DATA MINING AND BIOINFORMATIC – PROJECT1

Principal Component Analysis

Abhinav Kumar – akumar39
Sachin Kumar Kuppayya - skuppayy
Vinooth Rao Kulkarni - vinoothr

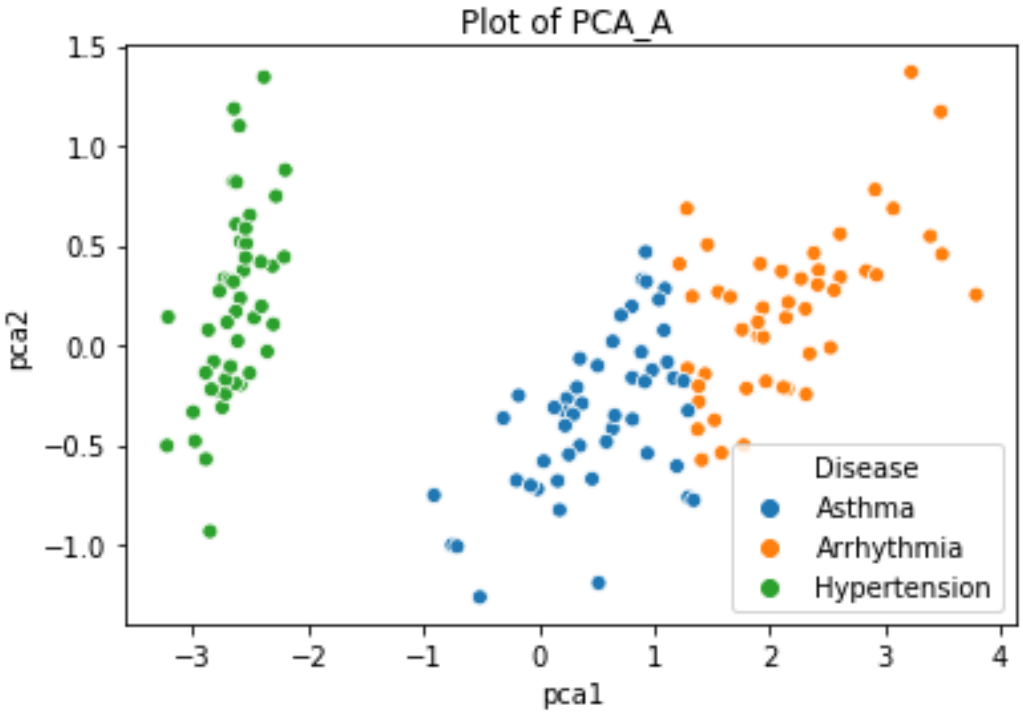
Principal Component Analysis is a dimension reduction technique that can be used to reduce a large set of variables to a small set that contains most of the information in the large set

PCA Algorithm

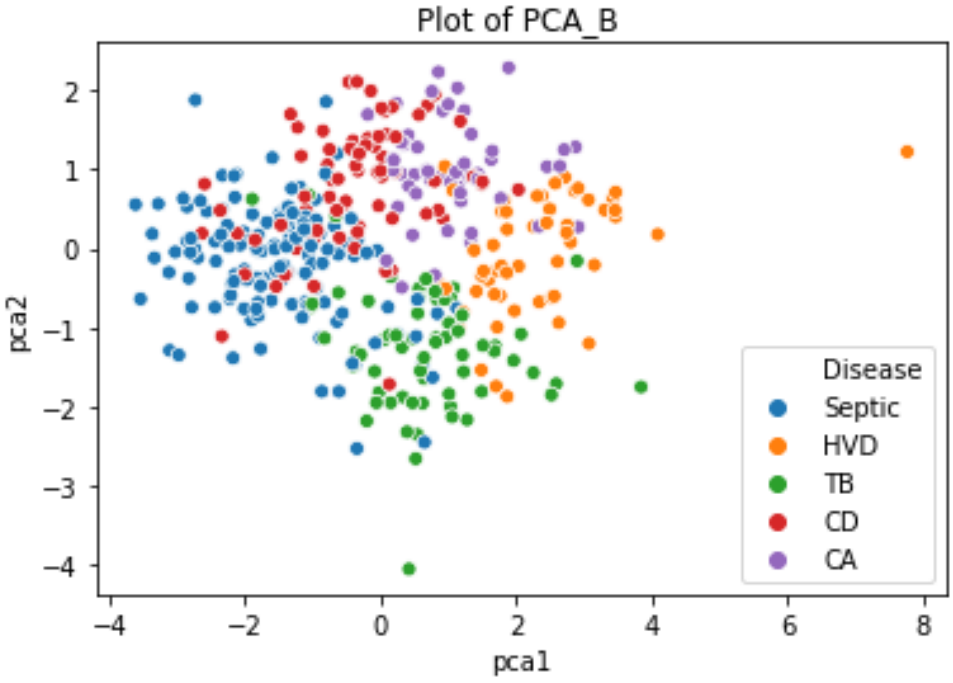
- Read the data as DataFrame
- Split the data to features and use it as a matrix
- Now we centre the features using mean. For this we subtract the mean of each column and subtract the value
- Next step we calculate the covariance of the matrix using np.cov. A covariance matrix is a calculation of covariance of a given matrix with covariance scores for every column with every other column, including itself.
- Using the covariance matrix, we calculate the Eigen vectors and Eigen values. The eigenvectors represent the directions or components for the reduced subspace of B, whereas the eigenvalues represent the magnitudes for the directions.
- Now we sort the Eigen matrix using Eigen values in descending order
- Now these vectors represent the principal components.
- Based on the number of dimensions required we take the top vectors representing the dimensions
- Now using these new dimension's, we create a plot for different diseases
- Repeat this for different datasets given and plot the graph using seaborn package

Scatter plots of implemented PCA for different datasets

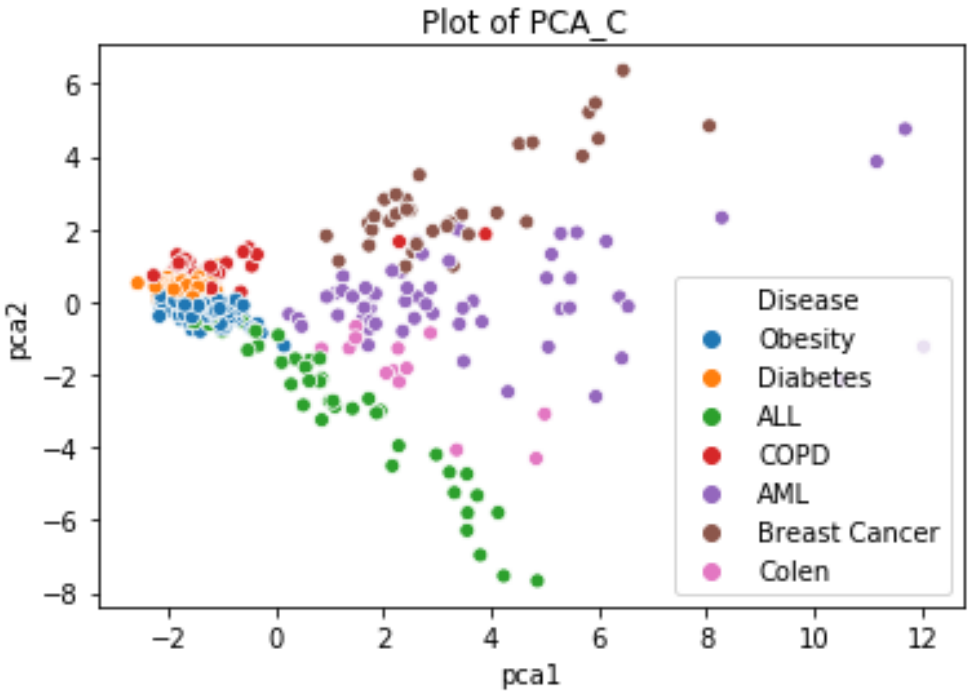
Plot 1: This plot is for PCA_A dataset



Plot 2: This plot is for PCA_B dataset

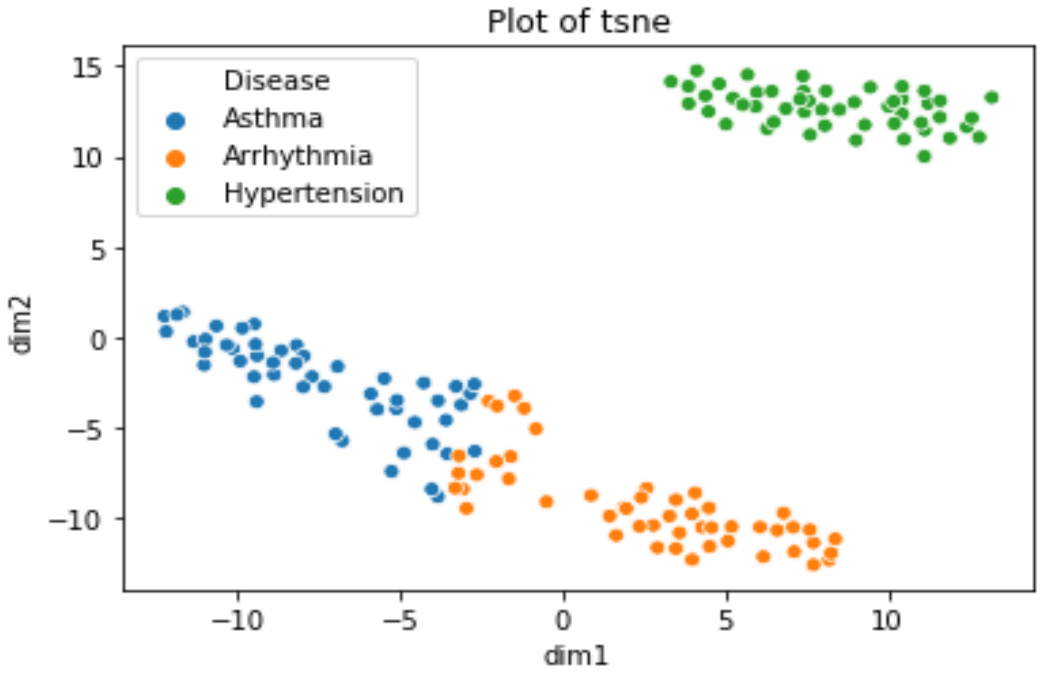


Plot 3: This plot is for PCA_C dataset

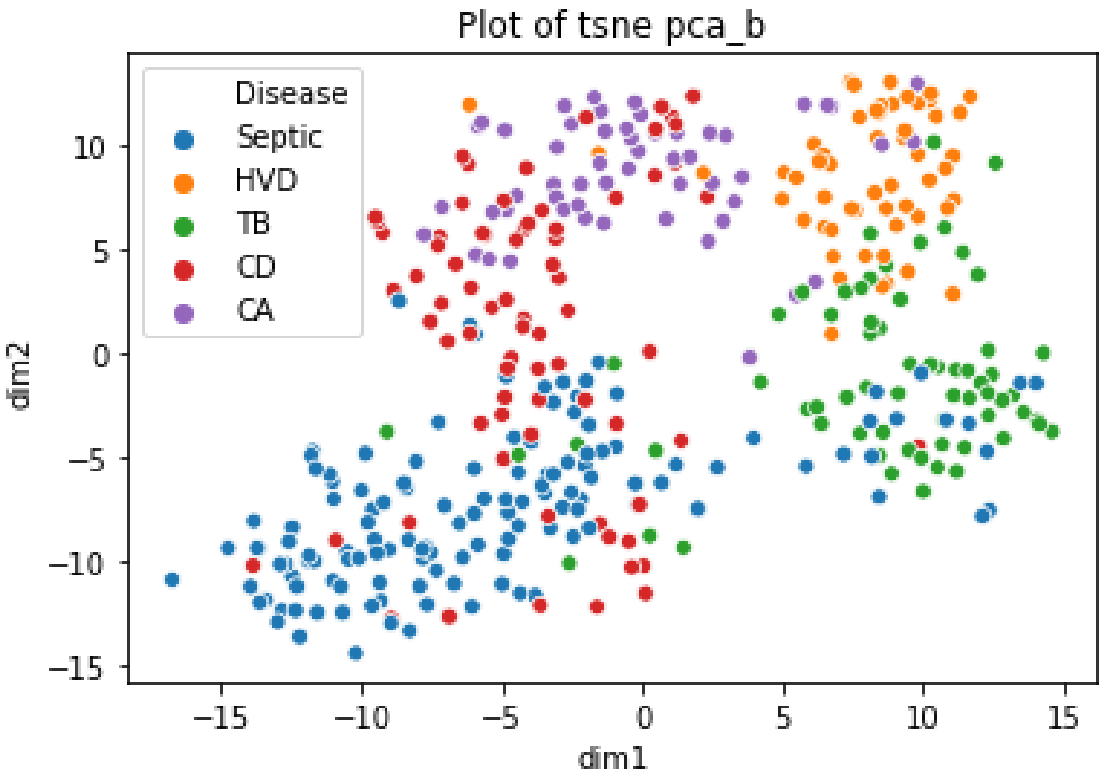


Scatter plots of t-SNE for different datasets

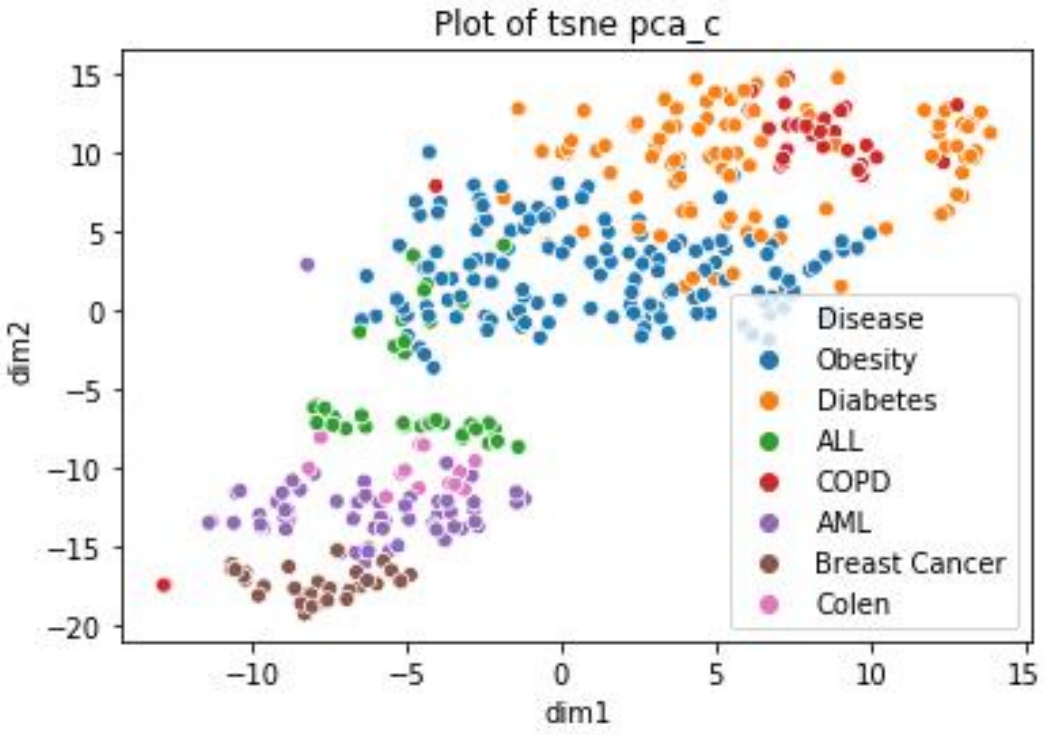
Plot 4: This plot is for t-SNE PCA_A dataset



Plot 5: This plot is for t-SNE PCA_B dataset

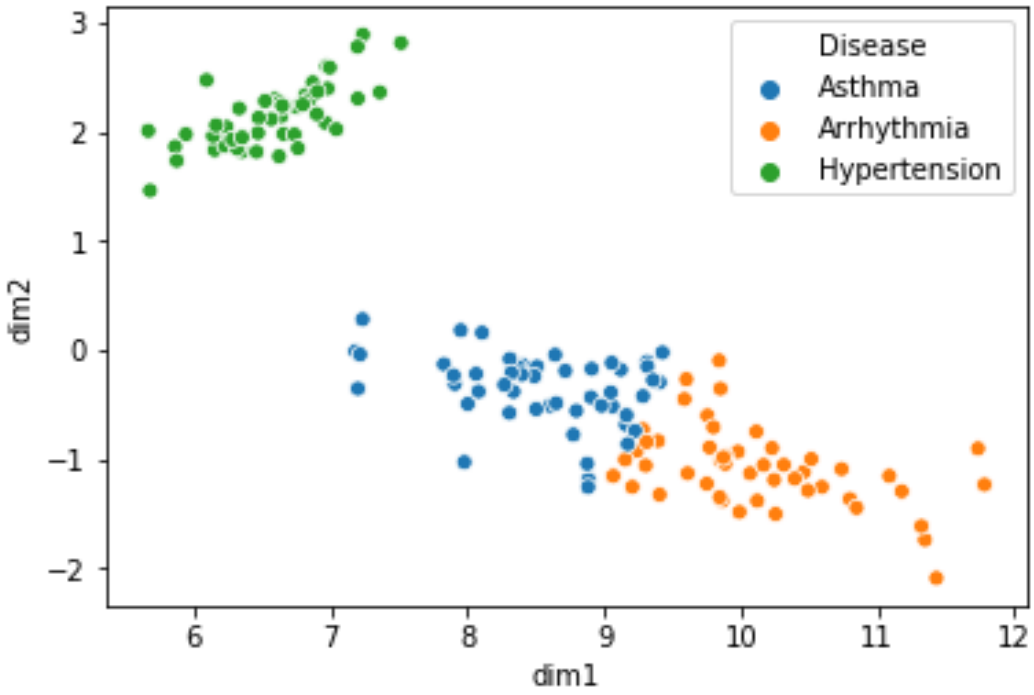


Plot 6: This plot is for t-SNE PCA_C dataset

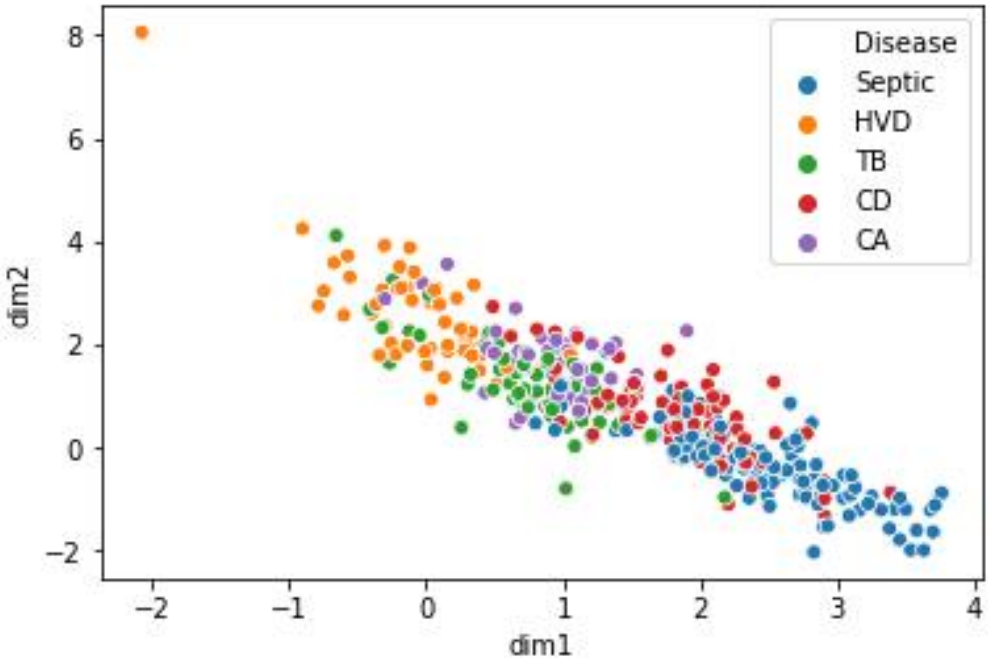


Scatter plots of SVD for different datasets

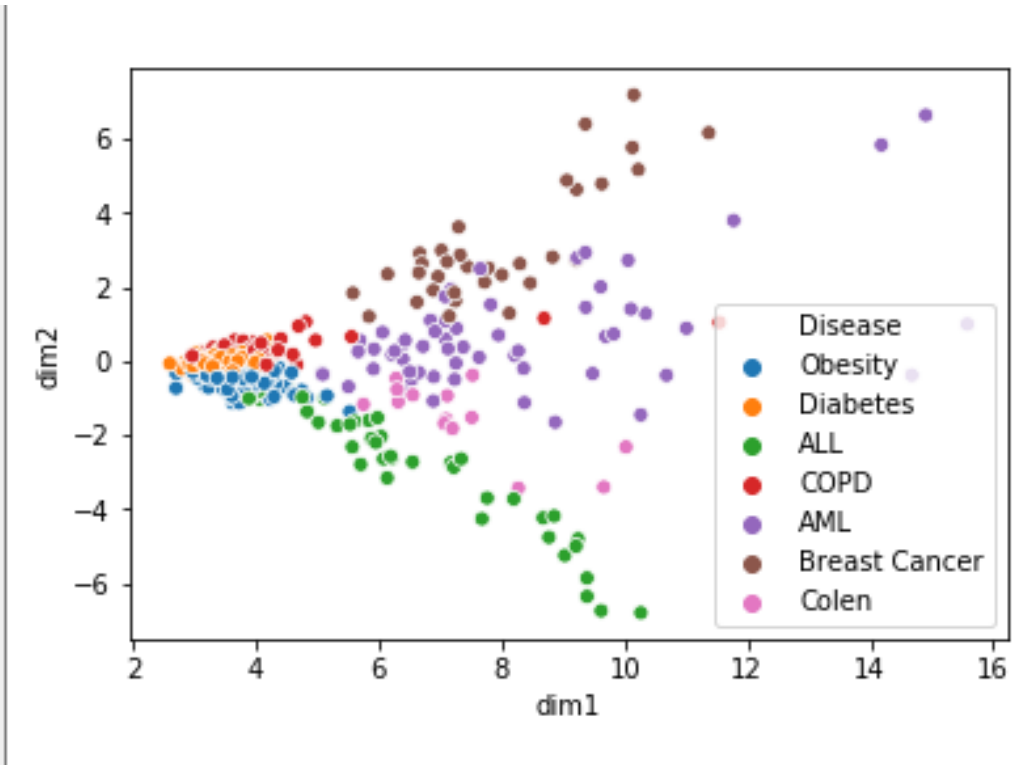
Plot 7: This plot is for SVD PCA_A dataset



Plot 8: This plot is for SVD PCA_B dataset



Plot 9: This plot is for SVD PCA_C dataset



Inference:

1. PCA and SVD plots are very similar
2. T-sne plots are quite different from the PCA plots
3. T-SNE is non linear so it can capture the structure of trickier manifolds
4. PCA is linear dimension reduction technique