Machine Intelligence and Expert Systems

Computer Assignment 4

K- Nearest Neighbours

Name: Anand Jhunjhunwala
Roll Number: 17EC35032

Aim: Define and implement the function to return k-Nearest Neighbours with k=1, 3, 5 & 7 on cancer dataset.

Code Specification:

Train_data_fraction = 0.85
Distance matrix:

1) Euclidean Distance
2) Normalized Euclidean Distance
3) Cosine Similarity

Value of K used: 1, 3, 5, 7

Results:
Note: Change in seed value will result in change in training and test data due to different reshuffling.

For seed = 421
Accuracy:

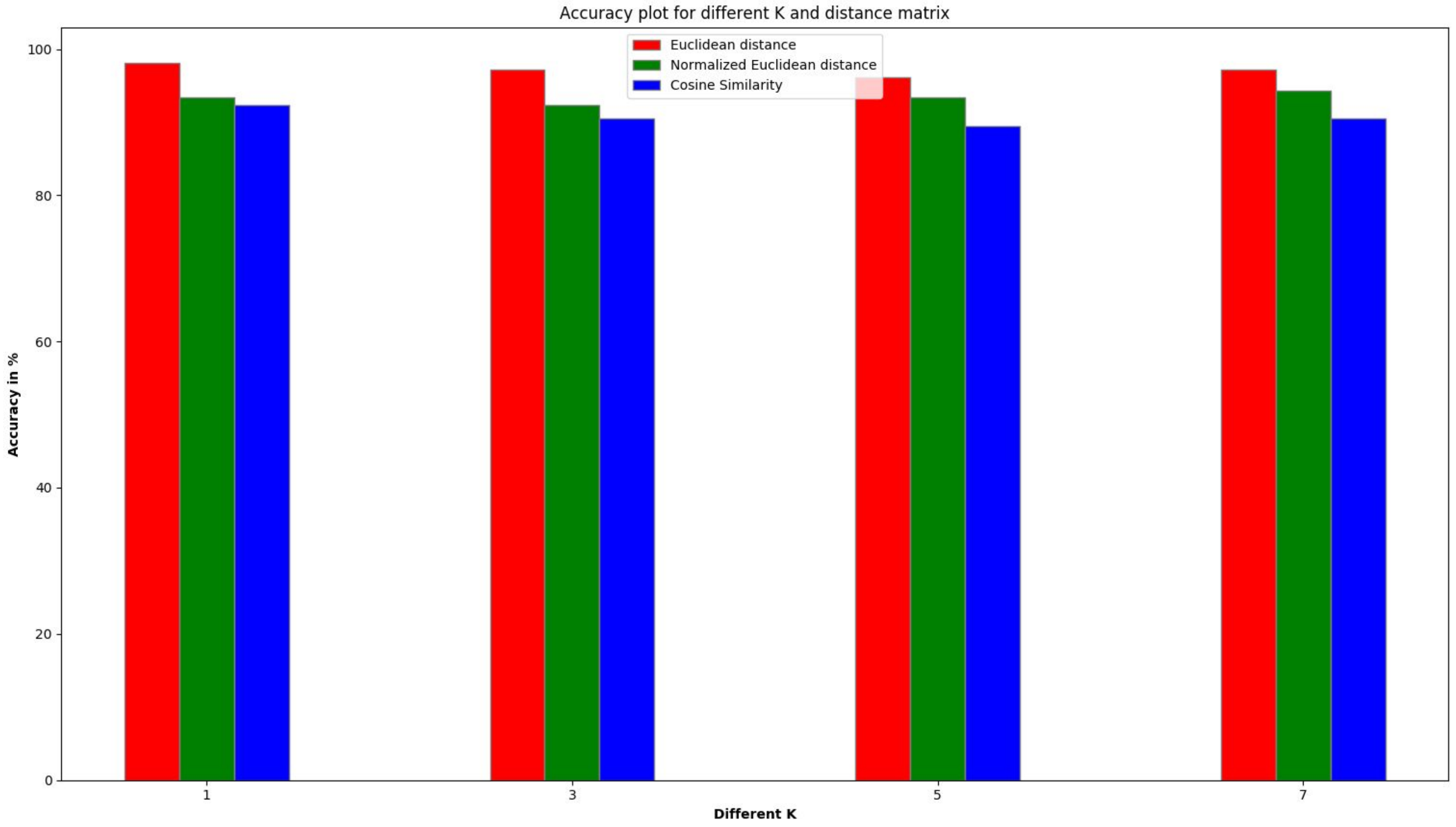
Distance Matrix	K=1	K=3	K=5	K=7
Euclidean	98.10%	97.14%	96.19%	97.14%
Norm Euclidean	93.33%	92.38%	93.33%	94.29%
Cosine Similarity	92.38%	90.48%	89.52%	90.48%

For seed = 300
Accuracy:

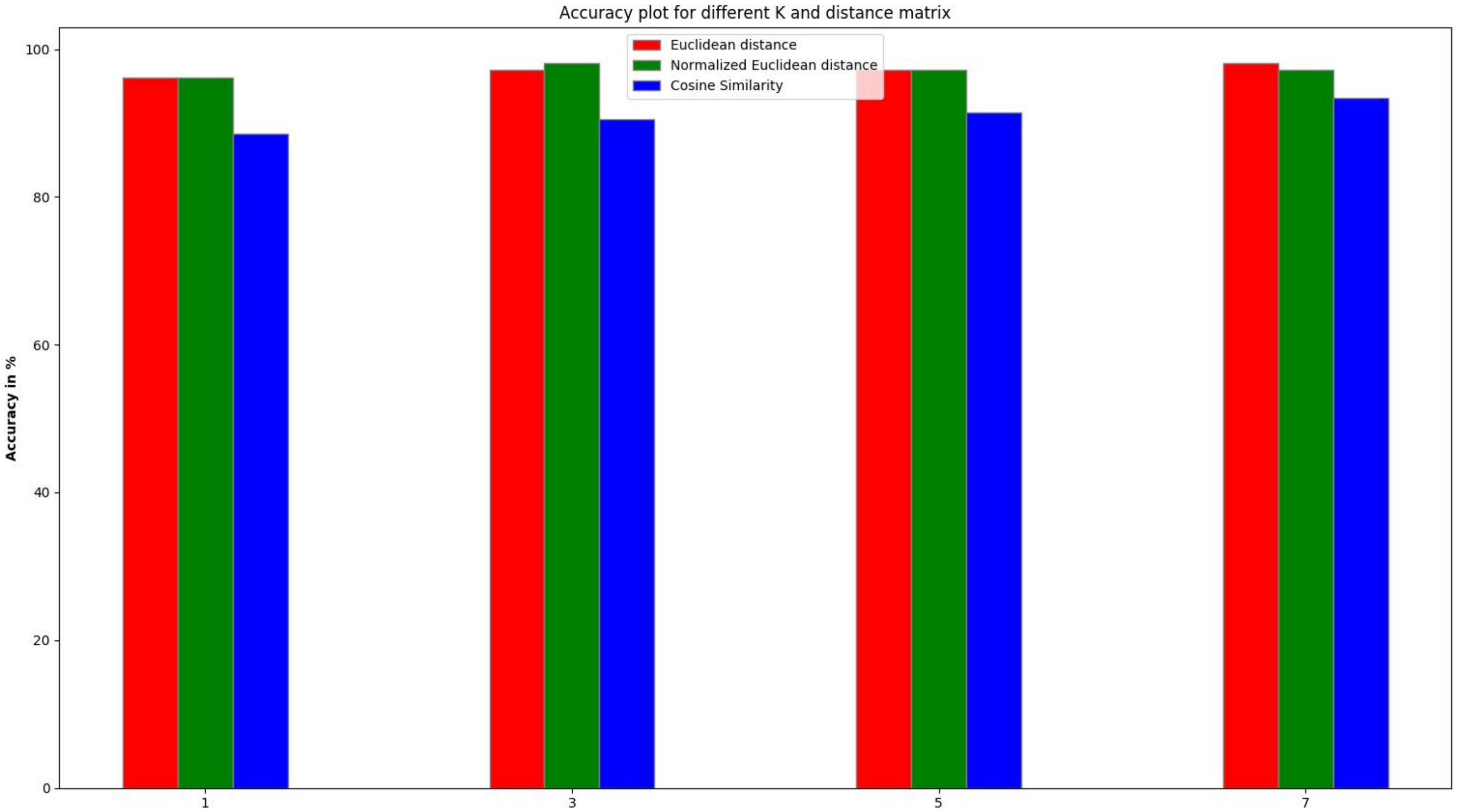
Distance Matrix	K=1	K=3	K=5	K=7
Euclidean	96.19%	97.14%	97.14%	98.10%
Norm Euclidean	96.19%	98.10%	97.14%	97.14%
Cosine Similarity	88.57%	90.48%	91.43%	93.33%

Bar Graphs:

For seed: 421



For seed: 300



Different K

Discussion

- From the result I observed that the accuracy in case of euclidean and normalized euclidean matrix are very close to each other indicating both being a good candidate for distance function.
- In the case of cosine similarity I observed that accuracy is significantly less for both seed value this is due to the fact that in data set, there are some data points with value of **bare_nuclei** as -99999 which results in a -ve cosine similarity value, which in turn makes it farthest to test data, this may affect the accuracy.
- Accuracy results are almost similar for each K value.
- I used PriorityQueue to reduce the time of classification from O(K*n) to O(nlogn)