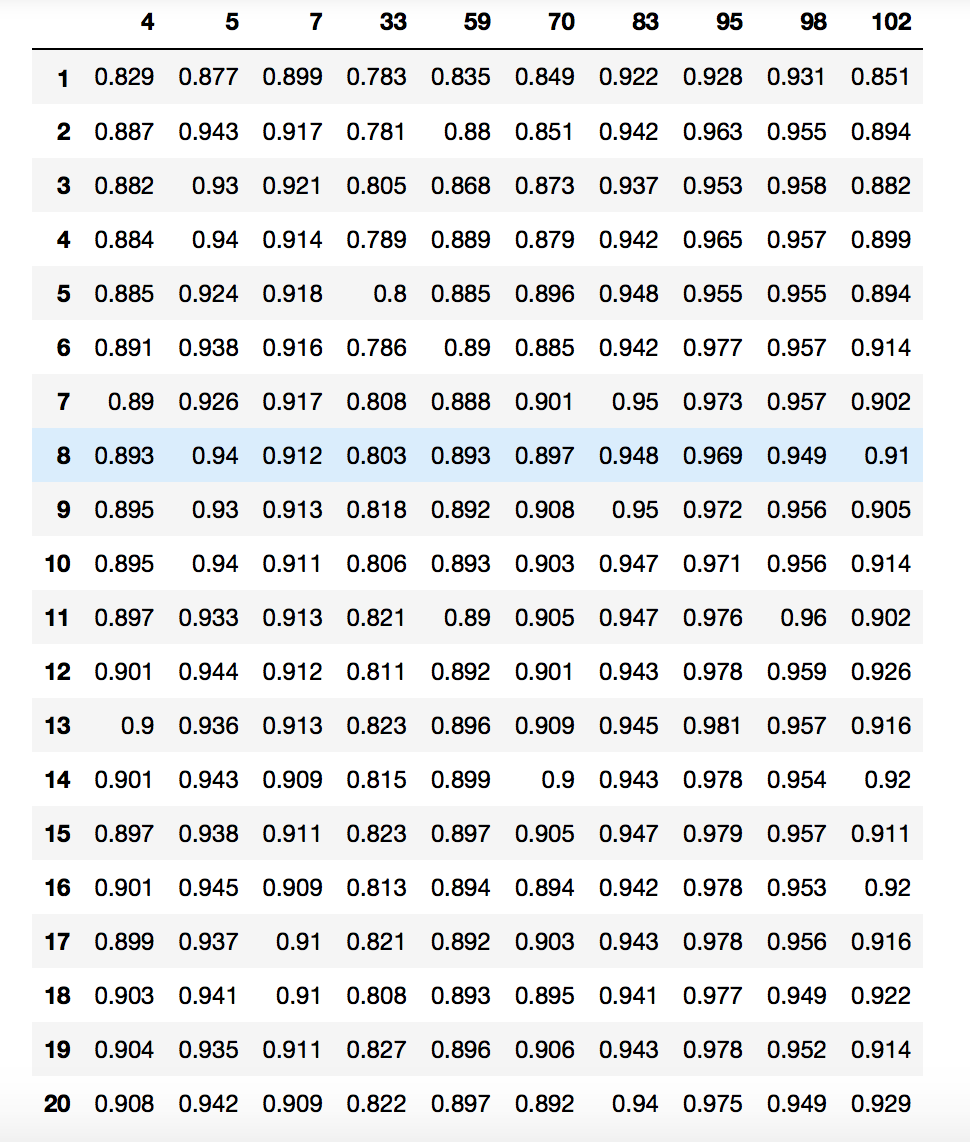
# K-Nearest-Neighbors and Naive Bayes

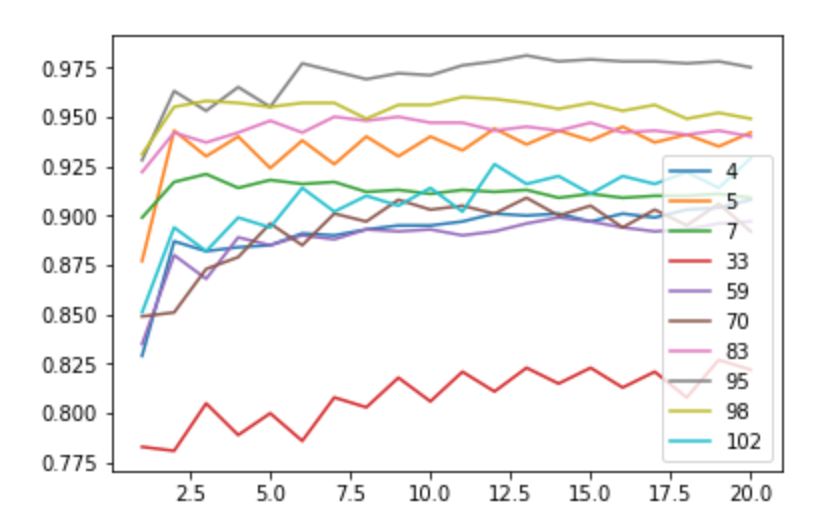
## Classification with kNN

Finding the value of k for each label was done by sweeping across all labels we were concerned with and all values of k from 1 to 20. Training (on the train set) and predicted (on the validation set) based on each label and then scoring the predictions. This yielded the matrix shown here:



This matrix shows the score (0 – 1 with 0 being all predictions are wrong and 1 being all predictions are correct) of each label with each value of k. (the rows represent values of k and the columns represent the id of the labels).

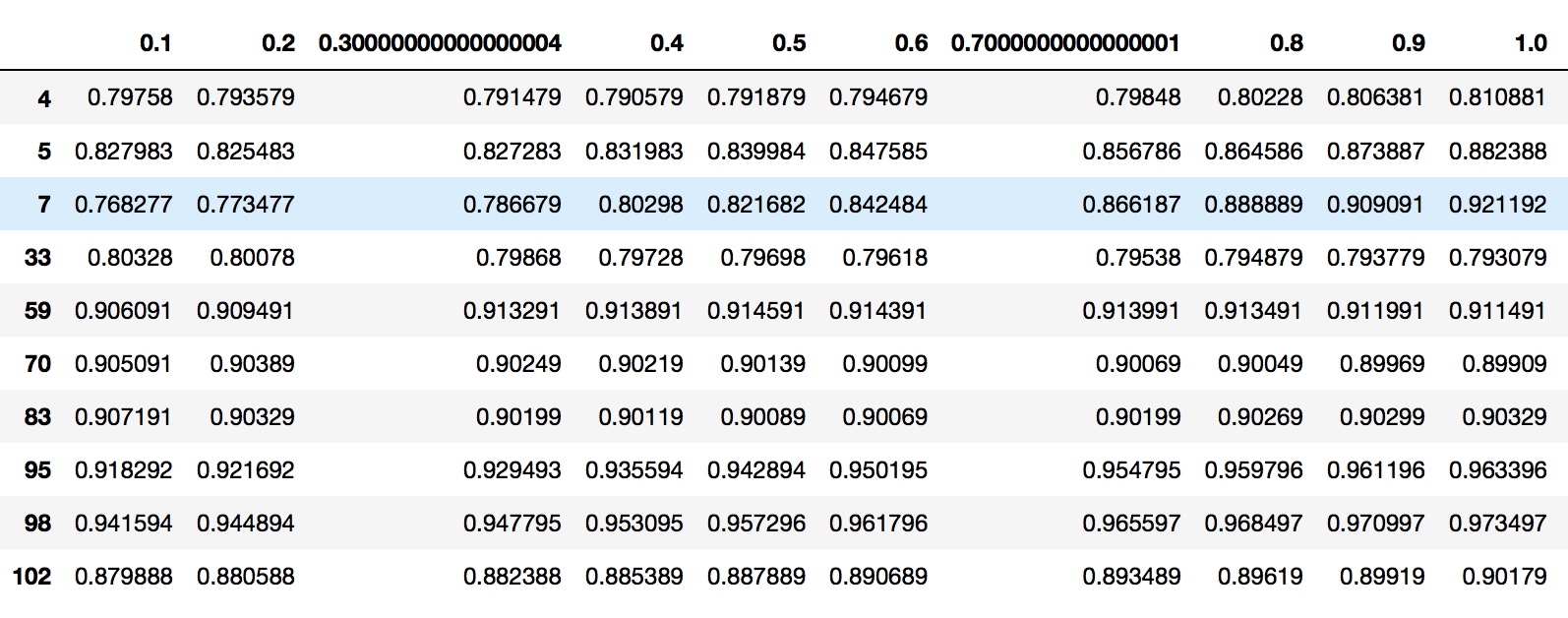
For each label, a graph of the score vs the value of k was then produced using df.plot().

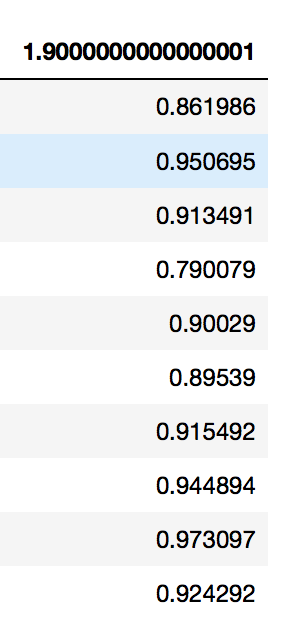
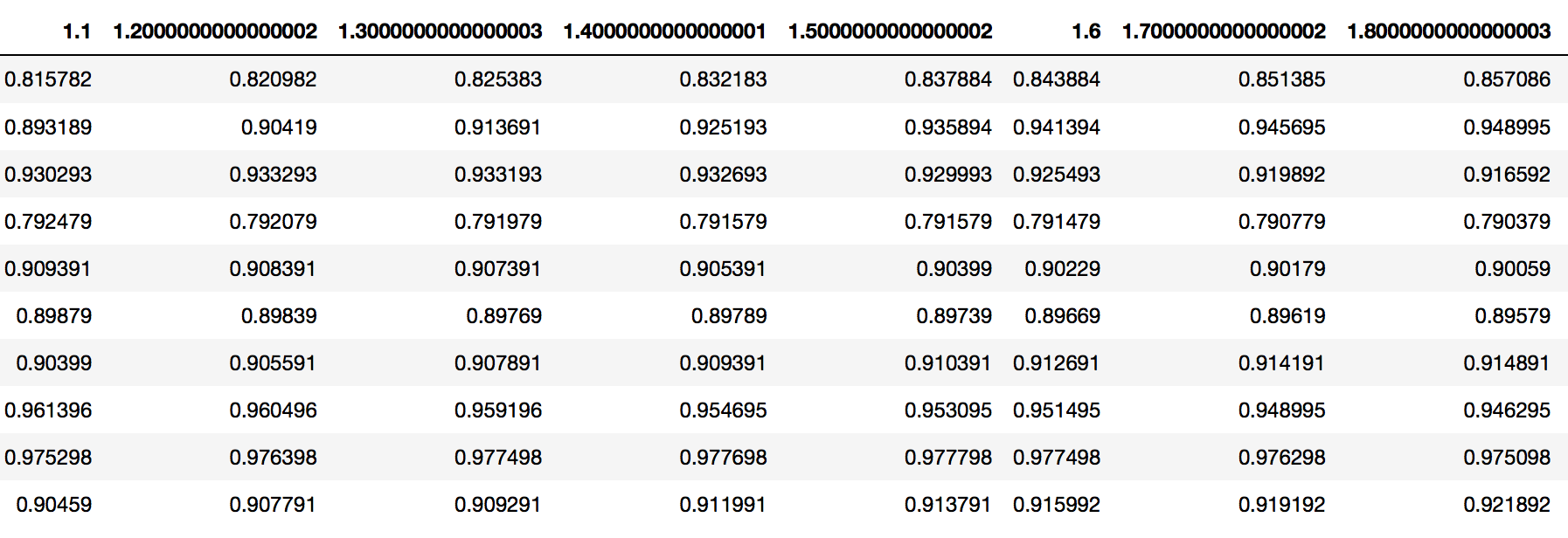


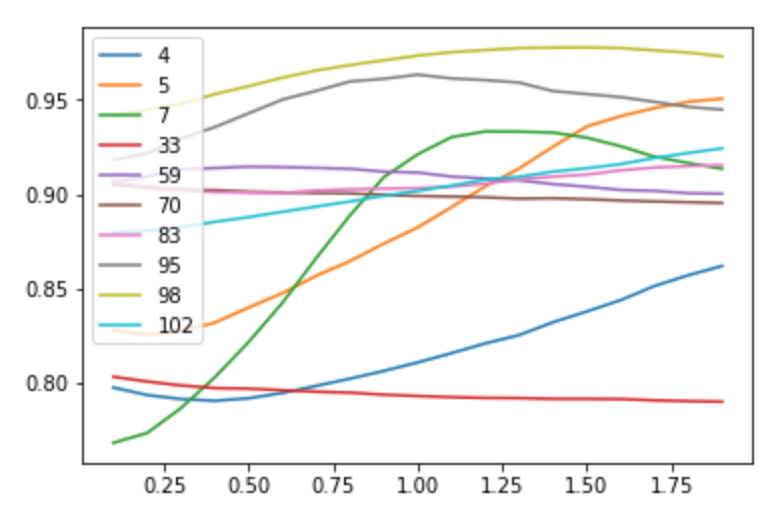
From this graph and table, an appropriate value of k was found for each class. These were then recorded in the ‘ks.txt’ file.

## Classification with NB

For the Nayve Bayse Bernoulli Classification, the parameter alpha was scanned for. Performing a similar test as in the kNN Classification. All labels were scanned across for values of alpha ranging from 0.1 to 1.9 in increments of 0.1. The compiled scores for each model are found in the table below:







This graph shows the relationship between the score as the value of alpha goes from 0.1->1.9. This clearly shows that there can be found a “best value” of alpha at roughly 1.25. This is therefore what was taken for the default value of alpha in the Naïve Bayes Bernoulli classifier.

## Data Analysis

Based on the parameter estimations done in the previous two sections it can be clearly seen that the label with id 33 is the hardest to classify with both classifiers. This can be because the data points with label 33 are