

## Programming assignment 2

Implement nearest k nearest neighbor for two datasets:

1. The test dataset provided
  1. Three different error rates (mis classifications)
  2. Non
  3. Small
  4. Medium
2. The abalone dataset
3. Logistics:
  1. The programming assignment will be due by the end of class on Monday, April 18th
  2. The code and confusion matrices (in a readable format)
  3. You may work with one other person on the programming projects. Go to gradescope and register your partner (or say you are working alone) along with the programming language you have selected
    1. Java, Python or Julia

Problem Description:

1. Download the dataset from: <https://archive.ics.uci.edu/ml/datasets/Abalone>
2. This tarball contains 3 artificial datasets from Piazza
  1. The last column in the dataset is the label
  2. Gzipped tar ball
3. There is also a readme file
  1. The attribute descriptions are your features
  2. For example Sex is the first feature in the observation
- Implement K-Nearest Neighbor to classify each file
- Vary K to achieve highest result
- Use any necessary variable calculation methods to transform the features that we have presented in class
4. Next you must test your algorithm
  1. Note the 'artificial datasets'. What results should you expect from each? If your results vary from what is expected, please explain when you show your confusion matrices

Some helpful hints for the assignment

1. Design techniques will be very helpful because you will be running the same set of algorithms, outputting the results for many runs on many files
2. Please have fun.
  1. The algorithm should be straight forward
  2. This is more about getting all of the techniques in place in order to run training algorithms