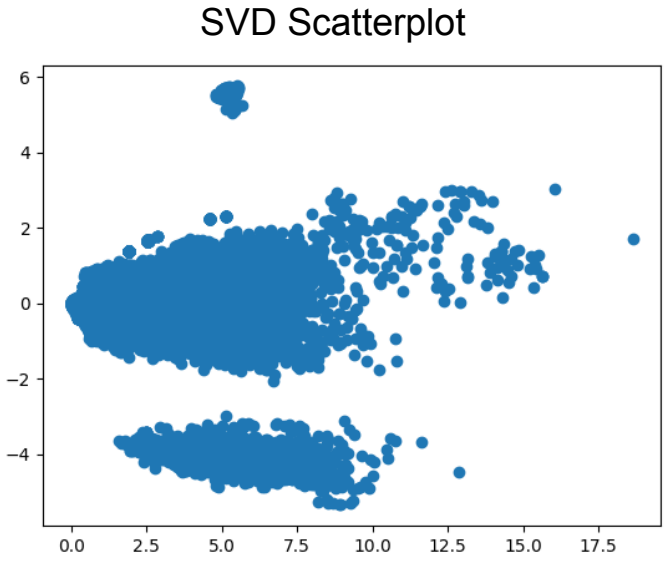
Project 3: Spam Email Classification

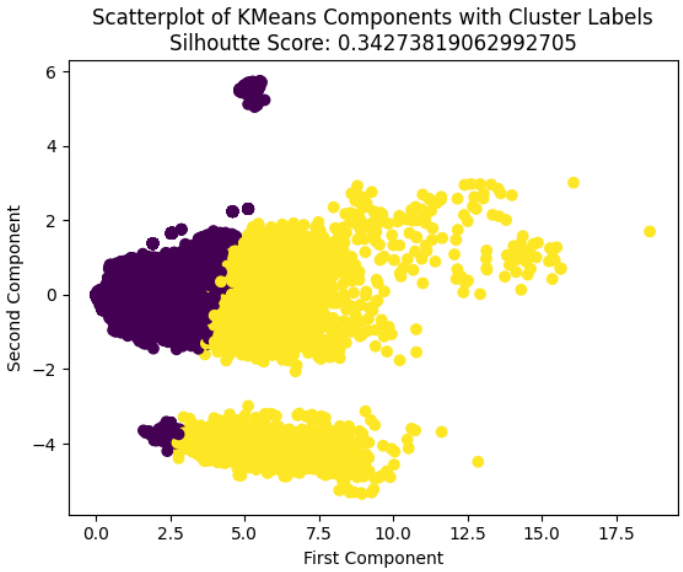
For this project we were given ~60 thousand emails with a label for spam indication, to address, from address, subject, and body. Using these emails, we were to use cluster methods to understand how clustering works. This was paired with classification models to see how well we could identify spam versus non-spam emails. Unfortunately I was not able to use the full 60 thousand emails and in my clustering I had to reduce them down to about 20 thousand, and then in the following classification modeling, to about 1000. This was due to the intense cpu and ram usage on my computer that would cause it to bluescreen otherwise.

Initially clustering the data using SVD we can see a few overall chunks. There are three islands spread out very nicely. The middle can even be considered a cluster in itself since it pulls apart from the main island.



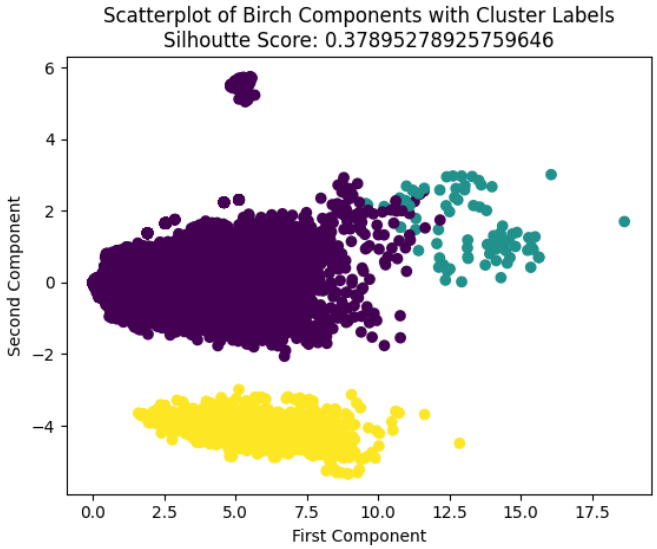
Graph 1: SVD Scatterplot of the first and second component

While graph 1 is great to look at, it's not that helpful at seeing how the cluster methods will separate this information. There were alot of issues with the cluster methods, primarily the way they colored the islands. KMeans in graph 2 was the first I used with a silhouette score of .34 and poor identification as it did not label the same way we see the clusters split from each other. This meant trying other cluster algorithms until I stumbled onto the Birch algorithm in graph 3



Graph 2: Scatterploit of KMeans Components with Cluster Labels

As you can see in graph 3, the birch plot looked great! Although it isn't perfect, it performed the best out of the other graphs I had made, including Agglomerative Clustering, DBSCAN, OPTICS, Gaussian Mixture, BisectingKMeans. This doesn't mean they couldn't possibly work, they do have some other parameters that I did not fully experiment with, but after some time doing so I found that the Birch algorithm worked great out of the box. That being said, out of the box looking for the 2 main clusters didn't happed easily, it wasn't until I changed it to 3 components did they color well. That explains the far right points that are a turquois color.



Graph 3: Scatterplot of Birch Components with Cluster Labels

After that was the classification which was the most difficult part. There are errors. I'm honestly not entirely sure where, but there were lots of issues getting arrays to match up with each other. I took Decision Tree Classifier, KNearest Neighbors, and Random Forest as my three models to use. This was because I'm the most familiar and comfortable using.

Table 1 shows the vector parameters I used lined them up with their accuracy score, as well as ROC AUC. KNN did the worst overall in each category. Random Forest and Decision Tree classifiers did the best, and comparatively very little difference in scores. The difference between binary = True for the count vectorizer didn't do too much, but there definitely is a small favor towards a binary = False. This is why I continued to use binary = False when incorporating the TfidfTransformer, which I'm not sure if I applied correctly. The scores were slightly different, but not by as much.

Vector Parameter Changes and Scores	Decision Tree	KNN	Random Forest
Count Vectorizer(binary = True)	0.98517579639305	0.94837912119542	0.99085354222947
ROC AUC	0.97759242892915	0.81394535277655	0.96370967741935
Count Vectorizer()	0.992	0.99088720250802	0.99395765338511
ROC AUC	0.98158786271910	0.96717115922816	0.97926793342171
Count Vectorizer() TfidfTransformer	0.99101550232913	0.99088720250802	0.99293794369845
ROC AUC	0.98101708646339	0.96717115922816	0.97523567535719

Table 1: Vector Parameter Changes and Their Score for Each Model

If I had to go back I'd definitely look at Tfid a little more to understand how to manipulate it and try for more variety in the classifications. Also get more ram for a bigger data set because it became incredibly frustrating to constantly restart the kernel which is another place where I think I could get more accurate scores from.

I'm sorry I'm getting this paper to you so late, but thank you so much for being an amazing professor the semester. I'm actually really sad you won't be teaching the next class. You taught me so much and made everything easy to understand. As well as a gave me all the patience and care I asked for when completing these assignments.