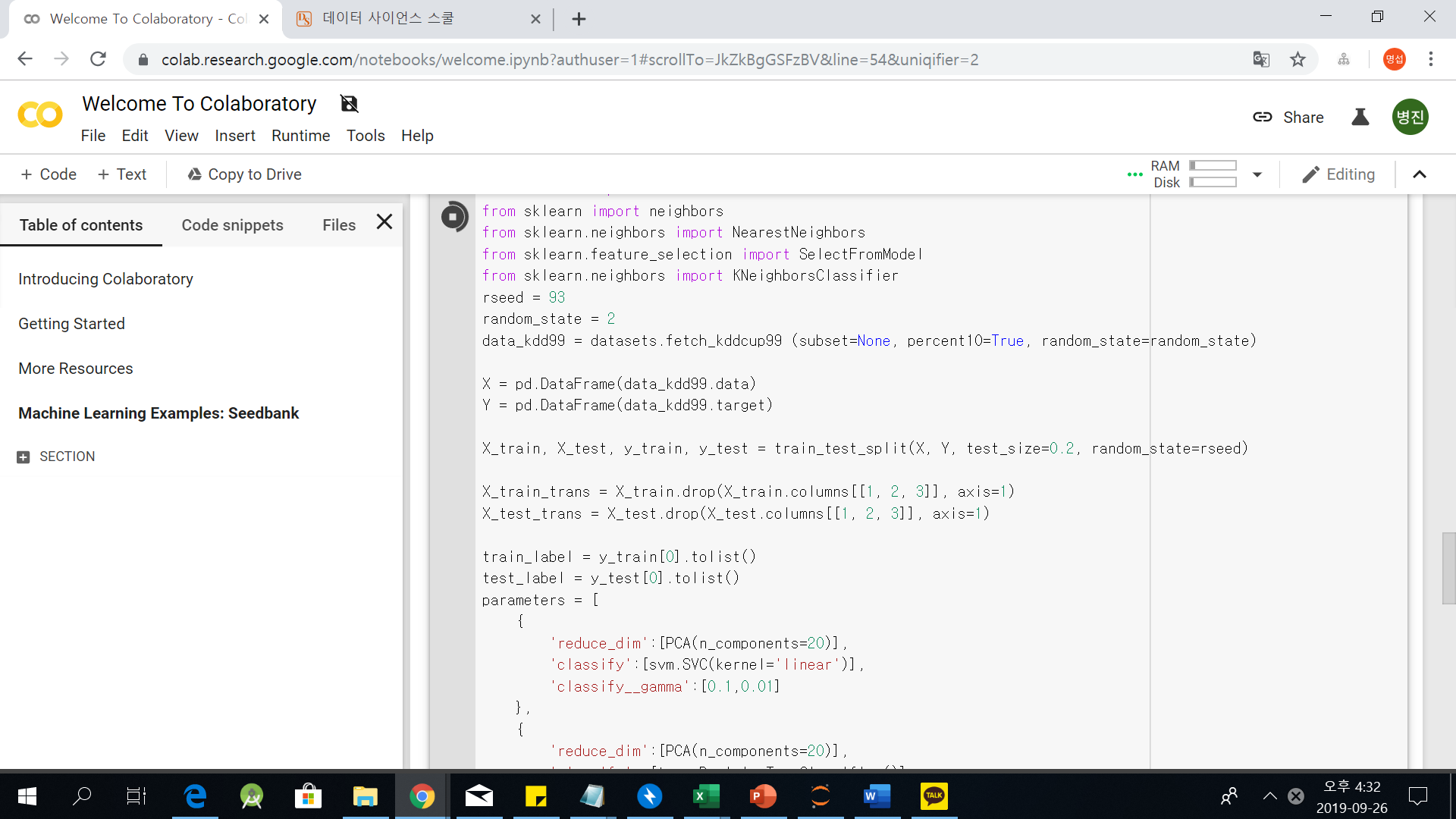
KDD99를 이용한 parameter tuning

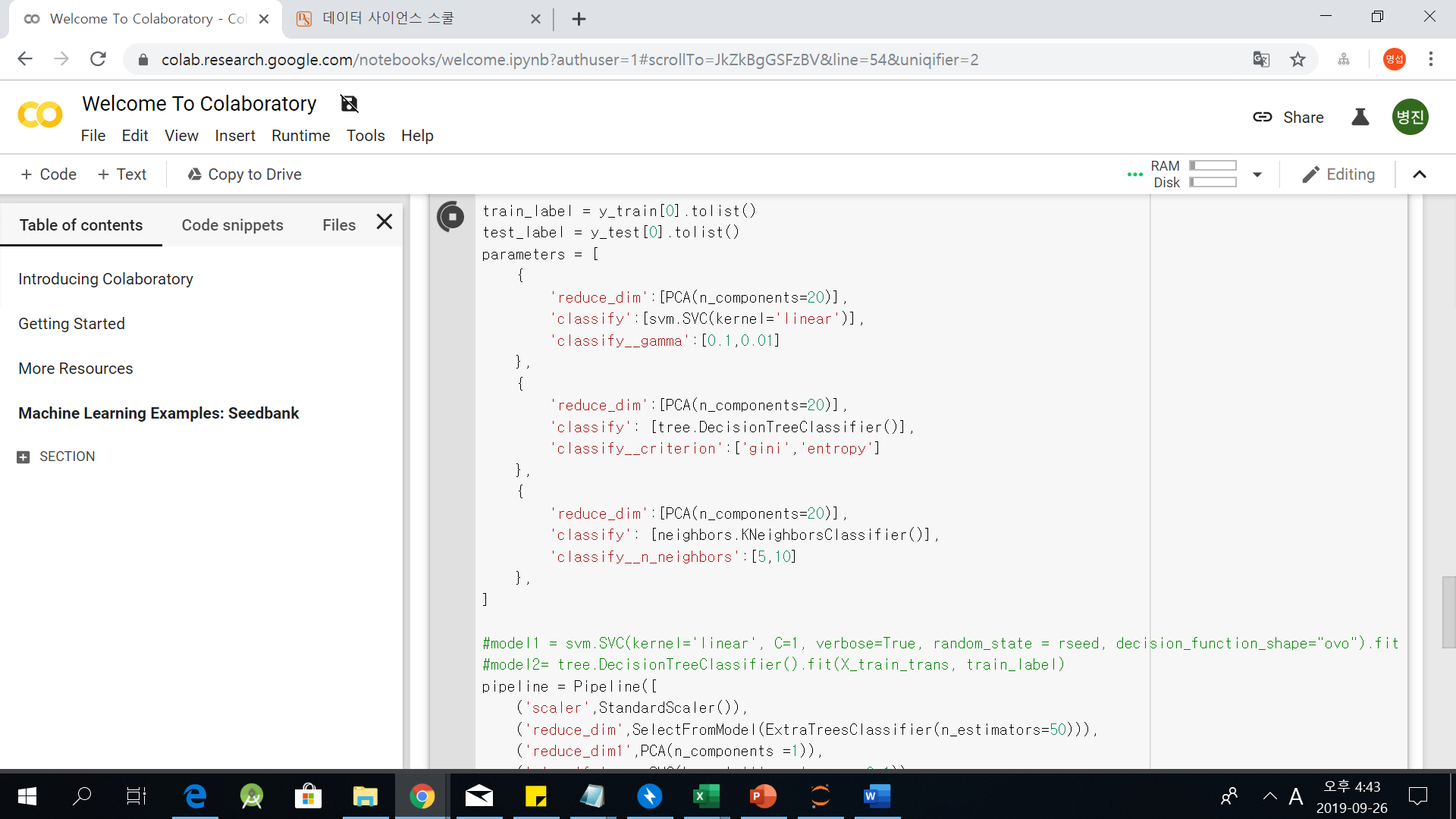
201520991

곽명섭

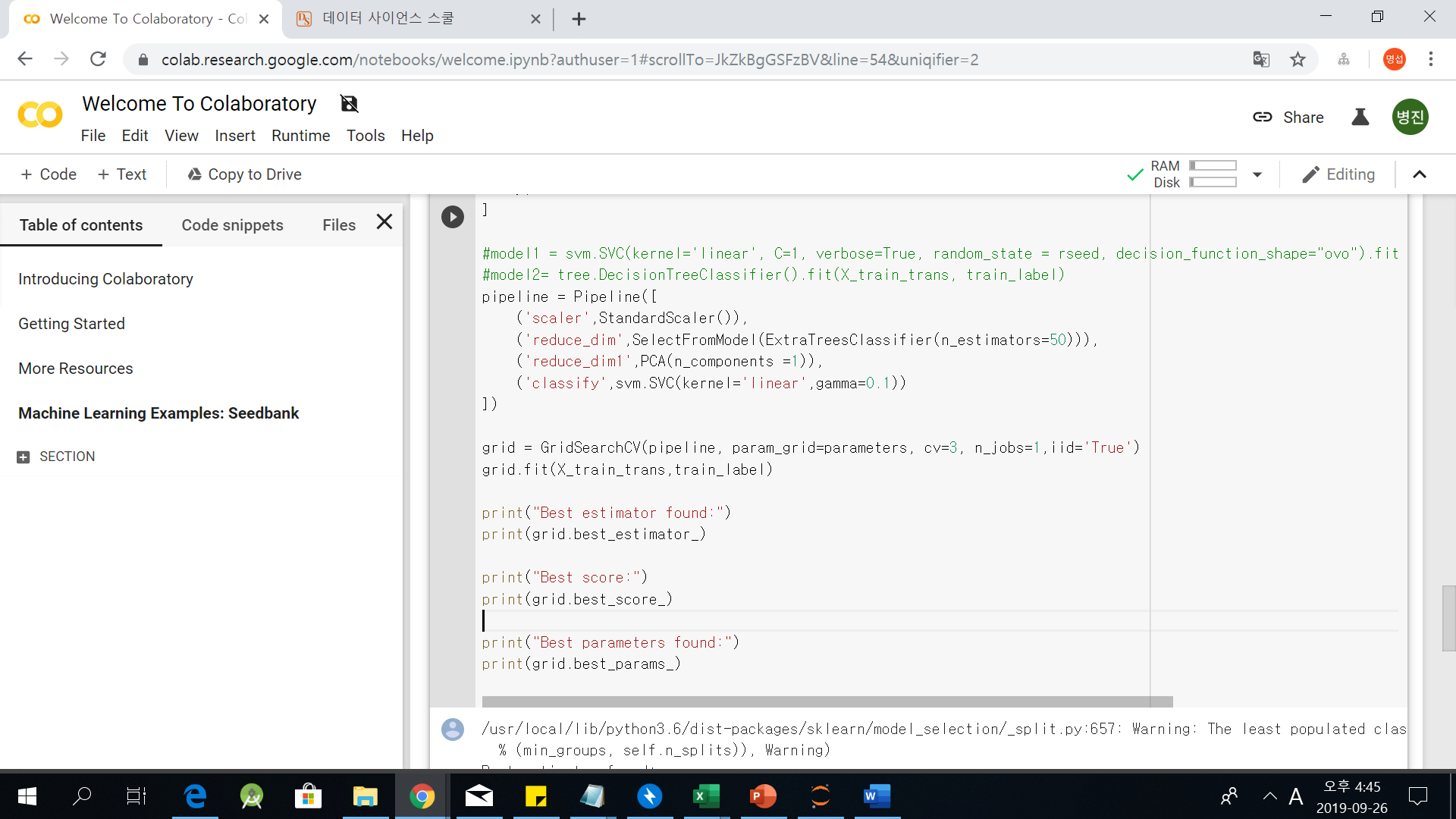
오늘 실습시간에는 KDD99 dataset을 이용하여 parameter tuning을 하는 실습을 해보았습니다.



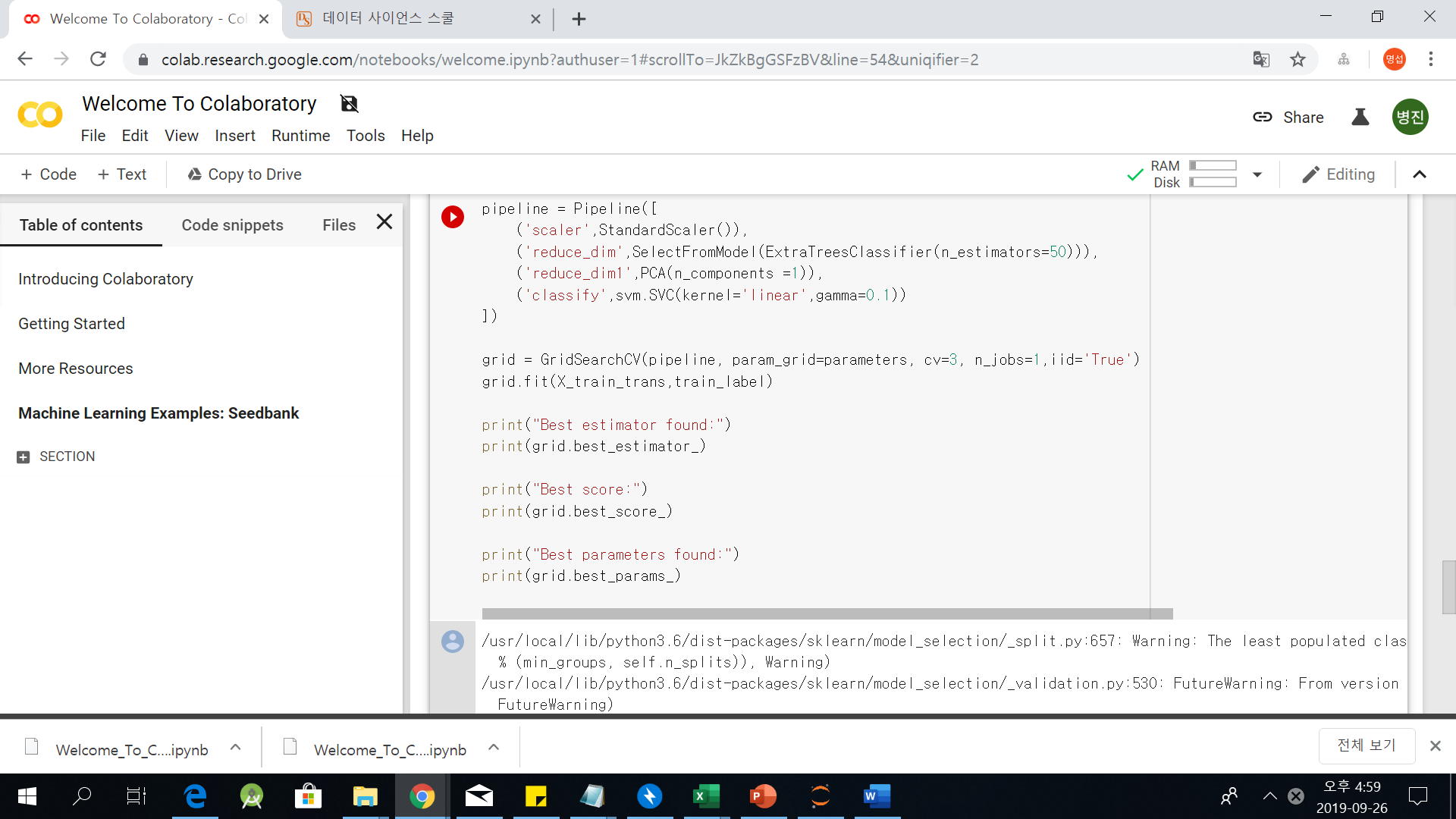
저번 과제에서와 같이 kdd99 데이터를 불러와서 앞의 3줄 char형 자료들을 drop하고 데이터를 가공하는 과정이다. 그리고 label들을 train\_label과 test\_label에 list로 저장해주는 과정이다.



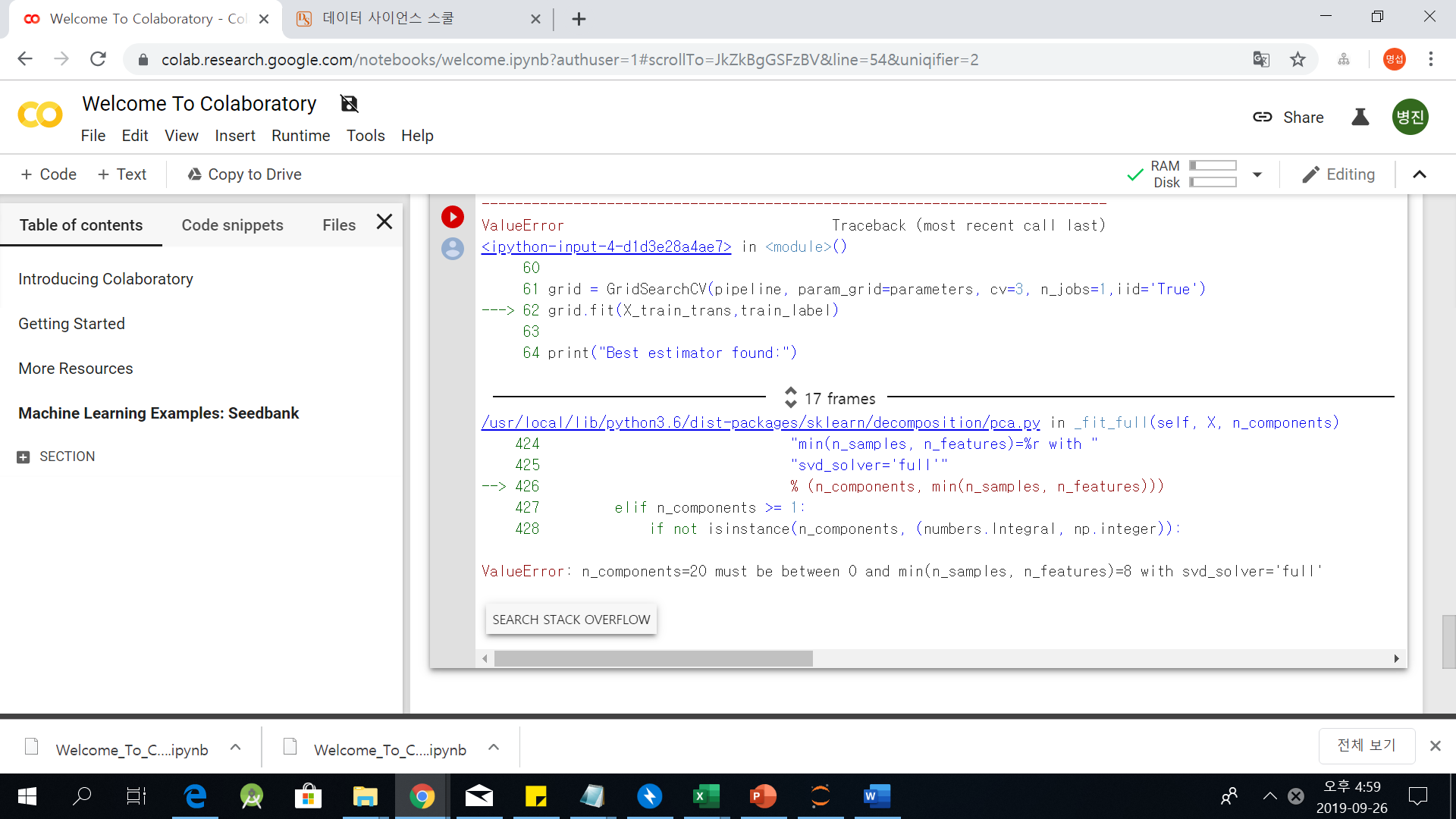
Parameters 부분인데 이 부분이 이해하기가 가장 어려운 부분이었습니다.



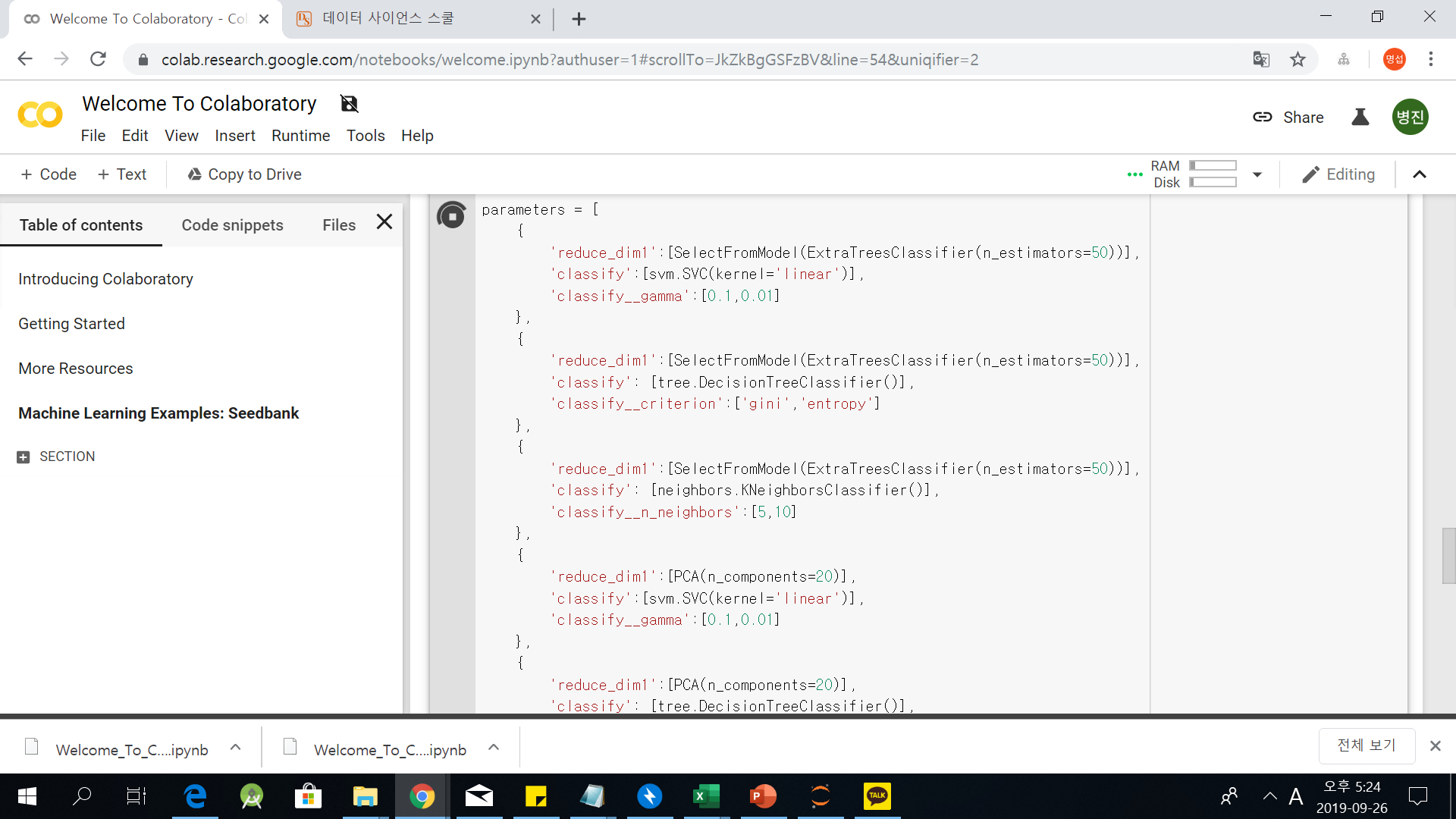
제가 이해한 바로는 우선 pipeline에서 sclaer, reduce\_dim, reduce\_dim1, classify 등의 default값으로 standardscaler, extratreeclassifier, pca, svm으로 설정을 해놓고 그 다음에 parameters에 있는 것들을 하나씩 돌면서 값을 바꿔가며 실행해 가는 과정인 것으로 이해하였습니다. 그리고 나서 그 중에서 가장 좋은 parameter을 출력을 해주는 것이 밑의 출력 부 코드 부분입니다

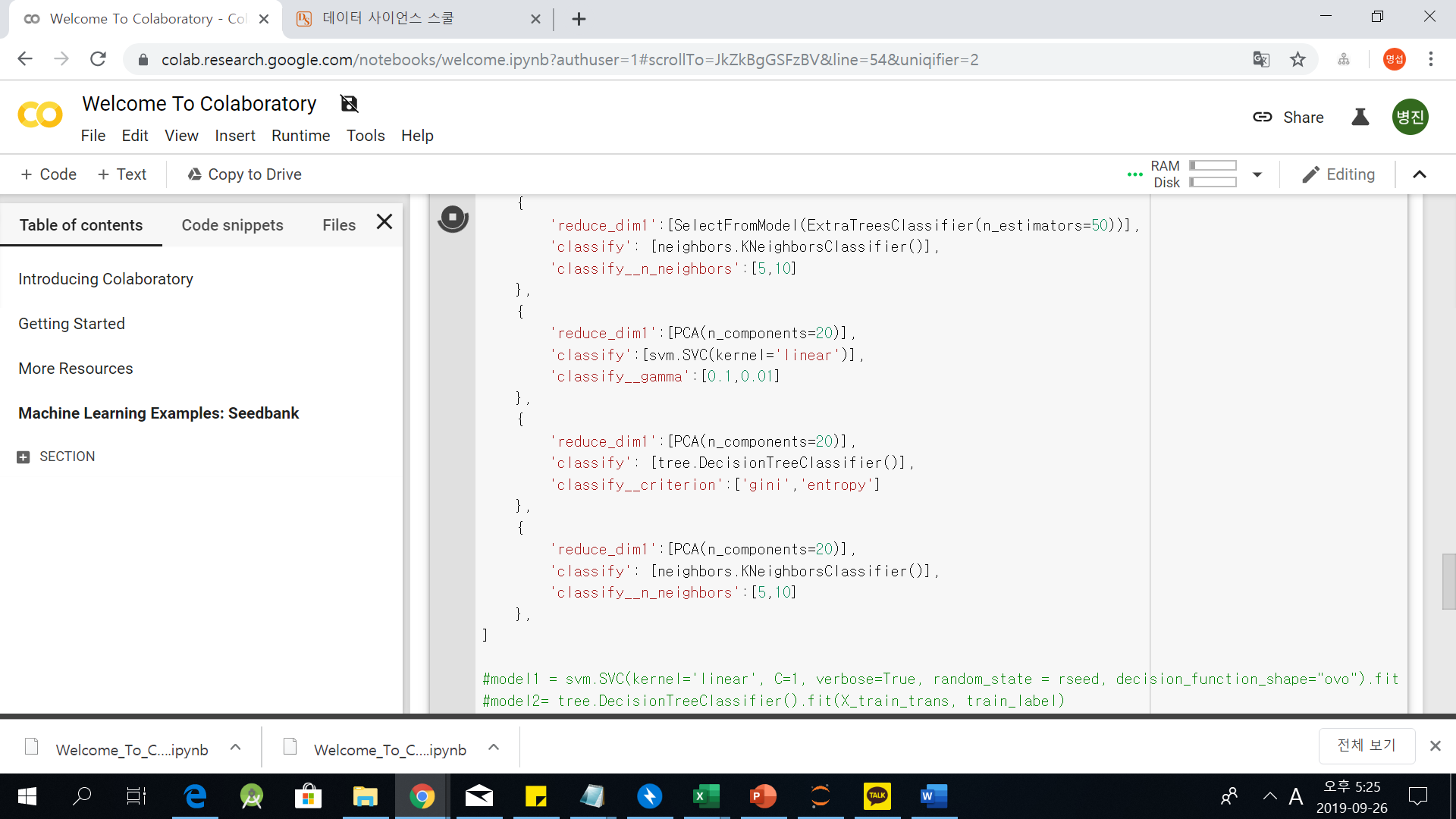


그래서 우선 ppt에 있는데로 PCA, Tree-based feature selection을 둘 다 사용하고 n\_components = 20이랑 n\_estimators = 50을 둘 다 값으로 주고 출력을 시도 해보았는데 밑의 사진과 같은 value error가 나왔습니다.

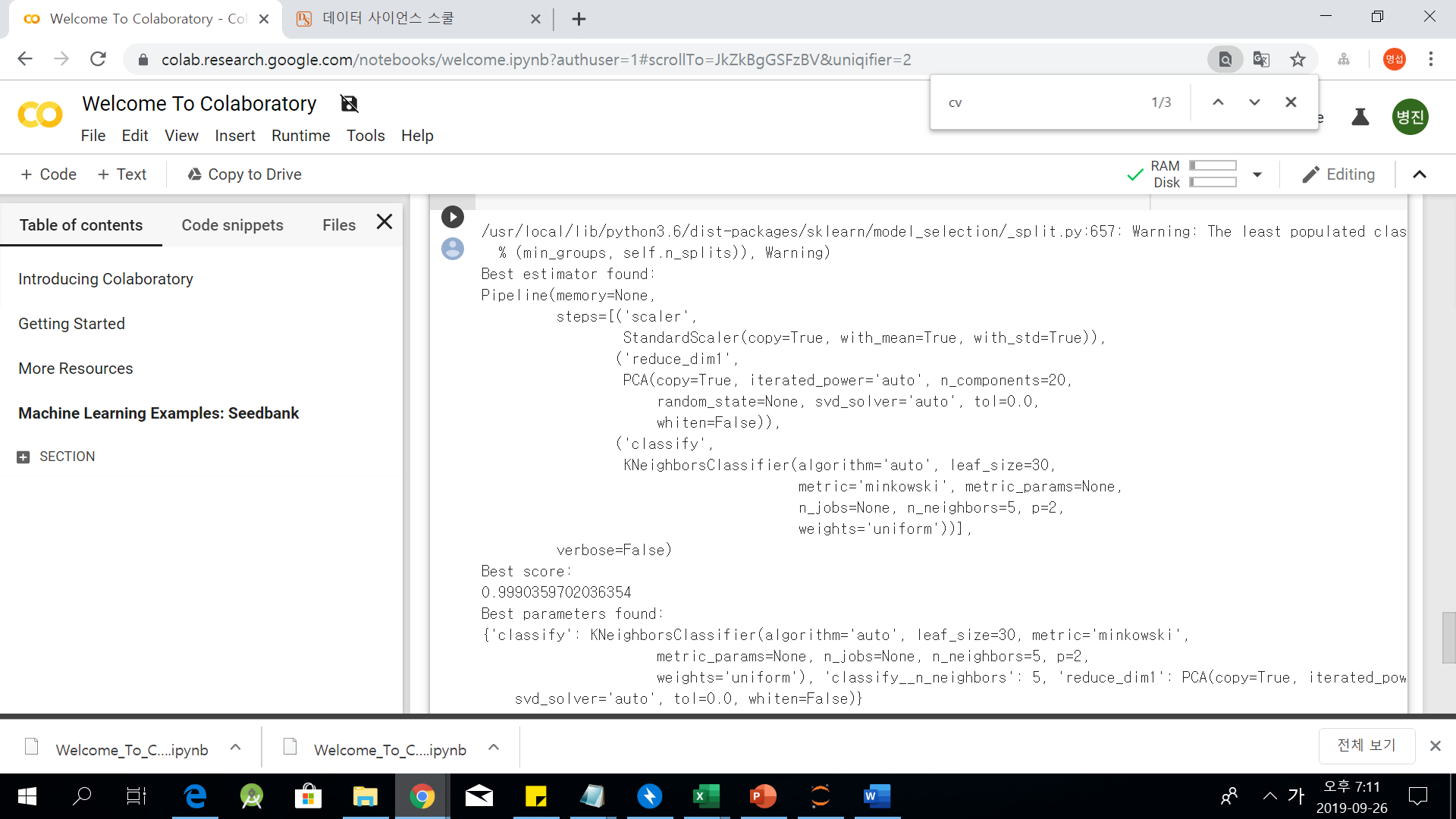


아마도 이런 value error가 뜨게 된 이유는 Tree-based feature selection에서 n\_estimators = 50으로 차원 축소를 해준후에 PCA에서 n\_component=20으로 또 한번의 차원축소를 해주는 것이 차원이 20보다 작아서 불가능 하기 때문인 것 같았습니다. 그래서 ppt에 있는 조건을 맞추는 것이 불가능 할 것이라고 생각이 들어 그냥 저는 PCA에서 n\_component=20인것따로 Tree-based feature selection n\_estimators = 50 따로 적용을 하여 해보았습니다.





그리고 SVM, Decision tree, K-NN에 대해서 비교하고 각 model별 2개이상의 parameter를 대상으로 fitting을 하라는 ppt의 조건이 있었는데 저는 decisiontree와 svm과 knn 3개를 전부다 각각 해서 각각의 방법안에서 가장 좋은 best parameter을 찾아 주는건지 아니면 decisiontree, svm, knn중에서도 제일 좋은 방법까지도 찾아주는 건지에 대해서 약간 헷깔렸는데 그냥 후자의 방법을 채택을 하여 어떤 방식이 제일 좋을지 까지 찾아 주기 위해 classify값도 각각 넣어주었습니다. 그리고 각각의 방법의 best parameter도 2개이상을 비교 해주기 위해 svm에서는 gamma값을 0.1과 0.01을 비교해서 best parameter값을 찾아주려고 하였고 decisiontree의 criterion값을 gini와 entropy로 비교하여 best parameter값을 찾아주려고 하였고 knn에서는 n\_neighbors값을 5와 10을 비교하여 best parameter값을 찾아 주기 위해서 넣어주었습니다.



결과는 이렇게 나왔습니다 대략 보면 SVM, Decision tree와 Knn중에 가장 좋은 방법은 knn이고 가장 최적의 n\_neighbors값은 10과 5중에 5라는 결과가 나왔습니다.