**SD7502 Intelligent Systems Development**

**Project 1**

**Report**

**­­­Qiang Zhang**

**2173138**

**26 August 2021**

Contents

[1 Overview 3](#_Toc81743021)

[2 Multi-Classification of Digits with the EMNIST Dataset 4](#_Toc81743022)

[2.1 Load Digits Data 4](#_Toc81743023)

[2.2 Explore Digits Data 4](#_Toc81743024)

[2.3 Visualize Digits Data 6](#_Toc81743025)

[2.4 Hyperparameter Tuning 7](#_Toc81743026)

[2.5 Running Multiple Models to Find the Top 3 Algorithm 9](#_Toc81743027)

[2.6 Running Multiple Models to Find the Top 3 Algorithm 10](#_Toc81743028)

[2.6.1 SVC Classification Algorithm 10](#_Toc81743029)

[2.6.2 kNN Classification Algorithm 11](#_Toc81743030)

[2.6.3 SGD Classification Algorithm 11](#_Toc81743031)

[3 Multi-Classification of Letters with the EMNIST Dataset 12](#_Toc81743032)

[3.1 Load Letters Data 12](#_Toc81743033)

[3.2 Explore Letters Data 12](#_Toc81743034)

[3.3 Visualize Letters Data 14](#_Toc81743035)

[3.4 Hyperparameter Tuning 15](#_Toc81743036)

[3.5 Running Multiple Models to Find the Top 3 Algorithm 17](#_Toc81743037)

[3.6 Performance Testing the Top 3 Algorithm 18](#_Toc81743038)

[3.6.1 LinearSVC Classification Algorithm 19](#_Toc81743039)

[3.6.2 kNN Classification Algorithm 20](#_Toc81743040)

[3.6.3 SGD Classification Algorithm 20](#_Toc81743041)

[4 Conclusion 21](#_Toc81743042)

[5 References 22](#_Toc81743043)

# Overview

Based on the requirements, classify digits and letters for digits and letters images have been downloaded from EMNIST dataset (The EMNIST Dataset, 2019) and 1,000 samples have been randomly selected in this project. The classification algorithm is fine-tuned to achieve the best performance, and with six different algorithms is evaluated. Finally, three optimal algorithms are selected to classify all samples and record the results.

The EMNIST dataset is a standard benchmark for learning, classification, and computer vision system. It is a database of handwritten digits and letter images and labels. The image can be converted from the data set to 28 by 28 pixels picture. For handwritten digits dataset, there are 280,000 characters within 10 balanced classes (From 0 to 9). For handwritten letters dataset, there are 145,600 characters within 26 balanced classes (From a to z).

Emnist-digits.mat and emnist-letters.mat are also been downloaded from EMNIST official website. These are MATLAB format file dataset. I import the Scipy module through python language to load digits and letters data set.

In this project, there are six classification algorithms being evaluated. They are k-nearest neighbours’ algorithm, Support Vector Clustering algorithm, GaussianNB algorithm, LinearSVC algorithm, SGDClassifier algorithm and Decision Tree algorithm. Three optimal algorithms have been selected base on the mean accuracy and standard deviation which best match the digits and letters data set.

The test results have been analysed to obtain the optimal algorithm with the least time, high mean accuracy and low standard deviation for classifying EMNIST's digits and letters data set.

# Multi-Classification of Digits with the EMNIST Dataset

## Load Digits Data

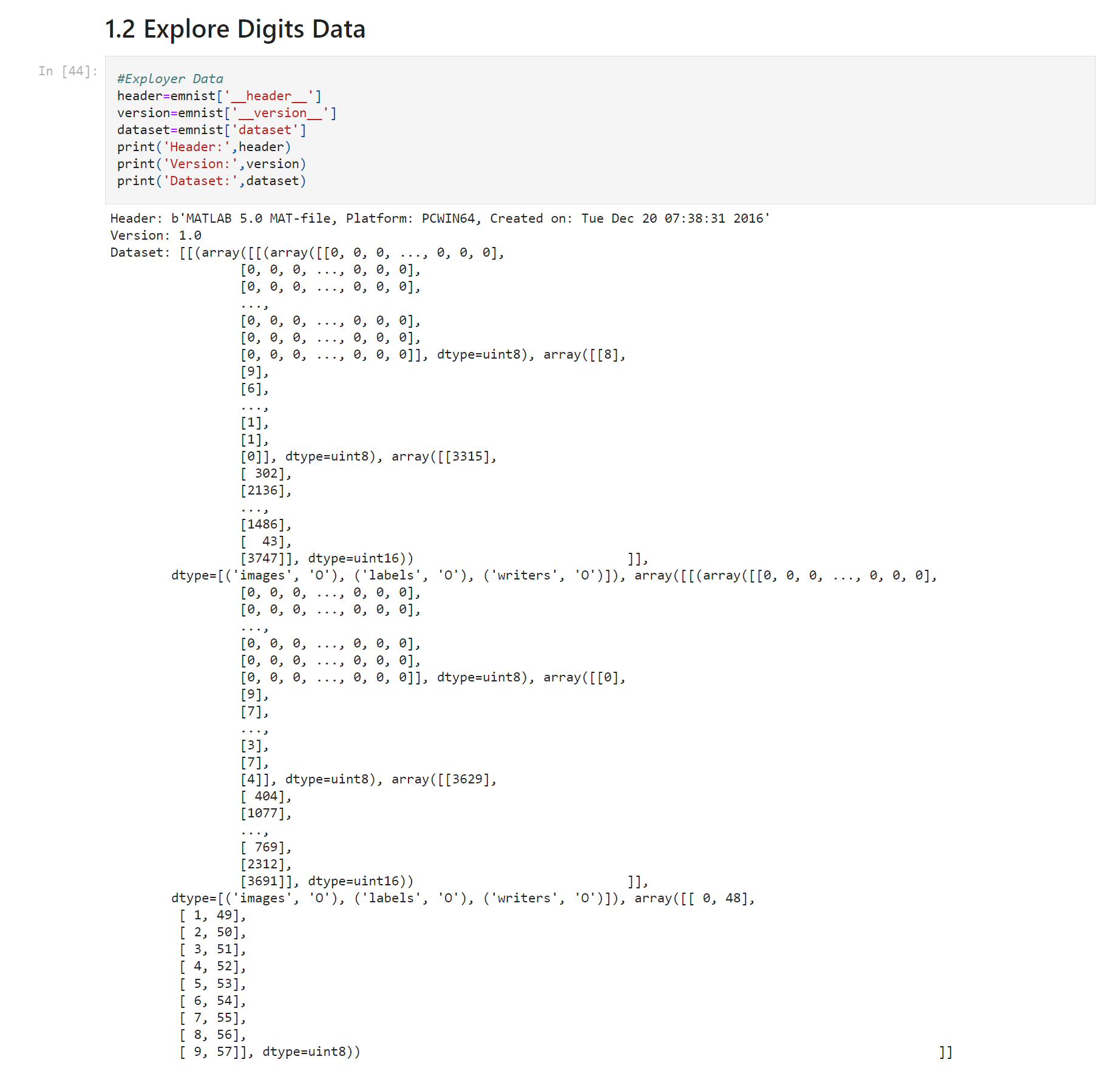
Load digits data from emnist-digits.mat file using SciPy open-source Python library.



Figure

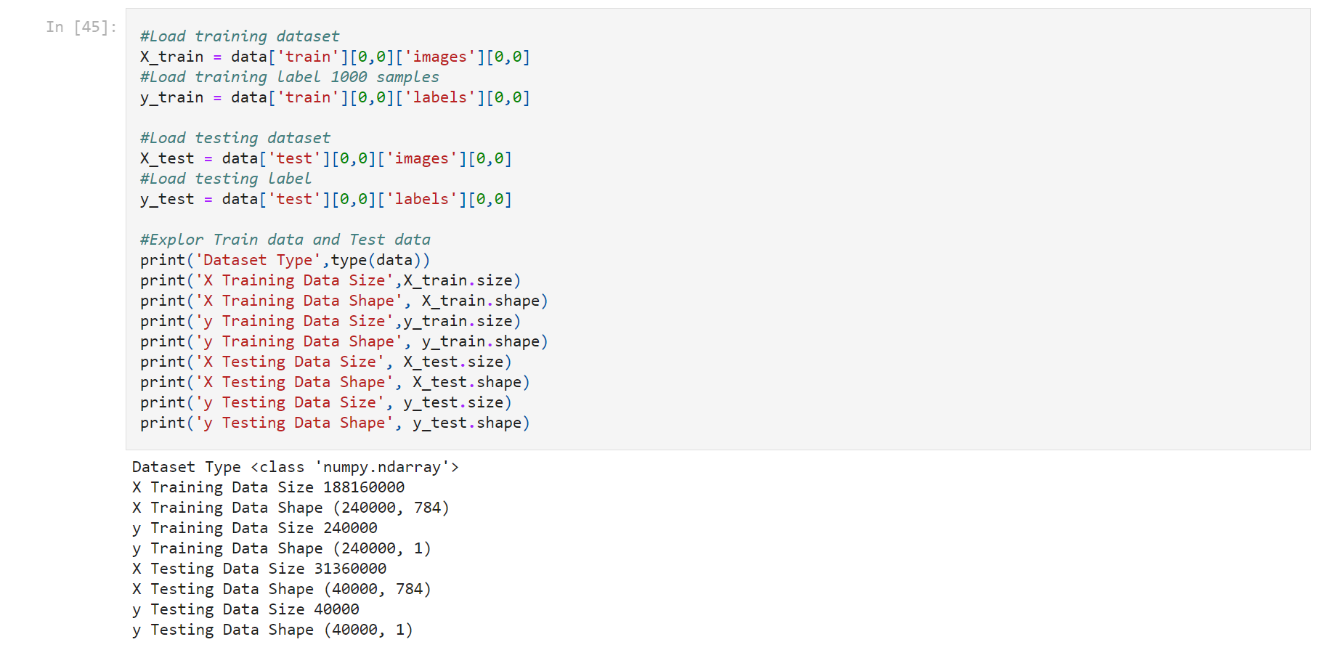
## Explore Digits Data

Display Header, Version and dataset of emnist-digits.mat MATLAB file.



Figure

According to figure 2, the emnist digits dataset splits into training dataset and testing dataset. Load training dataset and testing dataset as well as explore them.



Figure

X is image data and y is label data. The explore results is shown in the finger above.

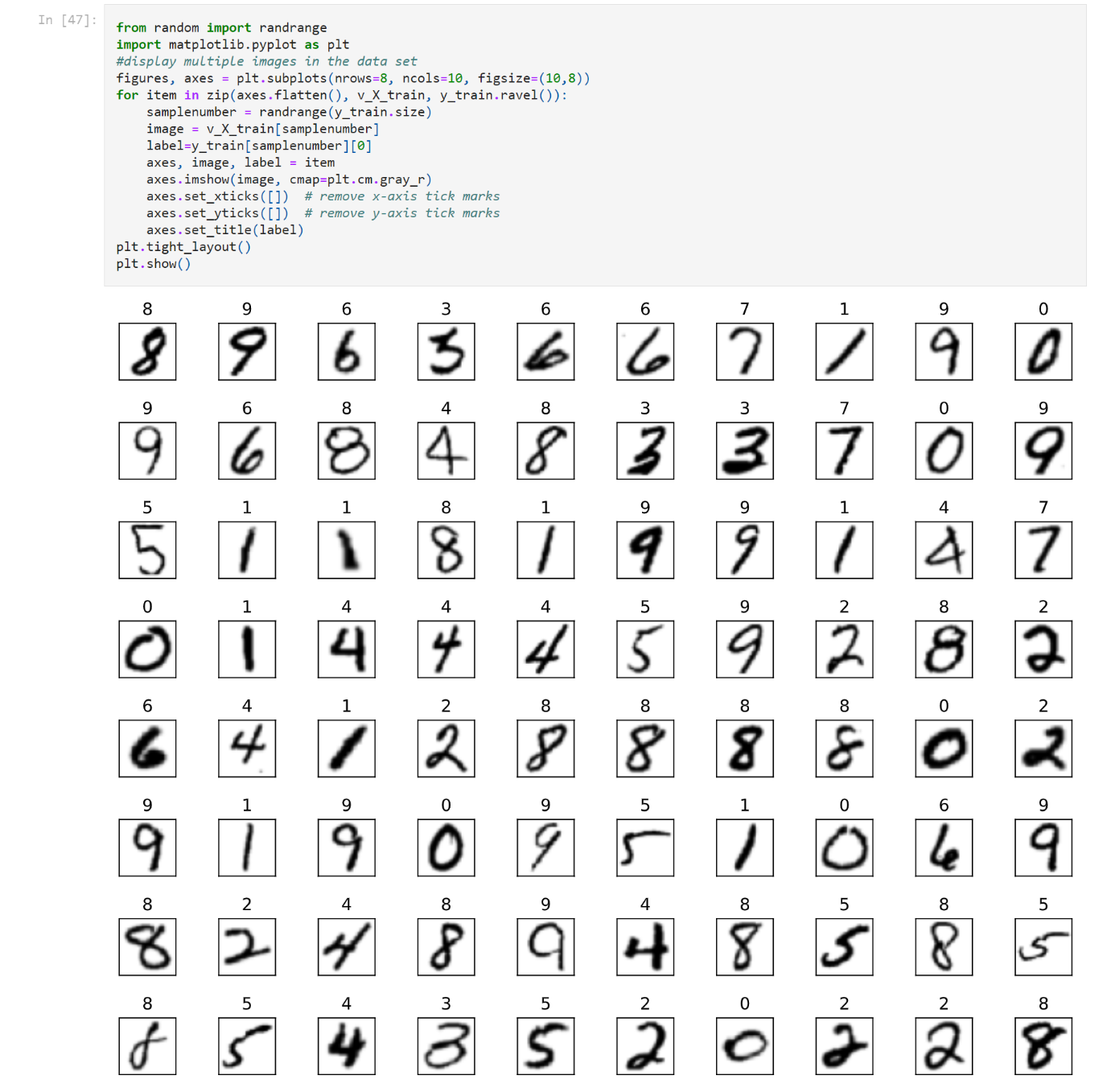
## Visualize Digits Data

Verify if the digits data loaded correctly.



Figure

Reshape image data from MATLAB to pyplot format and display image and label data shape in the finger 4 above.

By importing the matplotib.pylot library to randomly display image and label from digits dataset.

Figure

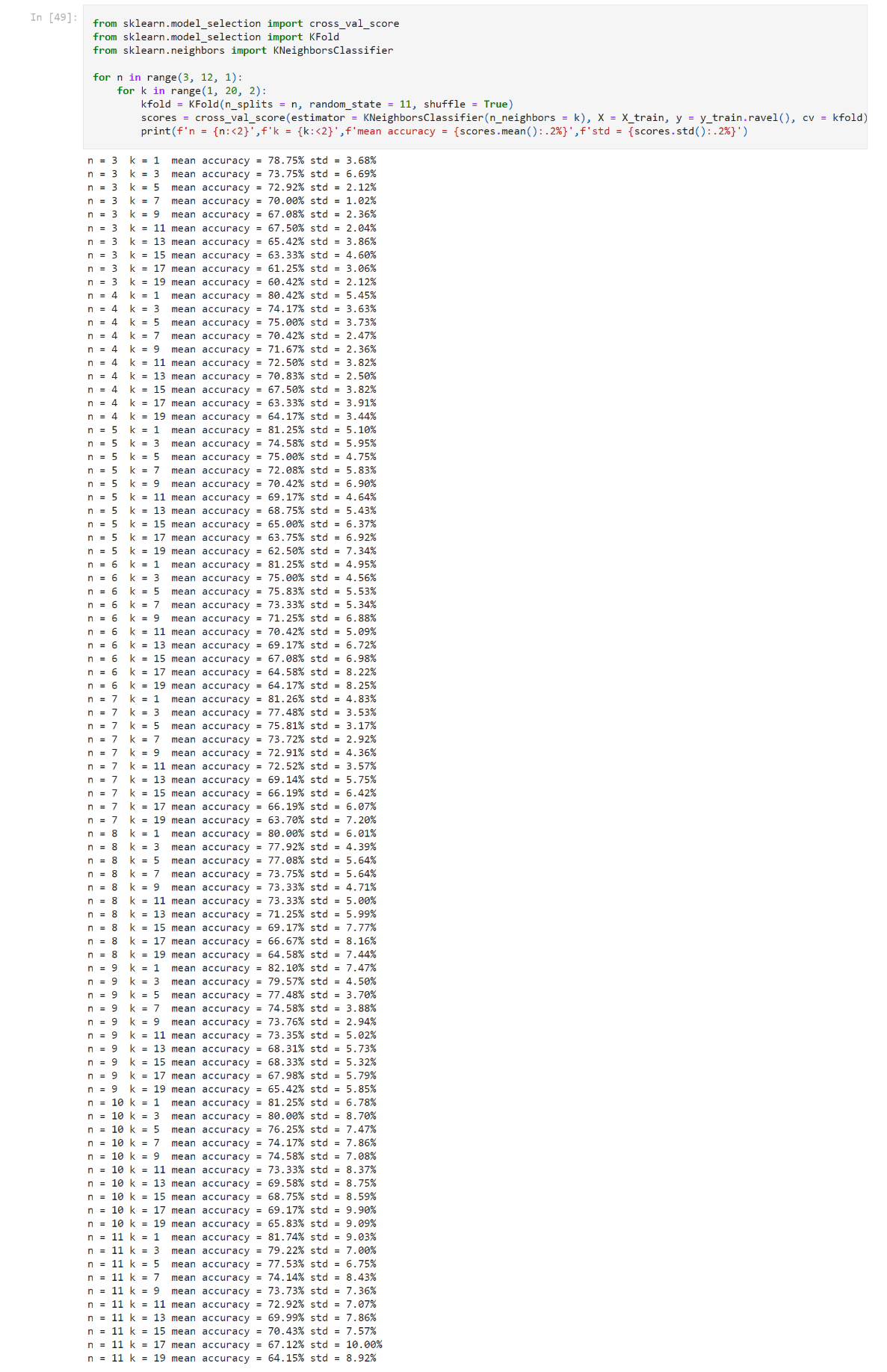
## Hyperparameter Tuning

Reload training and testing dataset with 1000 samples each to reduce test time.



Figure

Hyperparameter tuning is used to choose hyperparameter values that produce the best possible predictions. To determine the best value for k in the kNN algorithm, try different values of k then compare each estimator’s performance. To determine the best value for n in the K-Fold Cross-Validation, try different values of n then compare each estimator’s performance.

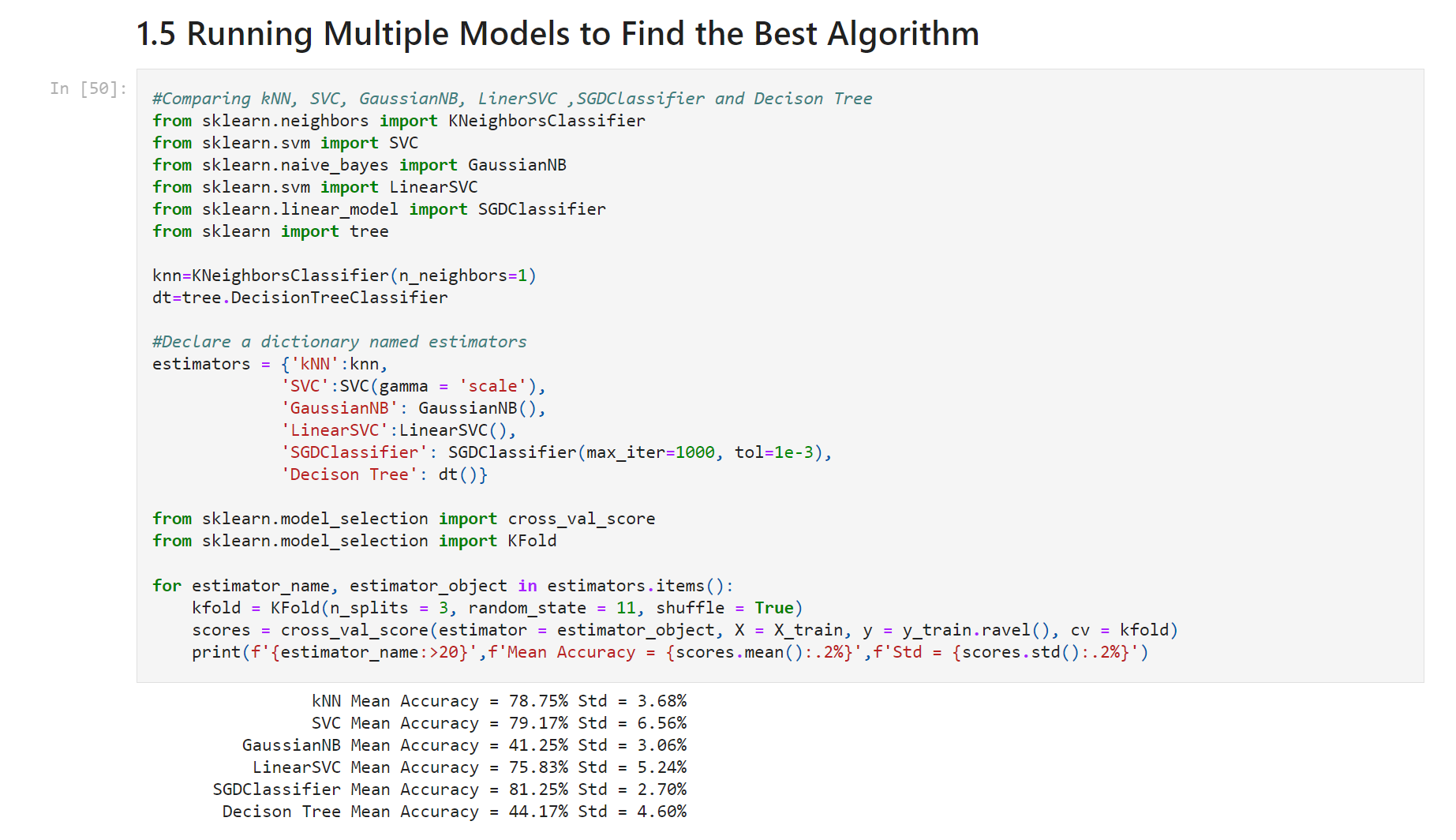


Figure

Finger 7 shows n=3 and k=1 that produce the best possible predictions.

## Running Multiple Models to Find the Top 3 Algorithm

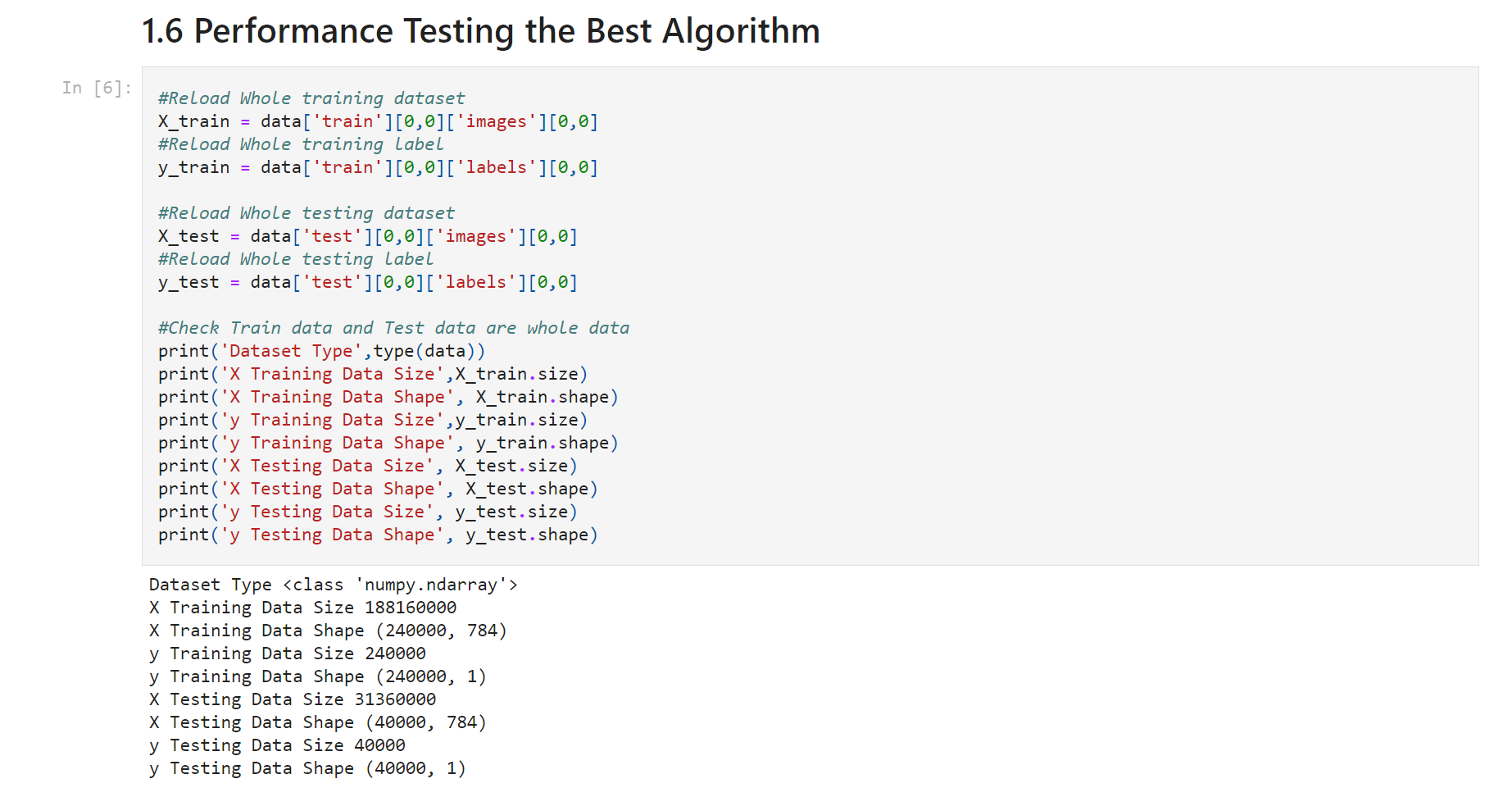
K-Fold Cross-Validation is performed on six classification algorithms. The six classification algorithms are k-nearest neighbours (kNN) algorithm, Support Vector Clustering algorithm, GaussianNB algorithm, LinearSVC algorithm, SGDClassifier algorithm and Decision Tree algorithm. Through calculation, the mean accuracy and standard deviation is obtained for each algorithm.



Figure

Figure 8 illustrates SVC, kNN and SGD classification algorithm have higher mean accuracy and lower standard deviation results.

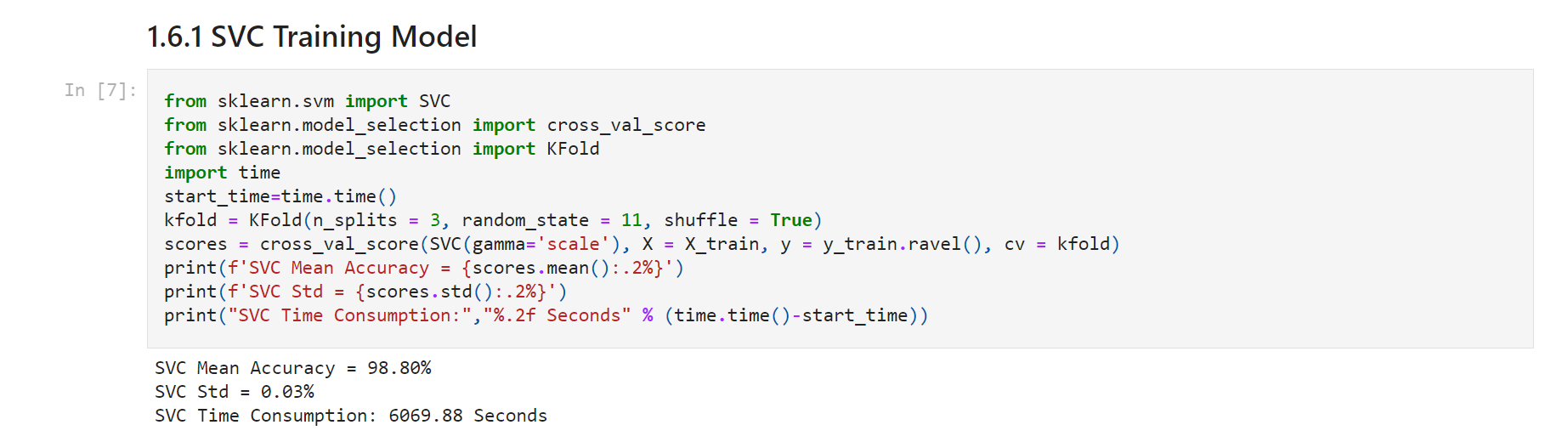
## Running Multiple Models to Find the Top 3 Algorithm

Reload all data sets and print data set information for verification.

Figure

### SVC Classification Algorithm

Use SVC classification algorithm to classify the EMNIST digits data set, and obtain the mean accuracy, standard deviation and time consumption scores.

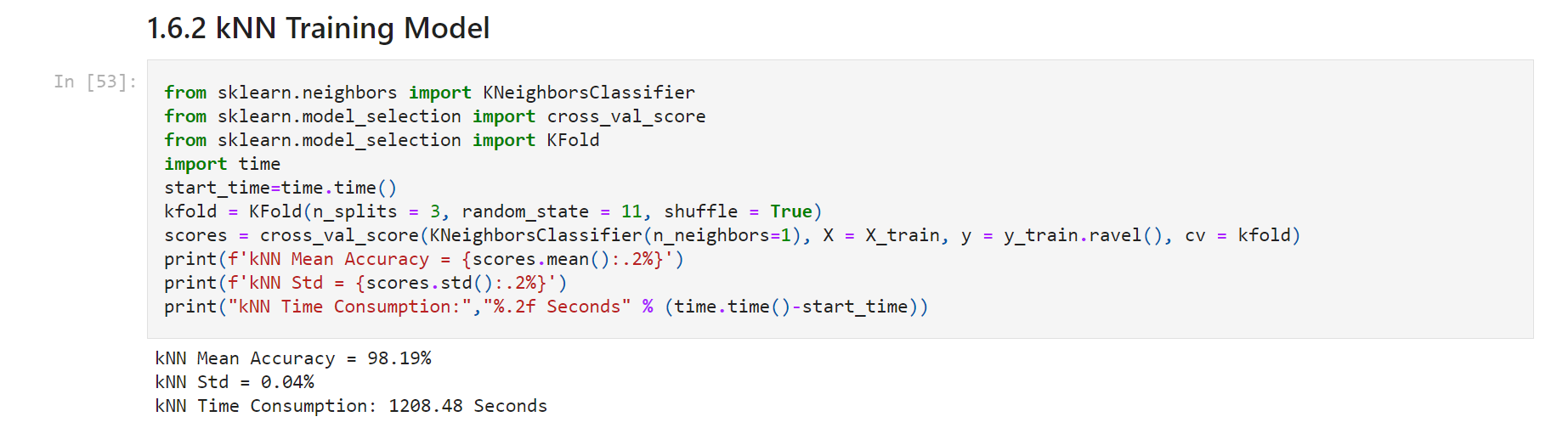


Figure

As can be seen from the above figure 10 results, the mean accuracy of the SVC classification algorithm is 98.8%, the standard deviation is 0.03%, and it takes 6069.88 seconds.

### kNN Classification Algorithm

Use the kNN classification algorithm to classify the EMNIST digits data set, and obtain the mean accuracy, standard deviation and time consumption scores.

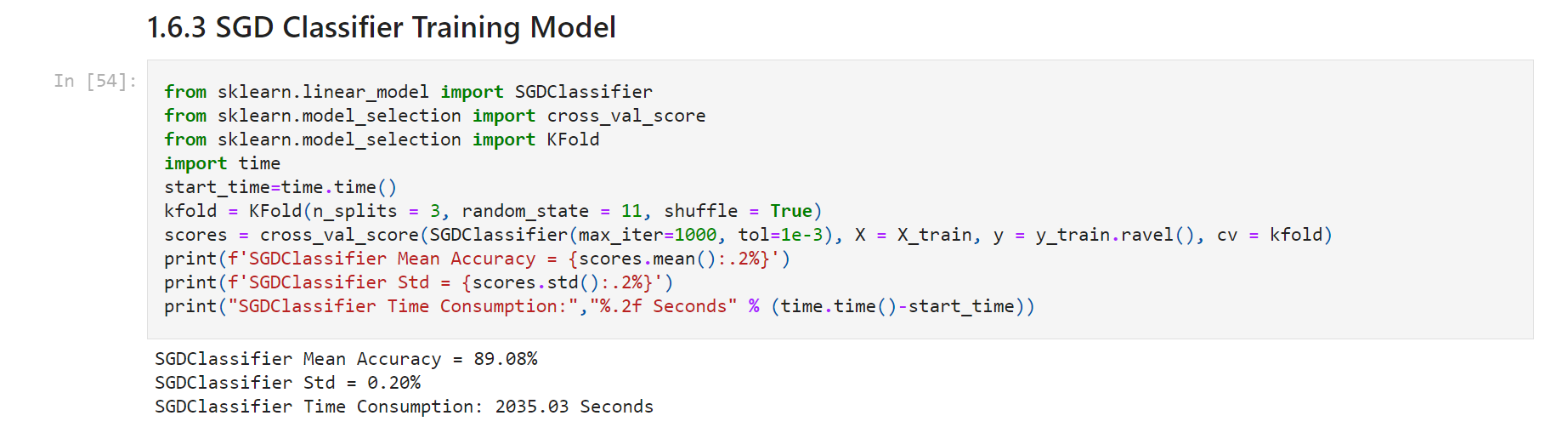


Figure

Figure 11 shows the mean accuracy of kNN classification algorithm is 98.19%, the standard deviation is 0.04%, and it takes 1208.48 seconds.

### SGD Classification Algorithm

Use the SGD classification algorithm to classify the EMNIST digits data set, and obtain the mean accuracy, standard deviation and time consumption scores.



Figure

As can be seen from the above figure 12 results, the mean accuracy of the SGD classification algorithm is 89.08%, the standard deviation is 0.20%, and it takes 2035.03 seconds.

# Multi-Classification of Letters with the EMNIST Dataset

## Load Letters Data

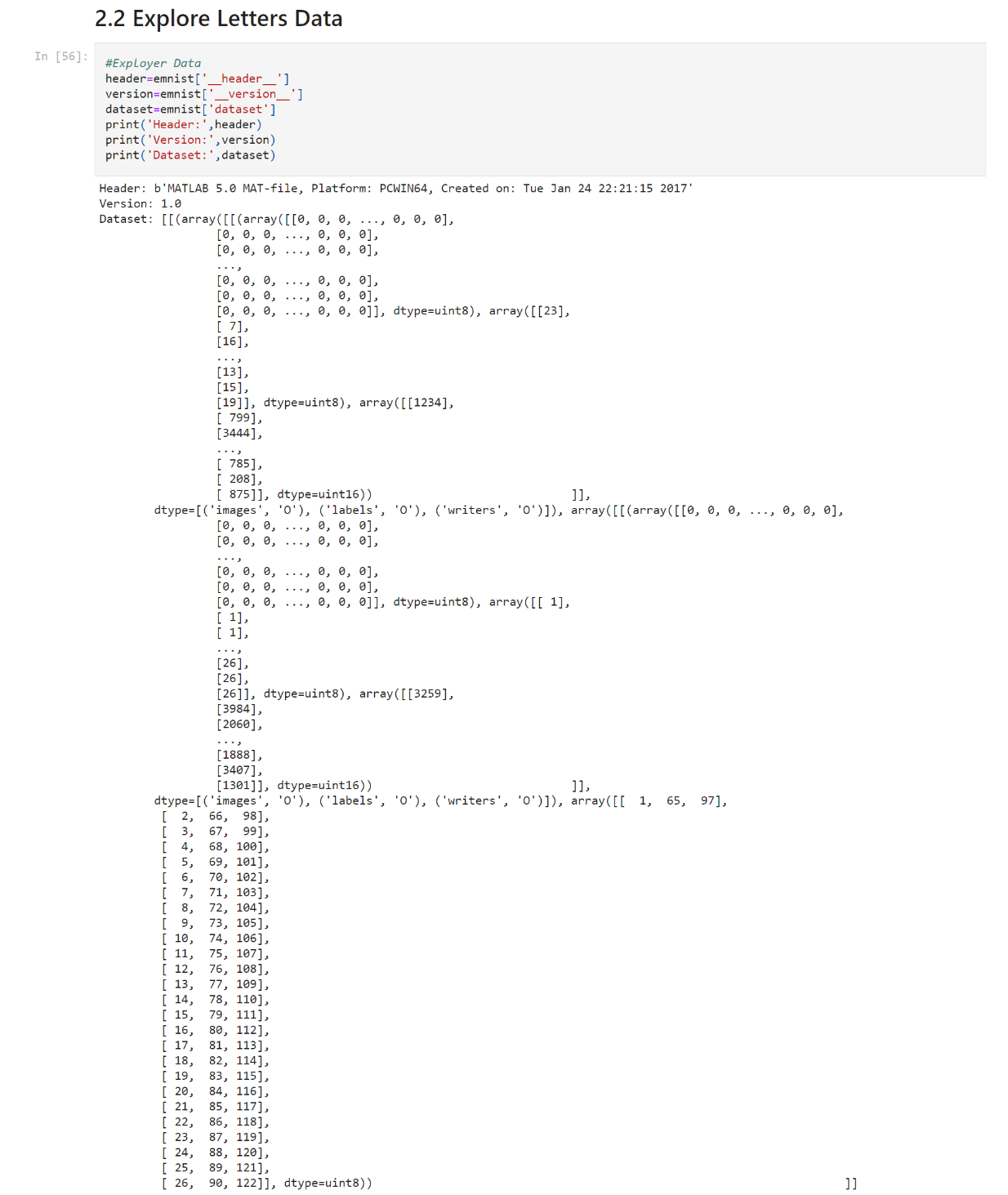
Load letters data from emnist-letters.mat file by using SciPy open-source Python library.



Figure

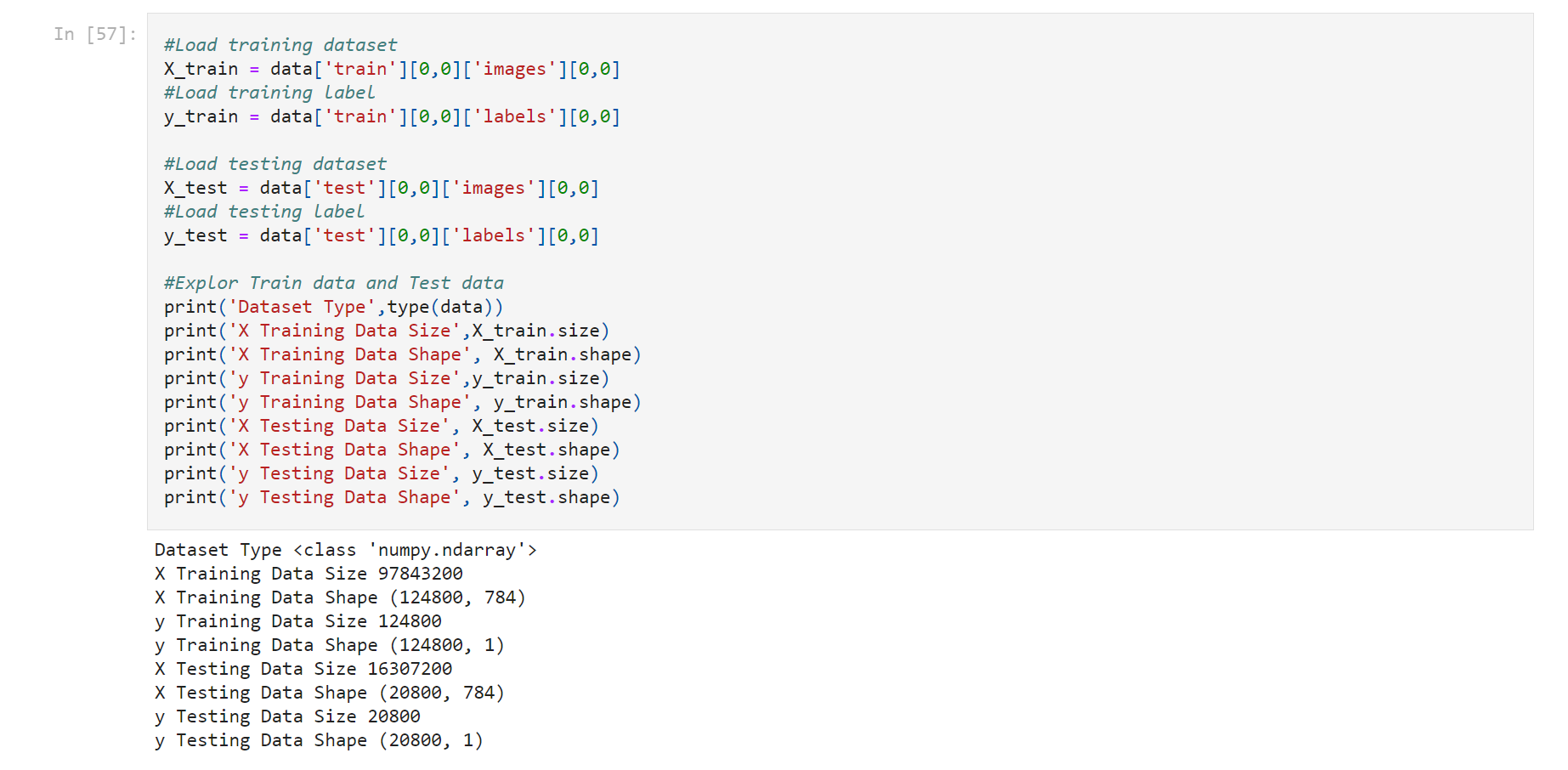
## Explore Letters Data

Display Header, Version and dataset of emnist-letters.mat MATLAB file.



Figure

According to figure 14, the emnist letters dataset is divided into training dataset and testing dataset. Load training dataset and testing dataset as well as explore them.



Figure

X is image data and y is label data. The explore results are shown in the finger above.

## Visualize Letters Data

Verify if the letters data loaded correctly.



Figure

Reshape image data from MATLAB to pyplot format and display image and label data shape in the finger 16.

Import the matplotib.pylot library to randomly display image and label from digits dataset.



Figure

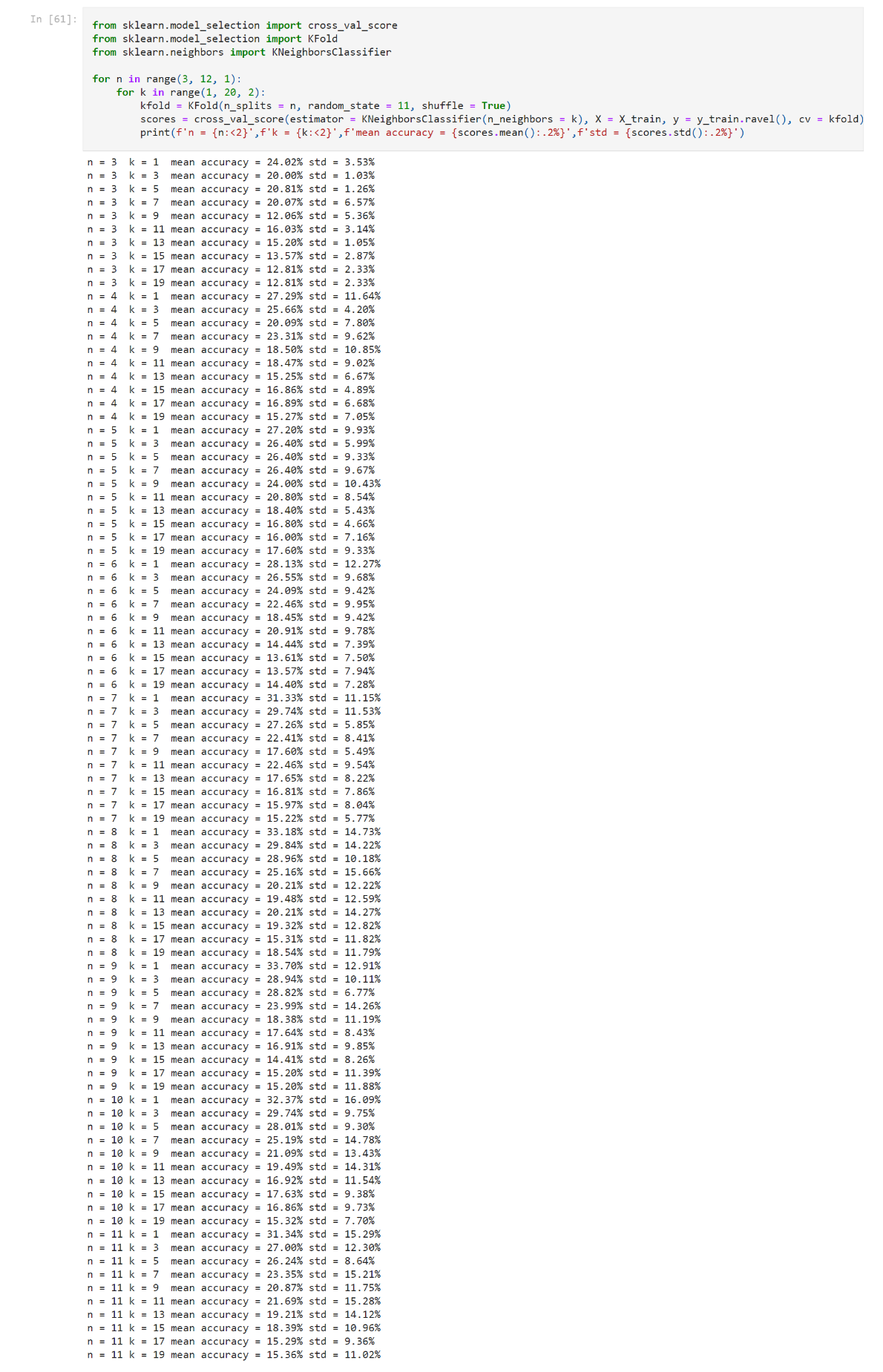
## Hyperparameter Tuning

Reload training and testing dataset with 1000 samples each to reduce test time.



Figure

Hyperparameter tuning is used to choose hyperparameter values that produce the best possible predictions. To determine the best value for k in the kNN algorithm, try different values of k then compare each estimator’s performance. To determine the best value for n in the K-Fold Cross-Validation, try different values of n then compare each estimator’s performance.

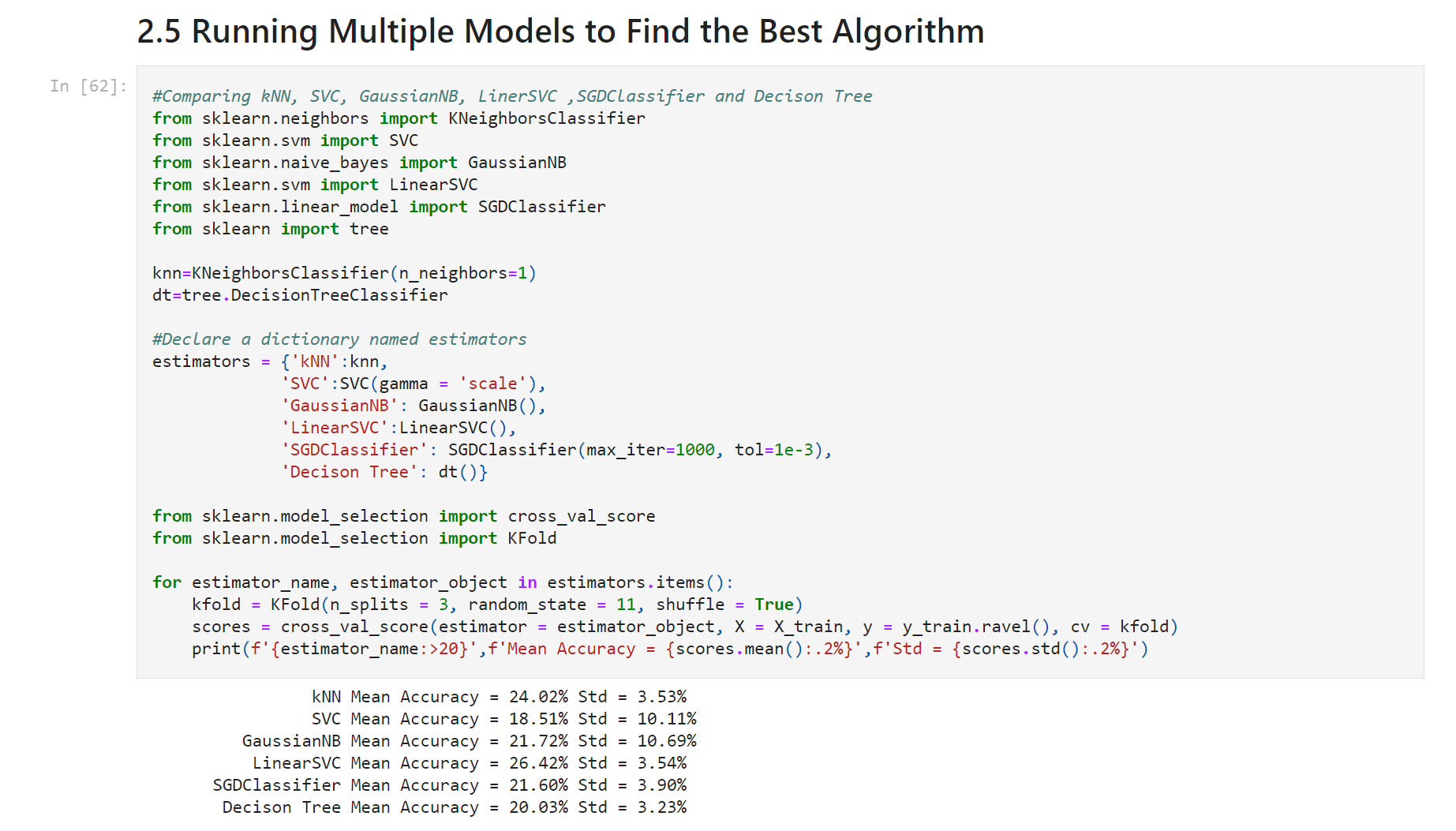


Figure

In finger 19, n=3 and k=1 produce the best possible predictions.

## Running Multiple Models to Find the Top 3 Algorithm

K-Fold Cross-Validation is performed on six classification algorithms. The six classification algorithms are k-nearest neighbours (kNN) algorithm, Support Vector Clustering algorithm, GaussianNB algorithm, LinearSVC algorithm, SGDClassifier algorithm and Decision Tree algorithm. Through calculation, the mean accuracy and standard deviation are obtained for each algorithm.

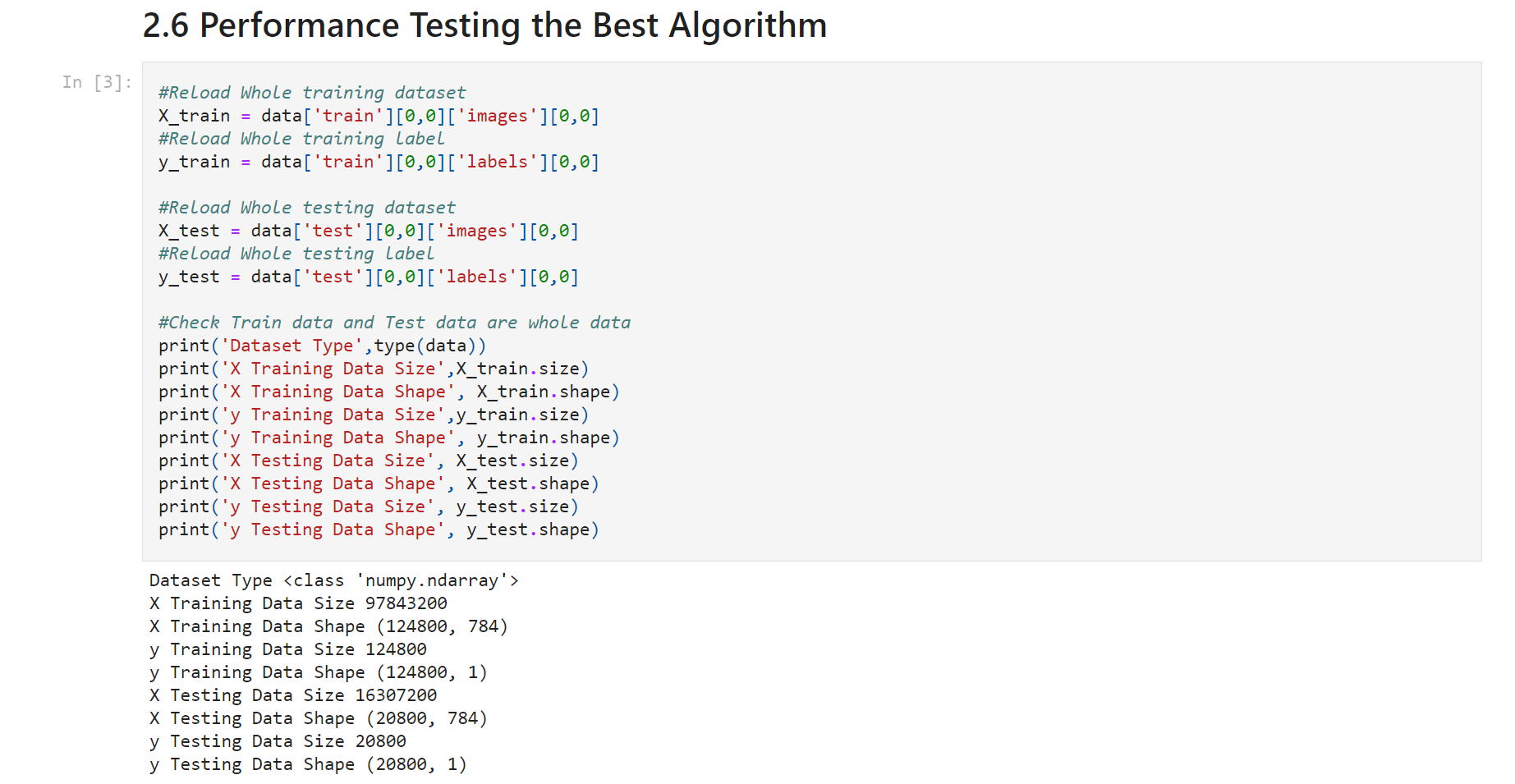


Figure

As can be seen from the finger 20 above, LinearSVC, kNN and SGD classification algorithm has higher mean accuracy and lower standard deviation results.

## Performance Testing the Top 3 Algorithm

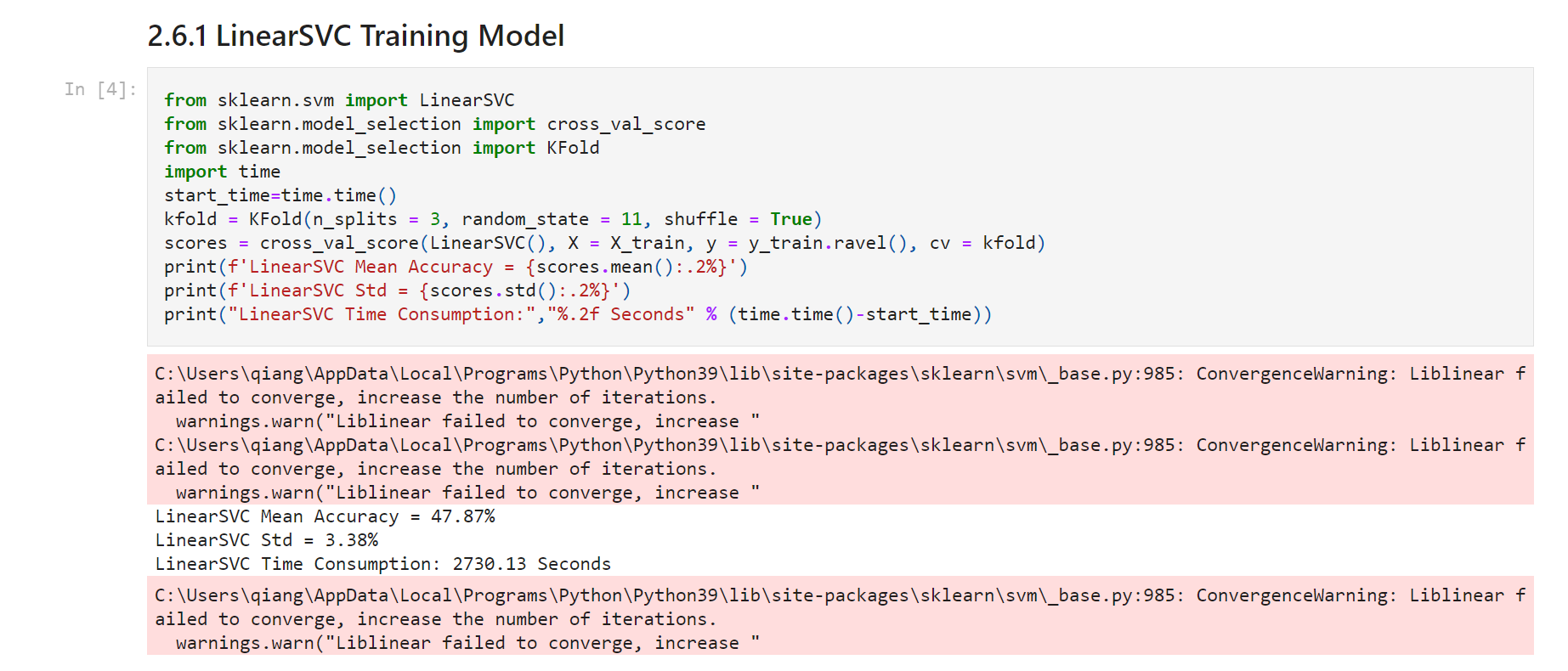
Reload all data sets and print data set information for verification.



Figure

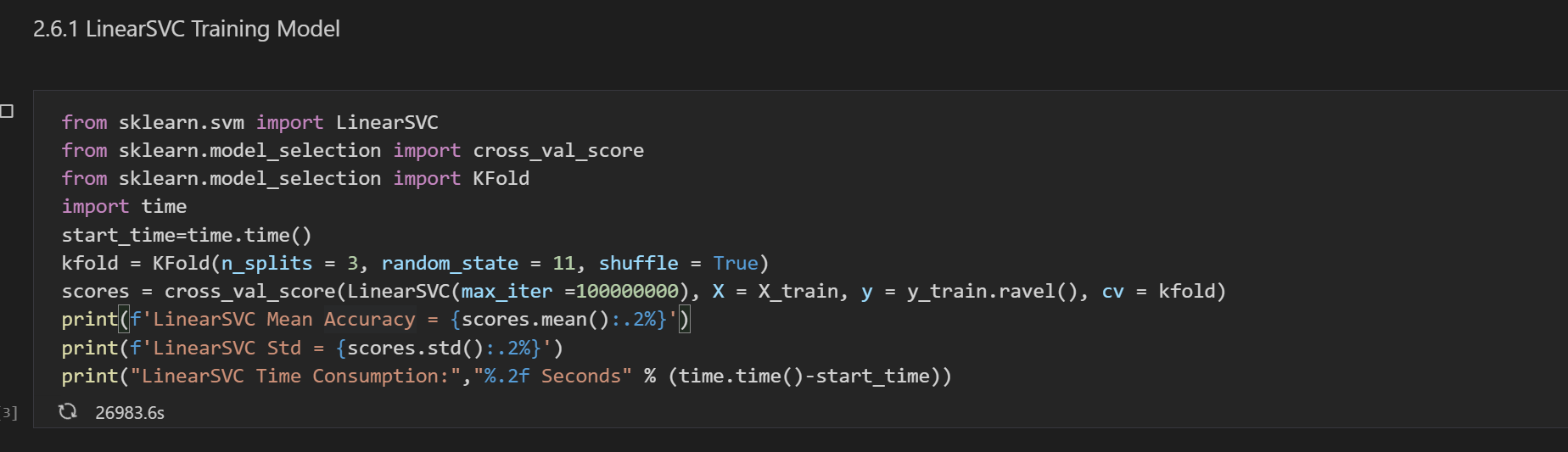
### LinearSVC Classification Algorithm

Use the SVC classification algorithm to classify the EMNIST digits data set, and obtain the mean accuracy, standard deviation and time consumption scores.



Figure

Figure 22 results shows the LinearSVC classification algorithm makes mistakes in classifying all data sets, because the default number of iterations is 1000 (Scikit learn) and it is not enough to run.

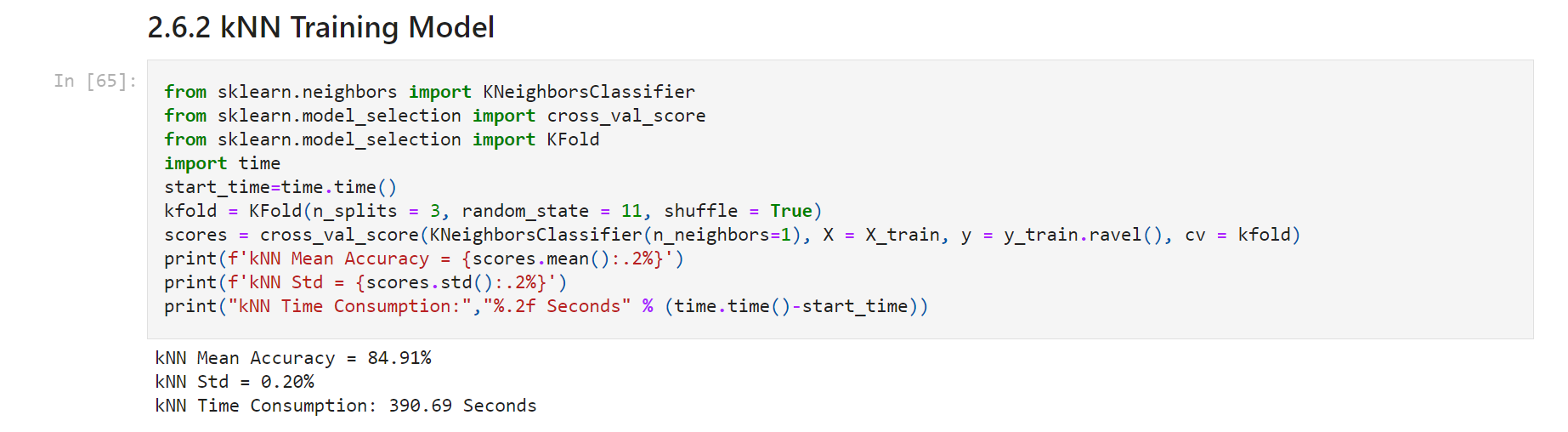


Figure

As can be seen from the above figure 23 results, modify maximum number to 10000000 performs 26983.6 seconds of operation, but it is left as incompleted.

### kNN Classification Algorithm

Use kNN classification algorithm to classify the EMNIST letters data set, and obtain the mean accuracy, standard deviation and time consumption scores.

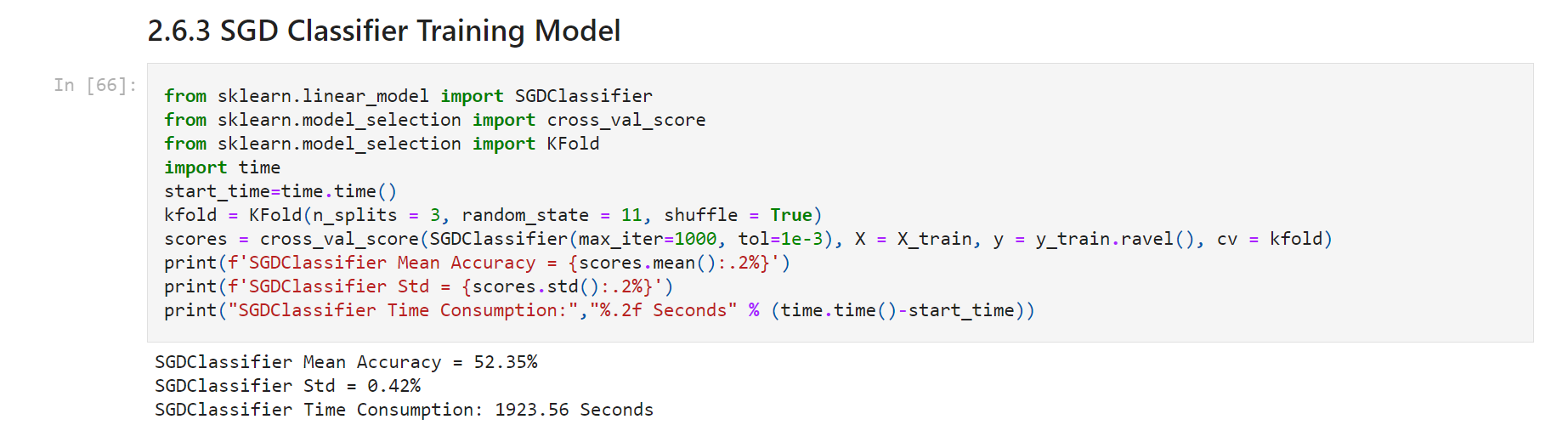


Figure

Figure 24 shows the mean accuracy of the kNN classification algorithm is 84.91%, the standard deviation is 0.20%, and it takes 390.69 seconds.

### SGD Classification Algorithm

Use SGD classification algorithm to classify the EMNIST letters data set, and obtain the mean accuracy, standard deviation and time consumption scores.



Figure

Figure 25 shows the mean accuracy of the SGD classification algorithm is 52.35%, the standard deviation is 0.42%, and it takes 1923.56 seconds.

# Conclusion

Different algorithms are used to classify EMNIST’s digits and letters datasets in this project. 1000 samples from six algorithms are extracted for classification, and three algorithms are selected with high mean accuracy and low standard deviation classification test of all samples.

K-Fold Cross-Validation is used for testing, and tune n splits to get the best test results. K is tuned to get the best test results for kNN classification algorithms.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Algorithm | Mean Accuracy | Standard Deviation | Time Consumption |
| Digits | SVC | 98.8% | 0.03% | 6069.88 Seconds |
| Digits | kNN | 98.19% | 0.04% | 1208.48 Seconds |
| Digits | SGD | 89.08% | 0.2% | 2035.03 Seconds |
| Letters | LinearSVC | N/A | N/A | N/A |
| Letters | kNN | 84.91% | 0.2% | 390.69 Seconds |
| Letters | SGD | 52.35% | 0.42% | 1923.56 Seconds |

All test results are displayed from the above table. kNN algorithm takes the fastest time in digital dataset, nearly 1 time fast than SGD algorithm and 5 times faster than SVC algorithm, while the mean accuracy is only 0.61% lower than SVC algorithm and 0.1% for standard deviation.

Due to the failure of the linearSVC algorithm test for letters dataset, kNN algorithm is the fastest. It is 5 times faster than SGD algorithm, 32.56% higher than the SGD algorithm for accuracy, and 0.22% lower for standard deviation.

In summary, kNeighborsClassifier Classification algorithm has the advantages of fast calculation, high mean accuracy and low standard deviation.

# References

Scikit learn. *sklearn.svm.LinearSVC.* <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>

The EMNIST Dataset. (2019, March, 28). *The EMNIST Dataset.*  <https://www.nist.gov/itl/products-and-services/emnist-dataset>