CS464 Homework1 Report
Çerağ Oğuztüzün 21704147

Q3.1
MATCH: 501
FAIL: 27
ACCURACY: 94.88636363636364

Q3.2
3 th 8mer -> pos: 20 - 21
40 th 8mer -> pos: 316 - 317
58 th 8mer -> pos: 460 - 461
77 th 8mer -> pos: 612 - 613
129 th 8mer -> pos: 1028 - 1029
181 th 8mer -> pos: 1444 - 1445
194 th 8mer -> pos: 1548 - 1549
213 th 8mer -> pos: 1700 - 1701
293 th 8mer -> pos: 2340 - 2341
313 th 8mer -> pos: 2500 - 2501
318 th 8mer -> pos: 2540 - 2541
339 th 8mer -> pos: 2708 - 2709
340 th 8mer -> pos: 2716 - 2717
360 th 8mer -> pos: 2876 - 2877
364 th 8mer -> pos: 2908 - 2909
374 th 8mer -> pos: 2988 - 2989
445 th 8mer -> pos: 3556 - 3557
463 th 8mer -> pos: 3700 - 3701
465 th 8mer -> pos: 3716 - 3717
480 th 8mer -> pos: 3836 - 3837

Q3.3
Most confident 8mer with positive cleavage: APGTSDEN
at index: 360
with confidence: -17.313659730372596
Least confident 8mer with negative cleavage: RKHWYFCM
at index: 409
with confidence: -29.967623100215004
The sequences vary, which shows that the confidence of cleavage for each sequence varies mostly due to the sequence of the amino acids.

Q3.4

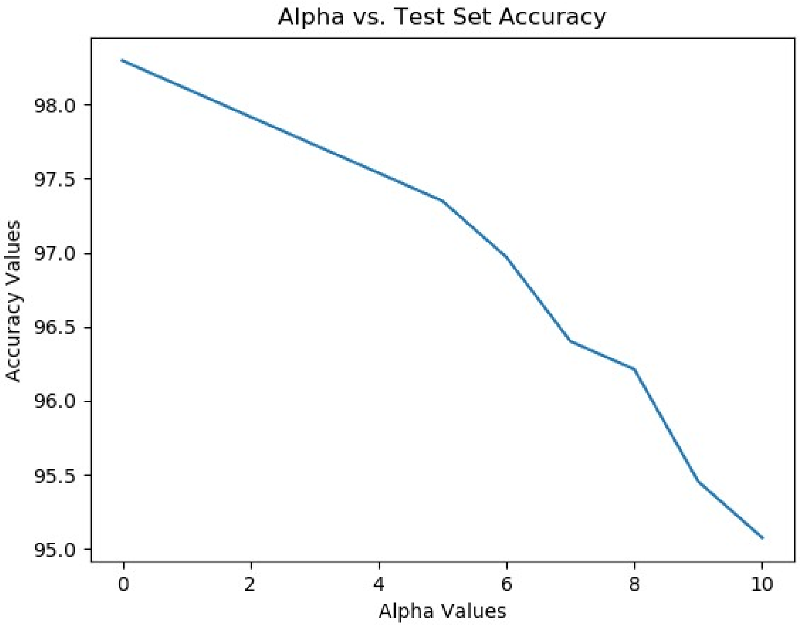

Figure1: Alpha vs. Test Set Accuracy Plot using whole data set

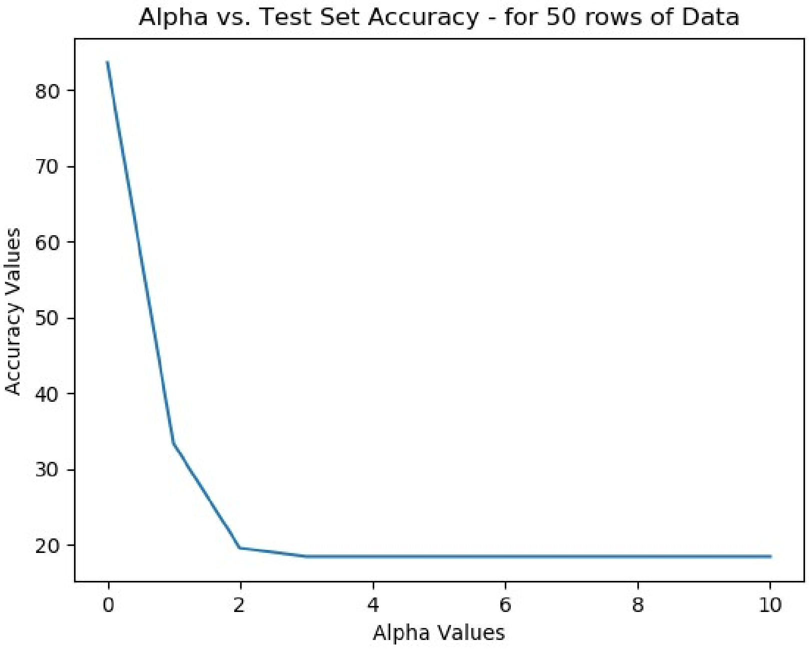

Figure2: Alpha vs. Test Set Accuracy Plot using first 50 rows of data

When the whole dataset is used MAP estimate decreases accuracy as can be seen from the first plot. However when we only used the first 50 rows of the dataset, we observe a dramatic accuracy decrease observed from the 2nd plot at alpha values between 0 and 1. Additive smoothing improves results by preventing conditional probability to be 0 for some data. When we alter the sample size the prior distribution changes and therefore we see a decrease in performance with respect to alpha values, this case can be seen from the 2nd plot where 50 rows of data were used.

Q3.5
k value that yielded the most accuracy: 158
with accuracy: 98.10606060606061
k value that yielded the most accuracy was 158 with accuracy 98.106% which is greater than the accuracy value obtained in part 3.1 which was 94.508%.

Q3.6

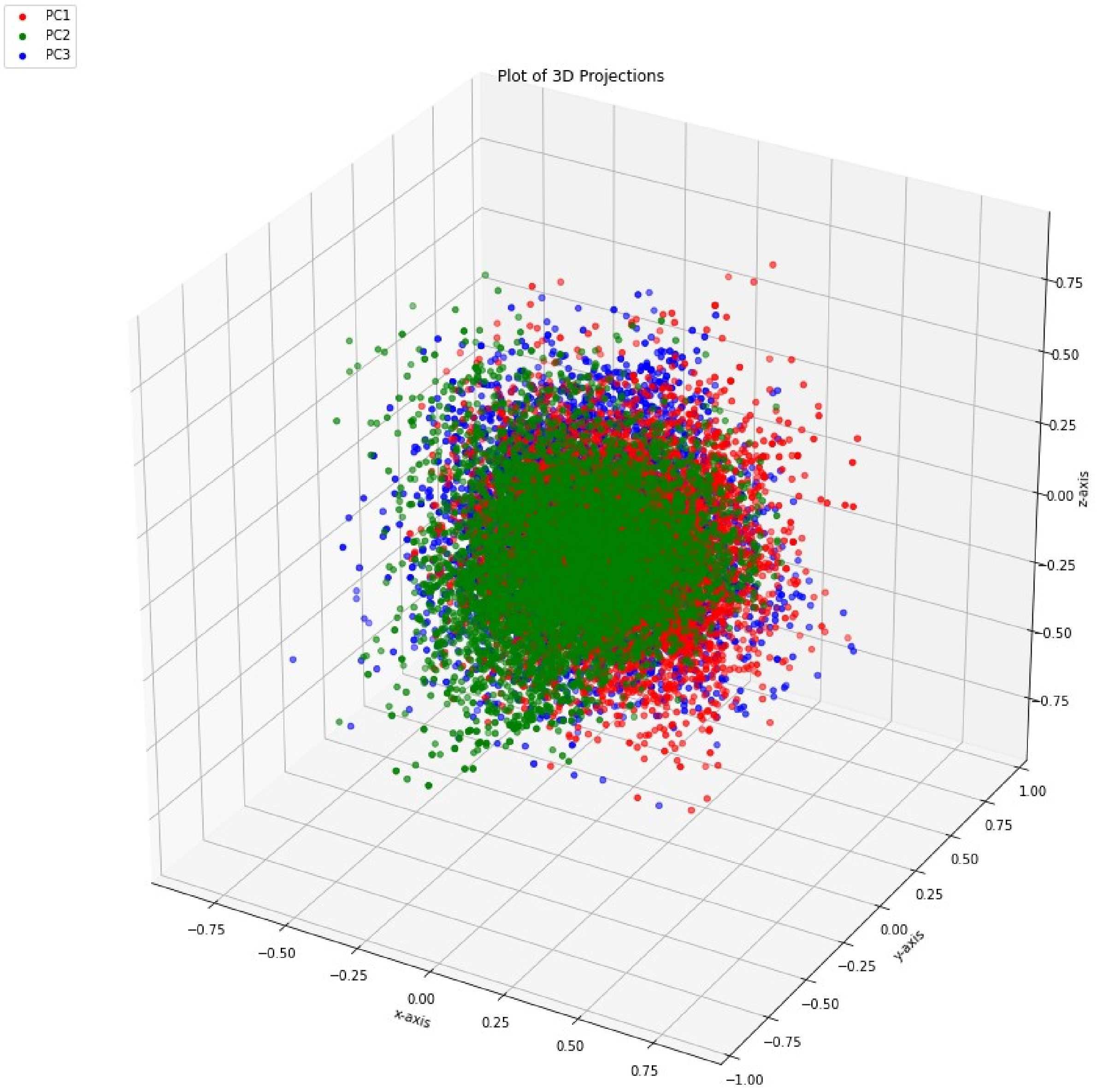

Figure3: 3D Plot of Projected Data

% of variance covered for whole data is: 0.06980266880746583
% of variance covered for PC1 is: 0.8484916243371633
% of variance covered for PC2 is: 0.6967519983703445
% of variance covered for PC3 is: 0.7181714868140181

Variance coverage for whole set is 6% of variance for original dataset. From the PC data the variances go up to 84%, it can be seen that variance is maximized for all 3 principle components. However, from the 3D Plot of projected data, I conclude that it is not feasible to use PCA regarding that the data is still not easily separable.