MAIS 202 - PROJECT DELIVERABLE 3

1. Final Training Results: I finally got the data preprocessing finished. Using Pandas, I was able to unite multiple csv files in a very specific way such that my vectors can now be made. I plan to run a PCA and find the most variant features, then extract those features from the training data, and then run naive bayes. Then, when an input is provided, it will be projected onto the eigenvectors, then pushed through the naive bayes.

I think this is one of the best methods considering the number of features per vector is near 400,000. So the PCA will drastically reduce the features considered for naive bayes.

I have already partitioned the data into a training and test set, about 70 30, and am excited to start the algorithm.

2. Final demonstration proposal: Because of how much time the preprocessing took me, I hope to make a website, but might run short on time. Worst case scenario, I will build a poster to present my project and algorithm. My website would run on react (that is what I am most comfortable with), and display nice graphs as well as a way for users to use the algorithm.

If all goes to plan, I hope to display a heat map as well as a visualization for the accuracy.

I will also include a very detailed write up explaining where the data was found. Why I chose the data that I did. Why I chose the ML pipeline that I did, and why I chose the project that I did.

I will make sure all my code is public so people can test it out.

With this deliverable, I will add the preprocessing code to my github. I used pandas, and it is not very complicated, but just so people working with similar data don’t go through the same trouble that I did.