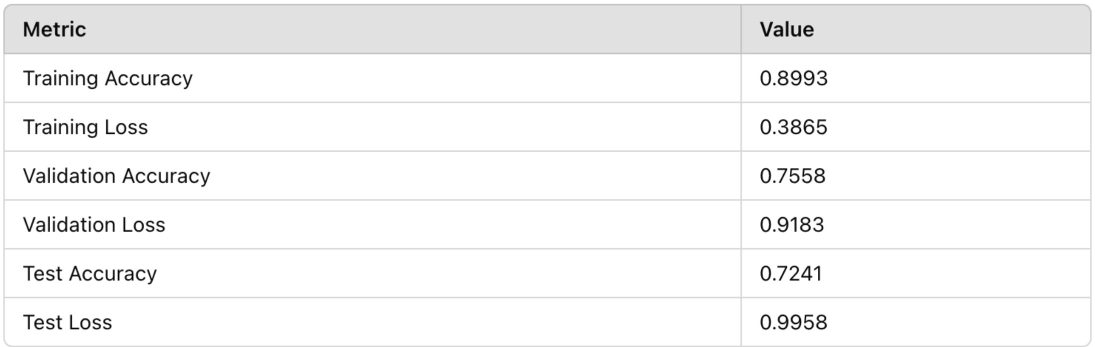
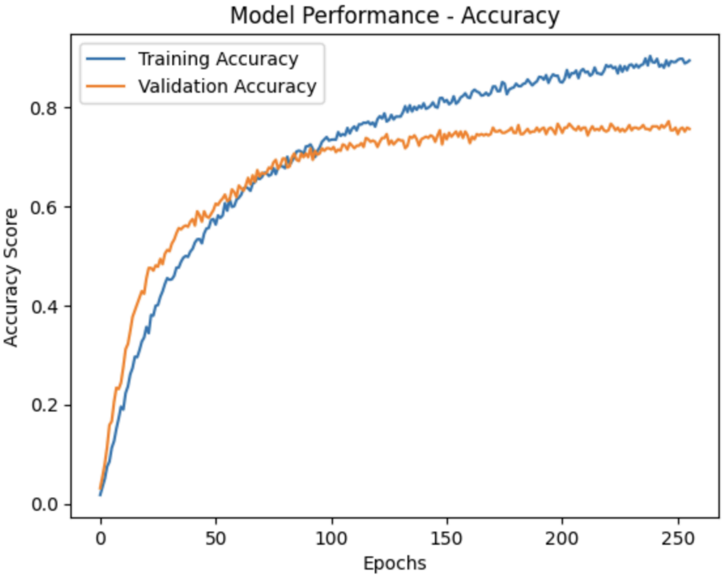
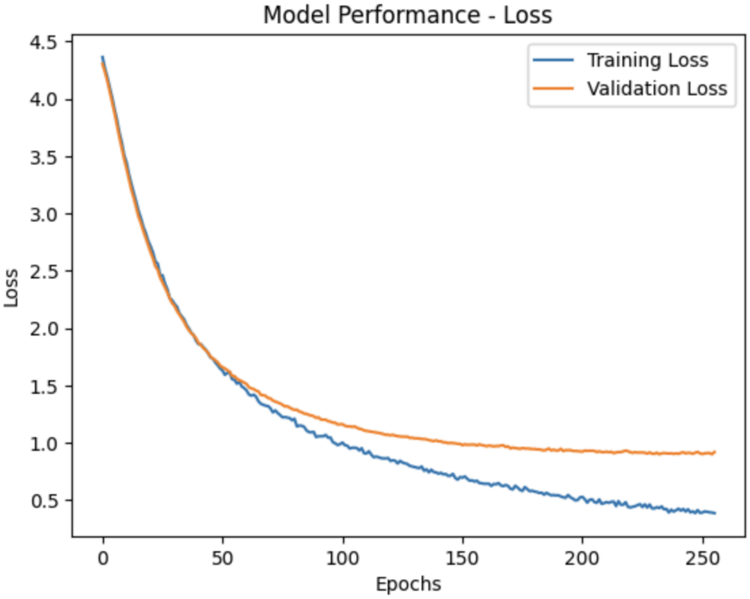
1) Training with minimum number of samples for each class: approx. 26%

- The validation set of 385 samples (5 samples each for all 77 classes) are drawn..
- Due to class imbalances, I first trained the models on the minimum number of samples (n_samples = 34) for each class from the training dataset.
    - The size of the training set = 2618
- The test sample set of 1/3rd of the testing set is also drawn from the testing set.
- The training and testing sets are both balanced.
- Then I trained a neural network model with an input layer of 64 neurons, a dropout layer of 0.3, and an output layer with softmax as an activation function for a multiclass classification task for num_class = 77 classes.
- The model's performance is shown below:

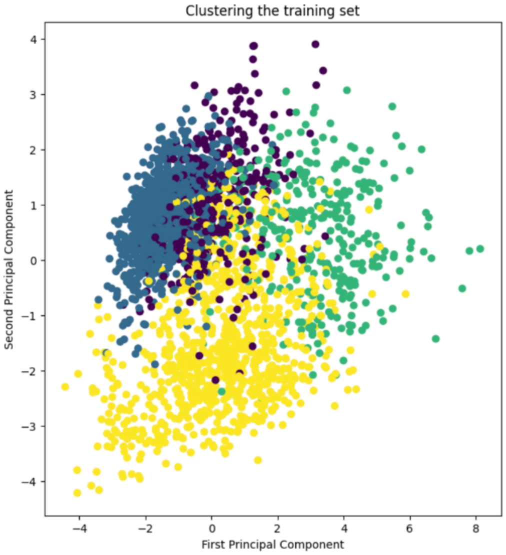


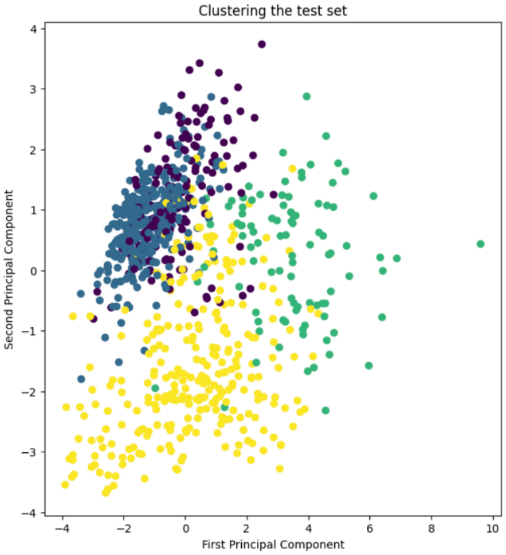
Metric	Value
Training Accuracy	0.8993
Training Loss	0.3865
Validation Accuracy	0.7558
Validation Loss	0.9183
Test Accuracy	0.7241
Test Loss	0.9958

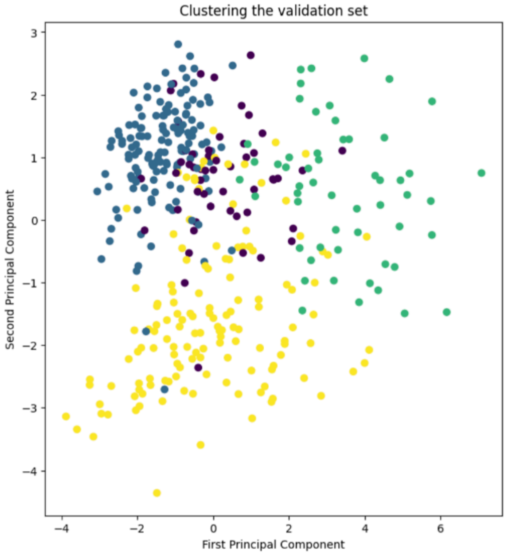




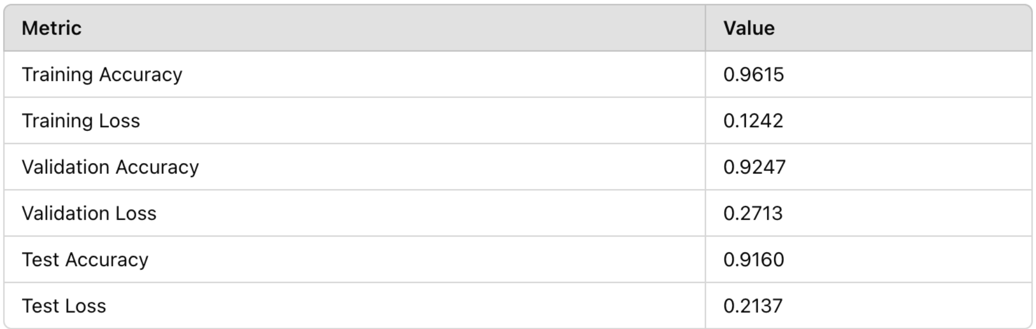
- Then I use hierarchical clustering (ward method) to form 4 clusters of the datasets. The clusters can be visualized as below along their first two principal components.



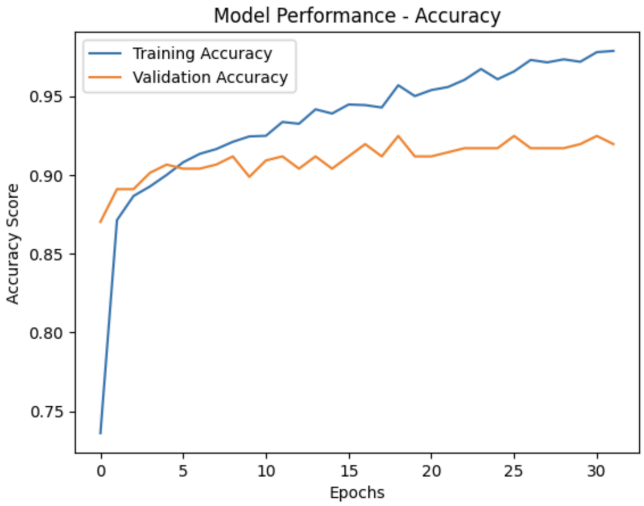


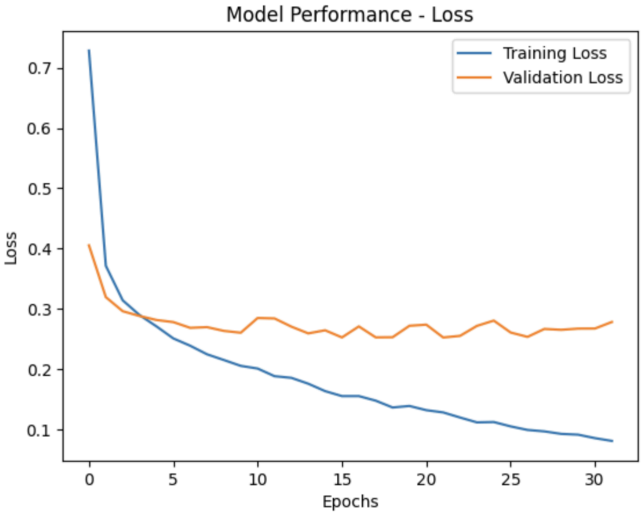


- Then I trained a neural network model with an input layer of 16 neurons and an output layer with softmax as an activation function for a multiclass classification task for num_class = 4 classes, for classifying the data into 4 distinct clusters.
- The model's performance is shown below:

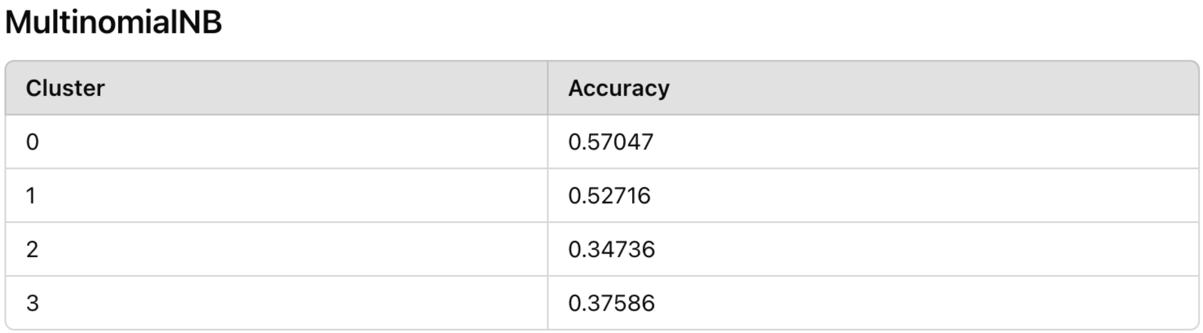


Metric	Value
Training Accuracy	0.9615
Training Loss	0.1242
Validation Accuracy	0.9247
Validation Loss	0.2713
Test Accuracy	0.9160
Test Loss	0.2137



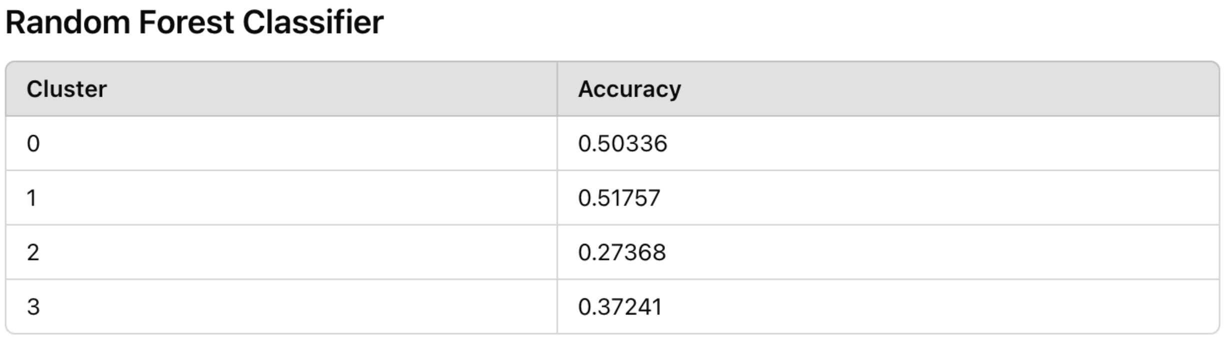


- Then I attempted classification of the labels within each individual cluster using MultiNomialNB and Random Forest Classifier:
- The initial performance of the models are shown below:



MultinomialNB

Cluster	Accuracy
0	0.57047
1	0.52716
2	0.34736
3	0.37586



Random Forest Classifier

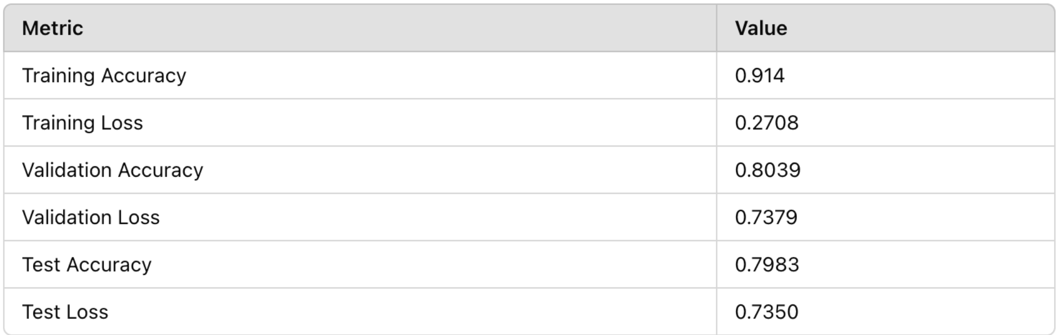
Cluster	Accuracy
0	0.50336
1	0.51757
2	0.27368
3	0.37241

- The model accuracy in the classification of the 77 labels within each of the 4 clusters is not well. I hypothesize that this is due to the homogeneity of the data within each cluster and lack of enough training samples.
-

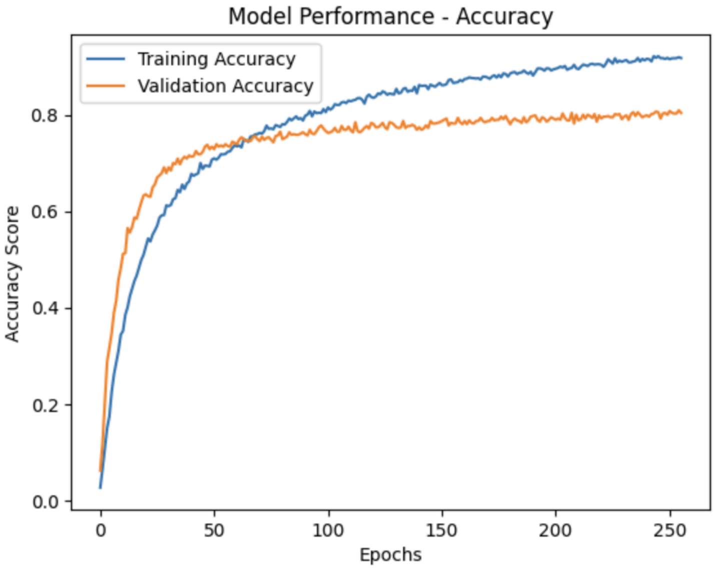
2) Increasing the training samples from 2618 to 5540: approx. 60%

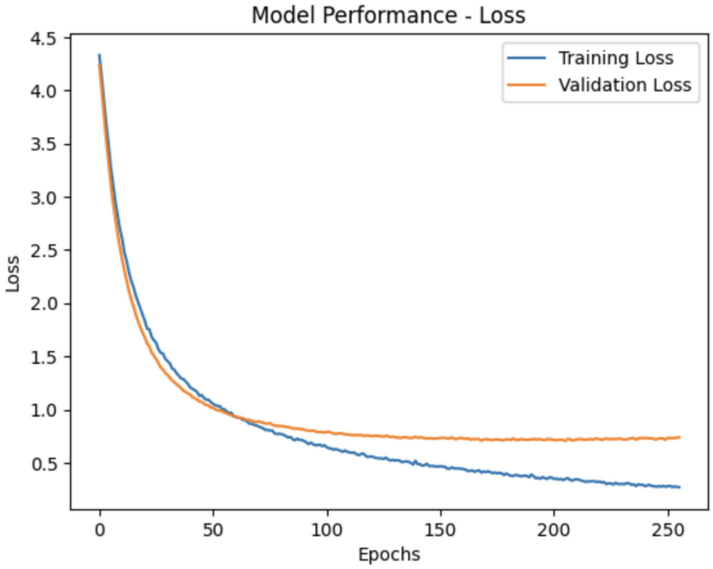
- I increased the number of training samples to 4089, and the performance of different models on the earlier classification tasks are shown below:
- Model’s performance on the classification of 77 classes increased for training, validation, and testing sets, as shown below:

-

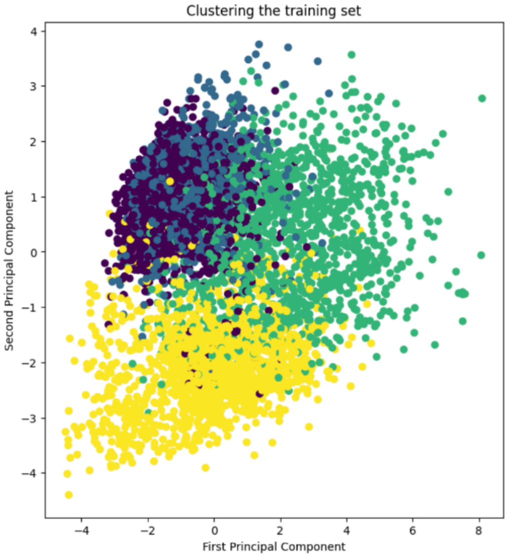


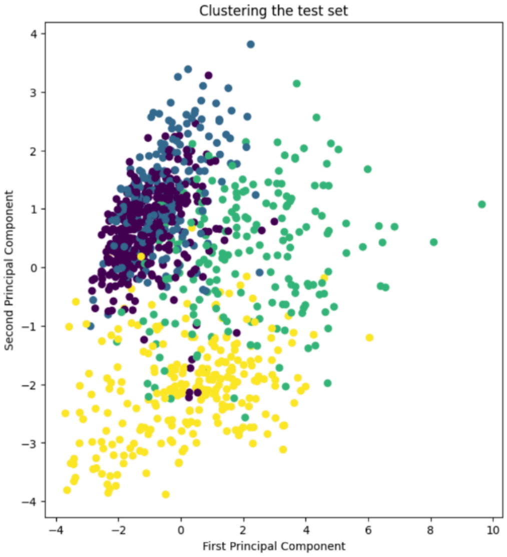
Metric	Value
Training Accuracy	0.914
Training Loss	0.2708
Validation Accuracy	0.8039
Validation Loss	0.7379
Test Accuracy	0.7983
Test Loss	0.7350

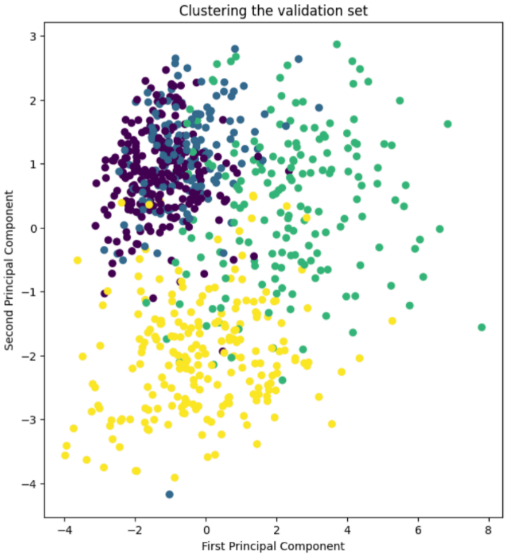




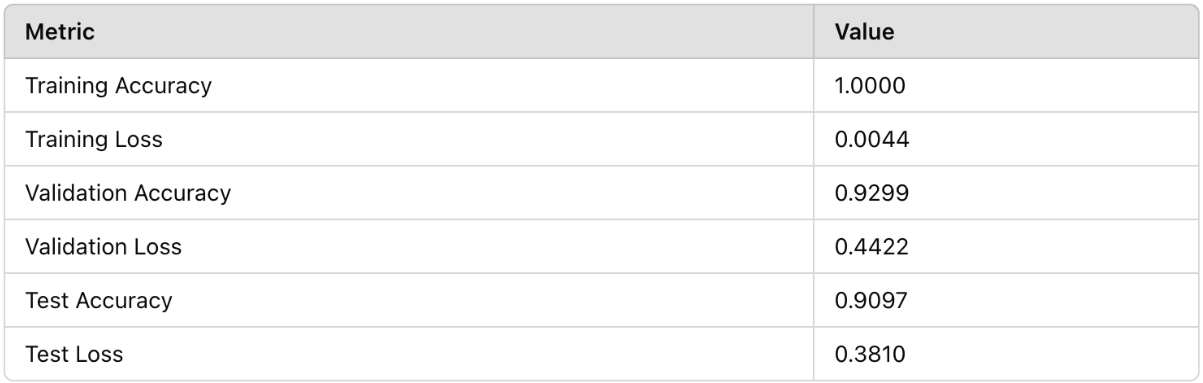
- Then I use hierarchical clustering (ward method) to form 4 clusters of the datasets. The clusters can be visualized as below along their first two principal components.





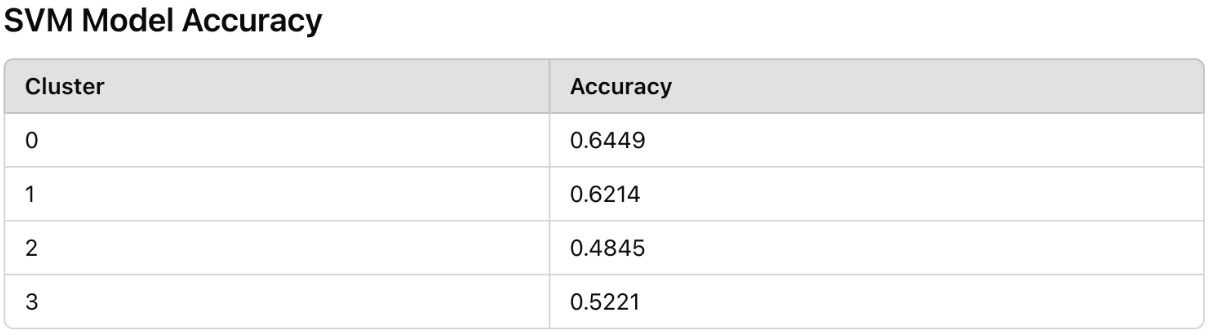


- The classification of the 4 clusters also increased as well, as shown below:



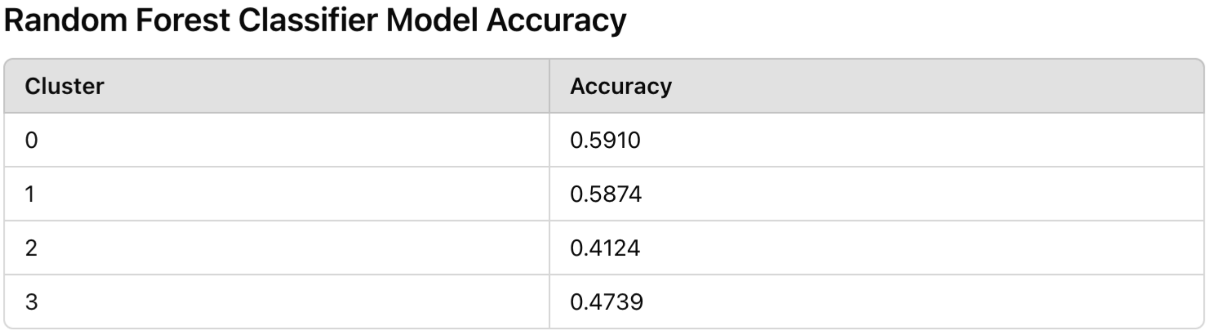
Metric	Value
Training Accuracy	1.0000
Training Loss	0.0044
Validation Accuracy	0.9299
Validation Loss	0.4422
Test Accuracy	0.9097
Test Loss	0.3810

- The performance of classification of the labels within the clusters didn't increase significantly. I alternatively also used SVM that resulted in slightly better performance.



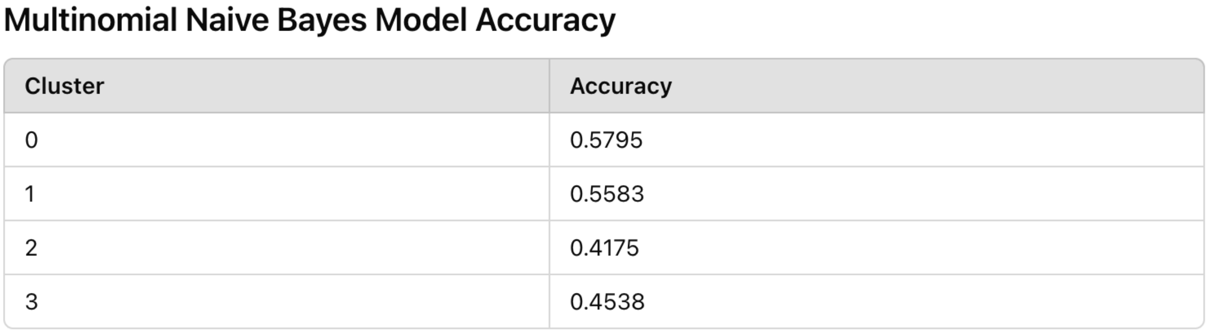
SVM Model Accuracy

Cluster	Accuracy
0	0.6449
1	0.6214
2	0.4845
3	0.5221



Random Forest Classifier Model Accuracy

Cluster	Accuracy
0	0.5910
1	0.5874
2	0.4124
3	0.4739

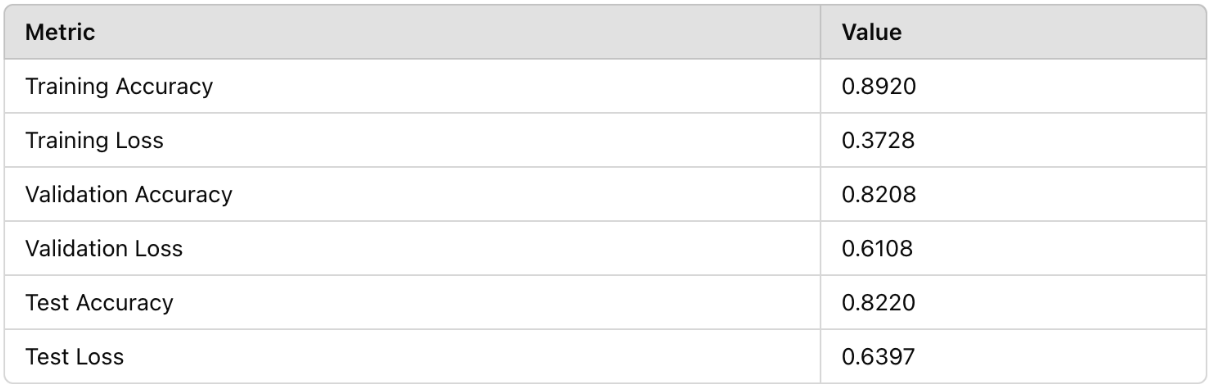


Multinomial Naive Bayes Model Accuracy

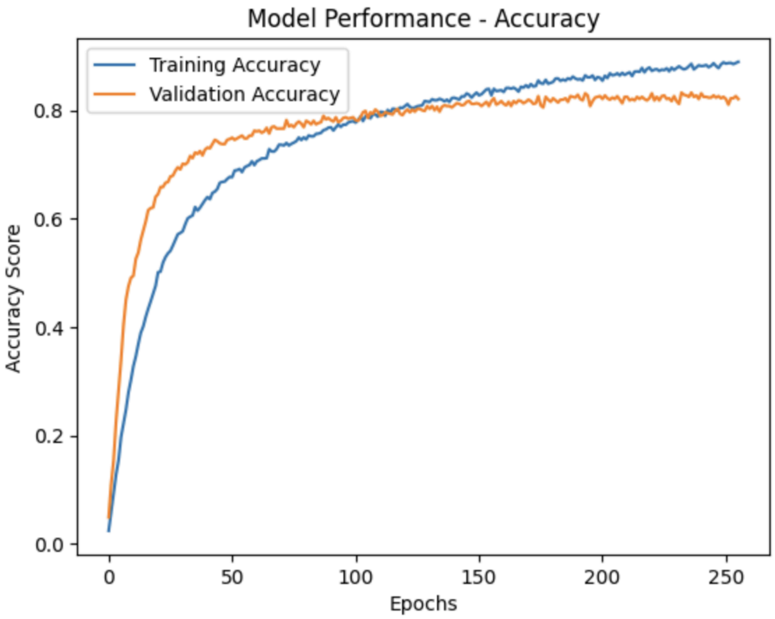
Cluster	Accuracy
0	0.5795
1	0.5583
2	0.4175
3	0.4538

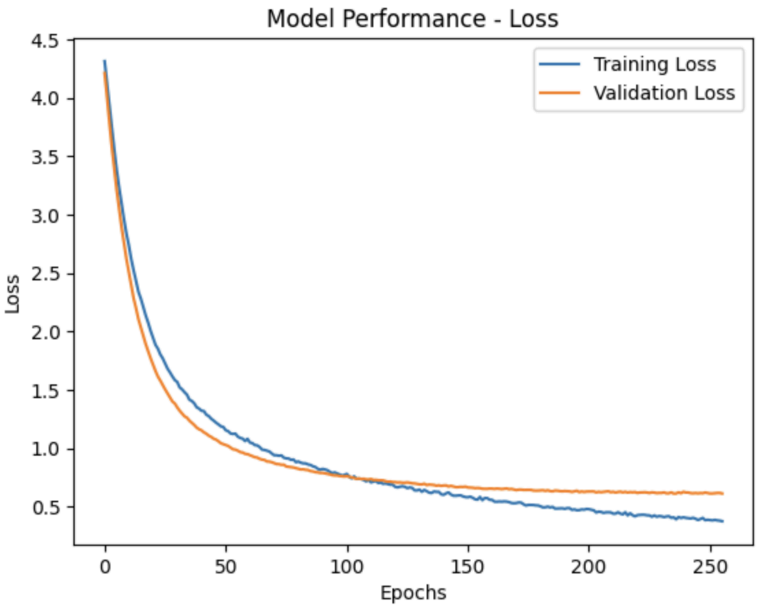
3) Increasing the training samples from 5540 to 7385: approx. 80%

- I increased the number of training samples to 7385, and the performance of different models on the earlier classification tasks are shown below:
- Model’s performance on the classification of 77 classes increased for training, validation, and testing sets, as shown below:

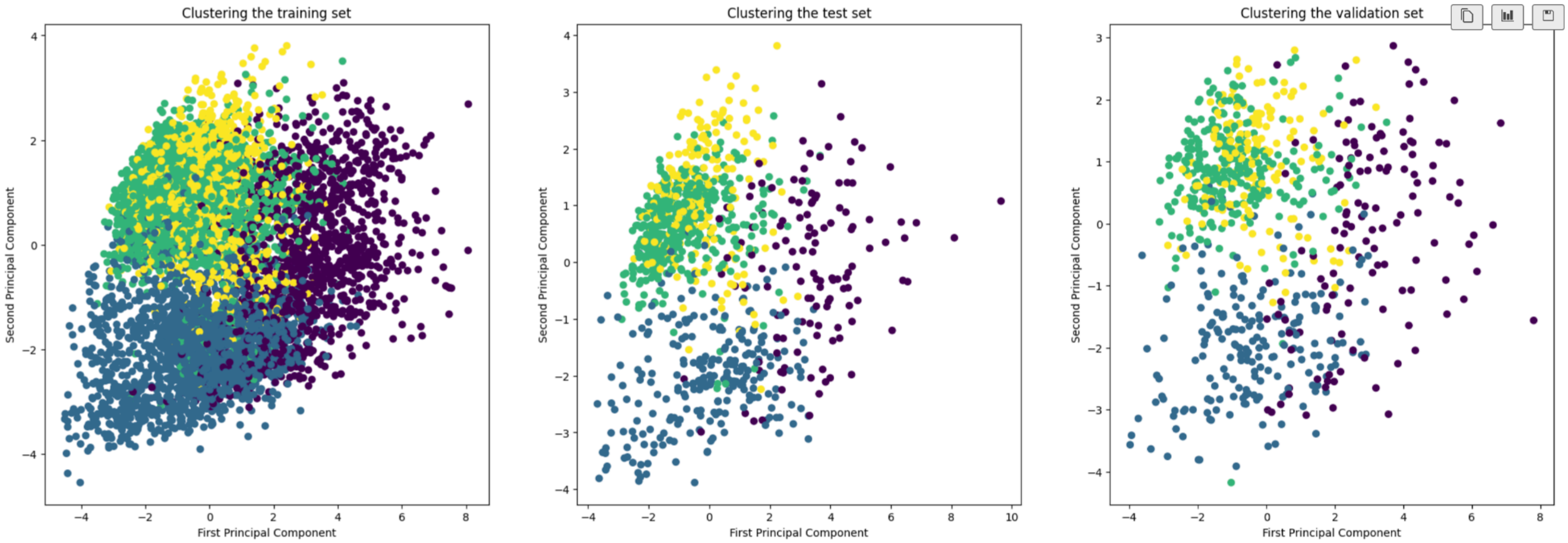


Metric	Value
Training Accuracy	0.8920
Training Loss	0.3728
Validation Accuracy	0.8208
Validation Loss	0.6108
Test Accuracy	0.8220
Test Loss	0.6397





- The clusters can be visualized as below along their first two principal components.



- The performance of classification of the labels within the clusters is also shown below:

Multinomial Naive Bayes

Cluster	Accuracy
0	0.41935
1	0.41036
2	0.53580
3	0.58823

Random Forest Classifier

Cluster	Accuracy
0	0.37097
1	0.42231
2	0.59753
3	0.65158

Support Vector Machine (SVM) using RBF kernel

Cluster	Accuracy
0	0.53225
1	0.54980
2	0.67160
3	0.68778