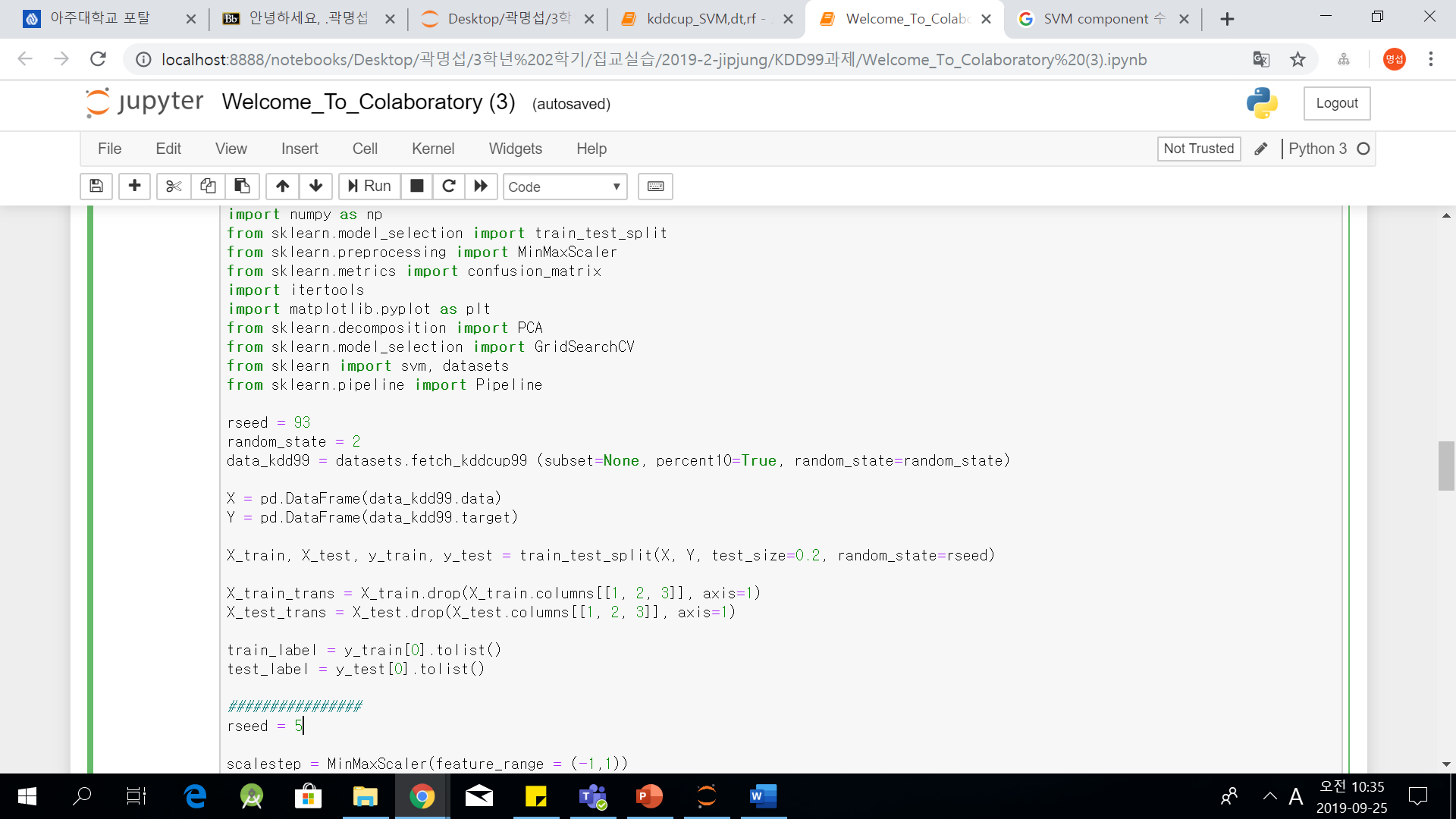
KDD99 data를 활용한 parameter tuning

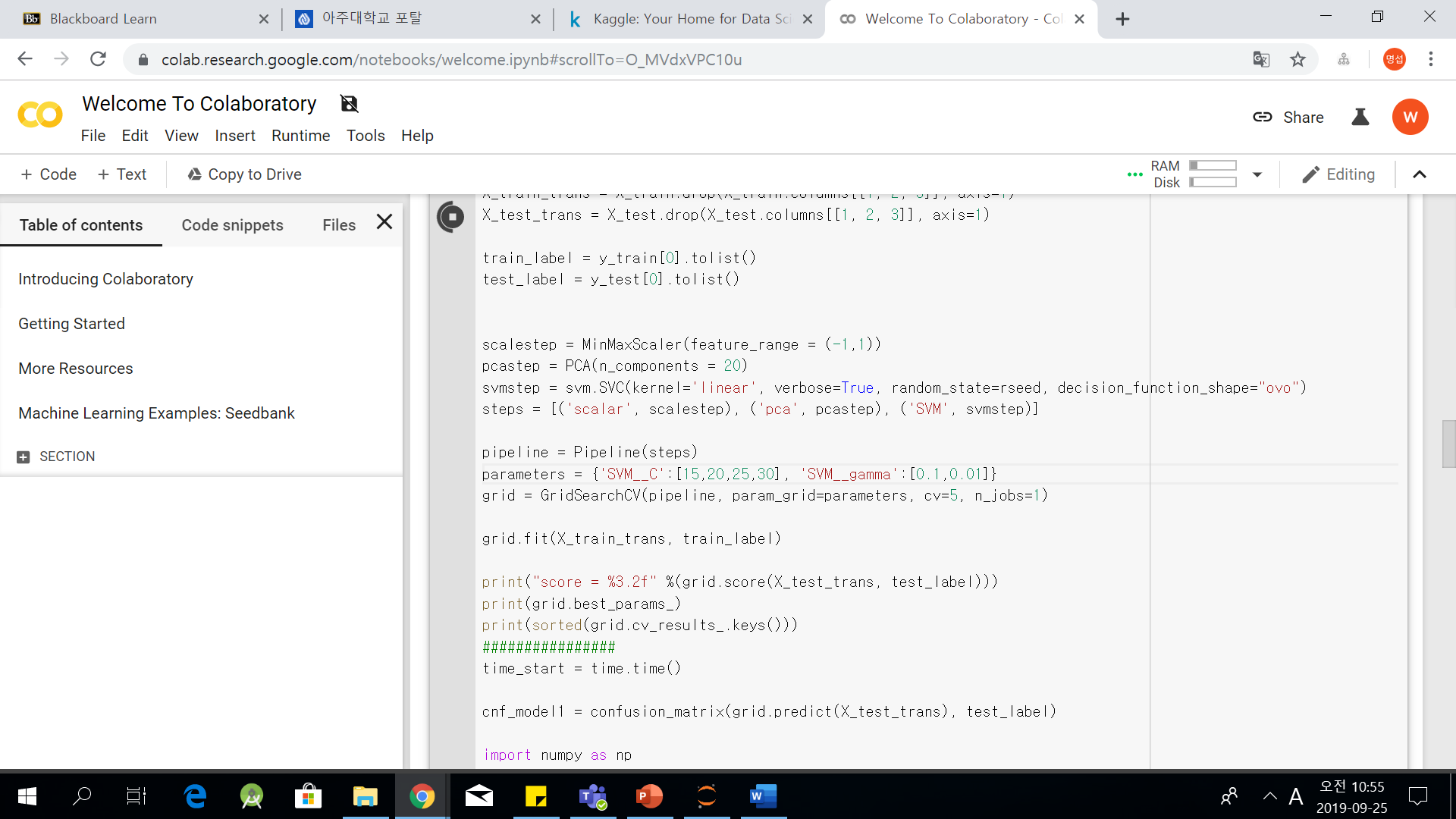
201520991

곽명섭

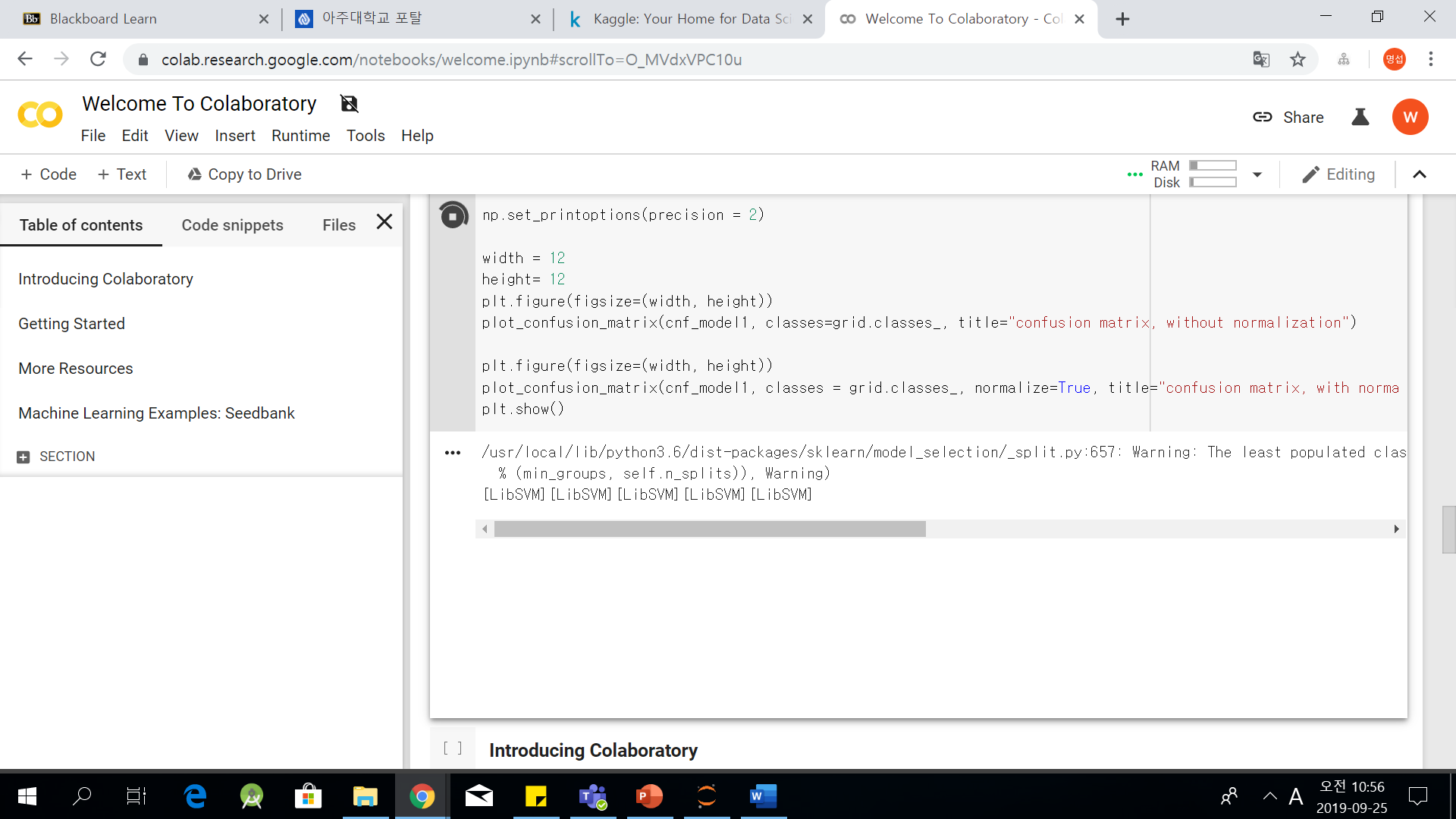
오늘 과제는 실습시간에 써보았던 KDD99 data를 활용하여 Test size = 0.2로 분할하여 Pipeline을 사용하여 SVM과 PCA를 사용하여 KDD99 데이터의 BEST SCORE과 Best parameter을 추출하고 Best parameter을 적용하여 confusion matrix를 띄우는 과제였습니다. 실습시간에 SVM과 PCA를 따로따로 KDD99 데이터에 적용해보는 것은 해보았기 때문에 코드를 합쳐서 과제의 조건에 맞게 변경하면 될 것이라는 생각이 들었고 해보았습니다.



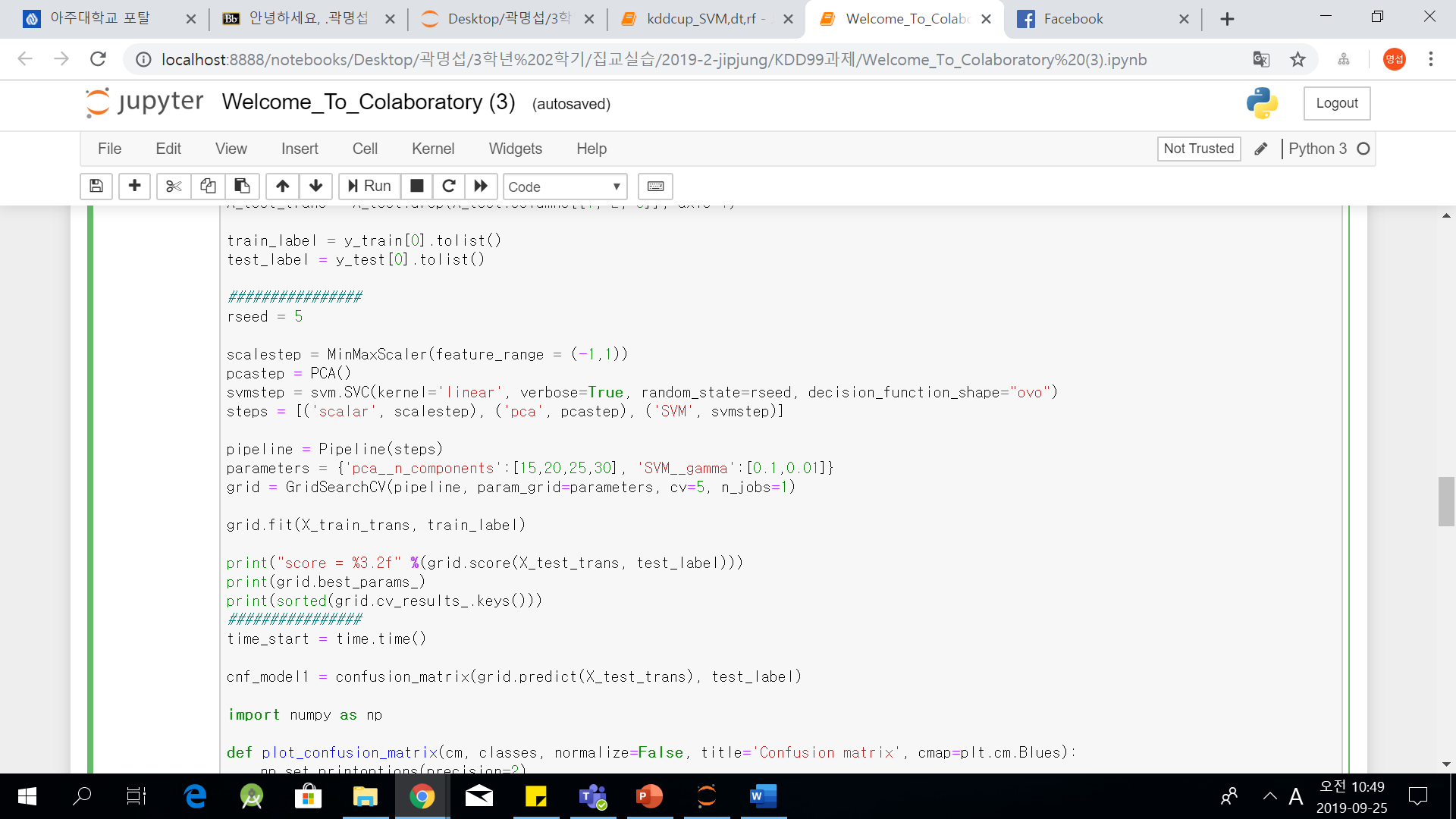
우선은 KDD99 데이터를 불러와서 ppt의 조건대로 test size를 0.2로 split을 한후에 KDD99에 있는 1,2,3열의 쓸모 없는 자료들을 drop해주는 과정입니다.



이렇게 data를 가공해주고 나면 이제 과제의 조건에 맞게 SVM과 PCA를 적용해주어야 하는데 위의 코드가 그 과정을 해주는 코드이다. 보면 pipeline을 통해서 svm과 pca를 둘 다 적용을 해주며 과제의 조건을 보면 SVM을 사용할 때 parameter에 # of Component: 15,20,25,30와 Range of ‘SVM\_gamma’: 0.1,0.01 이라는 조건이 붙어있어서 SVM의 component수를 설정해주는 parameter가 SVM\_\_C라고 생각을 하여 ‘SVM\_\_C’ = [15,20,25,30] 이런 식으로 코드를 놓고 colab의 gpu를 이용하여 돌려 보았지만

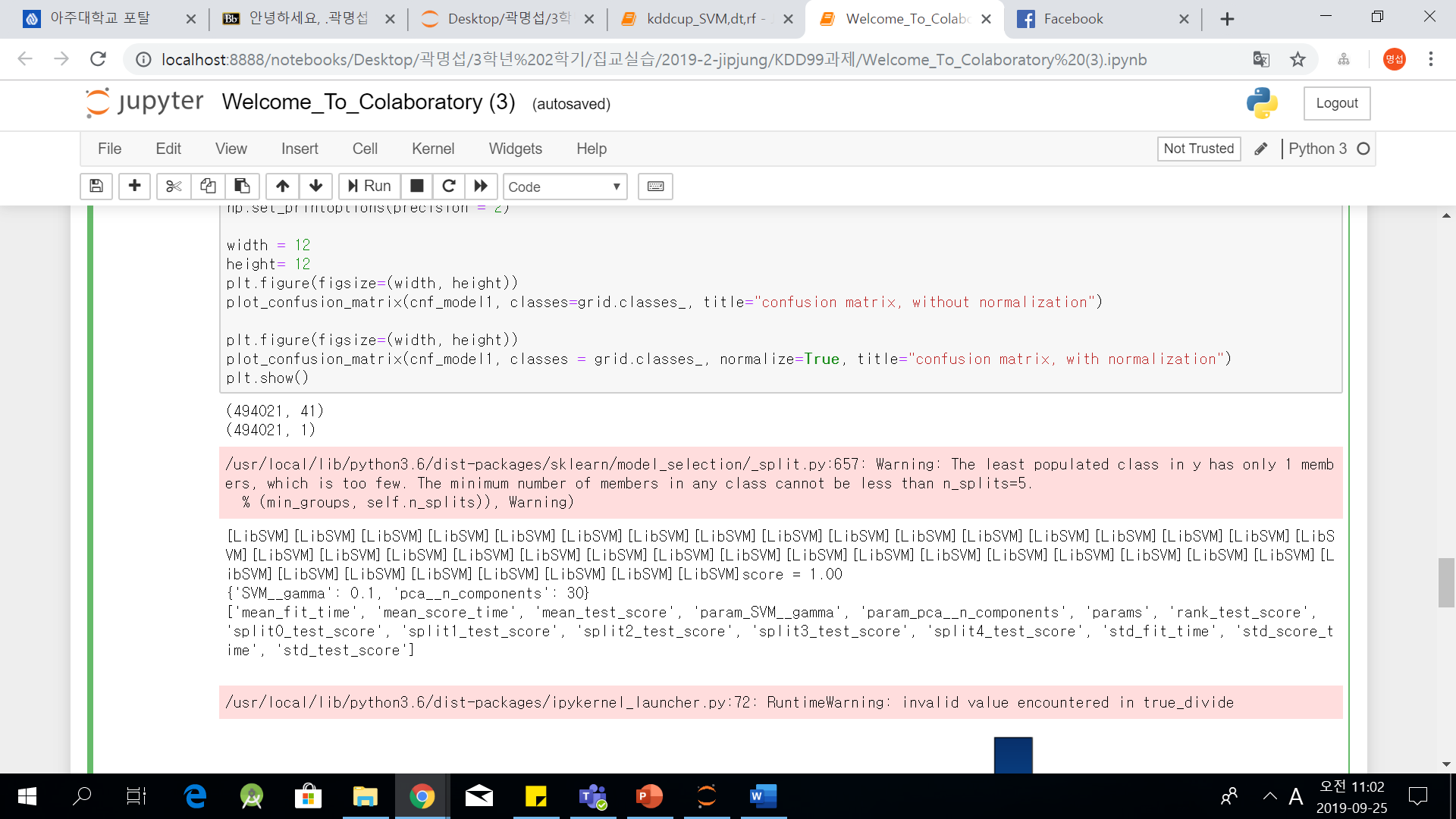


이런 식으로 LibSVM이 40개가 떠야 결과값이 뜨는데도 불구하고 몇시간을 돌려도 5개에서 멈춰 있을 뿐이어서 무엇이 문제인지 모르겠어서 SVM의 component라는게 도대체 무엇인지에 대해서 검색을 해서 찾아보았지만 도저히 무엇인지를 알 수가 없었고 SVM\_\_C라는 parameter는 아무리 구글링을 해보아도 componet값이 아니라 오버피팅과 언더피팅을 방지하기위한 error값이라는 말 밖에 나오지 않았고 default값이 1이라는 사실을 알게 되었습니다. 그런데 default값이 1인 것을 15,20,25,30으로 갑자기 확 올려주는 것은 조금 이상하다는 생각도 들고 component의 정의 랑은 뭔가 다른 느낌이 들었습니다. 반면에 pca에는 component의 수라는 것이 확실히 존재하는 것 같아 저는 ppt에 오타가 났다고 생각하여 pca의 n\_component값을 15,20,25,30으로 설정을 하였습니다. (혹시라도 제 추측이 아니라면 죄송합니다..)



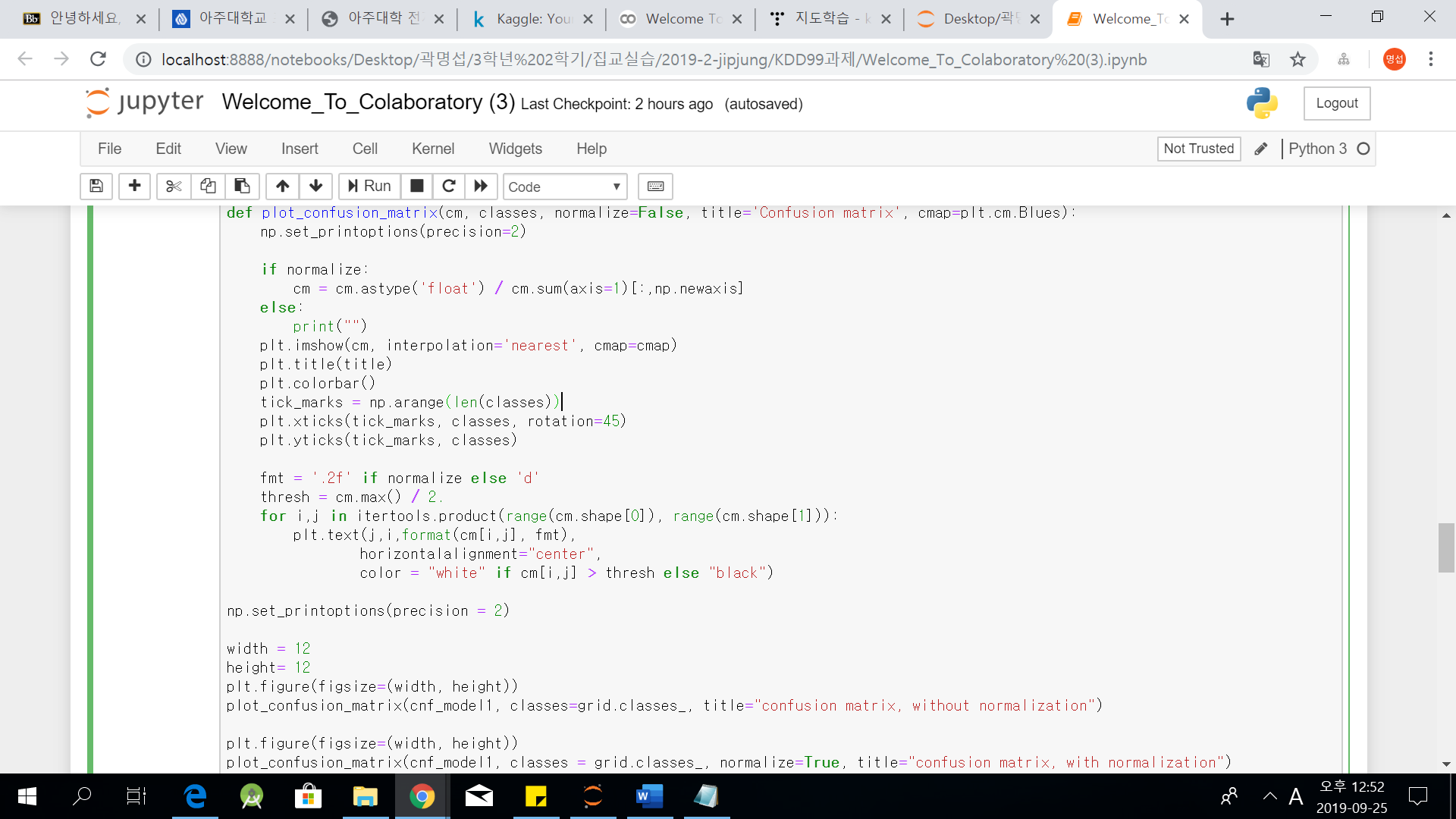
이렇게 설정을 해주니 결과값도 좀더 빠르게 나오기도하고 무언가 맞아 떨어지는 느낌이 들었습니다.

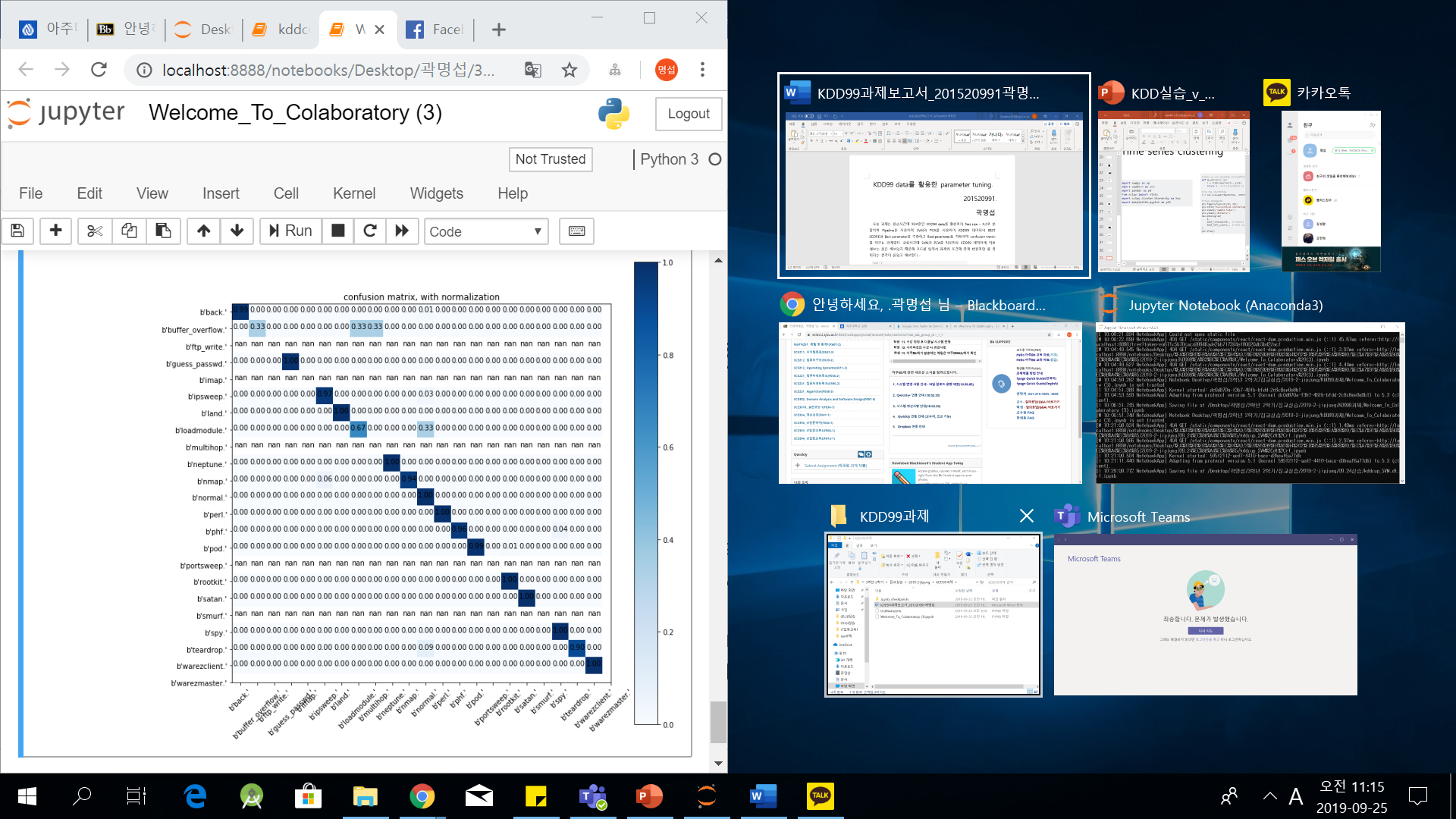
그리고 LibSVM이 40개가 다 돌고나니 Best parameter와 Best score가 추출이 되었고



Best score과 Best parameter는 이러하였습니다. Score도 최대치인 1.00이 나왔고 parameter도 잘 선별 된것같았습니다.

그리고 이제 matrix를 출력을 해주어야 하는데 matrix를 출력해주기 위해서 또 실습때 사용했던 출력 코드를 이용을 하였습니다 그 코드는 아래와 같습니다.

  
그리고 그 결과를 적용한 matrix plot은 밑의 그림과 같았습니다.



이번과제를 하며 component라는게 무엇인지 잘 모르겠어서 헤맸던 시간이 꽤나 길었는데 구글링을 하며 확실하지 않았던 것 들을 정리하다 보니 맞아떨어져 가는 느낌이 있어서 과제를 할 수 있게 되었던 것 같습니다. 제 생각이 맞는지 틀린지는 잘 모르겠지만 우선 제가 생각 하는 데로 해보았습니다.