Name: Onat Kaya – Student number: 24218 – CS 412 Homework 1 Report

In this task, we have the problem of classifying images from the MNIST dataset with correct labels. The images represent handwriting forms of digits and we are trying to predict the spesific digit by using decision tree and k-NN classifiers. Both the training and testing datasets are fetched from Keras.

The images are in 28x28 format, representing digits from 0 to 9; each pixel is a gray-level from 0-255.

The training dataset goes by the name "mnist_train.csv", has 60,000 samples and has a size of 104 MB. On the the other hand, the test dataset is called "mnist_test.csv", has 10,000 samples and has a size of 17.4 MB.

The 20% of the training data was used for validation and splitted in that fashion. Also, no cross-validation was used since we our data points are ample.

For pre-processing, normalization was done and re-shaping of data from 2D to 1D was made.

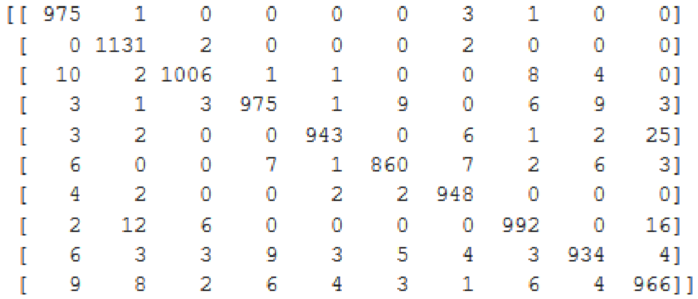
Validation accuracies for different approaches and parameters could be found in the table below:

```
 +-------------+------------------+-            +-------------+------------------+-----+

 |k-NN Classifier              |                |Decision Tree                 |

 |k-values  |Validation Accuracy|               |min_sample_splits |Validation Accuracy|

 | 1        | 97.40833%       |                 | 2             | 87.84167%       |

 | 5        | 97.50000%       |                 | 3             | 87.63333%       |

  +-------------+-------------------+-          | 5             | 87.89167%       |

                                                +-------------+-------------------+----
```

As a result, k-NN classifier with the parameter k = 5 was chosen.

Moving on, while testing the chosen classifier (k-NN, with the parameter k = 5) with the test data, resulted in producing a digit classification accuracy of 97.30000% on the test data. Another observation was that that k-NN classifier takes more time than the decision tree to compute.

Additionaly, you can see the confusion matrix to have a better understanding of the results in this homework.



```
[[ 975    1    0    0    0    0    3    1    0    0]
 [    0 1131    2    0    0    0    2    0    0    0]
 [   10    2 1006    1    1    0    0    8    4    0]
 [    3    1    3  975    1    9    0    6    9    3]
 [    3    2    0    0  943    0    6    1    2   25]
 [    6    0    0    7    1  860    7    2    6    3]
 [    4    2    0    0    2    2  948    0    0    0]
 [    2   12    6    0    0    0    0  992    0   16]
 [    6    3    3    9    3    5    4    3  934    4]
 [    9    8    2    6    4    3    1    6    4  966]]
```

Looking at the matrix, it could be clearly seen that the majority of the tests resulted on the diagonal of the matrix which tells us the chosen classifier has done a good job.