Dimensionality Reduction

Principle Components Analysis

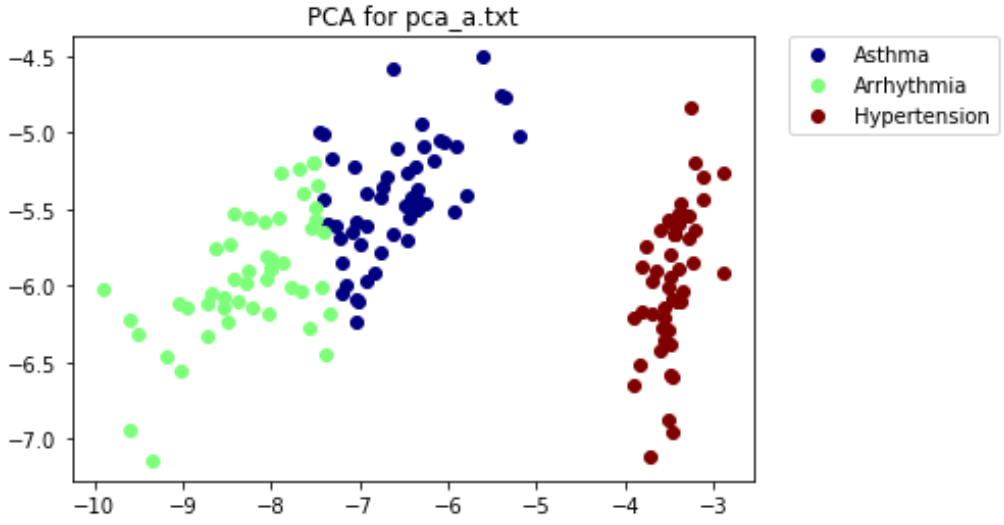
Members:

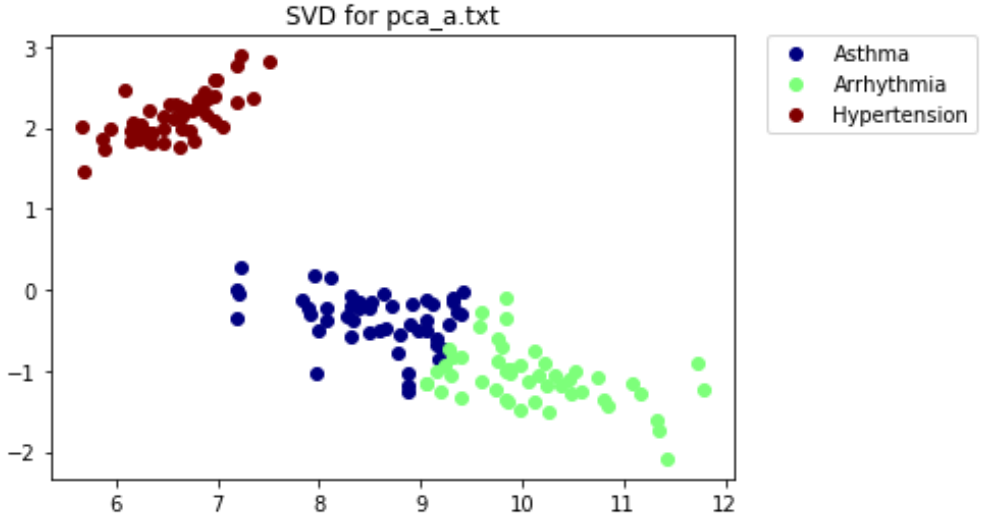
Jay Shah - 50205647

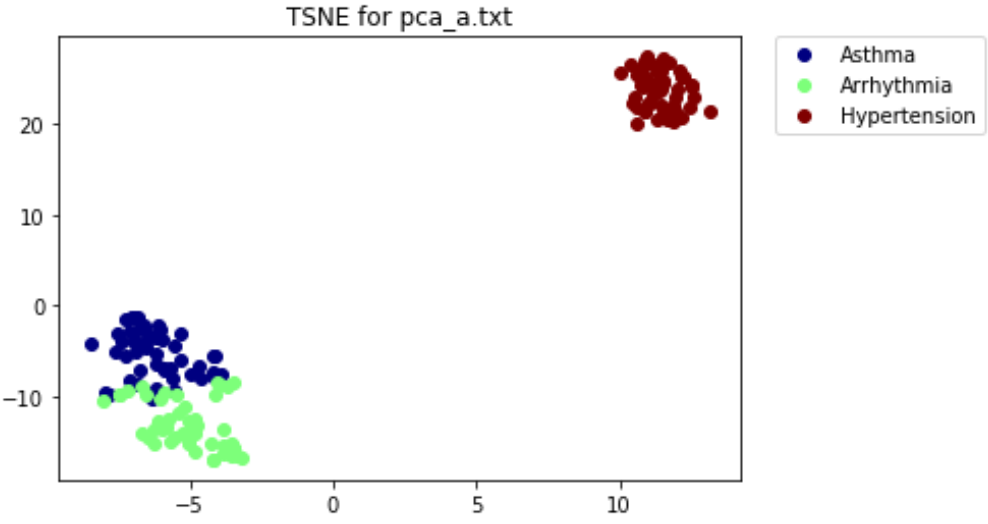
Yash Jain - 50206851

Akshay Shah - 50206543

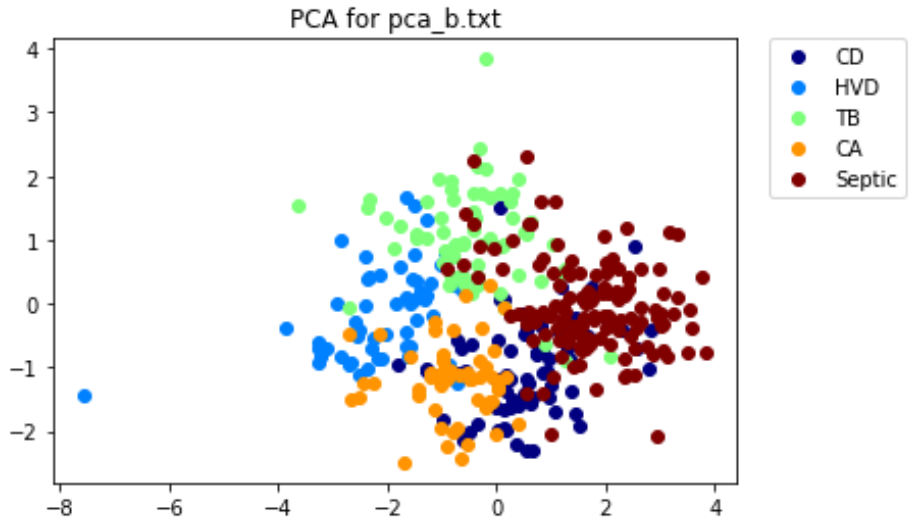
Plots for pca_a.txt

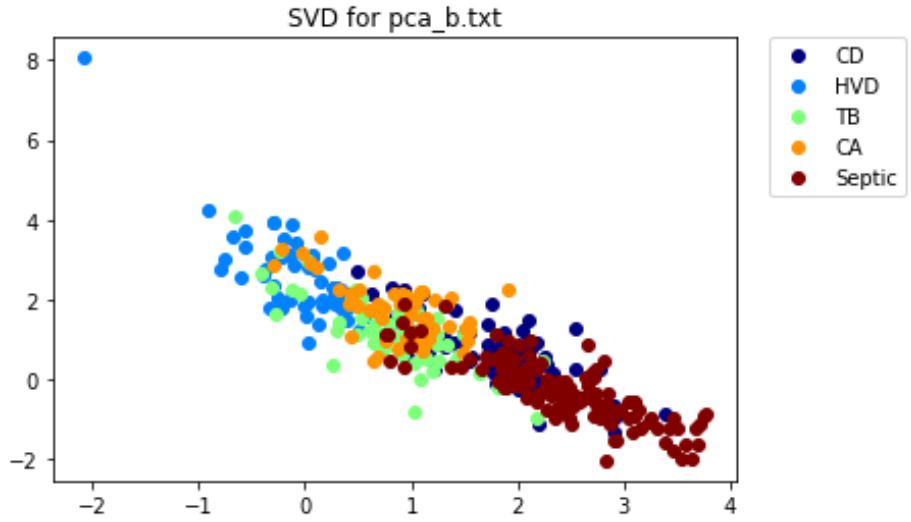


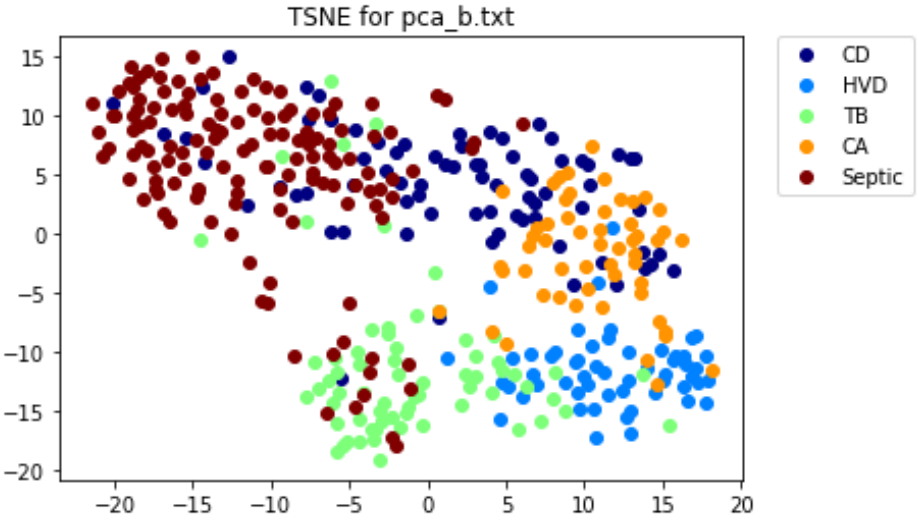




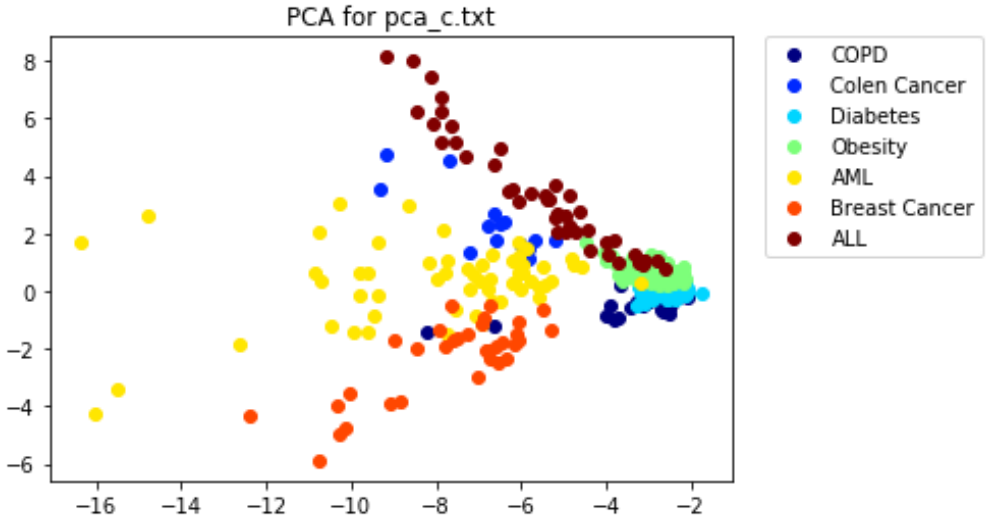
Plots for pca_b.txt

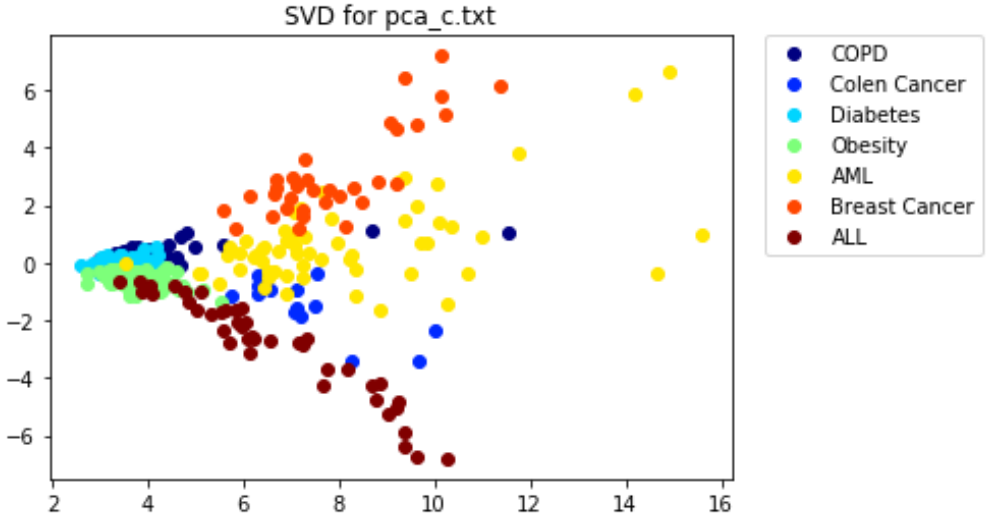


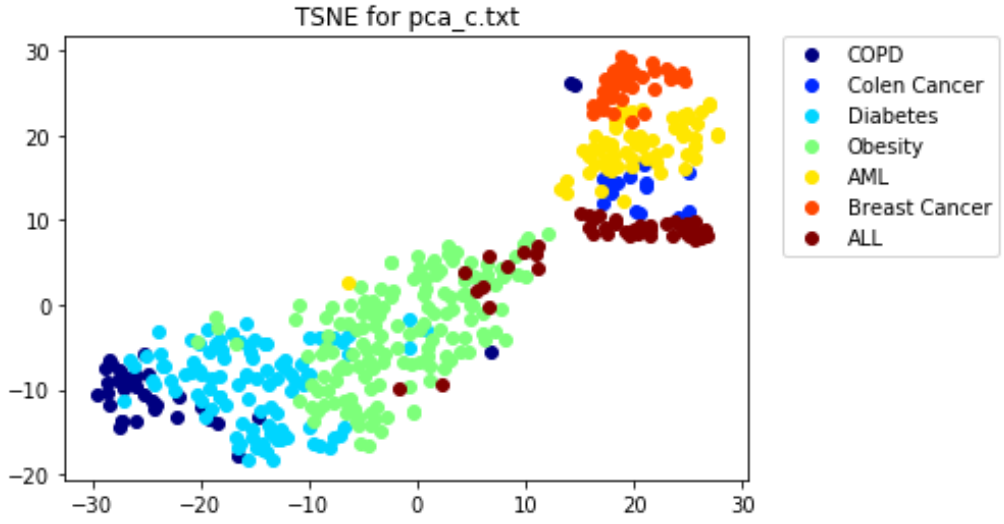




Plots for pca_c.txt







PCA implementation

Reading a file

File path is accepted as command line arguments. Values other than last column are stored as 2D list. This list is converted to np.array.

Calculating Adjusted Value

Mean $\bar{X}$ is calculated for each column by taking sum of all elements in column and dividing it by number of elements in that column. Adjusted value column is made for every column in 2D list, calculated as following: $X' = X - \bar{X}$

Calculating Covariance Matrix

Covariance is then calculated using adjusted values using the following: $S = \frac{1}{n-1} X' X'^T$.

Calculating 2 largest eigen vectors and eigen values

Using numpy.linalg.eig, eigen vectors and eigen values are calculated from covariance matrix. numpy.argsort gives list of positions of sorted list.

Transforming original data to 2 feature data

2 feature data are calculated by calculating dot product (numpy.dot) of original data and 2 maximum eigen vectors.

Plotting new 2-dimensional data on graph

Using matplotlib library, transformed data is plotted, with distinct colors for different classes.

Results

For pca_a.txt, all 3 algorithm clusters good according to classes.
For pca_b.txt, PCA performs best clustering, followed by TSNE and then SVD,
For pca_c.txt, TSNE performs better clustering than PCA and SVD. Clustering by PCA and SVD are similar.
So, we can conclude that performance of algorithm also depends on dataset.