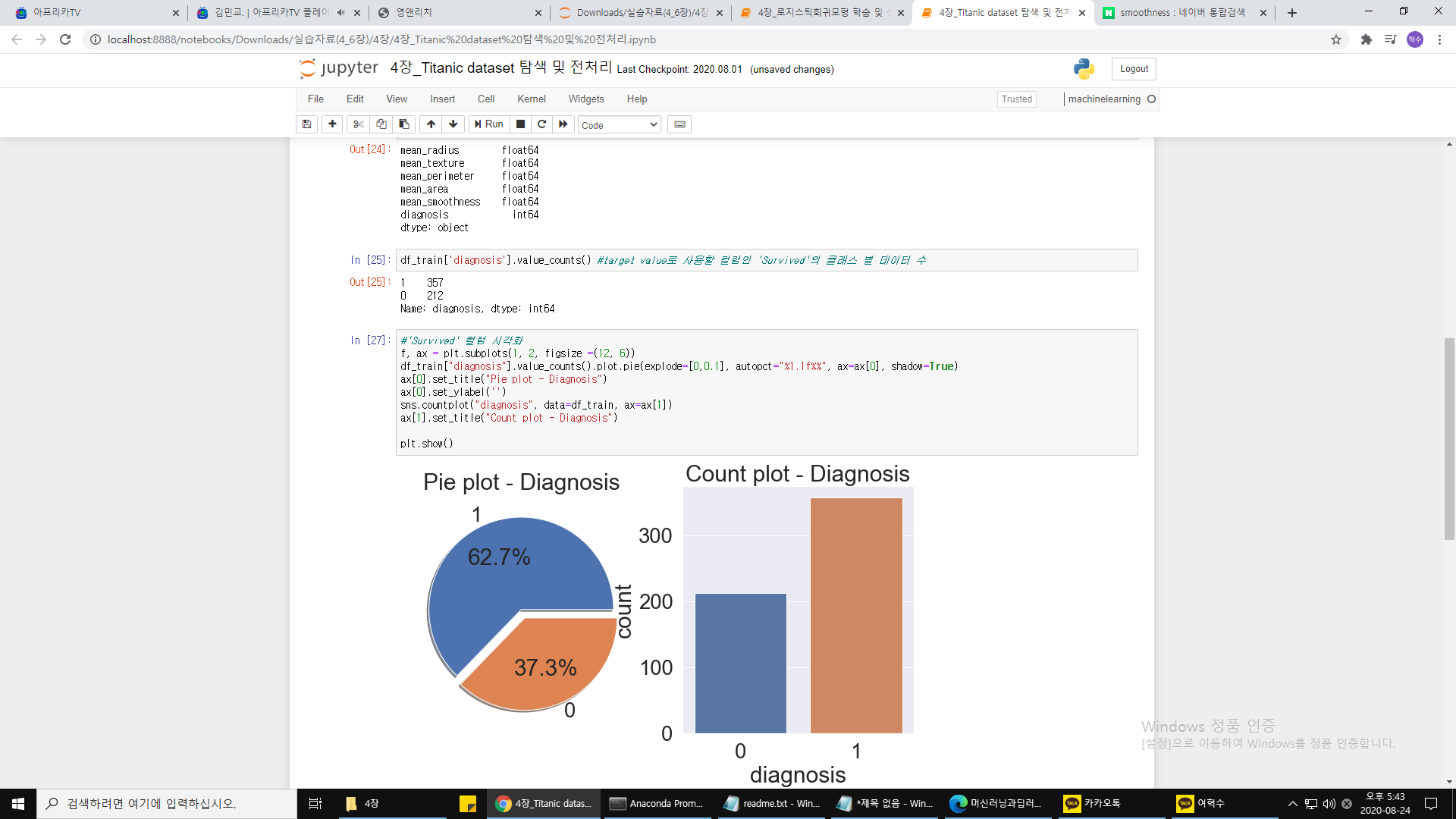
현재 의학계에서 암세포의 형태를 보고 1차적으로 암을 추정하고 세부 검사로 넘어간다. 그렇기 때문에 1차에 암세포를 정확하게 판단하는 것이 조기 치료에 중요한 역할을 할 것이다. 그래서 나는 암세포로 추정되는 세포의 형태를 보고 암 환자인지 아닌지를 판단할 수 있는지 그리고 얼마나 정확한지 확인해보기로 했다.

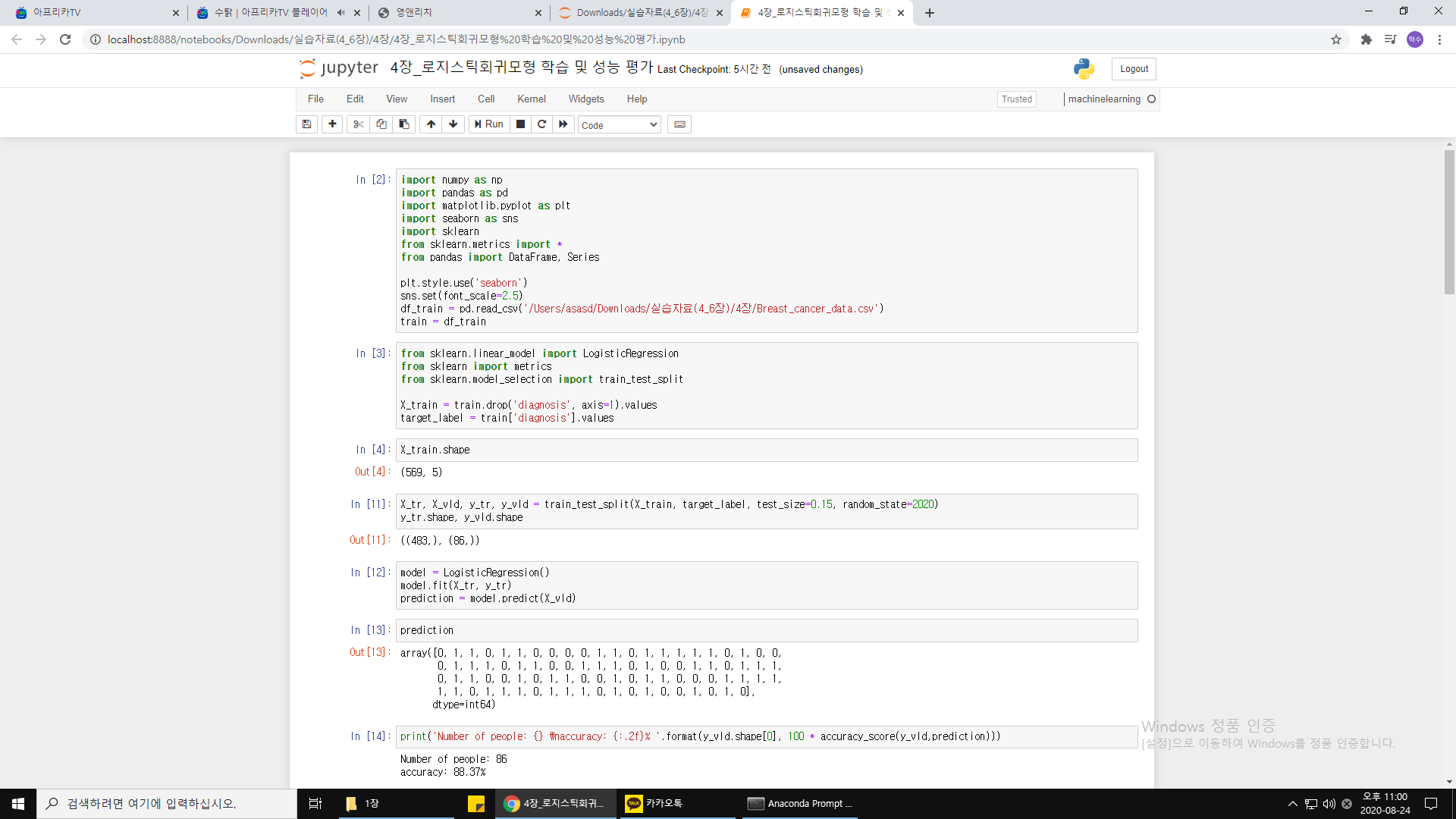
대장암, 간암 등 암은 인체 모든 부위에서 생길 수 있고, 부위마다 암세포의 형태가 다르다. 그래서 한 종류의 암만 뽑아서 탐색해야한다. 나는 유방암으로 추정되는 암세포의 형태와 암 진단 유무와 관련한 내용이 담긴 데이터를 캐글 사이트에서 가져왔다. 

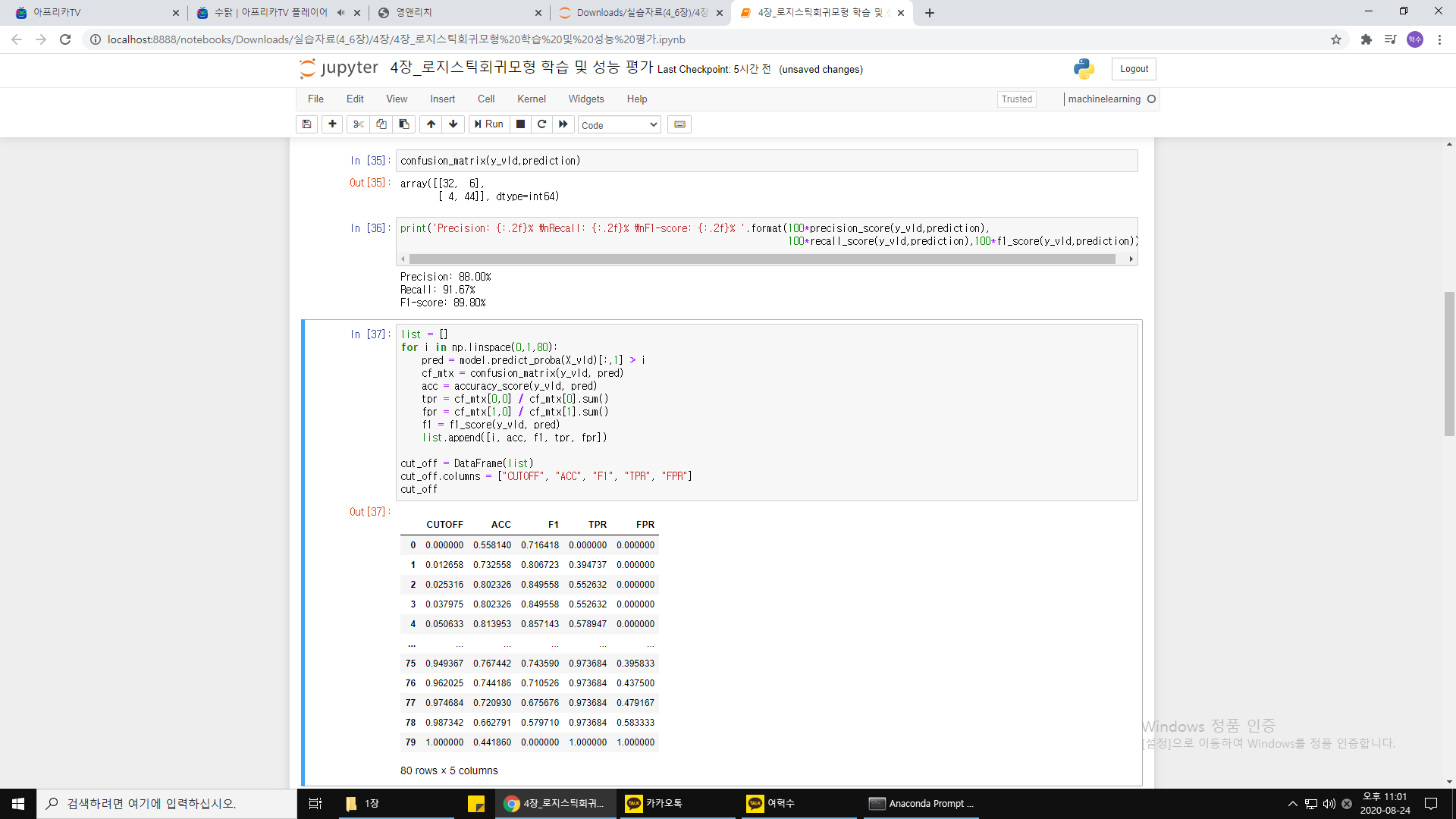
이 데이터에는 암을 판단할 수 있는 5가지 속성이 있다. Radius는 암세포의 반지름을 얘기한다. Texture는 정확히는 모르겠지만 암세포의 두께인 것 같다. Perimeter는 둘레, area는 넓이를 말한다. 마지막으로 smoothness는 평활도로 암세포의 표면이 얼마나 매끄러운지 혹은 울퉁불퉁한지를 수치화하여 나타낸 것이다.

암인지 아닌지 이렇게 크게 두 부류로 분류하기 때문에 2진 분류 모델을 사용하였다. 2진 분류 모델 중 하나인 로지스틱 회귀 알고리즘을 사용하면 분류의 정확도와 성능을 판단할 수 있기 때문에 적합한 방법이라고 생각했다.

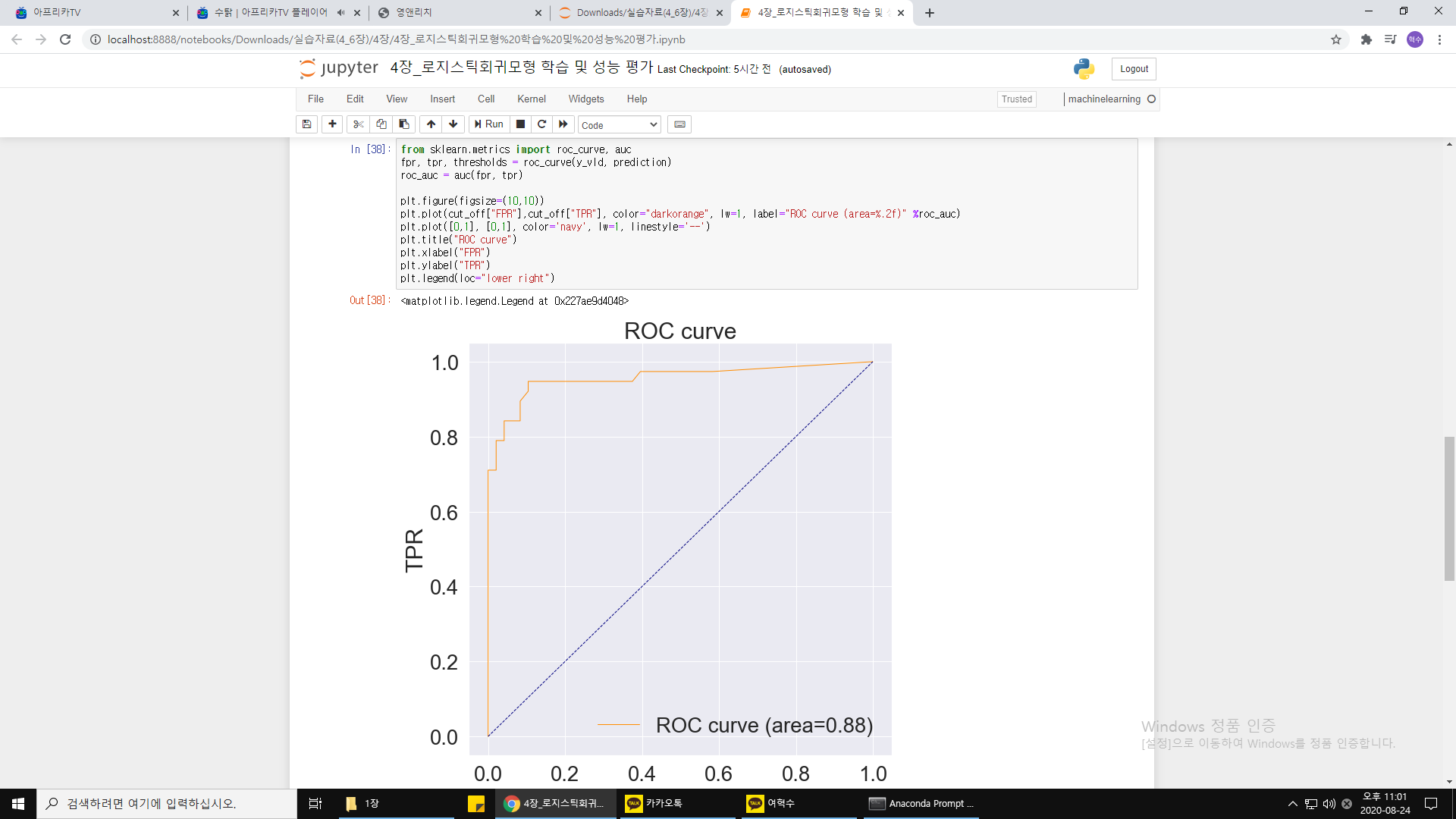


전반적인 해결과정은 4주차 실습의 로지스틱 회귀모형 알고리즘과 같다. Train set이 569개로 이루이루 있는 것을 볼 수 있고, 5개의 실수형 속성으로 암 진단 유무를 분류할 것을 알 수 있다.

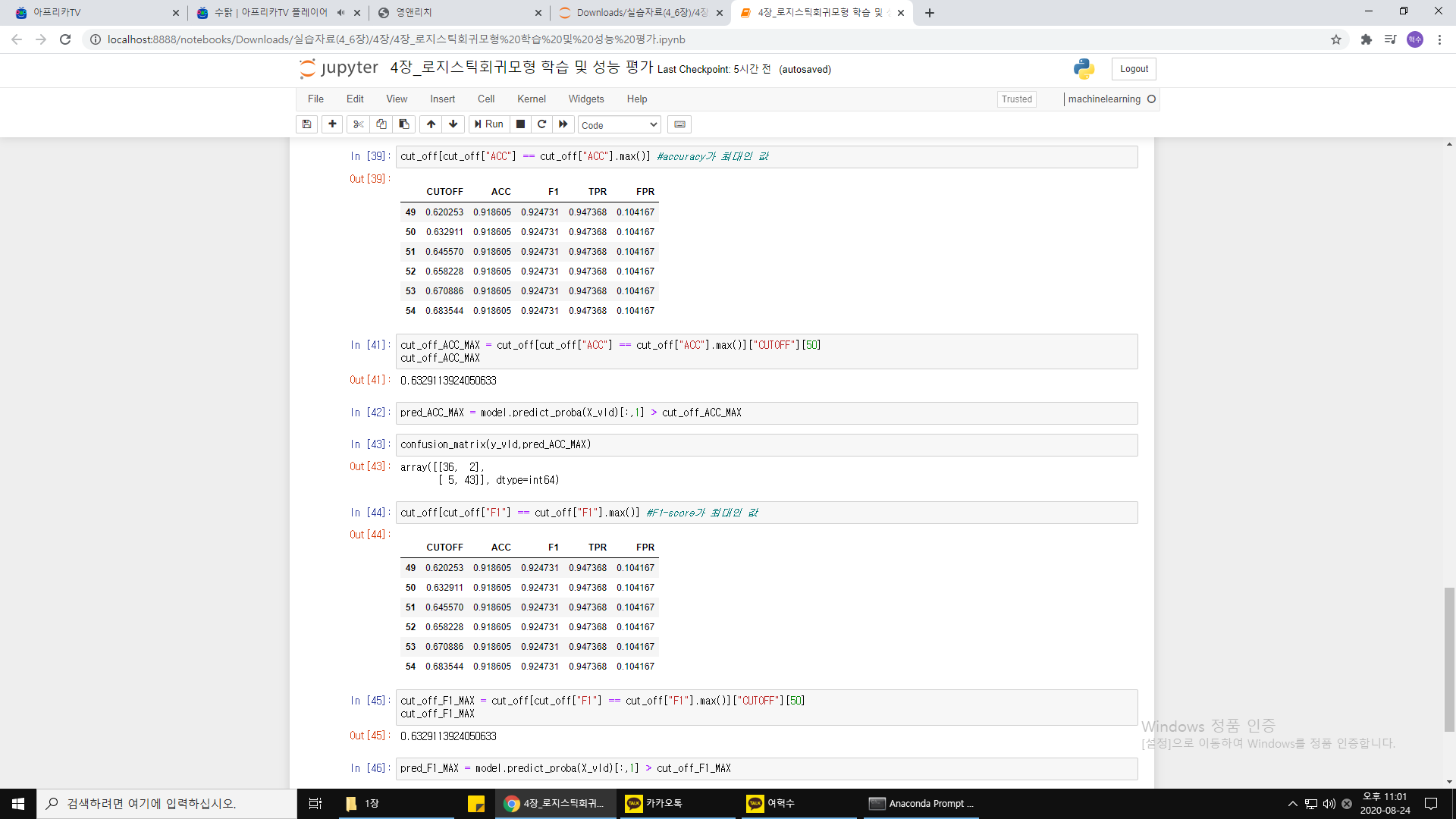
암으로 진단받은 케이스와 그렇지 않은 케이스의 비율을 두 개의 모형으로 나타낸 것이다.

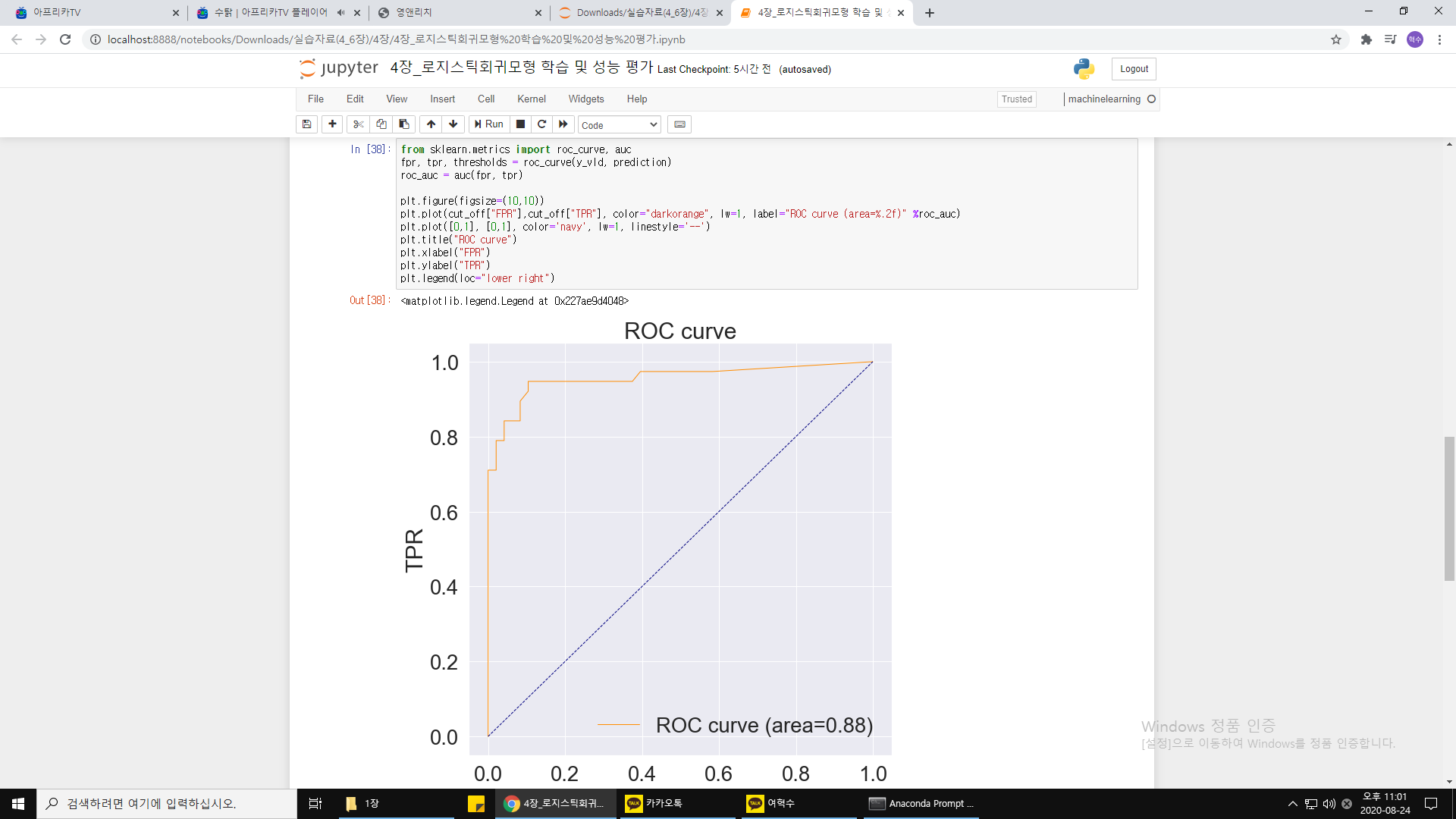
위 사진부터 본격적인 학습인데 우선 암 진단 유무를 분류하는 것이기 때문에 target label을 ‘diagnosis’ 속성으로 잡았다. 그리고 기존의 test set이 따로 없기 때문에 train set에서 15%을 validation set으로 설정하고 로지스틱 회귀 모형을 학습했다. 예측한 결과가 0과 1로 보여지고, 정확도는 약 88%로 나오는 것을 볼 수 있다.

다음으로 정밀도(precision)와 재현율(recall)을 구했는데 정밀도는 모델이 암이 맞다고 예측한 것 중에 실제로 그것이 참인 케이스들의 비율이고, 재현율은 반대로 실제로 암이 맞는 것 중 모델이 참이라고 예측한 케이스들의 비율이다. 이 두 값은 서로 trade off 관계인데 차이가 88% 대 91.67%로 차이가 크지 않기 때문에 두 값의 조화평균인 F1-score도 89.8%로 괜찮은 것을 알 수 있고, 성능이 나쁘지 않다는 것을 의미한다. Cut off value는 80으로 정했는데 train set이 크지 않아서 그냥 이렇게 설정했다.

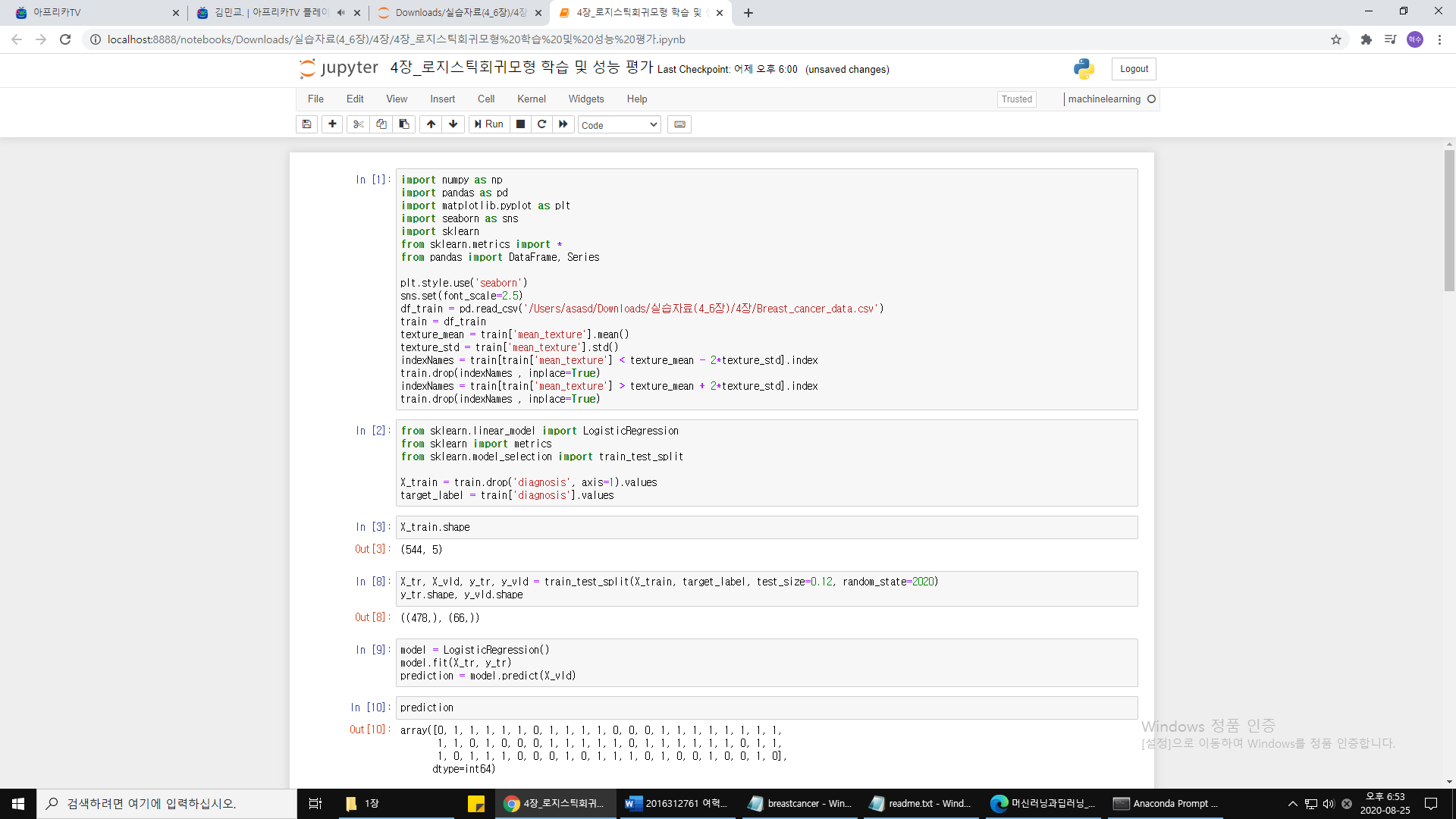


다음으로 의학에서 많이 사용되는 ROC curve이다. FPR와 TPR를 이용하여 그려지는데, 이 학습에서 FPR은 실제로 암이 아닌 것 중에서 모델이 암이라고 잘 못 예측한 것들의 비율이 될 것이고, TPR은 반대로 암으로 예측한 것들 중에서 정답을 맞춘 것의 비율이 될 것이다. 정말 속성이 5개밖에 없고, 정확도도 높기 때문에 ROC curve가 이상적인 곡선이 아닌 직선에 가깝게 그려지는 것을 볼 수 있다.

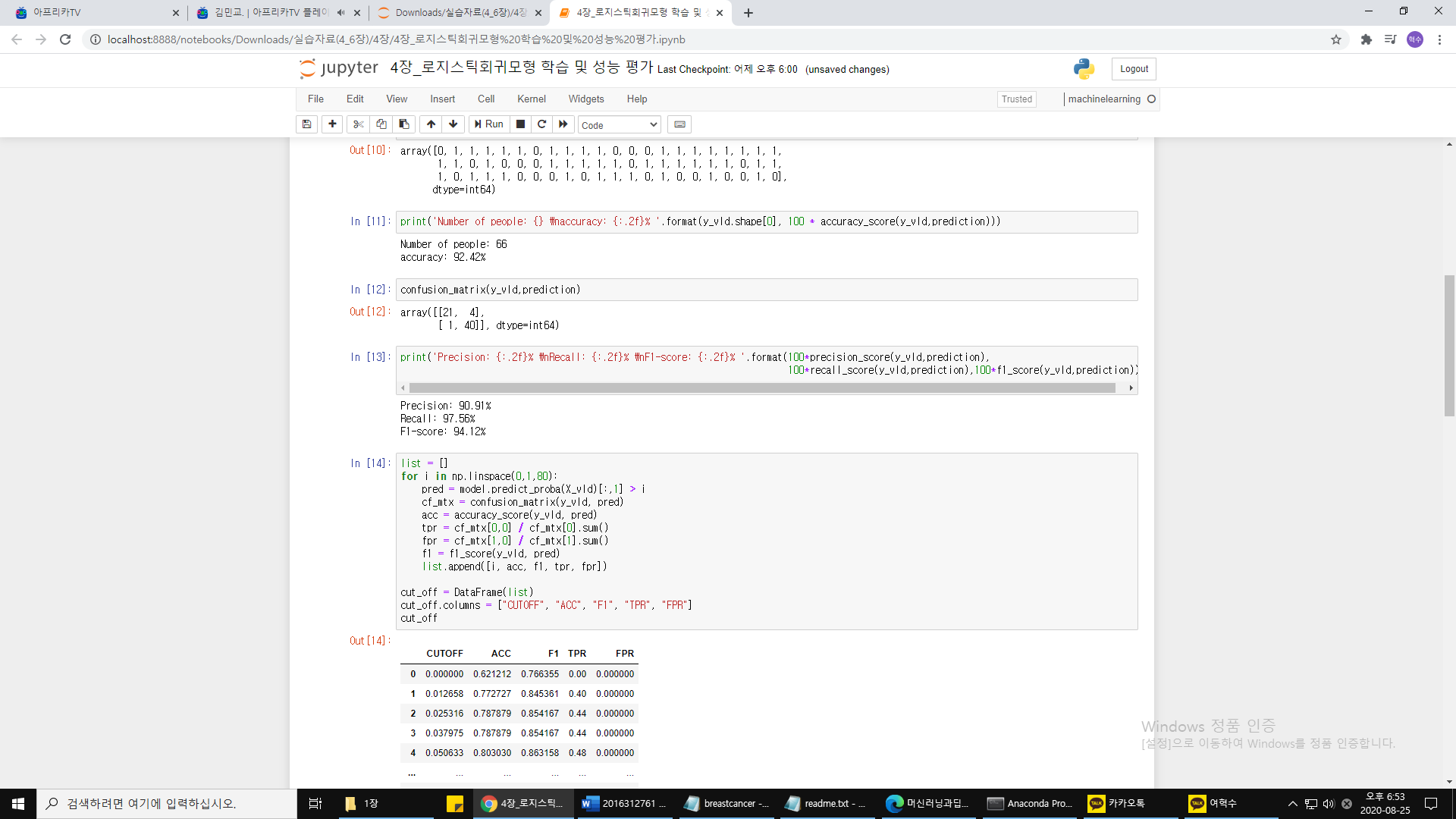




cut off value를 좀 작게 잡았기 때문에 여러 개의 max accuracy 케이스가 나왔고, F1-score도 마찬가지로 동일한 케이스들이 최대값을 가진다.



더 좋은 학습을 위하여 이상치 제거를 해보았다. Texture 속성에서 평균보다 2\*σ 크거나 작은 범위 안에 포함되지 않는 값은 이상치로 판단하였다. 그랬더니 569개의 케이스 중 544개의 케이스만 train set에 남았다.



그 결과 정확도는 약 92%로 올라갔고, 성능 역시 f1-score 94.12%로 향상된 모습을 보여주었다.

이상치에 대한 의학적인 연구가 잘 되어 있다면 이런 분류 방법은 현실에서 적용될 수 있을만큼훨씬 정확하고 좋은 성능을 보여줄 것 같다.