Machine Learning with Big Data
Assignment 2
M22MA003

Question 1 : Write a code to classify the test samples using the kNN algorithm and Jaccard similarity by varying the value of k in {1,2,3,4,5}. Report the classification accuracy and time required to classify all the test samples using one CPU core.

Solution 1 :-

- A metric used to compare two sets of data is called the Jaccard similarity. The size of the intersection between the two sets and the size of the union between the two sets are used to compute it.
- When comparing documents or phrases, Jaccard similarity is frequently employed in natural language processing and text mining. The two sets are typically the sets of distinctive terms found in each document or sentence in this context.
- Although Jaccard similarity is a quick and effective method of comparing data sets, it has significant drawbacks. It can be sensitive to the size of the sets, for instance, and it disregards the order of the elements in the sets.
- Jaccard similarity is a value between 0 and 1, where 0 indicates no similarity between the sets and 1 indicates complete similarity.

Steps Followed for implementation of the code solution:

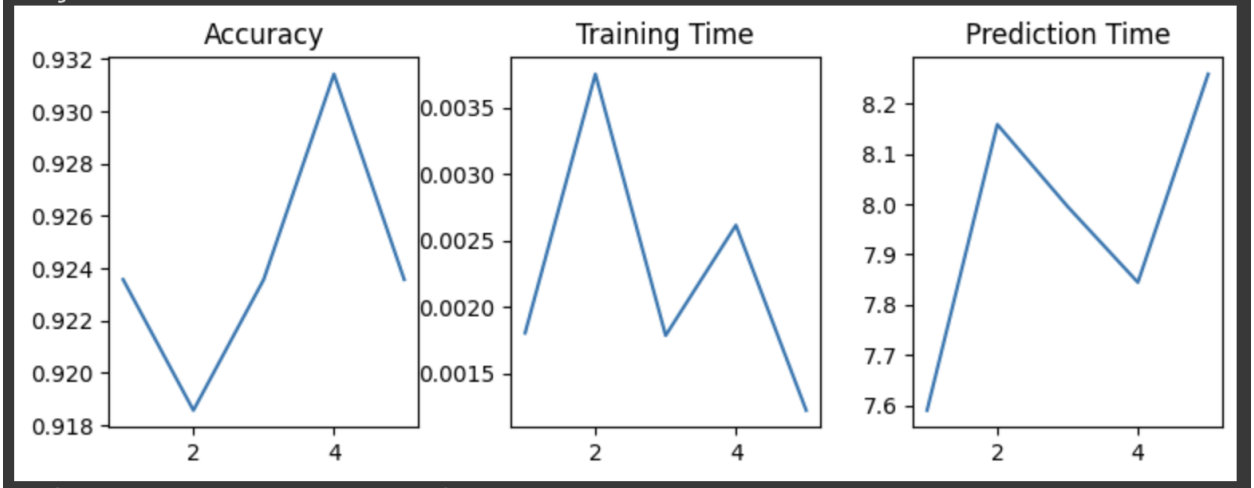
Using single CPU Core with the help of os library and setting the environ["NUMEXPR_NUM_THREADS"] = "1" and environ["OMP_NUM_THREADS"] = "1".

1. Import the required libraries. Libraries used are : os, sklearn, time, psutils, matplotlib, numpy.
2. Import the MNIST data using the datasets library. Preprocess the data to reshape it to obtain single np arrays with dimension 784 and binarise the data to get input data arrays as binary raw feature vectors.
3. Define jaccard_knn function to implement the knn algorithm using the jaccard_similarity as metric and calculate the accuracy (value ranges from 0 to 1).
4. Define the draw_plot function to draw the graphs between the varying value of k = {1,2,3,4,5} against three parameters:-
    - Accuracy
    - Training Time
    - Prediction Time
5. Define get_peak_memory function to run the code and obtain the peak memory in each call with a different size dataset.

Question1 Results:-

Number of Training samples : 10000 , Number of Test samples : 1400

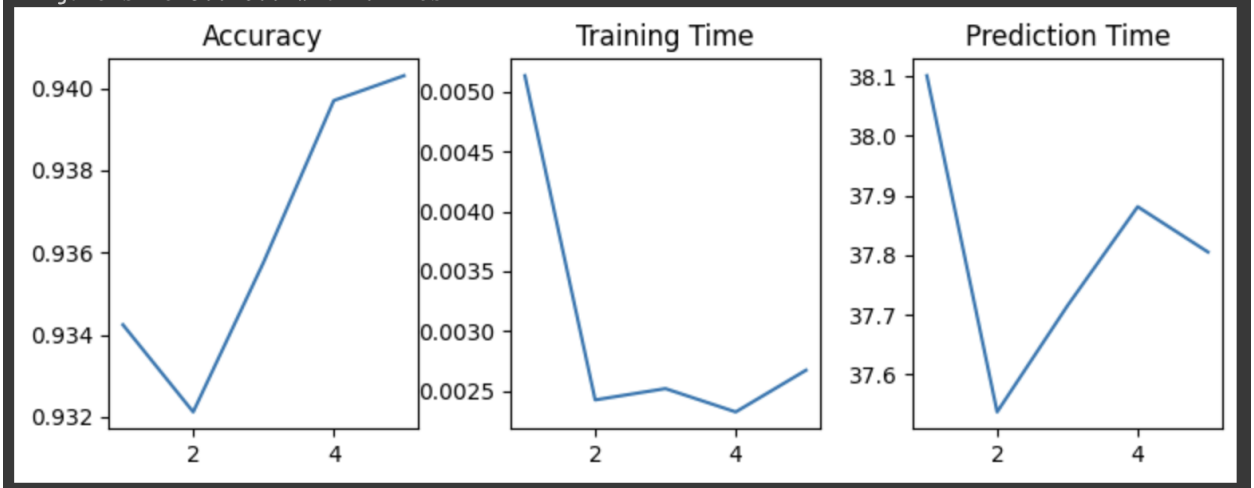
```
Number of Training samples : 10000 , Number of Test samples :  1400
K=1, Accuracy=0.9236, Training Time=0.0018, Prediction Time=7.5901
K=2, Accuracy=0.9186, Training Time=0.0038, Prediction Time=8.1583
K=3, Accuracy=0.9236, Training Time=0.0018, Prediction Time=7.9951
K=4, Accuracy=0.9314, Training Time=0.0026, Prediction Time=7.8444
K=5, Accuracy=0.9236, Training Time=0.0012, Prediction Time=8.2585
<Figure size 500x300 with 0 Axes>
```




```
Peak memory usage: 782.82421875 bytes
**********************************************************************************************
```

Number of Training samples : 20000 , Number of Test samples : 3300

```
Number of Training samples : 20000 , Number of Test samples :  3300
K=1, Accuracy=0.9342, Training Time=0.0051, Prediction Time=38.1009
K=2, Accuracy=0.9321, Training Time=0.0024, Prediction Time=37.5371
K=3, Accuracy=0.9358, Training Time=0.0025, Prediction Time=37.7151
K=4, Accuracy=0.9397, Training Time=0.0023, Prediction Time=37.8812
K=5, Accuracy=0.9403, Training Time=0.0027, Prediction Time=37.8052
<Figure size 500x300 with 0 Axes>
```

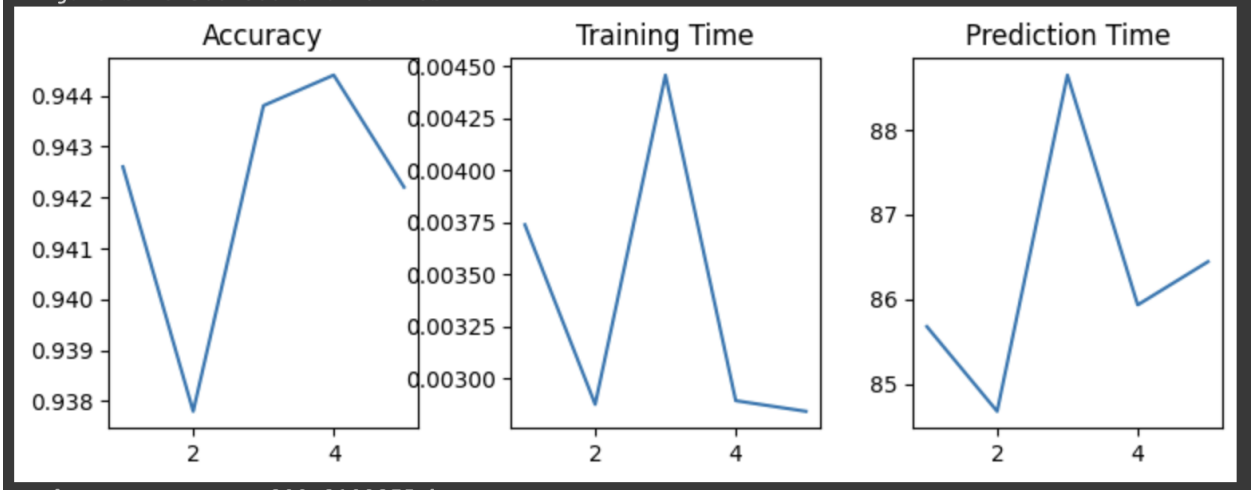



```
Peak memory usage: 791.63671875 bytes
**********************************************************************************************
```

Question1 Results :-

Number of Training samples : 30000 , Number of Test samples : 5000

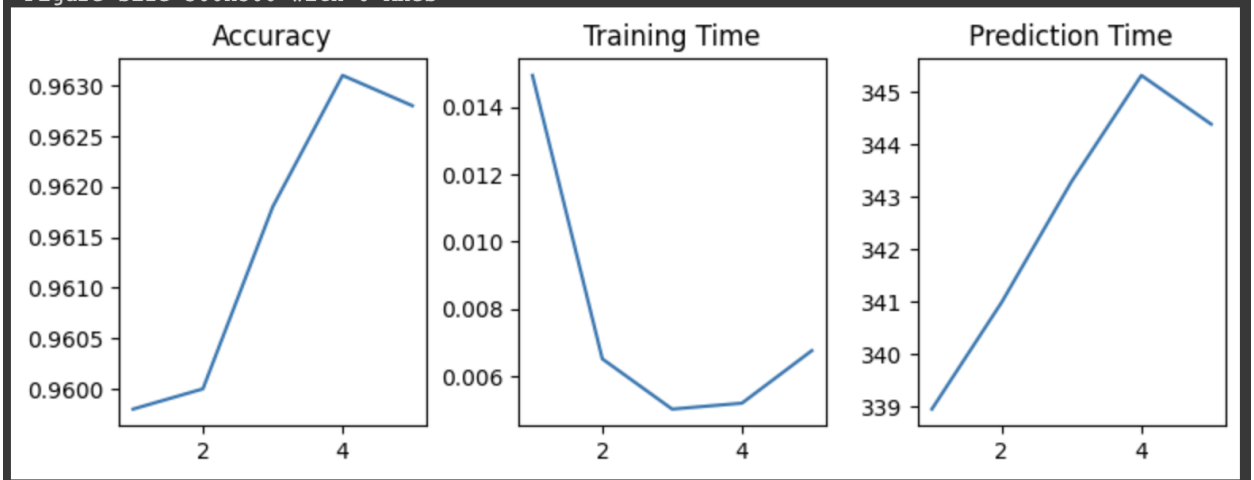
```
Number of Training samples : 30000 , Number of Test samples :  5000
K=1, Accuracy=0.9426, Training Time=0.0037, Prediction Time=85.6758
K=2, Accuracy=0.9378, Training Time=0.0029, Prediction Time=84.6745
K=3, Accuracy=0.9438, Training Time=0.0045, Prediction Time=88.6473
K=4, Accuracy=0.9444, Training Time=0.0029, Prediction Time=85.9318
K=5, Accuracy=0.9422, Training Time=0.0028, Prediction Time=86.4422
<Figure size 500x300 with 0 Axes>
```




```
Peak memory usage: 800.2109375 bytes
**********************************************************************************************
```

Number of Training samples : 60000 , Number of Test samples : 10000

```
Number of Training samples : 60000 , Number of Test samples :  10000
K=1, Accuracy=0.9598, Training Time=0.0149, Prediction Time=338.9475
K=2, Accuracy=0.9600, Training Time=0.0065, Prediction Time=340.9851
K=3, Accuracy=0.9618, Training Time=0.0050, Prediction Time=343.2857
K=4, Accuracy=0.9631, Training Time=0.0052, Prediction Time=345.3091
K=5, Accuracy=0.9628, Training Time=0.0068, Prediction Time=344.3788
<Figure size 500x300 with 0 Axes>
```




```
Peak memory usage: 803.08984375 bytes
**********************************************************************************************
```

[Solution1_Colab_Notebook_Link]

Question 2 : Using any publicly available code for LSH, classify the test samples using the kNN algorithm. Vary the length of the signature vector in {40,60} and 's' in {0.8, 0.9}. For each combination, run the experiment multiple times to calculate the average and standard deviation of classification accuracy and time required to classify all the test samples in all the set-ups using one CPU core.

Solution 2 :-
Observations on LSH Algorithm as metric in KNN algorithm:-
Firstly, few points on LSH algo:-

- A probabilistic approach for approximating closest neighbor search in high-dimensional spaces is called Locality Sensitive Hashing (LSH). It is frequently used in applications involving machine learning and data mining that call for effective similarity search.
- LSH functions by probabilistically assigning related data items to the same bucket. To do this, a series of hash functions that map the data points to a number of buckets are applied to the data points.
- LSH is a compromise between precision and effectiveness. The method may be tweaked to either provide lower accuracy with slower processing time, or more accuracy at the price of higher computational cost.

Few points on KNN:-

- KNN library inbuilt function - "fit" assigns labels to the training points.
- When the predict function is called KNN inherently takes one test sample at a time and finds distance with each train point. This increases the complexity of the algorithm.
- If we implement the bucketing system in the LSH algorithm then we would end up with a number of candidate pairs for each test sample. After saving the candidate pairs in a list, there is no way to execute the inbuilt KNN library in such a way that for each test sample, only the list items should be called.
- Hence, the time taken by KNN algo with LSH is huge compared to the inbuilt defined metrics.

Steps Followed for implementation of the code solution:

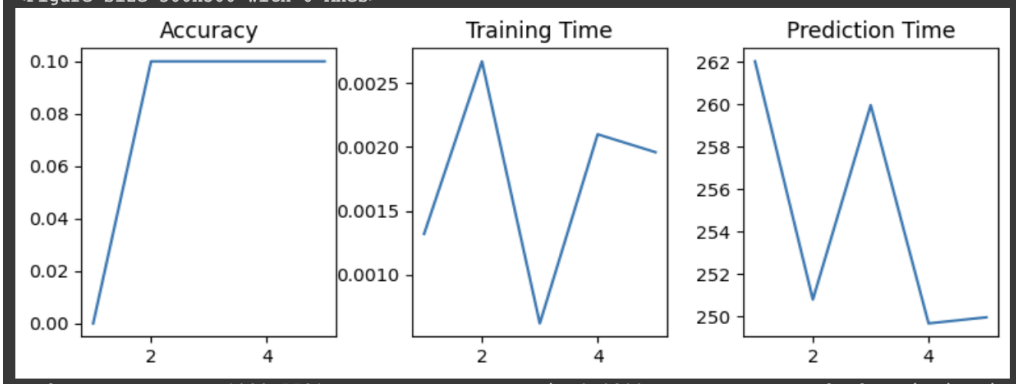
1. Import the required libraries. Libraries used are : datasketch, sklearn, time, psutils, matplotlib, numpy.
2. Import the MNIST data using the datasets library. Preprocess the data to reshape it and binarise the data to get input data arrays as binary raw feature vectors.
3. Define LSH functions with varying length of the signature vector in **{40,60}** and 's' in {**0.8, 0.9**}.
4. Function names are my_metric**40_s8**, my_metric**50_s8**, my_metric**60_s8**, my_metric**40_s9**, my_metric**50_s9**, my_metric**60_s9**. Three signature vector lengths are considered for execution [40, 50, 60]. Threshold is represented by s8 and s9 for 's' as [0.8, 0.9] respectively.
5. Define function to implement the knn algorithm using the customized LSH values and calling the respective methods in the metric parameter of KNN and calculate the accuracy (value ranges from 0 to 1).

6. Define the draw_plot function to draw the graphs and run for each combination of signature vector and “s” multiple times ,i.e., between the varying value of k = {1,2,3,4,5} against three parameters:-
    - Accuracy
    - Training Time
    - Prediction Time
7. Calculate the average accuracy and standard deviation of accuracy and display for each combination (6 combinations of signature vector and s, each with K values - 1 to 5).
8. Define get_peak_memory function to run the code and obtain the peak memory in each call with a different size dataset.

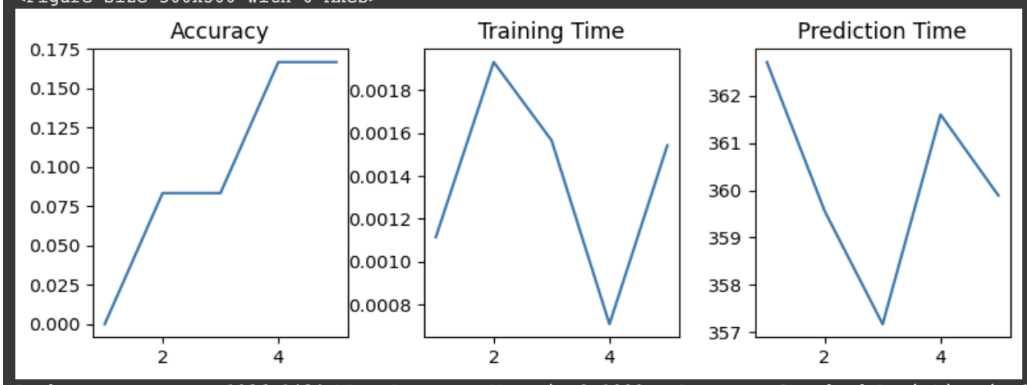
Results with multiple K values and displaying Accuracy Average and Standard Deviation:

Signature Vector Length : 40 and “s” value : 0.8 and K value 1 to 5 and diff subsets.

```
Number of Training samples : 1000 , Number of Test samples :  10
K=1, Accuracy=0.0000, Training Time=0.0013, Prediction Time=262.0300
K=2, Accuracy=0.1000, Training Time=0.0027, Prediction Time=250.7920
K=3, Accuracy=0.1000, Training Time=0.0006, Prediction Time=259.9648
K=4, Accuracy=0.1000, Training Time=0.0021, Prediction Time=249.6658
K=5, Accuracy=0.1000, Training Time=0.0020, Prediction Time=249.9539
<Figure size 500x300 with 0 Axes>
```



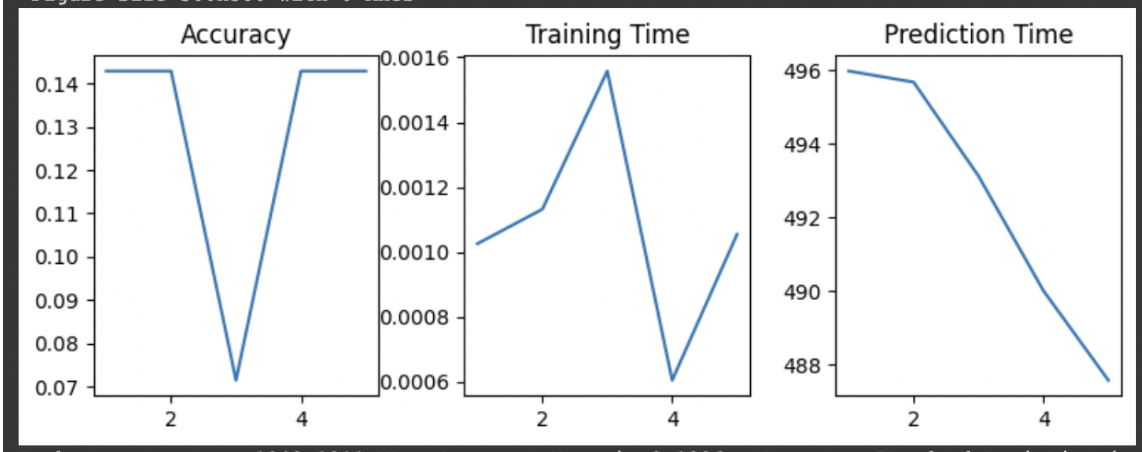

```
Peak memory usage: 1033.5781 MB , Accuracy Mean is 0.0800 , Accuracy Standard Deviation is 0.0400
**********************************************************************************************
Number of Training samples : 1200 , Number of Test samples :  12
K=1, Accuracy=0.0000, Training Time=0.0011, Prediction Time=362.7095
K=2, Accuracy=0.0833, Training Time=0.0019, Prediction Time=359.5550
K=3, Accuracy=0.0833, Training Time=0.0016, Prediction Time=357.1714
K=4, Accuracy=0.1667, Training Time=0.0007, Prediction Time=361.6003
K=5, Accuracy=0.1667, Training Time=0.0015, Prediction Time=359.8891
<Figure size 500x300 with 0 Axes>
```



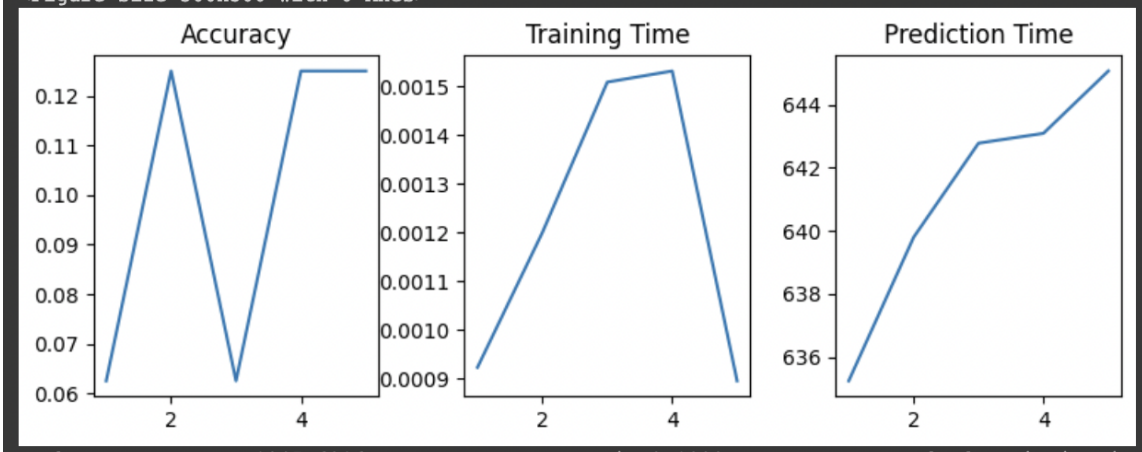

```
Peak memory usage: 1036.1484 MB , Accuracy Mean is 0.1000 , Accuracy Standard Deviation is 0.0624
**********************************************************************************************
```

Signature Vector Length : 40 and “s” value : 0.8 and K value 1 to 5 and diff subsets.

```
Number of Training samples : 1400 , Number of Test samples :  14
K=1, Accuracy=0.1429, Training Time=0.0010, Prediction Time=495.9703
K=2, Accuracy=0.1429, Training Time=0.0011, Prediction Time=495.6691
K=3, Accuracy=0.0714, Training Time=0.0016, Prediction Time=493.1237
K=4, Accuracy=0.1429, Training Time=0.0006, Prediction Time=490.0150
K=5, Accuracy=0.1429, Training Time=0.0011, Prediction Time=487.5882
<Figure size 500x300 with 0 Axes>
```



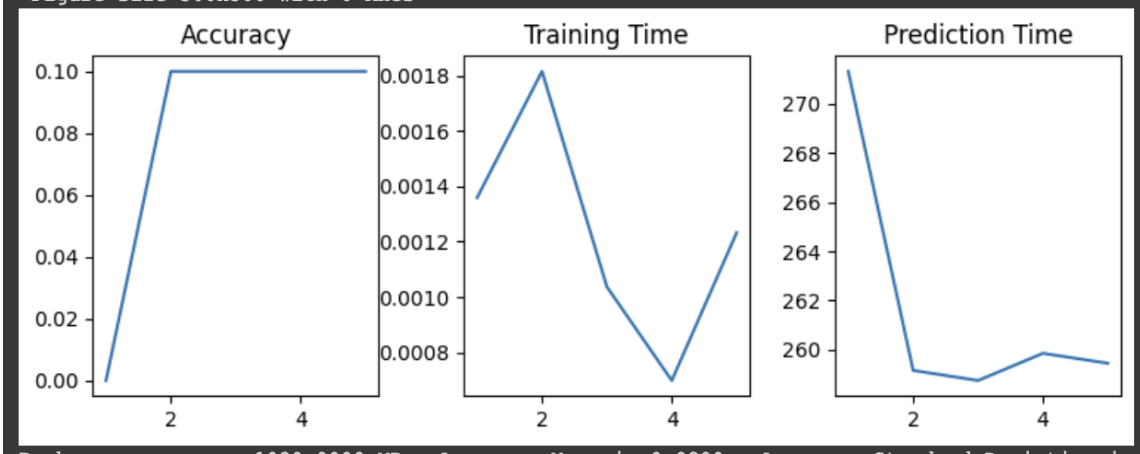

```
Peak memory usage: 1042.1211 MB , Accuracy Mean is 0.1286 , Accuracy Standard Deviation is 0.0286
**************************************************************************************************
Number of Training samples : 1600 , Number of Test samples :  16
K=1, Accuracy=0.0625, Training Time=0.0009, Prediction Time=635.2380
K=2, Accuracy=0.1250, Training Time=0.0012, Prediction Time=639.7958
K=3, Accuracy=0.0625, Training Time=0.0015, Prediction Time=642.7832
K=4, Accuracy=0.1250, Training Time=0.0015, Prediction Time=643.0942
K=5, Accuracy=0.1250, Training Time=0.0009, Prediction Time=645.0744
<Figure size 500x300 with 0 Axes>
```



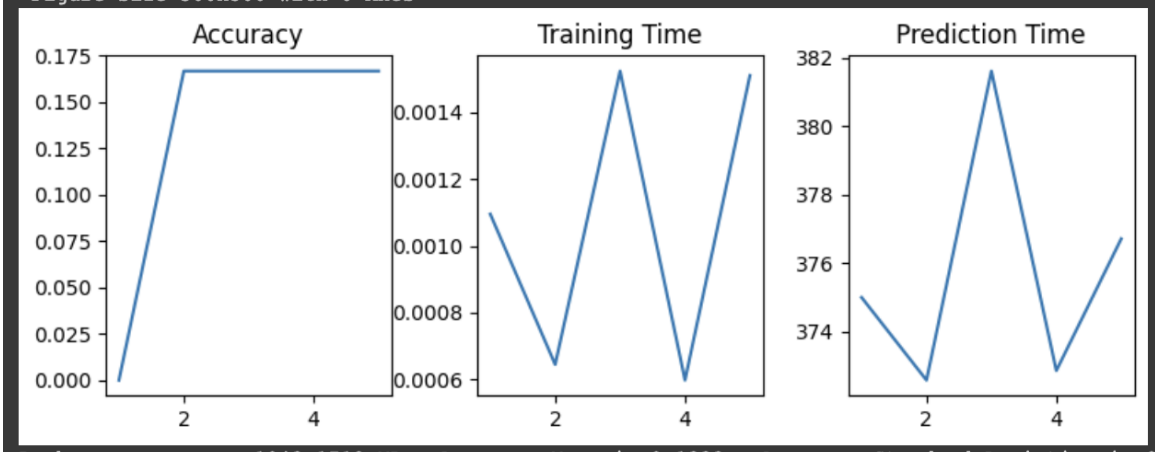

```
Peak memory usage: 1047.6016 MB , Accuracy Mean is 0.1000 , Accuracy Standard Deviation is 0.0306
**************************************************************************************************
```

Signature Vector Length : 50 and “s” value : 0.8 and K value 1 to 5 and diff subsets.

```
Number of Training samples : 1000 , Number of Test samples :  10
K=1, Accuracy=0.0000, Training Time=0.0014, Prediction Time=271.3146
K=2, Accuracy=0.1000, Training Time=0.0018, Prediction Time=259.1555
K=3, Accuracy=0.1000, Training Time=0.0010, Prediction Time=258.7447
K=4, Accuracy=0.1000, Training Time=0.0007, Prediction Time=259.8508
K=5, Accuracy=0.1000, Training Time=0.0012, Prediction Time=259.4464
<Figure size 500x300 with 0 Axes>
```



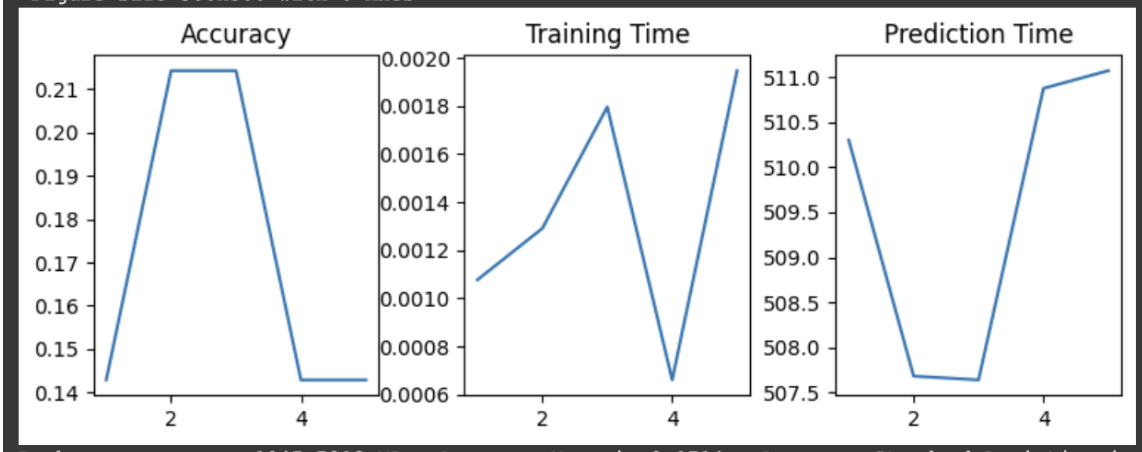

```
Peak memory usage: 1030.0000 MB , Accuracy Mean is 0.0800 , Accuracy Standard Deviation is 0.0400
**********************************************************************************************************
Number of Training samples : 1200 , Number of Test samples :  12
K=1, Accuracy=0.0000, Training Time=0.0011, Prediction Time=374.9874
K=2, Accuracy=0.1667, Training Time=0.0006, Prediction Time=372.5681
K=3, Accuracy=0.1667, Training Time=0.0015, Prediction Time=381.6045
K=4, Accuracy=0.1667, Training Time=0.0006, Prediction Time=372.8473
K=5, Accuracy=0.1667, Training Time=0.0015, Prediction Time=376.7017
<Figure size 500x300 with 0 Axes>
```



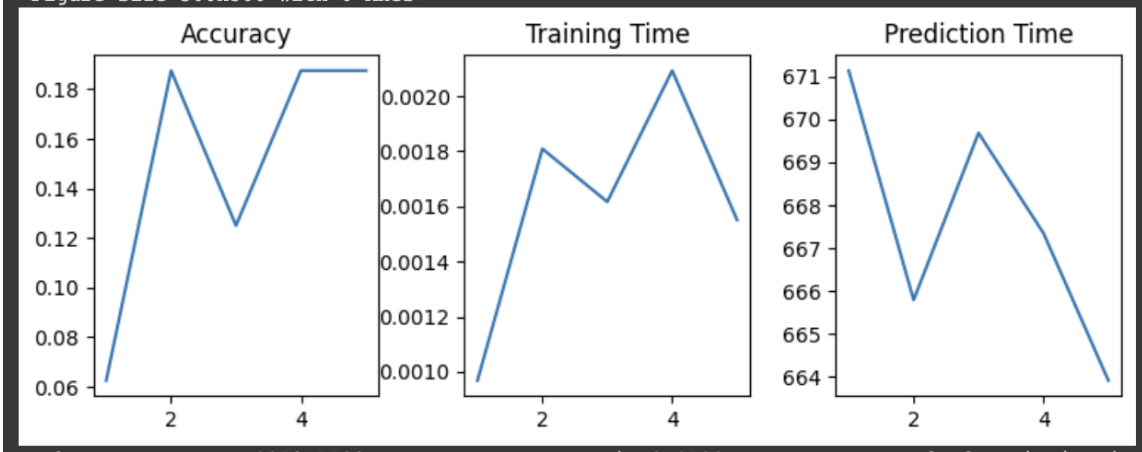

```
Peak memory usage: 1043.1719 MB , Accuracy Mean is 0.1333 , Accuracy Standard Deviation is 0.0667
**********************************************************************************************************
```

Signature Vector Length : 50 and “s” value : 0.8 and K value 1 to 5 and diff subsets.

```
Number of Training samples : 1400 , Number of Test samples :  14
K=1, Accuracy=0.1429, Training Time=0.0011, Prediction Time=510.3022
K=2, Accuracy=0.2143, Training Time=0.0013, Prediction Time=507.6806
K=3, Accuracy=0.2143, Training Time=0.0018, Prediction Time=507.6385
K=4, Accuracy=0.1429, Training Time=0.0007, Prediction Time=510.8756
K=5, Accuracy=0.1429, Training Time=0.0019, Prediction Time=511.0725
<Figure size 500x300 with 0 Axes>
```



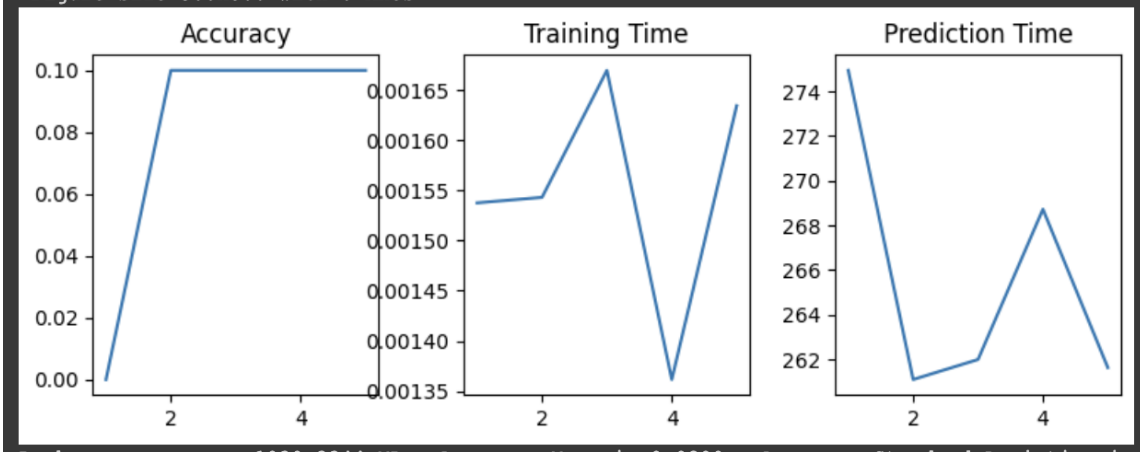

```
Peak memory usage: 1047.7812 MB , Accuracy Mean is 0.1714 , Accuracy Standard Deviation is 0.0350
**********************************************************************************************************
Number of Training samples : 1600 , Number of Test samples :  16
K=1, Accuracy=0.0625, Training Time=0.0010, Prediction Time=671.1316
K=2, Accuracy=0.1875, Training Time=0.0018, Prediction Time=665.7950
K=3, Accuracy=0.1250, Training Time=0.0016, Prediction Time=669.6807
K=4, Accuracy=0.1875, Training Time=0.0021, Prediction Time=667.3504
K=5, Accuracy=0.1875, Training Time=0.0016, Prediction Time=663.9202
<Figure size 500x300 with 0 Axes>
```



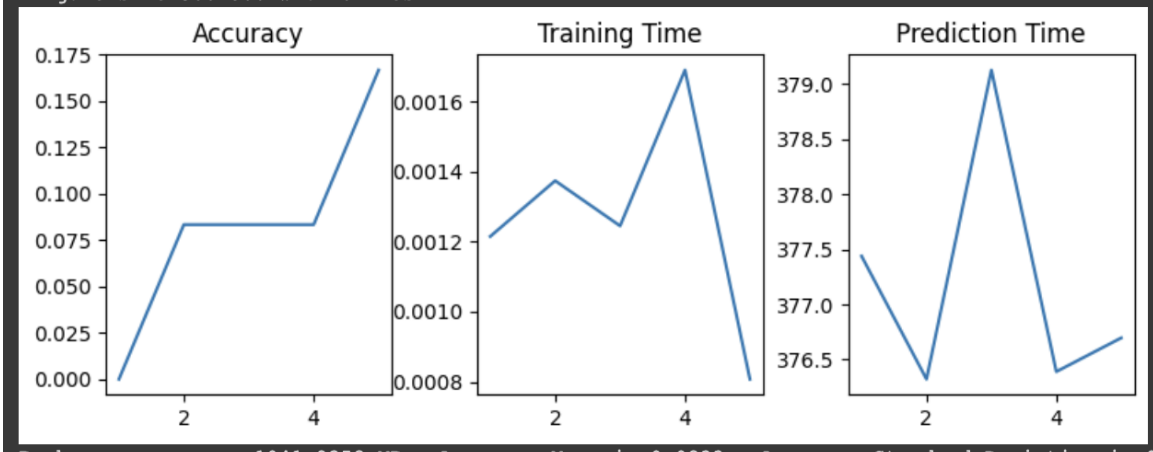

```
Peak memory usage: 1052.5508 MB , Accuracy Mean is 0.1500 , Accuracy Standard Deviation is 0.0500
**********************************************************************************************************
```

Signature Vector Length : 60 and “s” value : 0.8 and K value 1 to 5 and diff subsets.

```
Number of Training samples : 1000 , Number of Test samples :  10
K=1, Accuracy=0.0000, Training Time=0.0015, Prediction Time=274.9361
K=2, Accuracy=0.1000, Training Time=0.0015, Prediction Time=261.1029
K=3, Accuracy=0.1000, Training Time=0.0017, Prediction Time=261.9988
K=4, Accuracy=0.1000, Training Time=0.0014, Prediction Time=268.7227
K=5, Accuracy=0.1000, Training Time=0.0016, Prediction Time=261.6376
<Figure size 500x300 with 0 Axes>
```



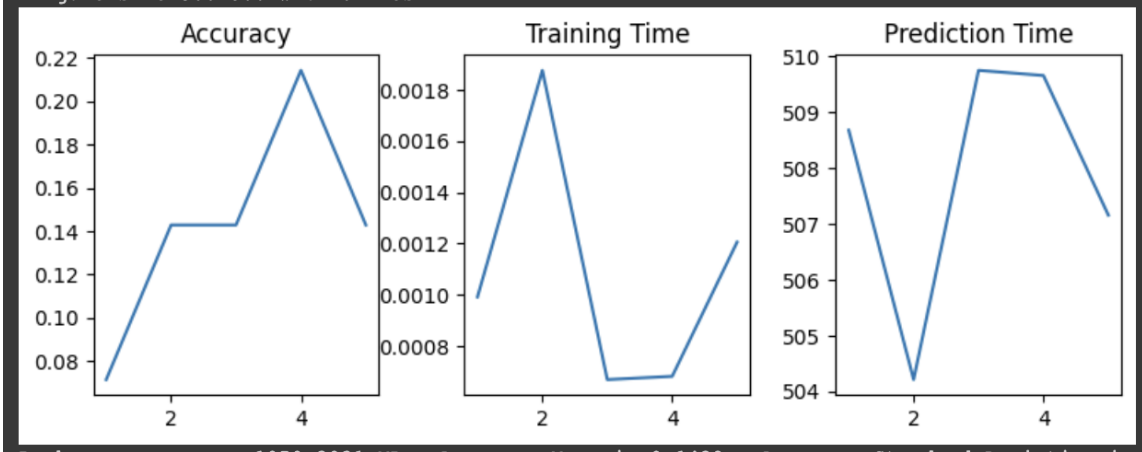

```
Peak memory usage: 1030.2344 MB , Accuracy Mean is 0.0800 , Accuracy Standard Deviation is 0.0400
**********************************************************************************************************
Number of Training samples : 1200 , Number of Test samples :  12
K=1, Accuracy=0.0000, Training Time=0.0012, Prediction Time=377.4386
K=2, Accuracy=0.0833, Training Time=0.0014, Prediction Time=376.3214
K=3, Accuracy=0.0833, Training Time=0.0012, Prediction Time=379.1284
K=4, Accuracy=0.0833, Training Time=0.0017, Prediction Time=376.3896
K=5, Accuracy=0.1667, Training Time=0.0008, Prediction Time=376.6964
<Figure size 500x300 with 0 Axes>
```



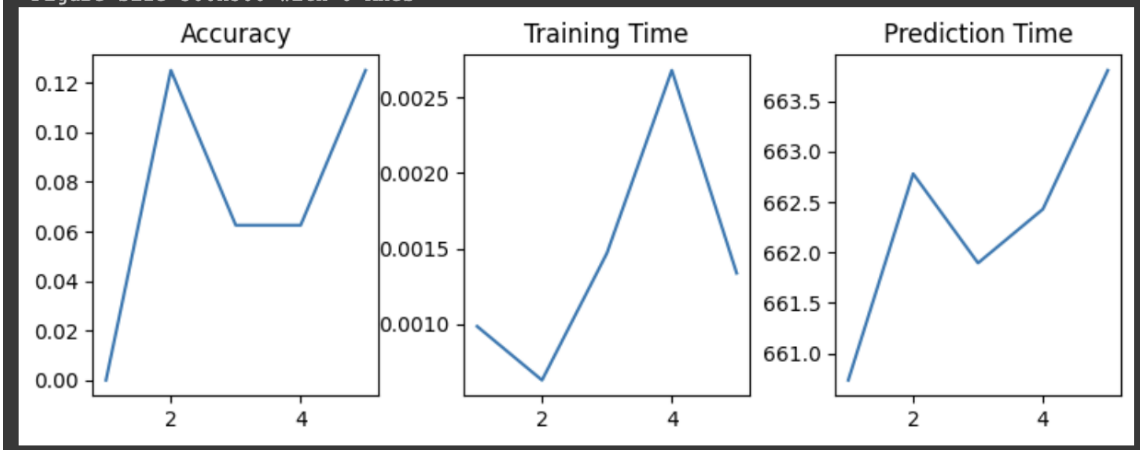

```
Peak memory usage: 1041.9258 MB , Accuracy Mean is 0.0833 , Accuracy Standard Deviation is 0.0527
**********************************************************************************************************
```

Signature Vector Length : 60 and “s” value : 0.8 and K value 1 to 5 and diff subsets.

```
Number of Training samples : 1400 , Number of Test samples :  14
K=1, Accuracy=0.0714, Training Time=0.0010, Prediction Time=508.6768
K=2, Accuracy=0.1429, Training Time=0.0019, Prediction Time=504.2140
K=3, Accuracy=0.1429, Training Time=0.0007, Prediction Time=509.7483
K=4, Accuracy=0.2143, Training Time=0.0007, Prediction Time=509.6557
K=5, Accuracy=0.1429, Training Time=0.0012, Prediction Time=507.1565
<Figure size 500x300 with 0 Axes>
```



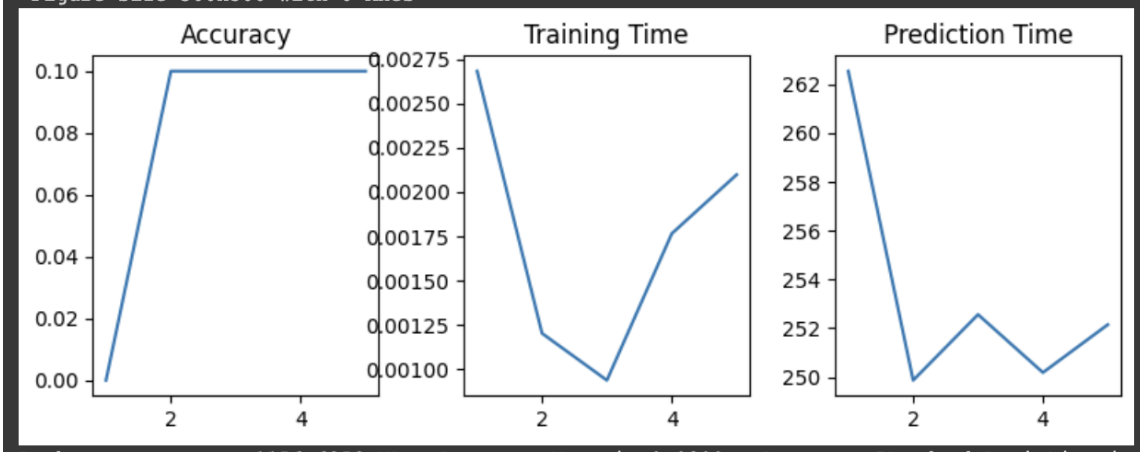

```
Peak memory usage: 1050.2031 MB , Accuracy Mean is 0.1429 , Accuracy Standard Deviation is 0.0452
**********************************************************************************************************
Number of Training samples : 1600 , Number of Test samples :  16
K=1, Accuracy=0.0000, Training Time=0.0010, Prediction Time=660.7336
K=2, Accuracy=0.1250, Training Time=0.0006, Prediction Time=662.7791
K=3, Accuracy=0.0625, Training Time=0.0015, Prediction Time=661.8946
K=4, Accuracy=0.0625, Training Time=0.0027, Prediction Time=662.4285
K=5, Accuracy=0.1250, Training Time=0.0013, Prediction Time=663.8041
<Figure size 500x300 with 0 Axes>
```



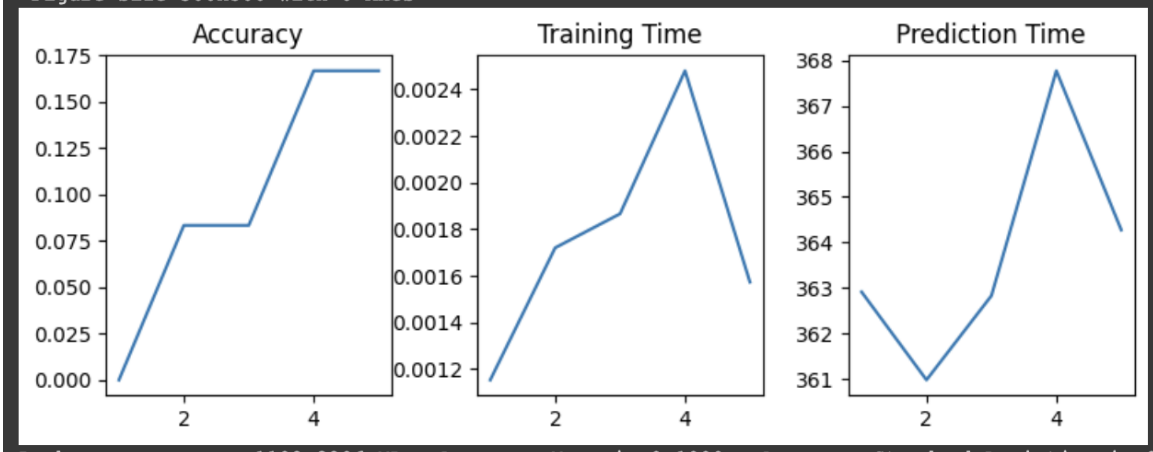

```
Peak memory usage: 1052.4883 MB , Accuracy Mean is 0.0750 , Accuracy Standard Deviation is 0.0468
**********************************************************************************************************
```

Signature Vector Length : 40 and “s” value : 0.9 and K value 1 to 5 and diff subsets.

```
Number of Training samples : 1000 , Number of Test samples :  10
K=1, Accuracy=0.0000, Training Time=0.0027, Prediction Time=262.5398
K=2, Accuracy=0.1000, Training Time=0.0012, Prediction Time=249.8638
K=3, Accuracy=0.1000, Training Time=0.0009, Prediction Time=252.5604
K=4, Accuracy=0.1000, Training Time=0.0018, Prediction Time=250.1809
K=5, Accuracy=0.1000, Training Time=0.0021, Prediction Time=252.1464
<Figure size 500x300 with 0 Axes>
```



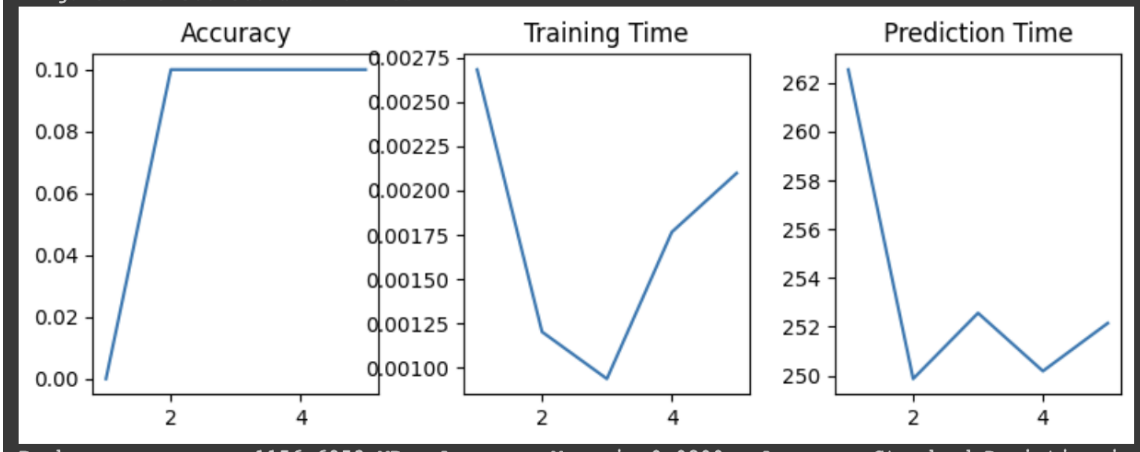

```
Peak memory usage: 1156.6953 MB , Accuracy Mean is 0.0800 , Accuracy Standard Deviation is 0.0400
**********************************************************************************************************
Number of Training samples : 1200 , Number of Test samples :  12
K=1, Accuracy=0.0000, Training Time=0.0012, Prediction Time=362.9151
K=2, Accuracy=0.0833, Training Time=0.0017, Prediction Time=360.9832
K=3, Accuracy=0.0833, Training Time=0.0019, Prediction Time=362.8318
K=4, Accuracy=0.1667, Training Time=0.0025, Prediction Time=367.7681
K=5, Accuracy=0.1667, Training Time=0.0016, Prediction Time=364.2754
<Figure size 500x300 with 0 Axes>
```



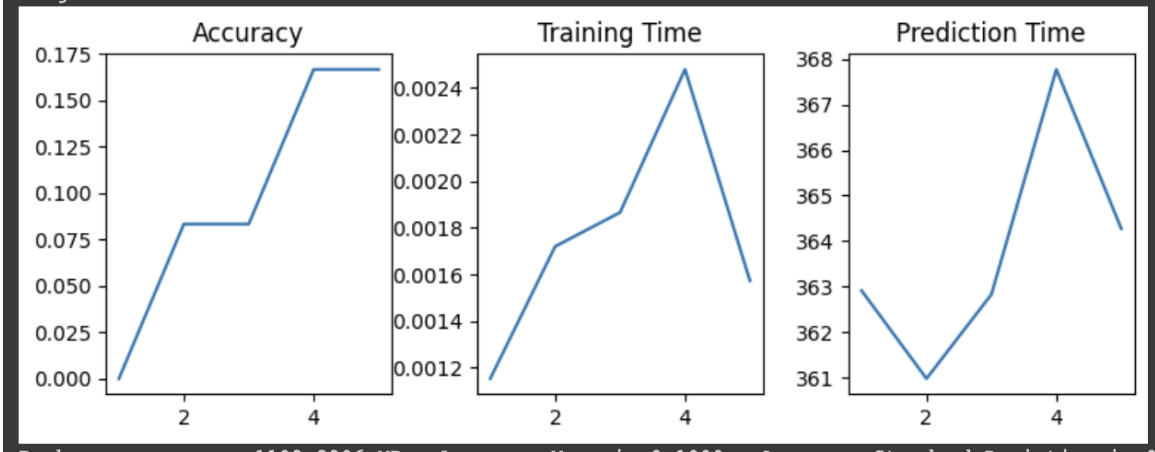

```
Peak memory usage: 1109.8906 MB , Accuracy Mean is 0.1000 , Accuracy Standard Deviation is 0.0624
**********************************************************************************************************
```

Signature Vector Length : 40 and “s” value : 0.9 and K value 1 to 5 and diff subsets.

```
Number of Training samples : 1000 , Number of Test samples :  10
K=1, Accuracy=0.0000, Training Time=0.0027, Prediction Time=262.5398
K=2, Accuracy=0.1000, Training Time=0.0012, Prediction Time=249.8638
K=3, Accuracy=0.1000, Training Time=0.0009, Prediction Time=252.5604
K=4, Accuracy=0.1000, Training Time=0.0018, Prediction Time=250.1809
K=5, Accuracy=0.1000, Training Time=0.0021, Prediction Time=252.1464
<Figure size 500x300 with 0 Axes>
```



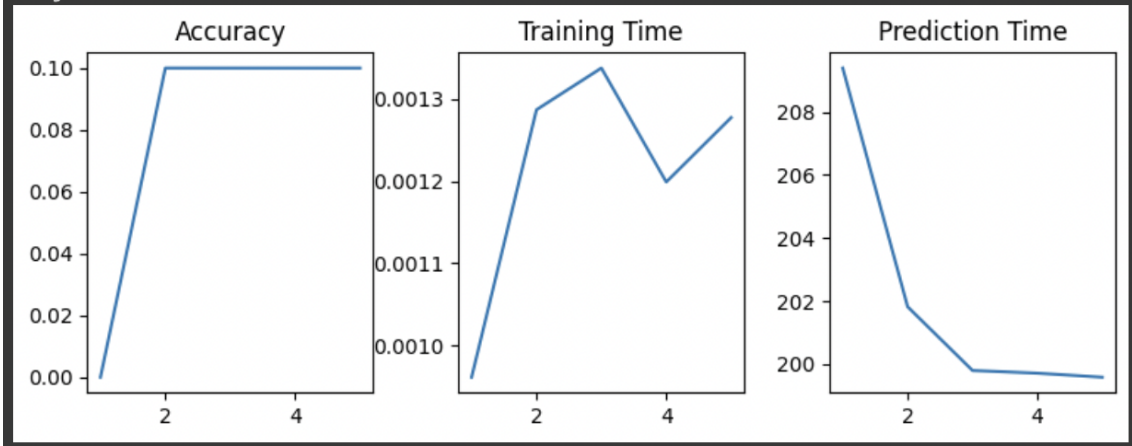

```
Peak memory usage: 1156.6953 MB , Accuracy Mean is 0.0800 , Accuracy Standard Deviation is 0.0400
**********************************************************************************************************
Number of Training samples : 1200 , Number of Test samples :  12
K=1, Accuracy=0.0000, Training Time=0.0012, Prediction Time=362.9151
K=2, Accuracy=0.0833, Training Time=0.0017, Prediction Time=360.9832
K=3, Accuracy=0.0833, Training Time=0.0019, Prediction Time=362.8318
K=4, Accuracy=0.1667, Training Time=0.0025, Prediction Time=367.7681
K=5, Accuracy=0.1667, Training Time=0.0016, Prediction Time=364.2754
<Figure size 500x300 with 0 Axes>
```



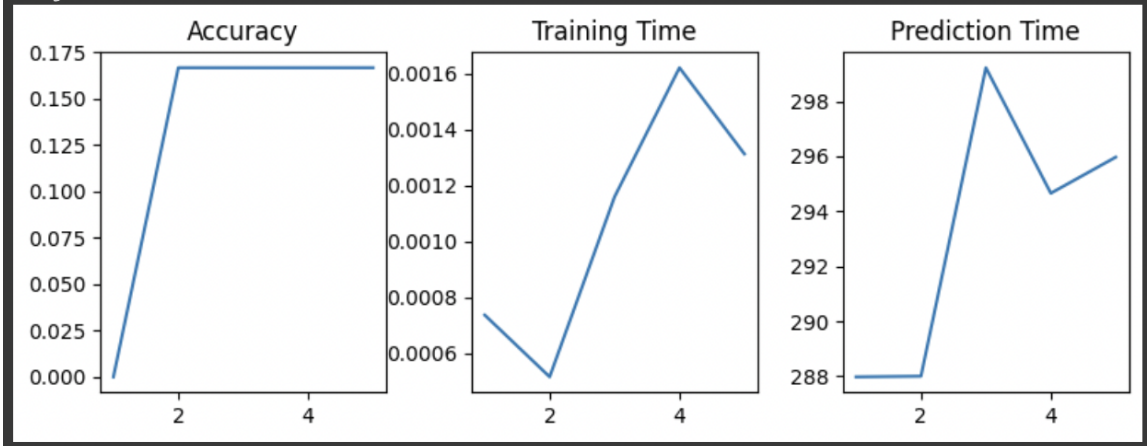

```
Peak memory usage: 1109.8906 MB , Accuracy Mean is 0.1000 , Accuracy Standard Deviation is 0.0624
**********************************************************************************************************
```

Signature Vector Length : 50 and “s” value : 0.9 and K value 1 to 5 and diff subsets.

```
Number of Training samples : 1000 , Number of Test samples :  10
K=1, Accuracy=0.0000, Training Time=0.0010, Prediction Time=209.3991
K=2, Accuracy=0.1000, Training Time=0.0013, Prediction Time=201.8149
K=3, Accuracy=0.1000, Training Time=0.0013, Prediction Time=199.7892
K=4, Accuracy=0.1000, Training Time=0.0012, Prediction Time=199.7042
K=5, Accuracy=0.1000, Training Time=0.0013, Prediction Time=199.5760
<Figure size 500x300 with 0 Axes>
```



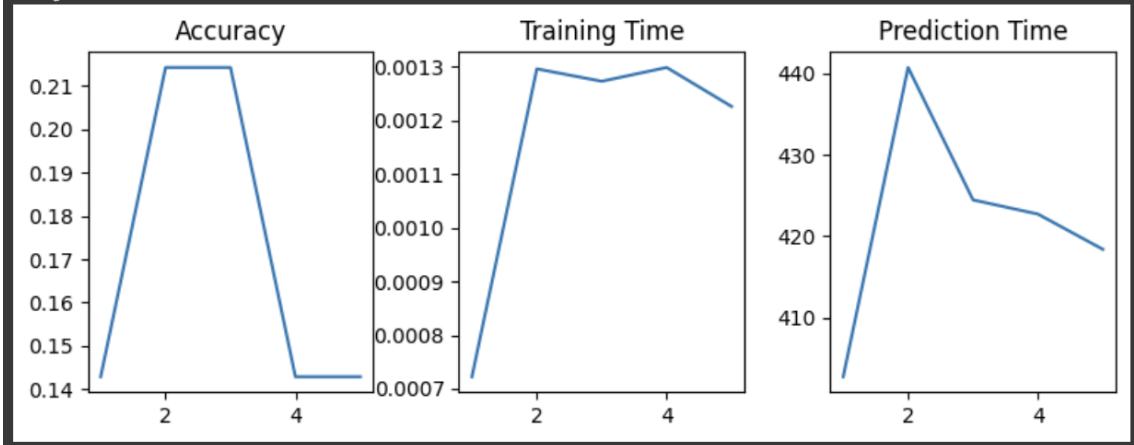

```
Peak memory usage: 1029.7930 MB , Accuracy Mean is 0.0800 , Accuracy Standard Deviation is 0.0400
**********************************************************************************************************
Number of Training samples : 1200 , Number of Test samples :  12
K=1, Accuracy=0.0000, Training Time=0.0007, Prediction Time=287.9797
K=2, Accuracy=0.1667, Training Time=0.0005, Prediction Time=288.0040
K=3, Accuracy=0.1667, Training Time=0.0012, Prediction Time=299.2189
K=4, Accuracy=0.1667, Training Time=0.0016, Prediction Time=294.6551
K=5, Accuracy=0.1667, Training Time=0.0013, Prediction Time=295.9666
<Figure size 500x300 with 0 Axes>
```



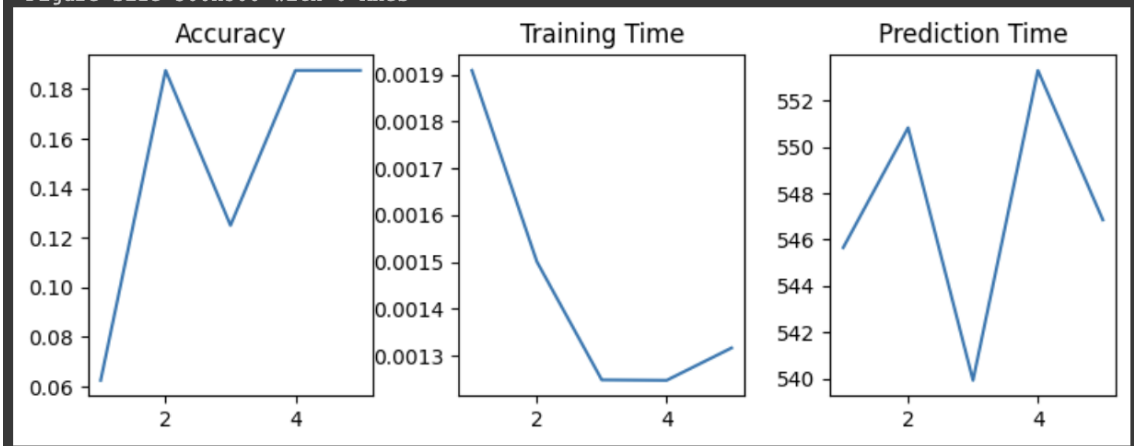

```
Peak memory usage: 1041.8438 MB , Accuracy Mean is 0.1333 , Accuracy Standard Deviation is 0.0667
**********************************************************************************************************
```

Signature Vector Length : 50 and “s” value : 0.9 and K value 1 to 5 and diff subsets.

```
**********************************************************************************************************************
Number of Training samples : 1400 , Number of Test samples :  14
K=1, Accuracy=0.1429, Training Time=0.0007, Prediction Time=402.7795
K=2, Accuracy=0.2143, Training Time=0.0013, Prediction Time=440.6854
K=3, Accuracy=0.2143, Training Time=0.0013, Prediction Time=424.4530
K=4, Accuracy=0.1429, Training Time=0.0013, Prediction Time=422.7125
K=5, Accuracy=0.1429, Training Time=0.0012, Prediction Time=418.3958
<Figure size 500x300 with 0 Axes>
```



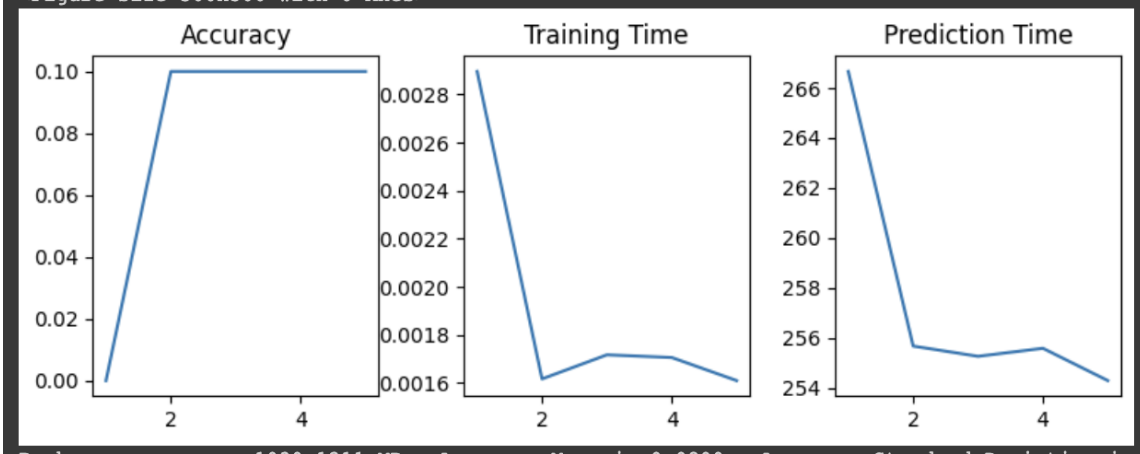

```
Peak memory usage: 1046.5586 MB , Accuracy Mean is 0.1714 , Accuracy Standard Deviation is 0.0350
**********************************************************************************************************************
Number of Training samples : 1600 , Number of Test samples :  16
K=1, Accuracy=0.0625, Training Time=0.0019, Prediction Time=545.6548
K=2, Accuracy=0.1875, Training Time=0.0015, Prediction Time=550.8295
K=3, Accuracy=0.1250, Training Time=0.0012, Prediction Time=539.9302
K=4, Accuracy=0.1875, Training Time=0.0012, Prediction Time=553.3029
K=5, Accuracy=0.1875, Training Time=0.0013, Prediction Time=546.8547
<Figure size 500x300 with 0 Axes>
```



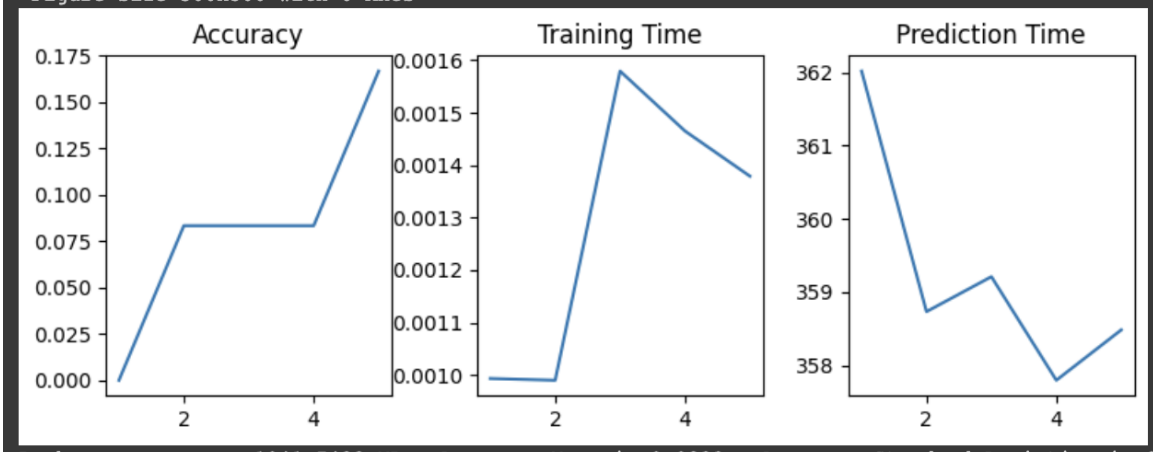

```
Peak memory usage: 1051.6484 MB , Accuracy Mean is 0.1500 , Accuracy Standard Deviation is 0.0500
**********************************************************************************************************************
```

Signature Vector Length : 60 and “s” value : 0.9 and K value 1 to 5 and diff subsets.

```
Number of Training samples : 1000 , Number of Test samples :  10
K=1, Accuracy=0.0000, Training Time=0.0029, Prediction Time=266.6620
K=2, Accuracy=0.1000, Training Time=0.0016, Prediction Time=255.6646
K=3, Accuracy=0.1000, Training Time=0.0017, Prediction Time=255.2535
K=4, Accuracy=0.1000, Training Time=0.0017, Prediction Time=255.5751
K=5, Accuracy=0.1000, Training Time=0.0016, Prediction Time=254.2814
<Figure size 500x300 with 0 Axes>
```



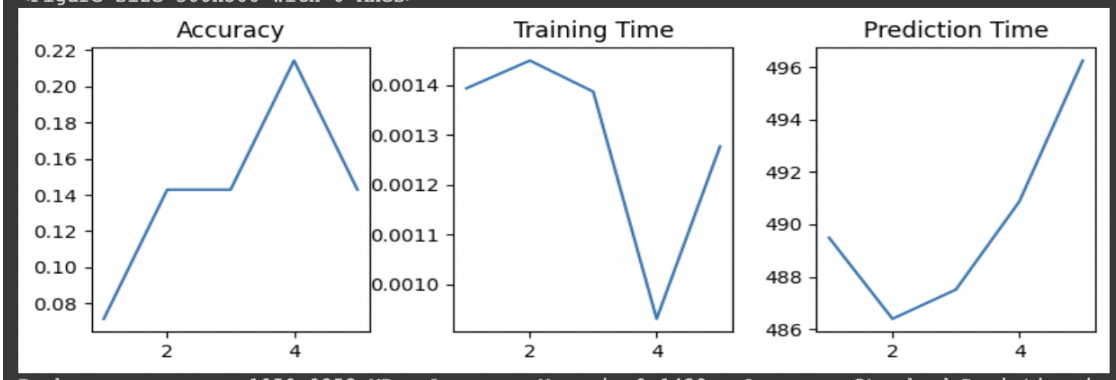

```
Peak memory usage: 1030.1211 MB , Accuracy Mean is 0.0800 , Accuracy Standard Deviation is 0.0400
**********************************************************************************************************
Number of Training samples : 1200 , Number of Test samples :  12
K=1, Accuracy=0.0000, Training Time=0.0010, Prediction Time=362.0164
K=2, Accuracy=0.0833, Training Time=0.0010, Prediction Time=358.7308
K=3, Accuracy=0.0833, Training Time=0.0016, Prediction Time=359.2079
K=4, Accuracy=0.0833, Training Time=0.0015, Prediction Time=357.7933
K=5, Accuracy=0.1667, Training Time=0.0014, Prediction Time=358.4813
<Figure size 500x300 with 0 Axes>
```



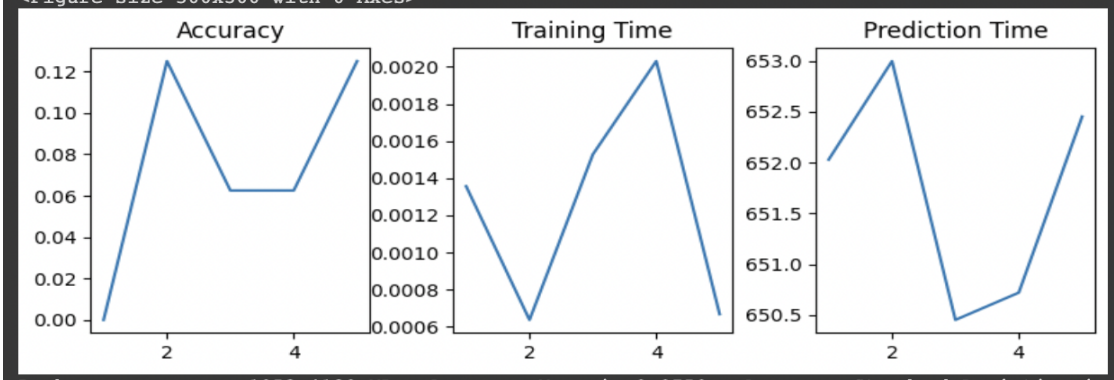

```
Peak memory usage: 1041.7422 MB , Accuracy Mean is 0.0833 , Accuracy Standard Deviation is 0.0527
**********************************************************************************************************
```

Signature Vector Length : 60 and “s” value : 0.9 and K value 1 to 5 and diff subsets.

```
Number of Training samples : 1400 , Number of Test samples :  14
K=1, Accuracy=0.0714, Training Time=0.0014, Prediction Time=489.4894
K=2, Accuracy=0.1429, Training Time=0.0014, Prediction Time=486.3960
K=3, Accuracy=0.1429, Training Time=0.0014, Prediction Time=487.5108
K=4, Accuracy=0.2143, Training Time=0.0009, Prediction Time=490.8646
K=5, Accuracy=0.1429, Training Time=0.0013, Prediction Time=496.2440
<Figure size 500x300 with 0 Axes>
```




```
Peak memory usage: 1050.0352 MB , Accuracy Mean is 0.1429 , Accuracy Standard Deviation is 0.0452
**********************************************************************************************************
Number of Training samples : 1600 , Number of Test samples :  16
K=1, Accuracy=0.0000, Training Time=0.0014, Prediction Time=652.0275
K=2, Accuracy=0.1250, Training Time=0.0006, Prediction Time=652.9968
K=3, Accuracy=0.0625, Training Time=0.0015, Prediction Time=650.4524
K=4, Accuracy=0.0625, Training Time=0.0020, Prediction Time=650.7203
K=5, Accuracy=0.1250, Training Time=0.0007, Prediction Time=652.4517
<Figure size 500x300 with 0 Axes>
```




```
Peak memory usage: 1052.4180 MB , Accuracy Mean is 0.0750 , Accuracy Standard Deviation is 0.0468
**********************************************************************************************************
```

[Solution2_Colab_Notebook_Link]

Question 3 : Compare (1) and (2) in terms of classification accuracy, time required, and peak RAM required.

Solution 3 :

Comparison between the two:-

In terms of Time Required : Since Jaccard_KNN (Question1) uses all the inbuilt library functions and the functions are optimized for the numpy arrays the time taken for the subsets and even for the whole dataset is very low whereas the time taken for the LSH_KNN

(Question2) is quite high for a small subset of the data. The reason for such behavior is that **Knn inbuilt library** inherently takes **only one test sample at a time** and finds the distance with all the labeled training points.

Let t be the number of test samples and x be the number of training samples
To find similarity value and classify 1 test point, the number of pairs is equal to (1*x).
MNIST dataset contains 10K test samples and 60K training samples, the pairs will be (10K * 60K) = **600K pairs.**
We calculated the time taken for a subset with 1500 training points and 20 test points then 30000 pairs created and this takes 12 minutes to execute knn (when numpy is used) and 15-20 minutes (when tensor- pytorch is used, reason being that there is no direct inbuilt KNN library for tensor data).
To find 1-nearest neighbor with x=1500 and t=20, time taken is 12 minutes.
To find 1-nearest neighbor with x=60K and t=10K, time taken will be **1000 hours**.

In terms of Accuracy : The Jaccard_KNN has shown a maximum of 96 percent accuracy whereas the LSH_KNN accuracy is stuck at 10 percent even with various combinations of signature vectors with similarity threshold values 0.8 and 0.9. The reason for such behavior is the training data used was a small sample size thus resulting in deficiency of variation in training samples. Hence, the testing data points could not be classified properly.

In terms of Peak RAM required : The peak RAM requirement of Jaccard_KNN is higher than that of LSH_KNN. The reason for this behavior is that the LSH using MinHash algorithm and the amount of memory requirement for two sample distance finding is lower in case of Jaccard_KNN than the LSH_KNN.

Summary Table:-

Question Number	Combination of Signature Vector and S-value	Max Dataset Size	Max Accuracy	Max Prediction Time	Peak RAM
Jaccard_KNN	NA	(60K, 10K)	93.6%	345 sec	803 MB
LSH_KNN	(40, 0.8)	(1600,16)	10%	645 sec	1047 MB
LSH_KNN	(50, 0.8)	(1600,16)	10%	670 sec	1052 MB
LSH_KNN	(60, 0.8)	(1600,16)	10%	644 sec	1052 MB
LSH_KNN	(40, 0.9)	(1600,16)	10%	645 sec	1120 MB
LSH_KNN	(50, 0.9)	(1600,16)	10%	546 sec	1051 MB
LSH_KNN	(60, 0.9)	(1600,16)	10%	652 sec	1052 MB

Other Observations:-

- Lesser the training time used, less will be the prediction time usually.
- Lesser the training time, higher is the accuracy usually.

References:-

http://ekzhu.com/datasketch/minhash.html
https://github.com/ekzhu/datasketch
https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/
https://pypi.org/project/datasketch/