Husna Manalai
Machine Learning
Homework 1
k-nearest Neighbors Algorithm

Results:

For when k= 1 the error rate was: 0.09866666666666667
For when k= 3 the error rate was: 0.08666666666666667
For when k= 5 the error rate was: 0.088
For when k= 7 the error rate was: 0.09466666666666666

The best k value was k= 3 because it had the lowest error rate.

I used the same test data and train data for 1, 3, 5, and 7, which was mnist_train.csv for the train data, mnist_test.csv for the test data and 6500 training points and 1500.

The digits that the algorithm most commonly confused were 4 (it misclassified 4 as 9 around 10 times), 2 (it misclassified 2 as 1 around 7 times), and 7 (it misclassified 7 as 1 around 10 times), etc.

To improve the performance:

One way we could improve the performance would be to use an Advanced Distance Metric such as Mahalanobis distance for the kNN algorithm. Unlike the Euclidean distance, which treats all dimensions equally, the Mahalanobis distance takes into account the variance of each dimension and the covariance[1] between dimensions.

This can be particularly useful in digit classification, where certain pixel positions might be more significant than others in determining the similarity between digits.

To compute the Mahalanobis distance, you first need to calculate the covariance matrix of the dataset. This matrix provides insights into the data's variance and the covariance between pairs of features.

[1] A metric indicating how two variables move in relation to each other and the degree to which they vary together.

Cited Work:
https://www.statisticshowto.com/mahalanobis-distance/