

3. Accuracy is 0.955

4. Accuracy is 0.969

5. The accuracy is different because uses a slightly different subset of training data. The first run trains off 80% of the data and tests 20%. The second one trains off 20% and tests 80%. The second run, although it has less training, did slightly better overall, but still pretty comparable.

6. Potential Issues with:

- How the data was collected- the collection of digits in digits.csv are from multiple sources. The difference in collecting these images can be attributed to different scanners, resolutions, etc.
- Issues presumably stem from handwriting differences demographically speaking: (right handed vs left, adult vs children, large vs small sample size)
- It seems to have issues with numbers that have loops or curves like how 3, 8, 9 got misread as 7, (9, 1, 1) and 5. The as the map averages the strength of color that would otherwise be blank (the loop/curved part) making the program think it is often another number.

7. Done

8. (a)- we chose to set  $k=3$ , because its a small odd number. A single neighbor ( $k=1$ ) would be overfit, and any larger values would risk blending similar digits.

(b)- the model achieved 96.6%, classifying almost all the digits. compareLabels showed the errors made between visually similar digits (such as 8 and 9), therefore  $k=3$ 's performance was the most reliable.

9. (a) No. The best  $k$  differed across random seeds. Seed: 8675309 gave  $k = 3$ , while seed: 5551212 and seed: 42 both gave  $k = 1$ . This small variation happened because each random split changes which samples appeared in the set, slightly affecting which  $k$  performed the best.

(b) We chose  $k = 3$  as the best overall value because, As aforementioned, using three neighbors provides a slightly more stability while keeping similarly high. And although  $k = 1$  performed very well in two splits,  $k = 3$  achieved the highest accuracy (0.993) for at least one seed.

10. Printed output from the terminal below:

--- Step 10: Training and testing final k-NN model ( $k = 3$ ) ---

Predicted 354 labels using  $k = 3$ .

Correct: 349 out of 354

Accuracy: 0.986

First 10 incorrect predictions:

row 46 : predicted 4   actual 9

row 220 : predicted 1   actual 8

row 258 : predicted 1   actual 6

row 326 : predicted 1   actual 9

row 330 : predicted 6   actual 5