Project 1 Write Up

Task A

For Task A, I split the selected 'ABCDE' data using a 30 train/9 test split. The predicted values are below with the errors highlighted in red:

y_test: [1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5]

KNN: [1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 4, 1, 2, 2, 2, 5, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 3, 5, 5, 5, 5, 3, 5, 3, 3, 5]

Centroid: [1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 4, 2, 2, 2, 2, 5, 2, 3, 3, 5, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 2, 4, 4, 4, 3, 5, 5, 5, 5, 3, 5, 5, 5, 5]

Linear Regression: [-1, 1, 0, 1, 1, 0, 1, 1, 1, 2, 2, 5, 1, 1, 4, 3, 4, 3, 4, 3, 3, 4, 2, 3, 3, 3, 2, 5, 6, 5, 6, 2, 6, 4, 6, 4, 6, 5, 4, 5, 3, 7, 5, 5, 5]

SVM: [1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 5, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 3, 5, 5, 5, 5, 5, 5, 5, 5, 5]

Task B

For Task B, I ran 5-fold cross-validation against KNN, Centroid, Linear Regression and SVM classifiers with a split of 5 train/5 test samples on 10 classes. The reported accuracies are below:

Classifier	CV #1	CV #2	CV #3	CV #4	CV #5	Average
KNN	1.0	1.0	1.0	1.0	1.0	1.0
Centroid	0.89	1.0	1.0	1.0	1.0	0.98
Linear Regression	0.56	0.56	0.67	0.78	0.56	0.63
SVM	1.0	1.0	1.0	1.0	1.0	1.0

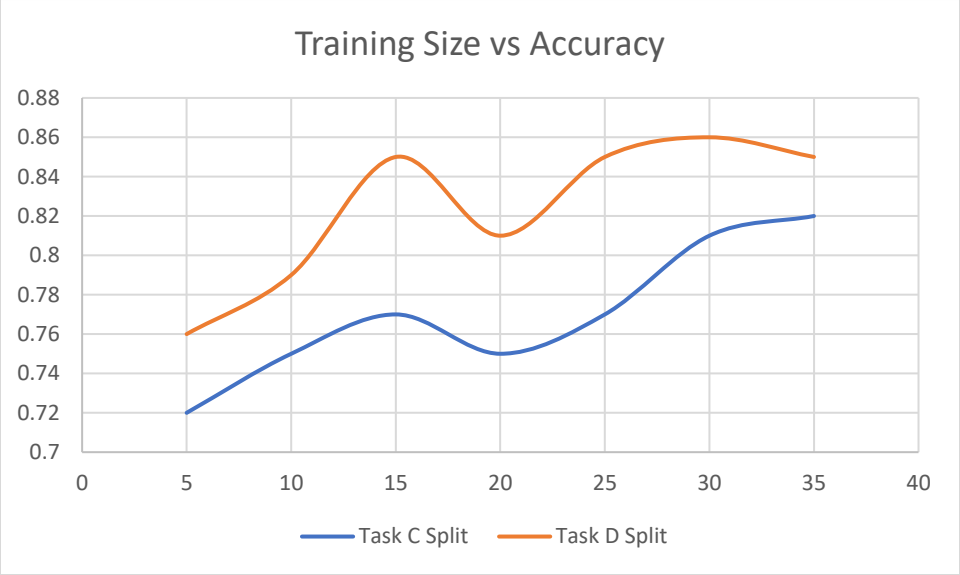
From these numbers, I observed that KNN and SVM had perfect accuracy followed closely by Centroid. Linear Regression had a surprisingly low accuracy that I was not expecting. This low number could be attributed to the code not being implemented correctly but I was able to increase the average accuracy by 7% by increasing the number of training samples.

In summary, it appears that linear regression may just not be a model that is as accurate as the other classifiers or the implementation of the code needs to be improved.

Task C and D

After using the dataHandler to generate training and test splits for the 7 different scenarios, the following accuracies were obtained by running my Centroid model on each of the splits:

Train Size	Task C	Task D
5	0.72	0.76
10	0.75	0.79
15	0.77	0.85
20	0.75	0.81
25	0.77	0.85
30	0.81	0.86
35	0.82	0.85



What I observed from the Task C split is that in general, the accuracy went up as more and more training samples were made available to the model. This makes sense because it is generally true that the more a model can learn from training data, the better it will perform against test data.

This trend was also observed when performing the Task D splits. The interesting thing to note is that compared to Task C, my centroid model performed a lot better in comparison on its 10 splits versus the splits. This leads credence to the idea that the more the model can learn from training data, the better it generally performs.