NAME: Venkata Sai Siva Pruthvi. Ustepalle S-ID: 800957126

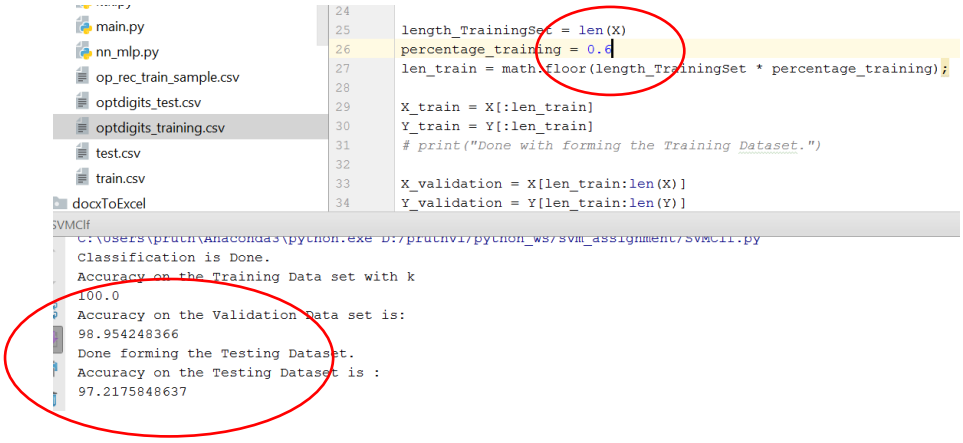
(Worked together with Chaitanya Sri Krishna Lolla)

Digit Recognition dataset

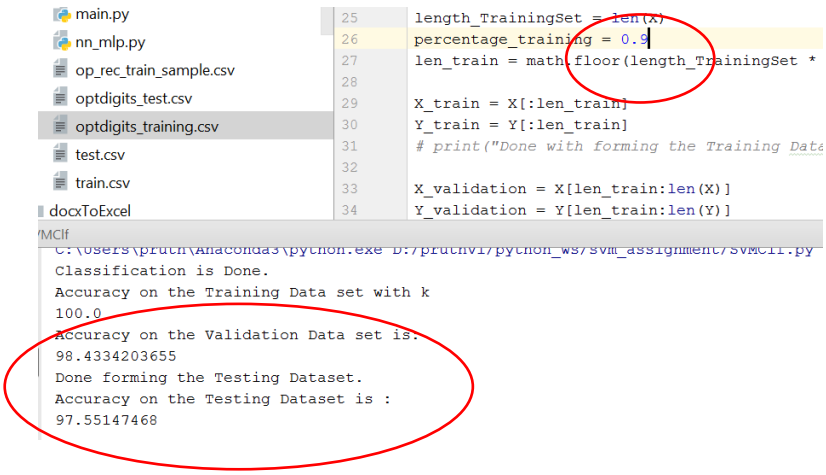
- RESULTS:

Splitting the given data into 2 parts to perform cross validation (as a step for pruning the data).

Training on anything below 70% for cross validation which is resulting in a 98% accuracy on training set and 97.21% accuracy on the validation data.



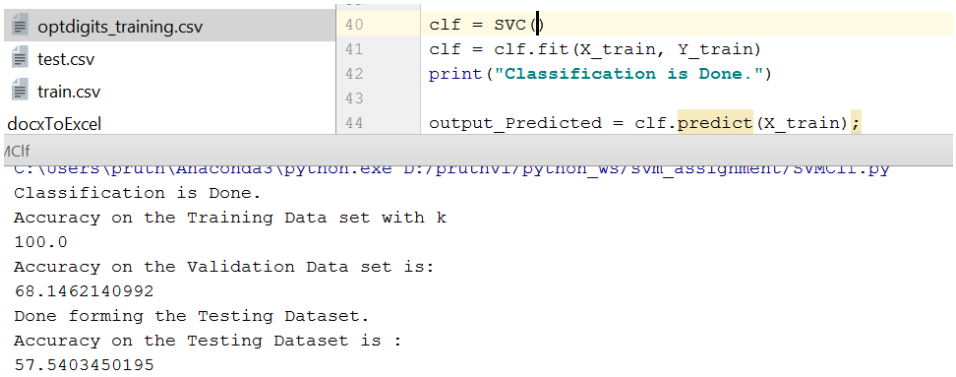
But on cross validating it with a 90-10 split of training data it is resulting in a better fit as it can be observed in the accuracy on testing data:



Above accuracy is a resultant of the classifier with all the default values provided by the sklearn classifier,

With kernel=”poly”.

With default SVM parameters given by the classifier this model it gives a poor classification.



Amazon review dataset

- Result:

Using nltk library services to lemmatize, tokenize the reviews along with removing the punctuation from the review texts as a part of preprocessing steps as in done in the previous assignment.

Pre-processing the text and making it into two sets of positive and negative reviews.

Picking top 5000 of the most commonly occurring words and assigning them numerical values to pass as parameters for the decision tree classifier method.

Used TF-IDF to build a matrix of numerical representation of the words in each sentence.

And the same procedure is repeated on the test dataset and the accuracy is measured.

- Used a normal SVM () method and checked for the accuracy of model :

```
  (0, 2158)     1
  (0, 2896)     1
  (0, 2764)     1
  (0, 1749)     1
  (0, 4414)     1
  (0, 2264)     1
  (0, 4463)     1
  (0, 254)      1
  (0, 2271)     1
  (0, 4385)     1
Accuracy :  0.764279399974
27948
[('the', 152974), ('it', 99372), ('i', 924
Accuracy of test :  76.2240447651 %
```

- Now I ‘ve added the kernel=”poly” as I’ve observed that this is giving a same fit:

```
  (0, 2765)     1
  (0, 1748)     1
  (0, 4415)     1
  (0, 2265)     1
  (0, 4464)     1
  (0, 254)      1
  (0, 2272)     1
  (0, 4386)     1
ccuracy :  0.764279399974
7948
('the', 152974), ('it', 99372), ('i'
ccuracy of test :  76.2240447651 %
```