CSE-601
DATA MINING AND BIOINFORMATICS

Dimensionality Reduction & Association Analysis - PCA REPORT

By:
Naina Nigam: 50208030
Surabhi Singh: 50208675
Vanshika Nigam: 50208031

Principal component analysis (**PCA**) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components** .It is used for dimensionality reduction.

t-distributed stochastic neighbor embedding (t-SNE) is a machine learning algorithm for dimensionality reduction.It is a nonlinear dimensionality reduction technique that is used for embedding high-dimensional data into a space of two or three dimensions.

Singular-value decomposition (SVD) is a factorization of a real or complex matrix. It is the generalization of the eigendecomposition of a positive semidefinite normal matrix.

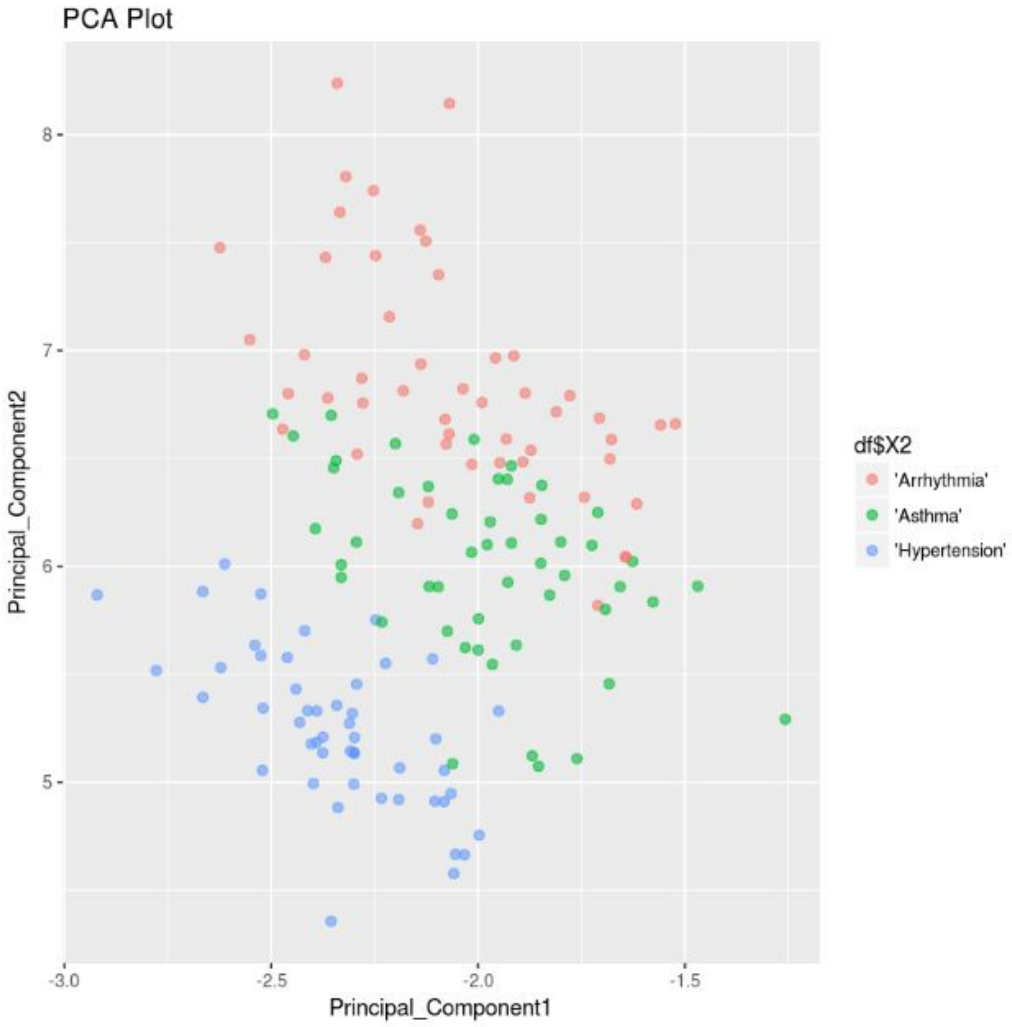
Implementation of PCA-

To implement PCA, we performed the following steps-

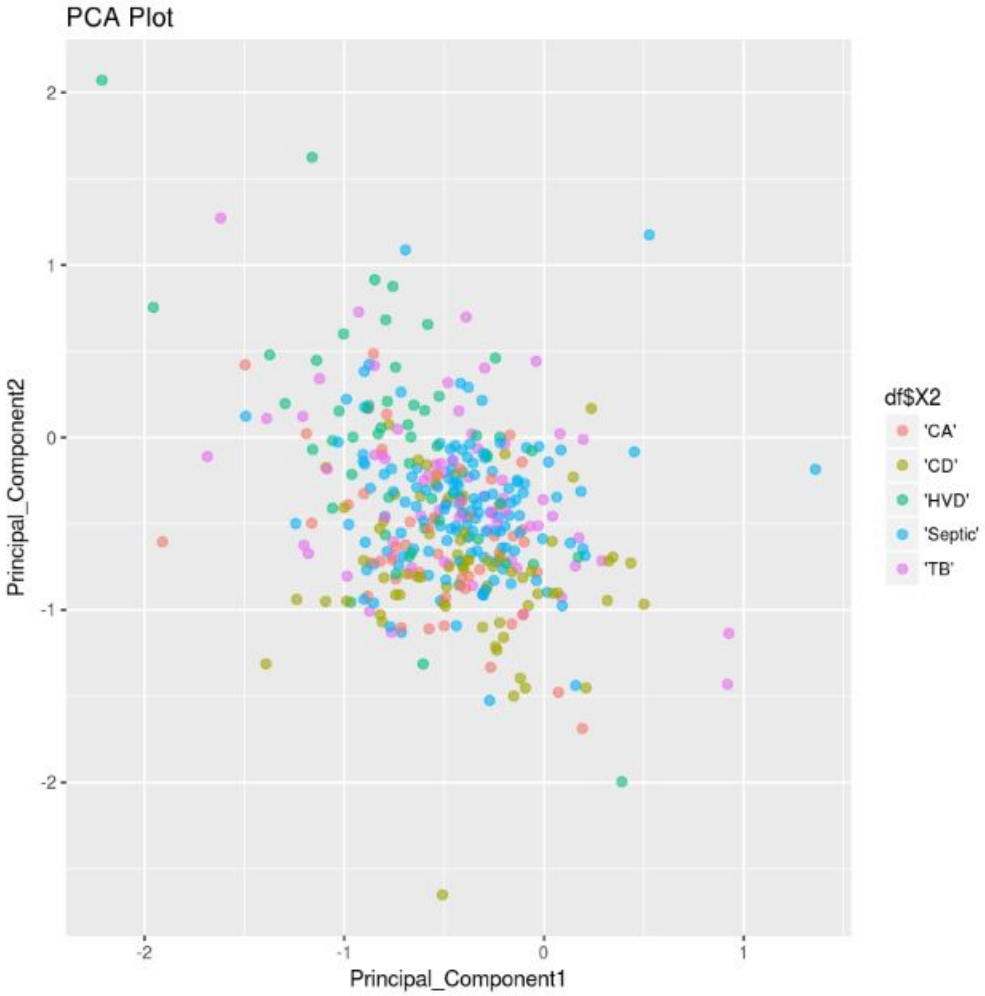
1. Loaded the entire data into ***initial*** matrix which was separated by \t
2. We stored the last column i.e the column containing the disease name in ***t*** and stored the entire data without the last column in a matrix named ***matrix***
3. Next we computed the mean of the matrix and stored it in ***mean***
4. Then we calculated the normalized matrix by subtracting matrix from mean and stored it in *normalize*
5. Next, we calculated the covariance matrix of the above and stored it in ***covmat*** and calculated the eigenvectors and values and stored in ***eigvecs*** and ***eigvals*** respectively.
6. We arranged the ***eigvecs*** in an order corresponding to the their eigen values sorted in decreasing order.
7. We extracted the first two columns of above and stored it in ***w.***
8. Then we calculated the final result of pca by multiplying transpose of w i.e ***w_transpose*** and initial matrix and stored it in ***ans***
9. Then we stored this in a csv file which was loaded in jupyter and using R we plotted these results for the three files.

Plots for PCA

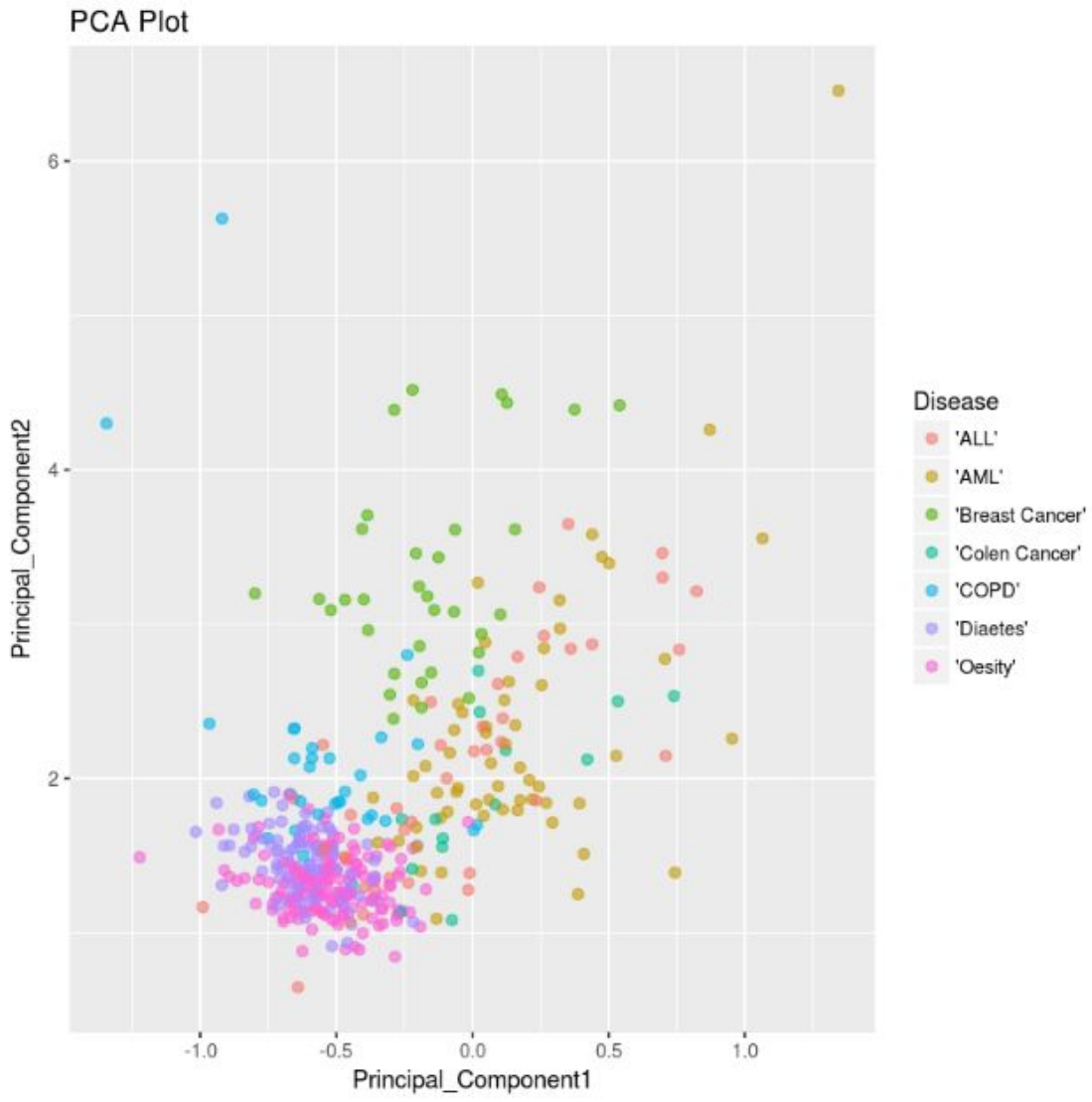
- For text file pca_a.txt



- For text file pca_b.txt

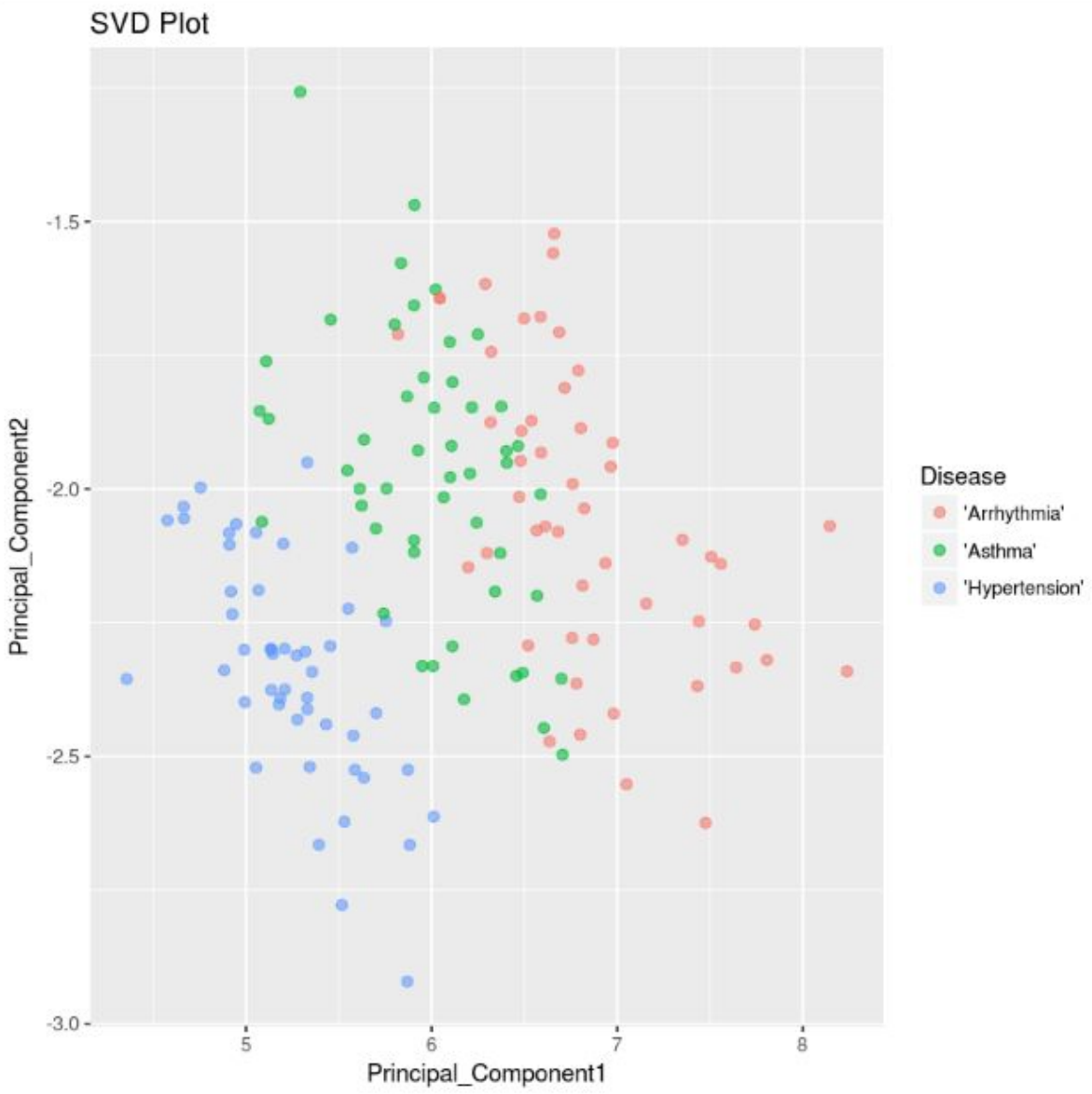


- For text file pca_c.txt

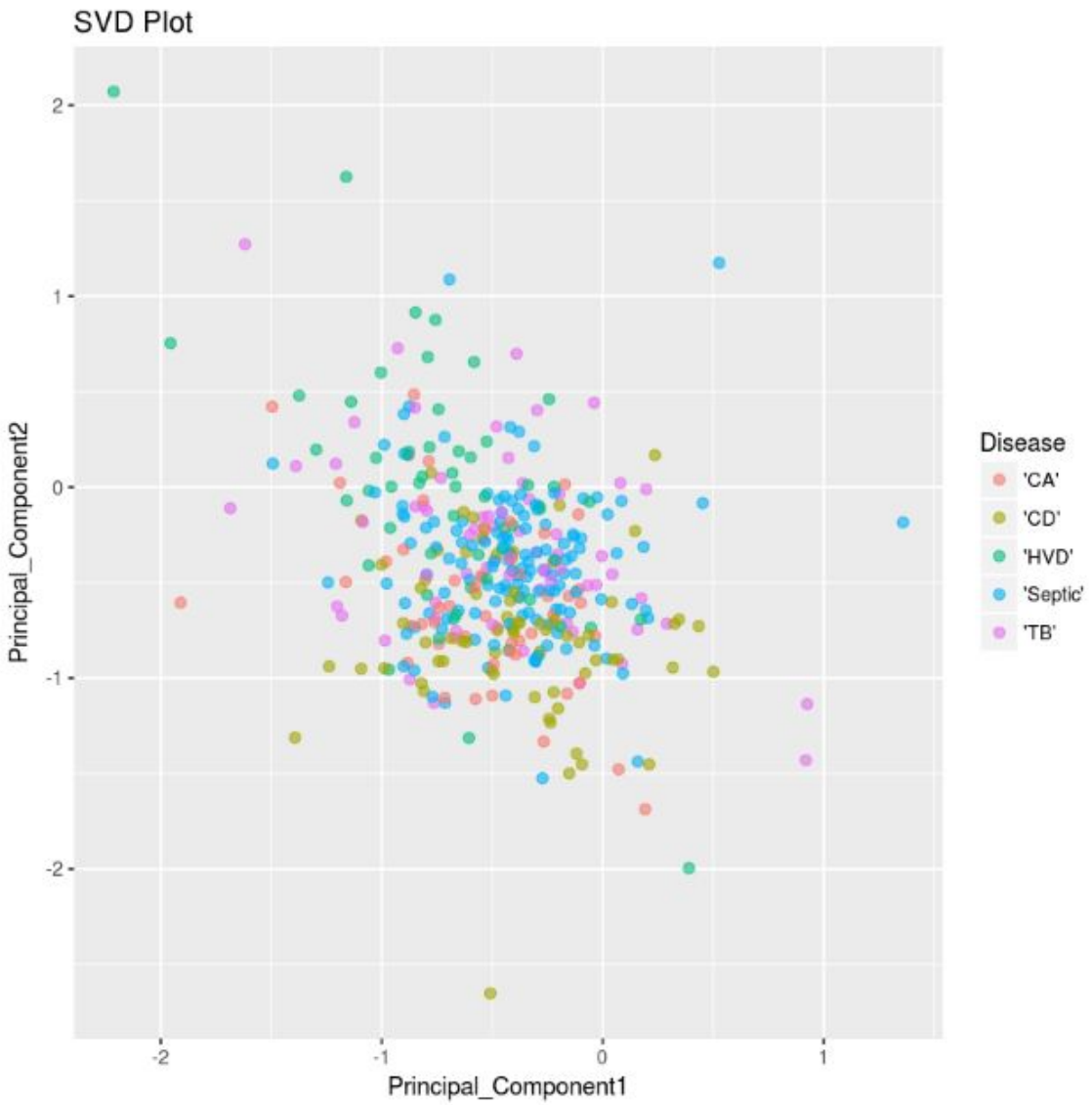


Plots for Single Value Decomposition

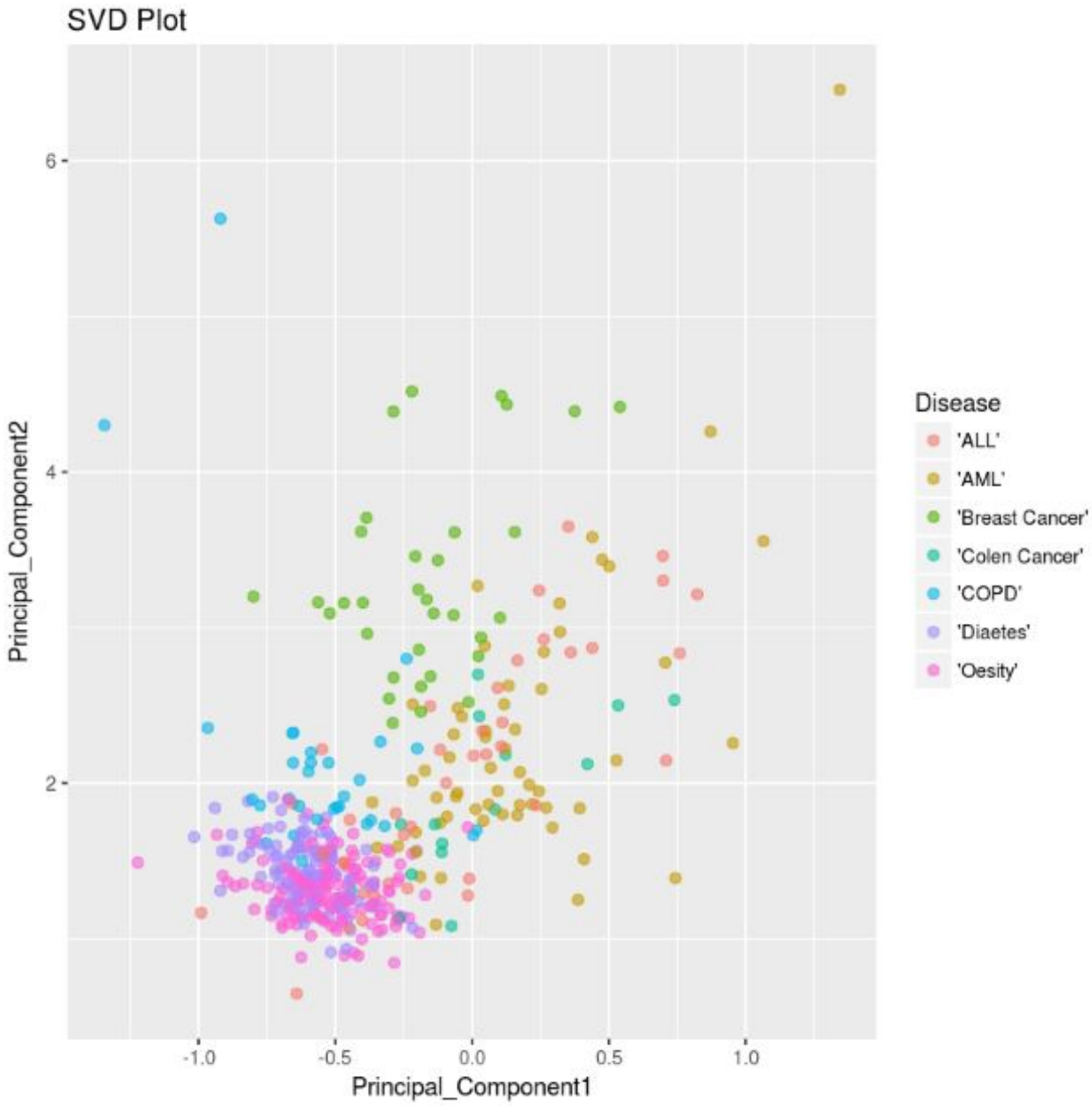
- For text file pca_a.txt



- For text file pca_b.txt

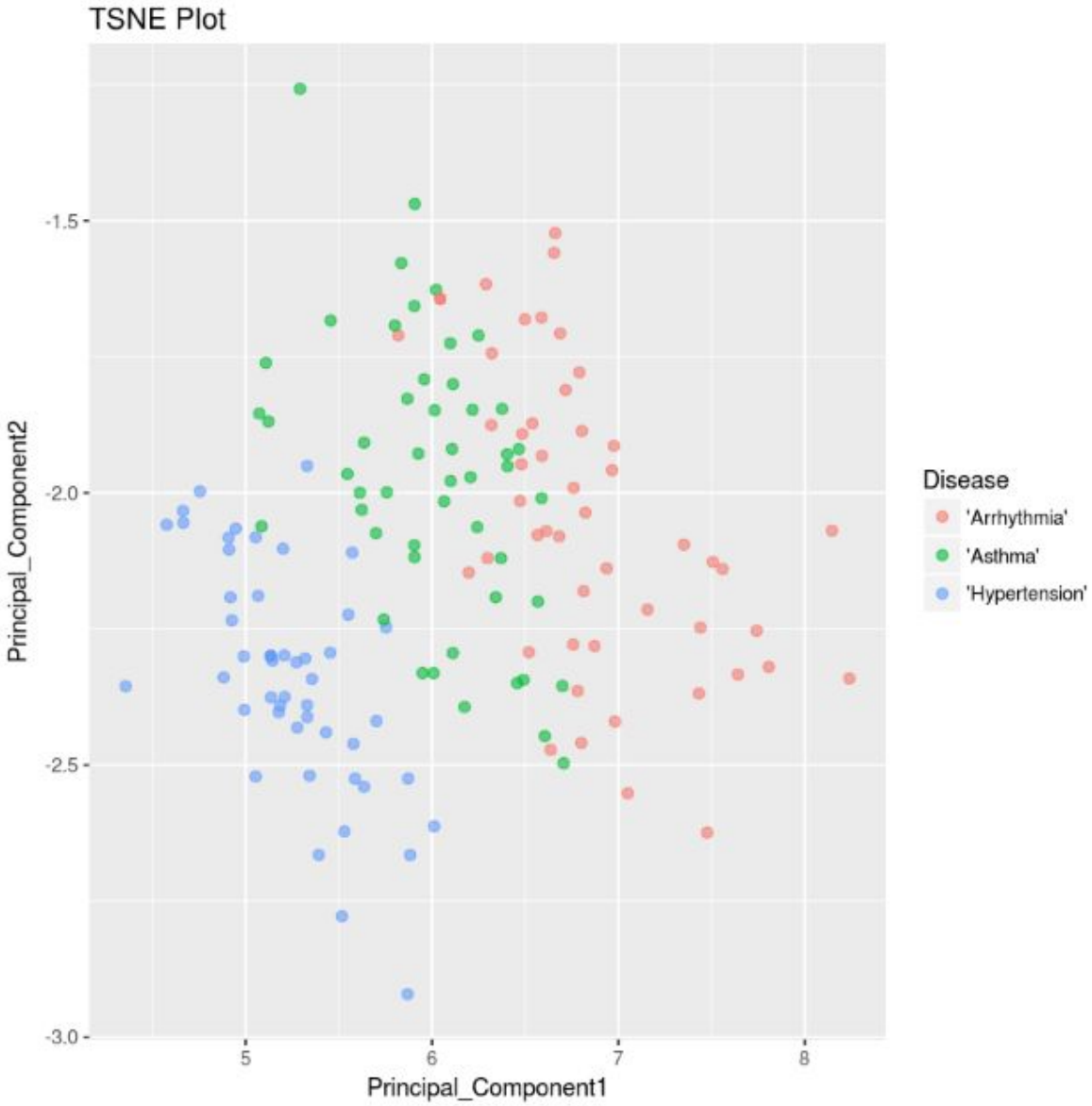


- For text file pca_c.txt

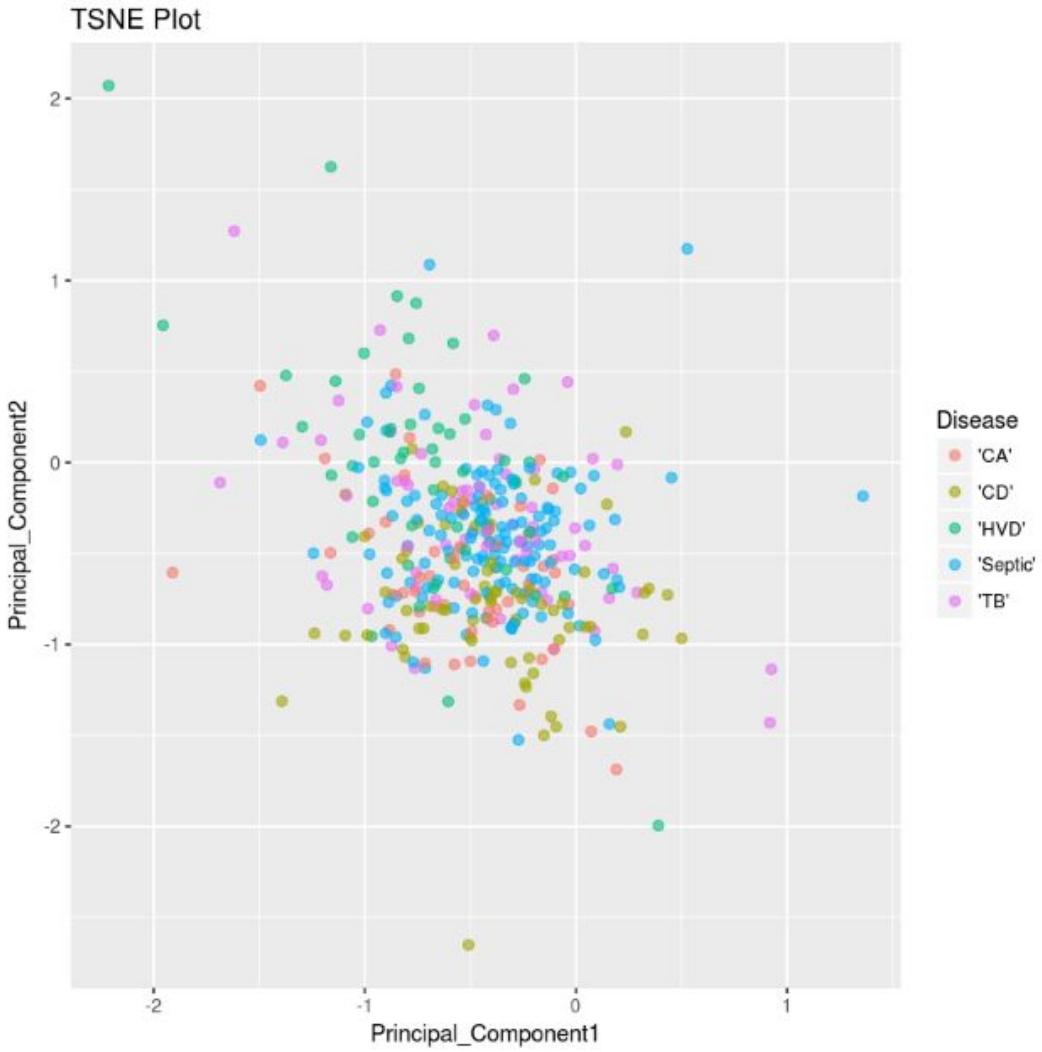


Plots for t- SNE algorithm

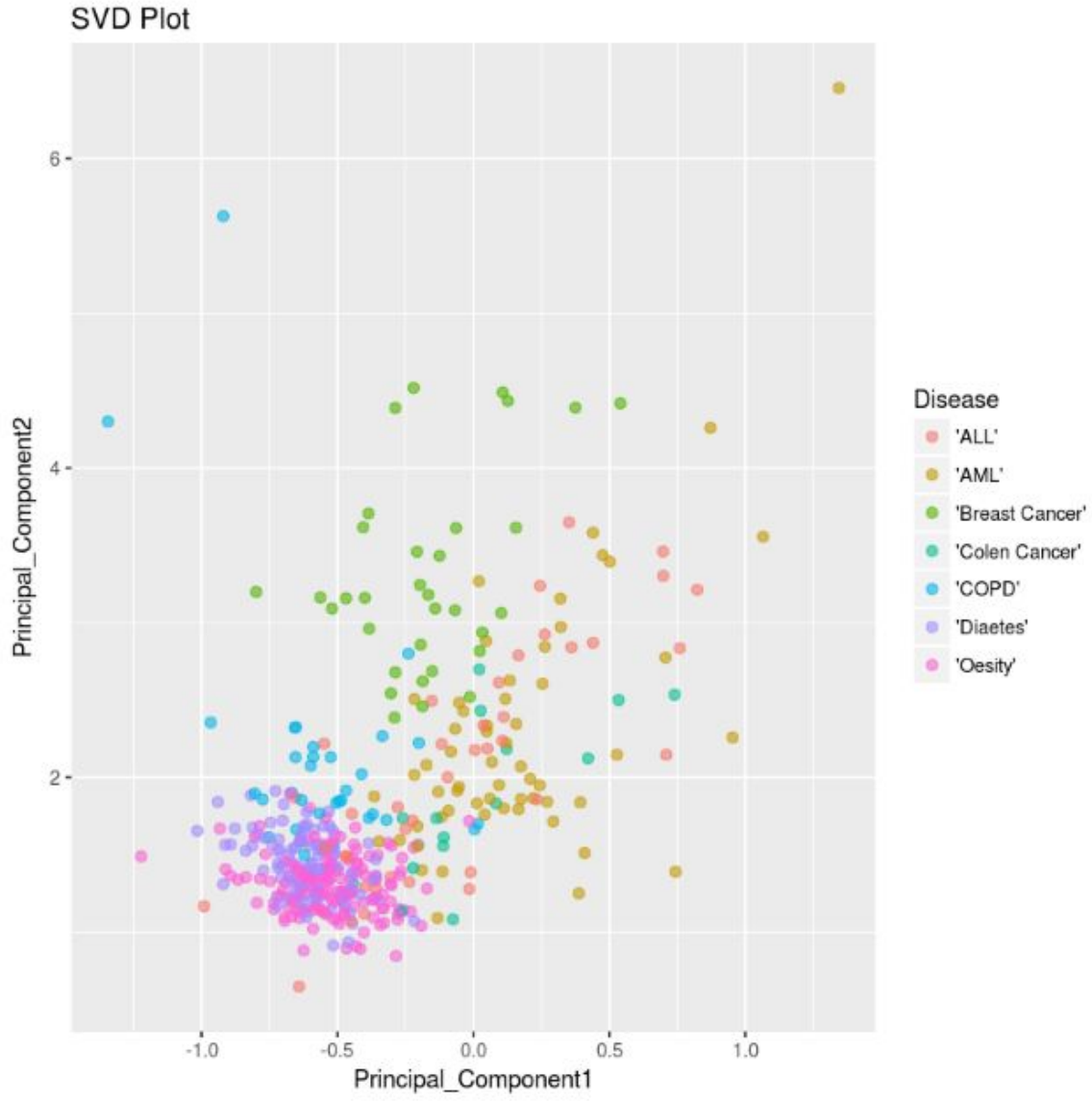
- For pca_a.txt



- For pca_b.txt


TSNE Plot
Principal_Component2
Principal_Component1
2
1
0
-1
-2
-2
-1
0
1
Disease
'CA'
'CD'
'HVD'
'Septic'
'TB'

- For pca_c.txt


SVD Plot
Principal_Component2
Principal_Component1
6
4
2
-1.0
-0.5
0.0
0.5
1.0
Disease
'ALL'
'AML'
'Breast Cancer'
'Colen Cancer'
'COPD'
'Diaetes'
'Oesity'

Analysis

PCA is parameter free. Given the data, it takes into consideration only the principal components. Whereas, t-SNE relies on the parameter, perplexity, early exaggeration, learning rate and so on.
PCA can be computed iteratively t-SNE is nonlinear.
From the plots, it is visible that for lower dimensions, PCA gives good results but for higher dimensional data, t-SNE gives better results.

In SVD there is use of left and right singular matrix along with a matrix with eigen values arranged in diagonal order. Although SVA used the PCA procedure but the plots of SVD are different from that of PCA as the matrix used in SVD was original and in PCA the normalized function was used.