In this statistical analysis, I compare several models after reducing dimensions using UMAP on the dataset collected through Reddit.

After vectorizing, I created a loop to compare different parameters for UMAP. The one that seemed to affect the performance most seemed to be the n_neighbors value so I experimented with several options. UMAP will unravel the dataset and project sparse data to a lower dimensional space. It is possible to observe the embeddings by plotting the result of the transformation.

After finding out that the model performs best at n_neighbors=35, I chose the best performing classifier, which turned out to be K nearest classifier. I used a gridsearch to look for the best parameters and fit the transformed dataset.

Once again, I re-trained the KNN classifier with the best parameter settings and used a OnevsRest classifier to prepare for plotting the ROC curve. The curve showed an average area of 97%!

However, this appeared to be a case of data leakage - therefore I inadvertently overfit the model. My suspicions were validated after trying out the model on a holdout dataset, which yielded extremely poor results.