Major project

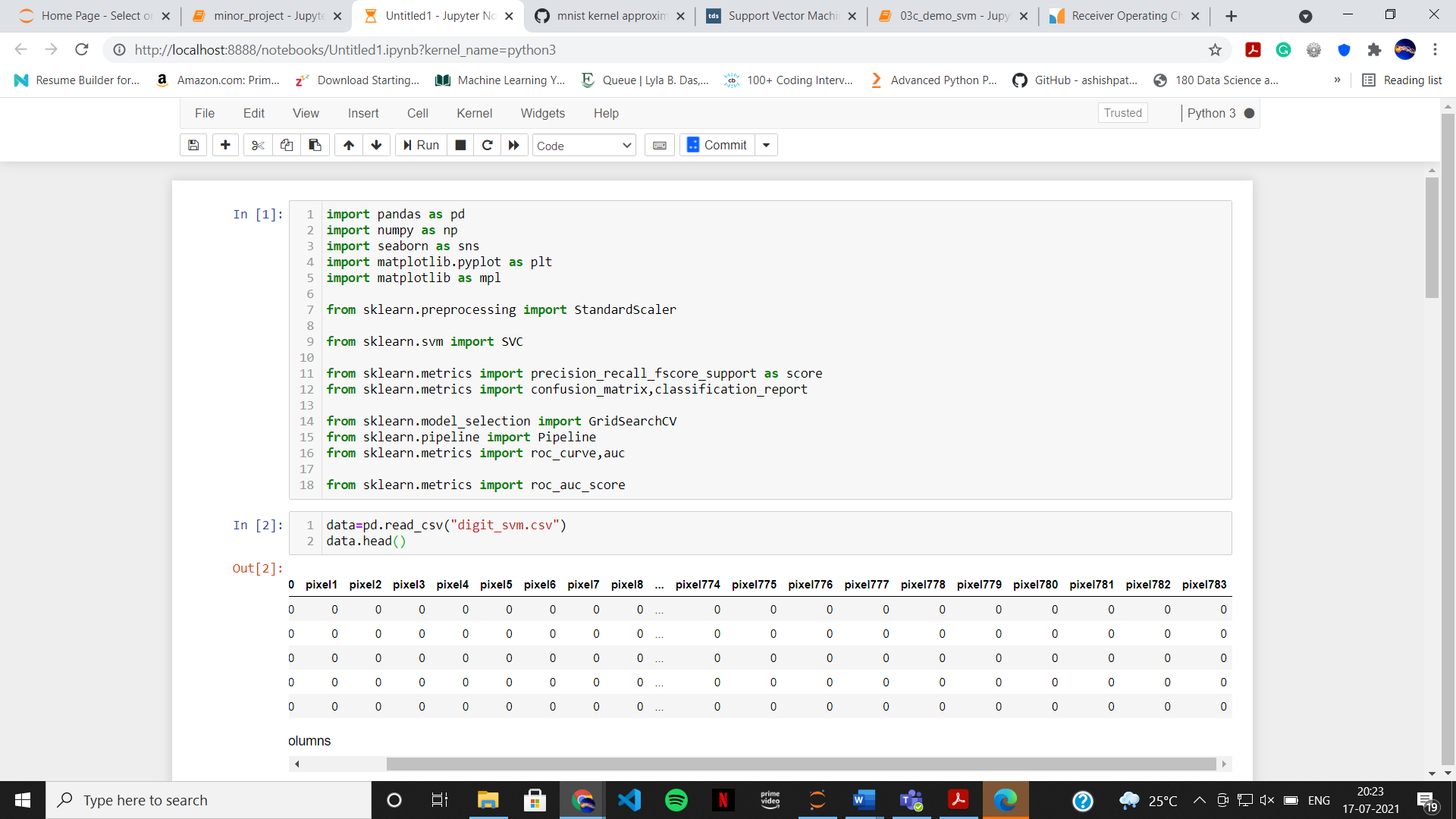
Problem statement:

Design a project from the MNIST dataset to identify digit classification using the SVM algorithm

Objective:

Develop a best-fit Support Vector Machine model to classify the Number from 0 to 9.

Packages and IDE:



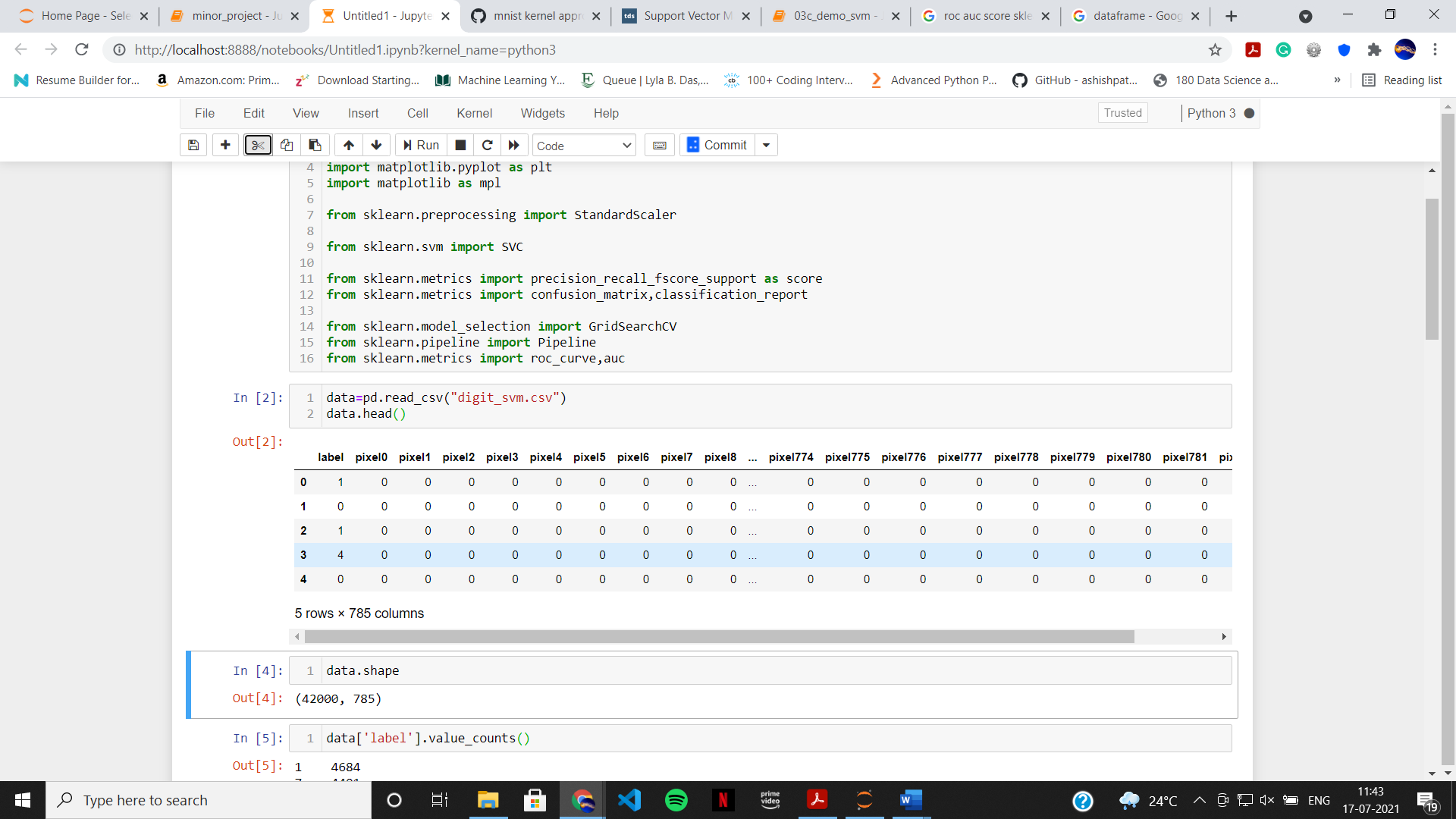
I have used various packages like Pandas, NumPy, Matplotlib, Seaborn for this project.

And the IDE used for this project Jupiter Notebook.

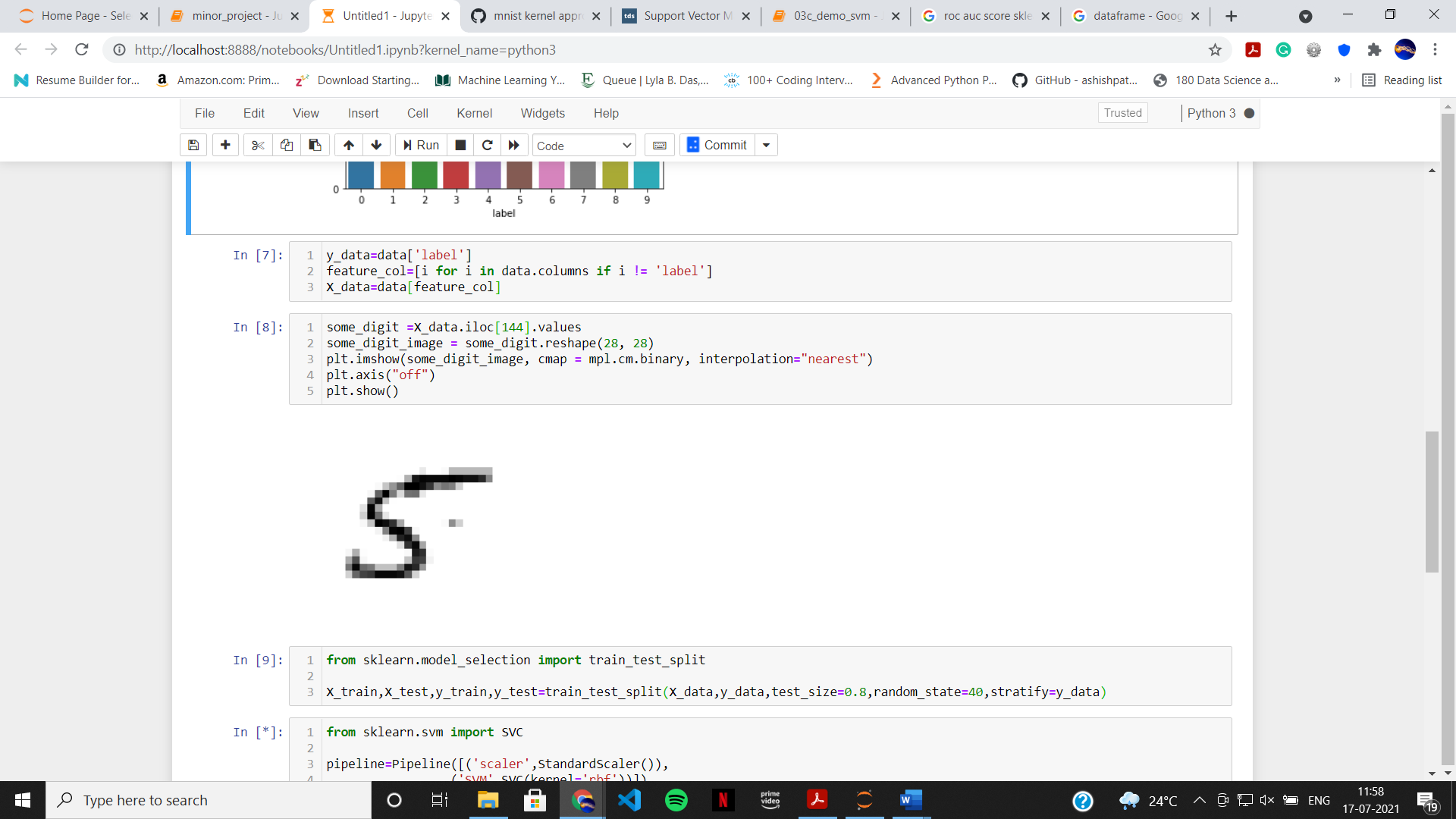
The data is based on greyscale image of hand written digits from 0 to 9, with 28-pixel height and 28-pixel width. Our goal is to classify numbers from 0 to 9 using support Vector Machine (SVM).

The best fit-model chosen on the basis of testing score and auc-roc score.

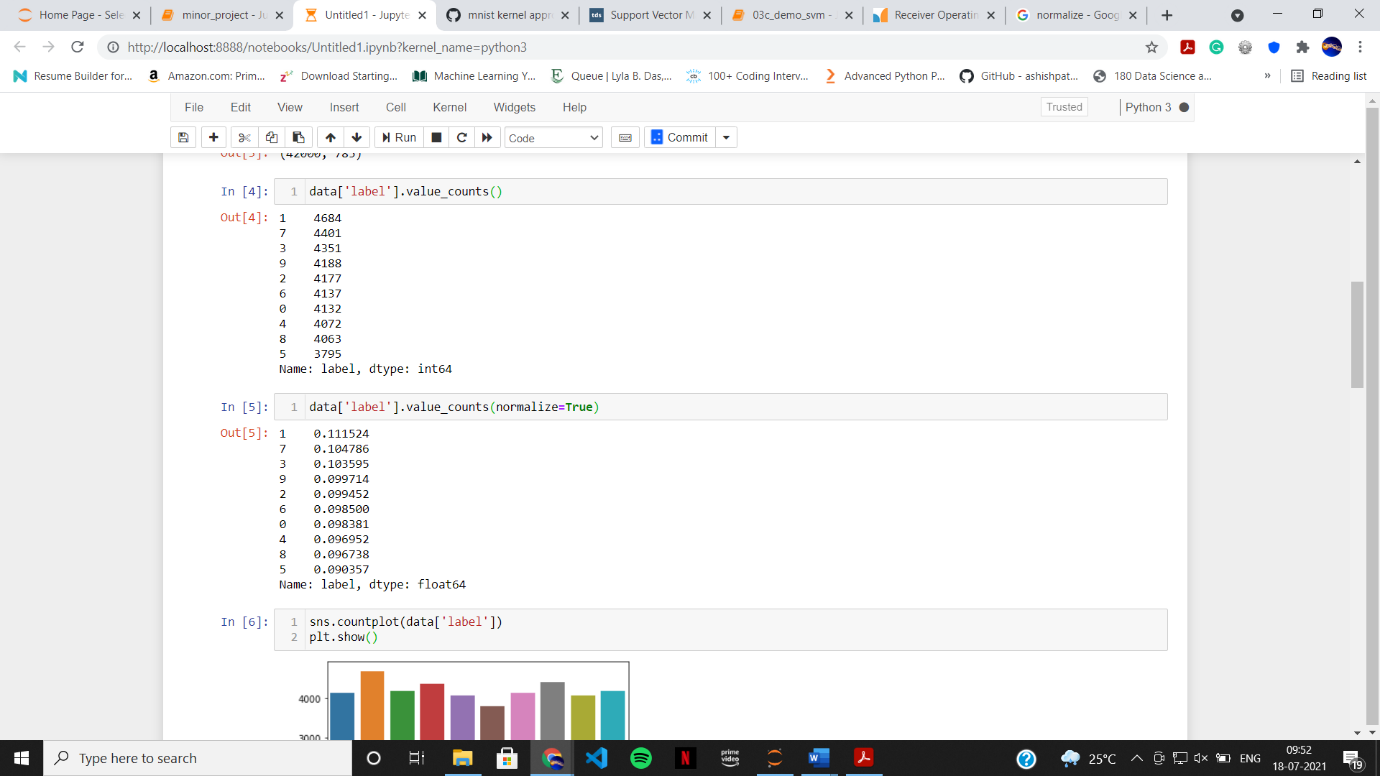
Analysis of data:

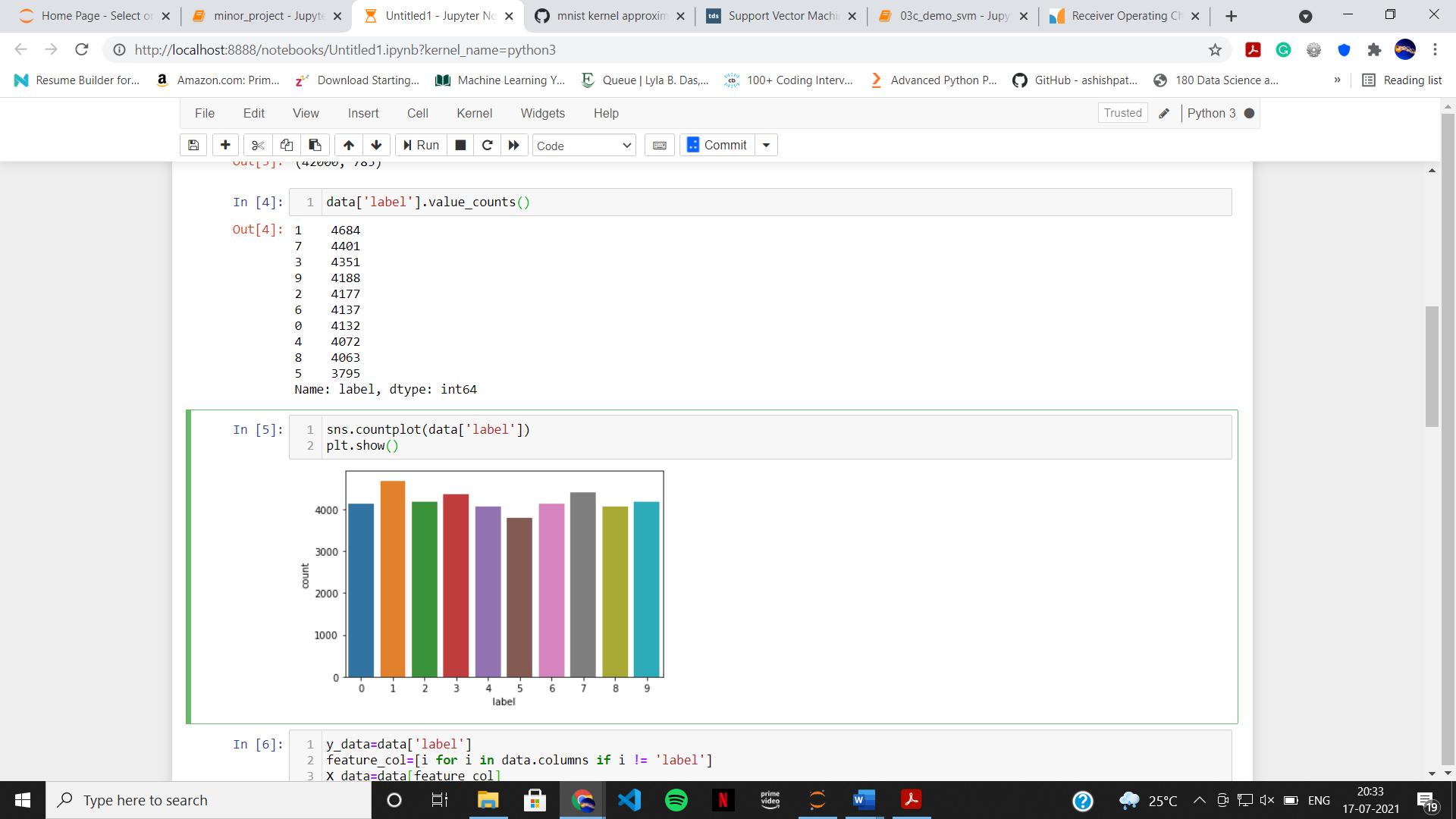
Sample of the data 

The shape of the data is (42000,785) that means every image has 784 features i.e. (28\*28). Sample of the image

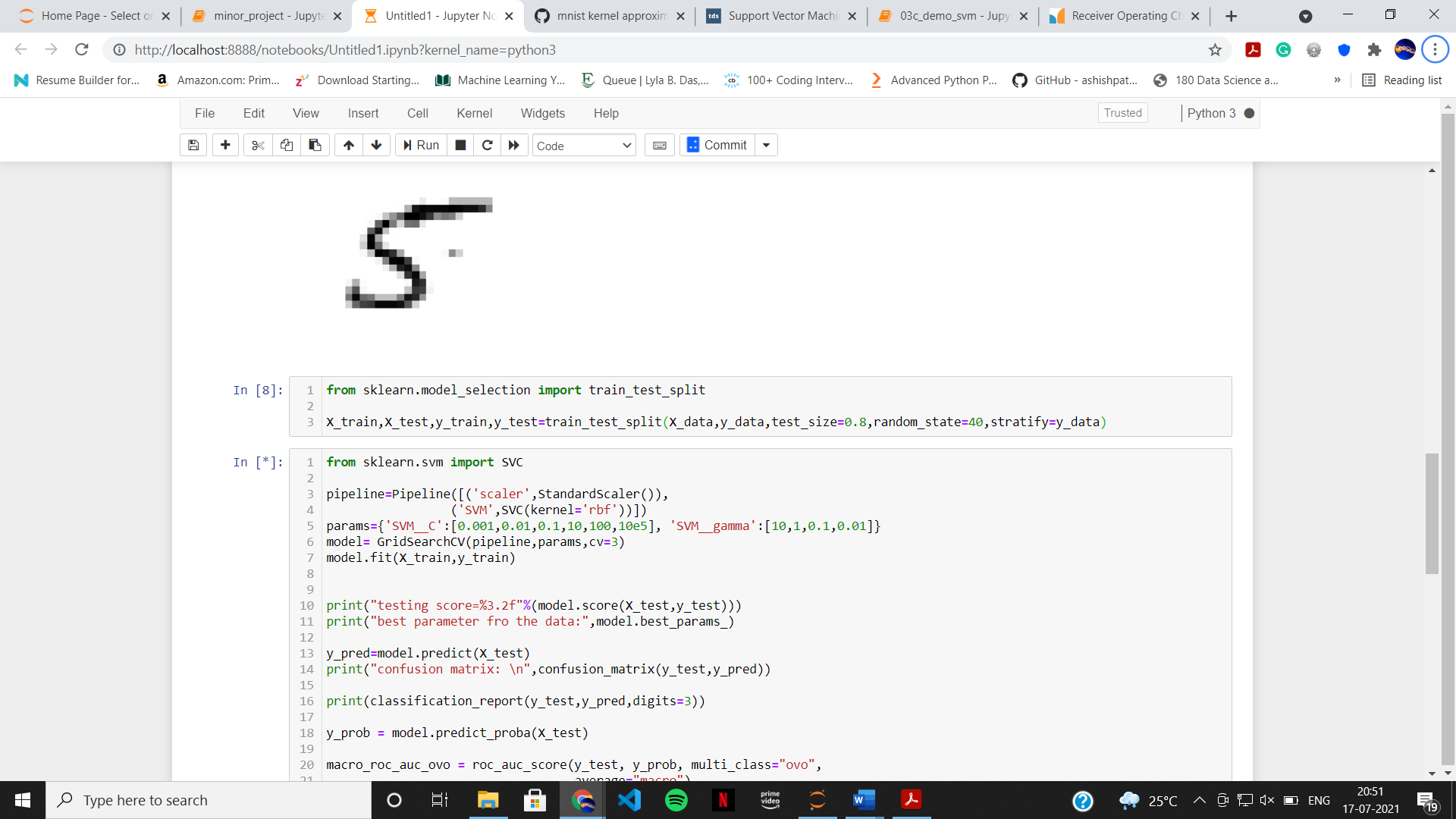


Before splitting the dataset into training set and test set, we need to check if dataset is biased to certain number.





It seems that dataset is little biased towards number 1, less to number 5.



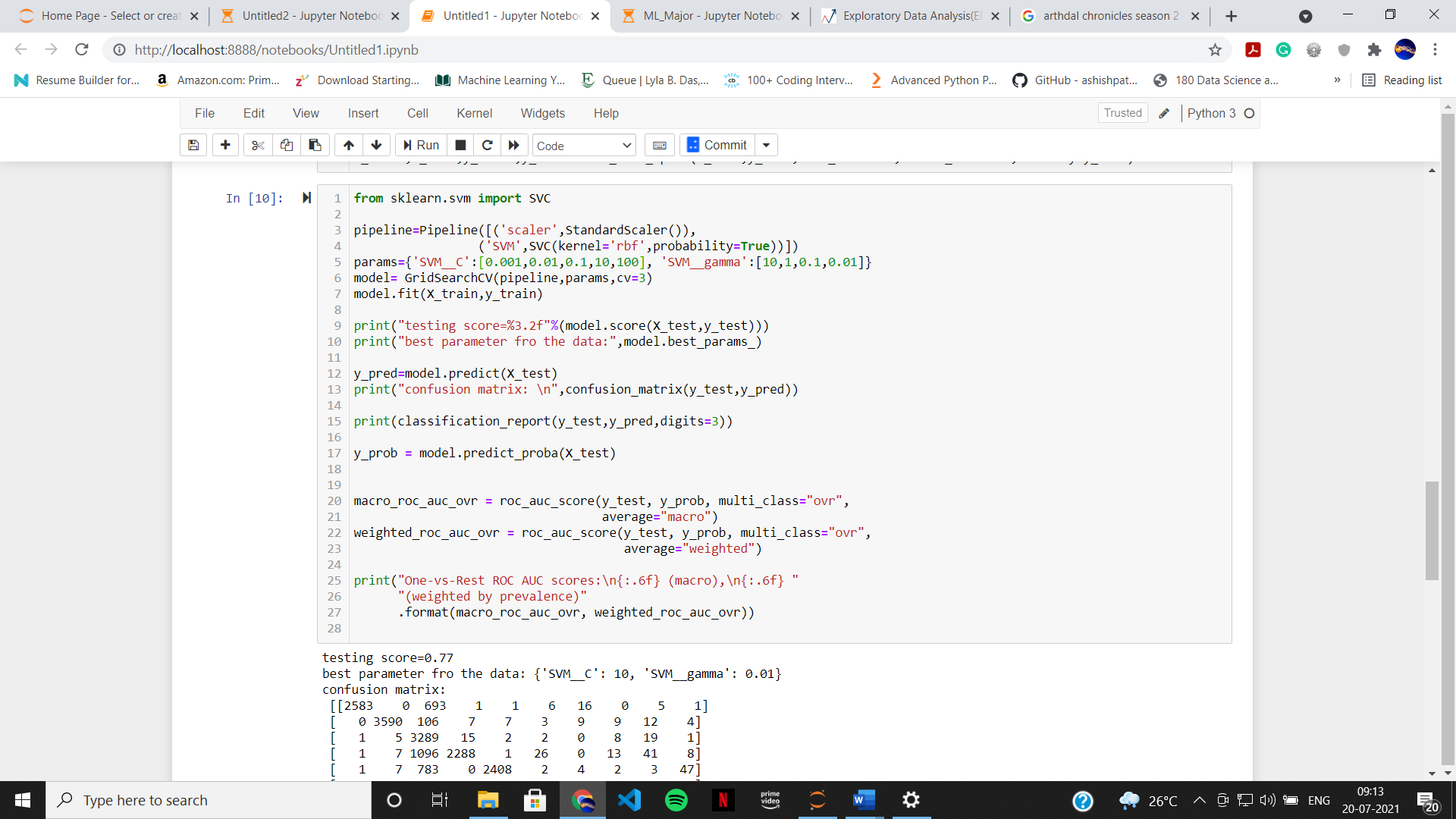
I used train\_test\_split function from sklearn library to split dataset into train\_set and test\_set. And used for for preserving the distribution.

Modelling machine learning algorithms

Pipeline and GridSearchCV from sklearn library to fit and tune the data.

Model 1: Using RBF kernel

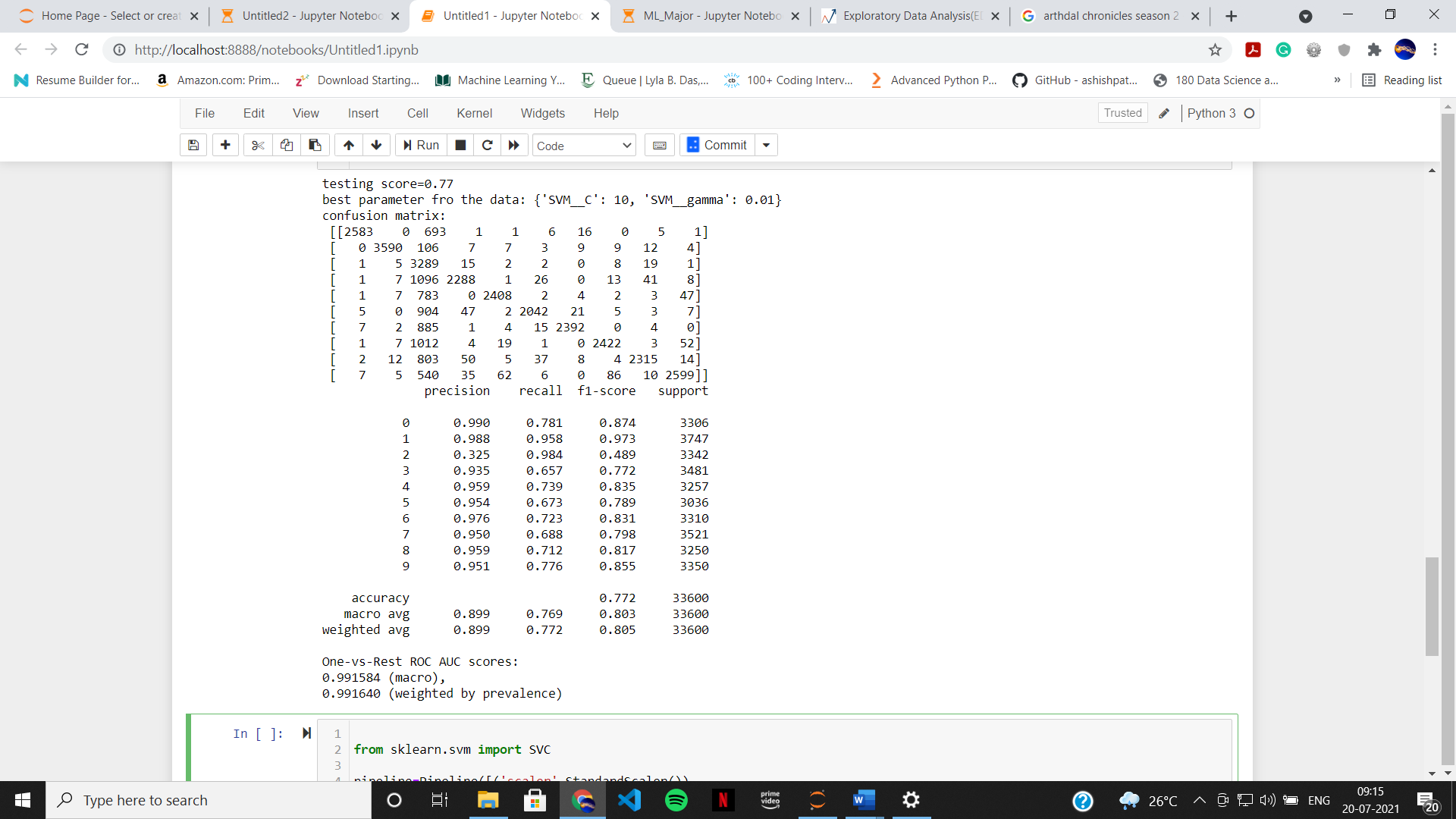
Code:



params parameter used for tuning the model with 5 different C values ([0.001,0.01,0.1,10,100]) and 4 gamma values ([10,1,0.1,0.01]).

Model trained with Training dataset

model.best\_params\_ gives best C and gamma values for the model



Best parameters for the data are {'SVM\_\_C': 10, 'SVM\_\_gamma': 0.01}

Accuracy of the model is 77.2%

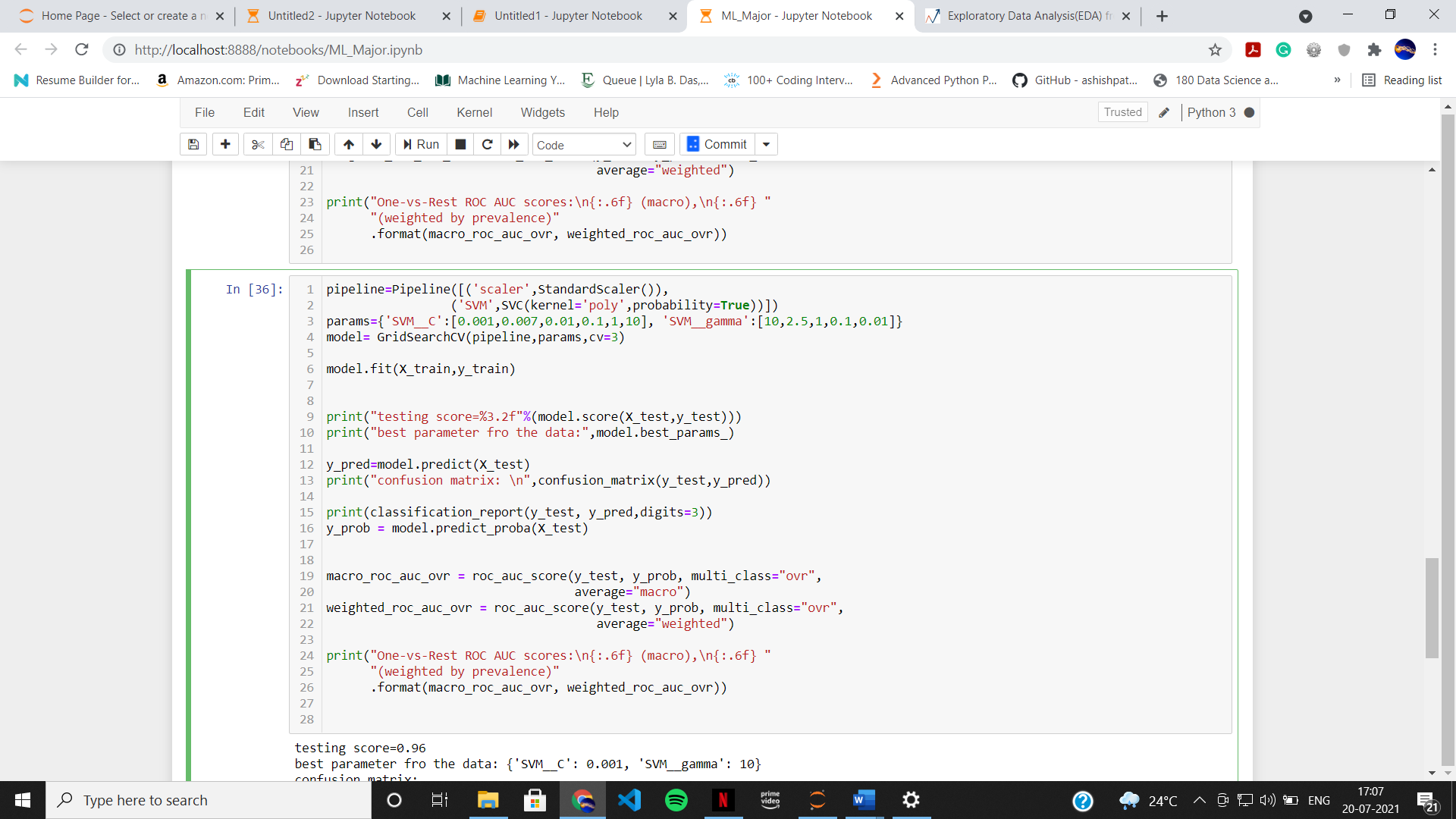
One-vs-Rest ROC AUC scores:

0.991584 (macro),

0.991640 (weighted by prevalence)

Model 2: Using Polynomial kernel

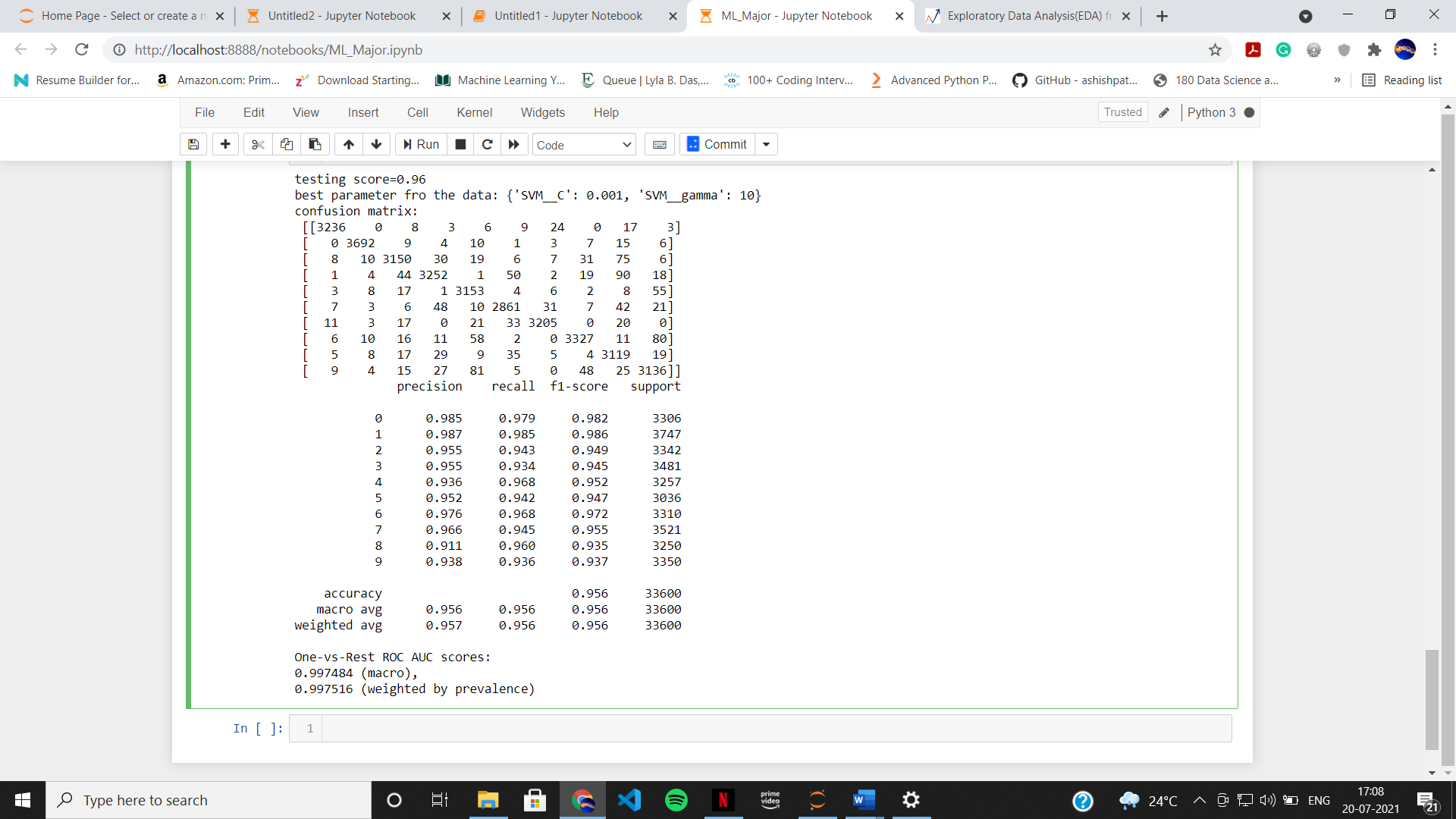
Code:



params parameter used for tuning the model with 6 different C values ([0.001,0.01,0.1,1,10,100]) and 4 gamma values ([10,1,0.1,0.01]).

Model trained with Training dataset

model.best\_params\_ gives best C and gamma values for the model



Best parameters for the data are 'SVM\_\_C': 0.001, 'SVM\_\_gamma': 10.

Accuracy of the model is 96%

One-vs-Rest ROC AUC scores:

0.997434 (macro),

0.997469 (weighted by prevalence)

Conclusion:

Based on Area under the curve (AUC) and accuracy score SVM with polynomial kernel gives best fit model with the parameter C=0.001, gamma=10.