EXPLORATORY DATA ANALYSIS FOR MACHINE LEARNING: HONORS PROJECT

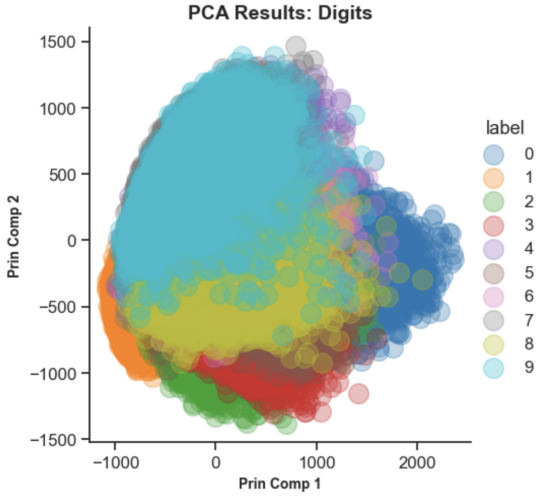
JACKSON WALTERS

This project is for the Exploratory Data Analysis for Machine Learning certification from IBM via Coursera. The long-term goal is to create a t-SNE plot to observe clustering of mental health disorders based on symptom data, as in the linked paper[1]. I would like to include additional data based on life factors such as income, housing status, and insurance status. The idea is that if someone has no income, is unhoused, and uninsured then they would be much more likely to be classified as having a mental illness. I want to go beyond correlation, and see clustering.

For now, I implemented code from here [2], and removed the SAS connection. I rewrote it in pure Python with sklearn. I did both PCA and t-SNE on the MNIST data, and plots are shown below.

All code is publicly available on GitHub [3]

FIGURE 1. PCA plot of MNIST handwritten digits data



[1]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6783390/
[2]https://medium.com/@violante.andre/an-introduction-to-t-sne-with-python-example-47e6ae7dc58f
[3]https://github.com/jacksonwalters/tsne-examples

FIGURE 2. t-SNE plot of MNIST handwritten digits data