DIMENSIONALITY REDUCTION DOCUMENTATION

Tariq Siddiqui - 50208470

Junaid Shaikh - 50208476

Kaushik Ramasubramanian – 50207352

PROJECT REQUIREMENTS:

We were asked to perform Dimensionality reduction on a dataset using various dimensionality reduction techniques such as principal component analysis, t-SNE and Singular value decomposition. It was also required to compare the plots obtained using above mentioned three techniques.

DIMENSIONALITY REDUCTION:

Quite often we come across dataset which has a large number of variables in the dataset. Therefore, it can be quite challenging to analyse and interpret these variables.

Dimensionality Reduction is a technique in order to handle the below mentioned issues with the dataset:

- If there are too many variables each and every variable needs to be explored.
- If every variable in the dataset has significance.
- If the variables are numeric and they have multi-collinearity, it would be difficult to identify those variables.

Dimensionality reduction refers to the process of converting a set of data having vast dimensions into data with lesser dimensions ensuring that it conveys similar information concisely. One of the major applications of this technique is **Image Processing.**

There are various advantages of Dimensionality reduction techniques. Some of them are mentioned below:

- It helps in data compressing and reduction in storage space.
- It improves the time taken to compute as the lesser number of dimensions leads to lesser computation time.
- It takes care of multi-collinearity that improves the model performance by removing the redundant features.
- Reduction in the number of dimensions in the data leads to ease in visualization using plots.

Some of the common methods of dimensionality reduction are mentioned below:

1. Principal Component Analysis (PCA)
2. t-SNE (non-parametric / non-linear)
3. Singular Value Decomposition (linear)

They are of two types – linear models and non-linear models.

PRINCIPAL COMPONENT ANALYSIS:

Principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. With fewer variables, visualization also becomes much more meaningful.

By using this technique we get new set of variables known as **principal components.** These components depicts most of the possible variation of original data after which each succededing component has the highest possible variance.

The second principal component tries to capture variance in the data that is not captured by the first component, thus making the former orthogonal to the latter. Thus when trying to reduce the dataset into two-dimensions only the top two Eigenvectors are selected as the principal components.

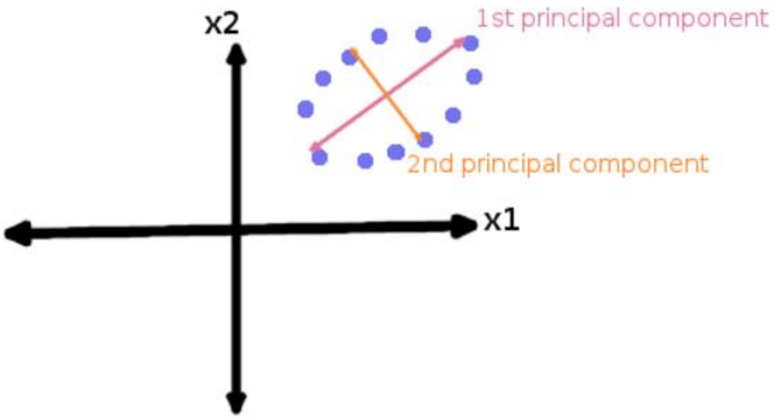

Figure 1: PRINCIPAL COMPONENT ANALYSIS

t-SNE (t-Distributed Schochastic Neighbour Embedding):

t-SNE is a non-linear dimensionality reduction algorithm used for exploring high dimensional data. It is used to overcome certain limitations observed in PCA. They are mentioned below:

- Since PCA is a linear algorithm it won't be able to interpret complex polynomial relationship between features.
- Linear dimensionality techniques emphasize on placing dissimilar data points far apart in the lower dimension representation.

Contrary to this, t-SNE uses probablity distributions based by traversing the neighbourhood graphs to find the structure of the data. t-SNE finds patterns in the data by identifying observed clusters based on similarity of data points with multiple clusters. Hence it is able to reduce high dimensional data into lesser dimensions.

SINGULAR VALUE DECOMPOSITION:

Singular value decomposition is used to remove the redundant features in the dataset. In case, a given dataset and large number of features, it becomes difficult to run the machine learning algorithm on that dataset as it consumes lot of memory and also leads to a large computation time.

The presence of redundant features causes multi-collinearity in linear regression. Thus, SVD helps in reducing the data to retain only the features which are most significant.

Given below are the plots that were obtained after executing **PCA, t-SNE and SVD** on the three datasets provided.

Plots for pca_a.txt file

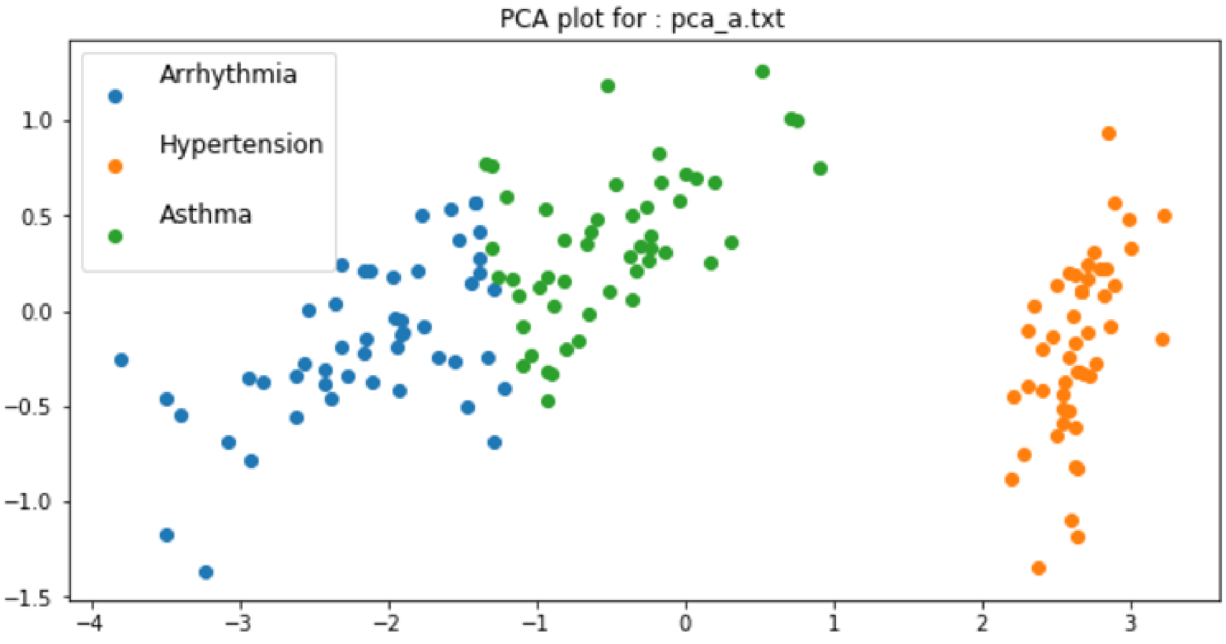

Figure 2: PCA plot for pca_a.txt

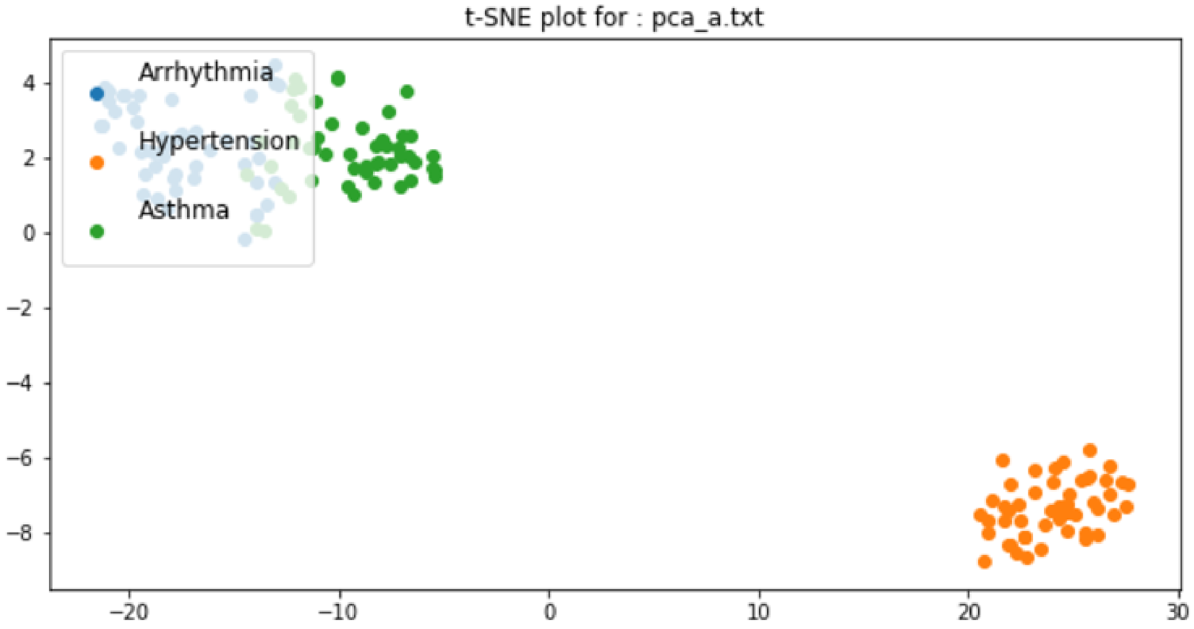

Figure 3: t-SNE plot pca_a.txt

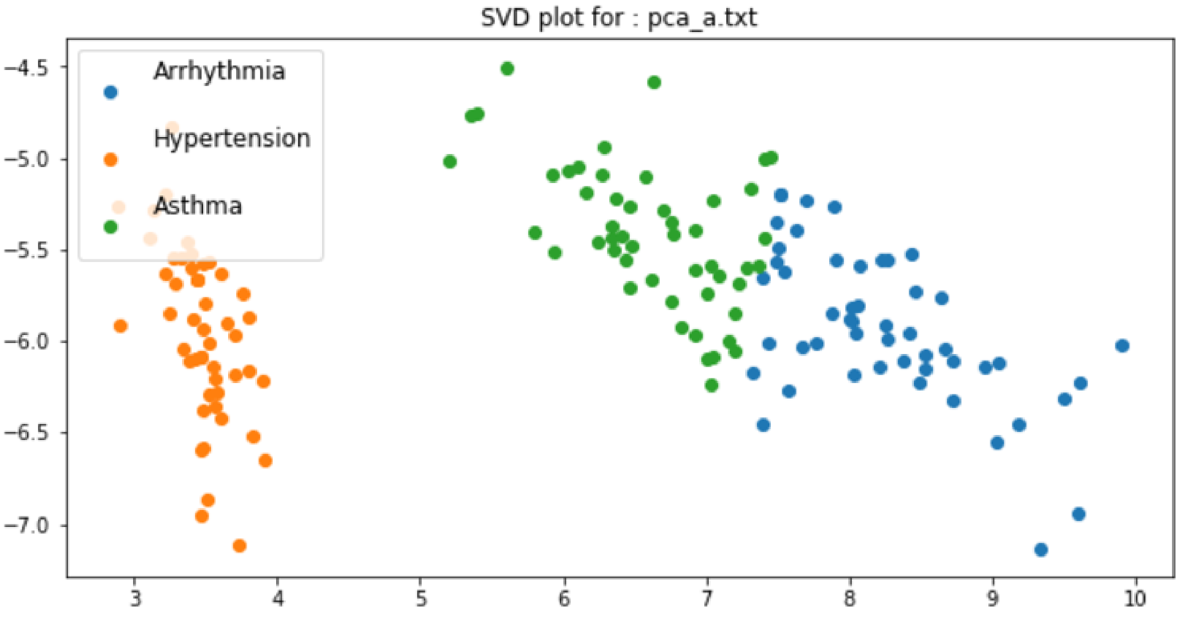

Figure 4: SVD plot for pca_a.txt

Plots for pca_b.txt file

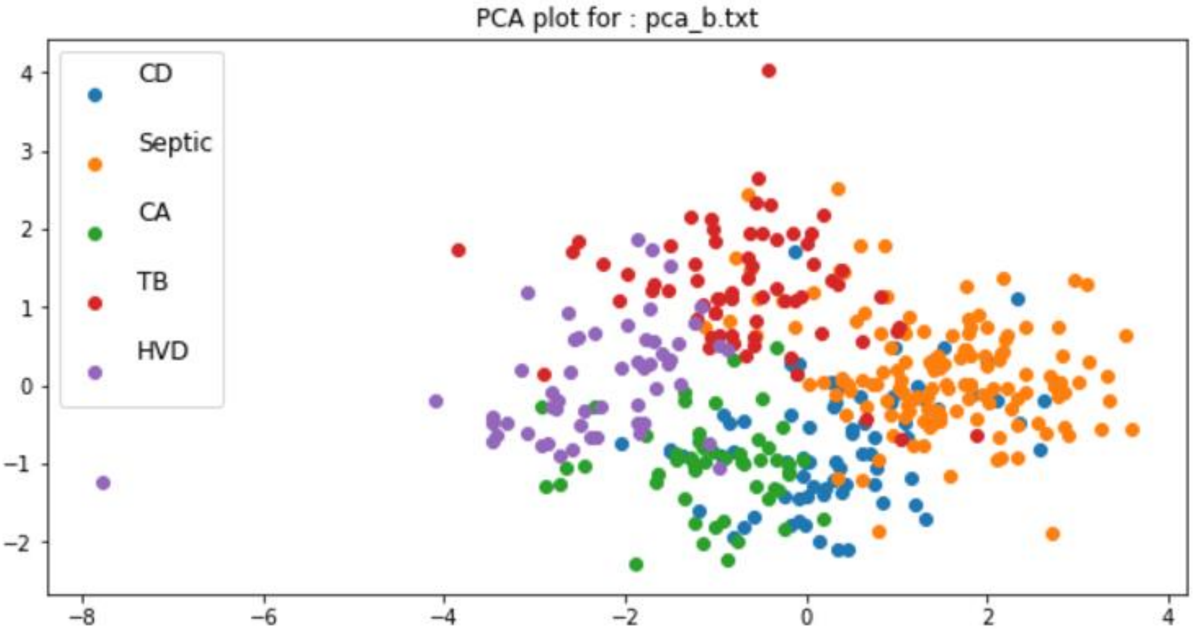

Figure 5: PCA plot for pca_b.txt

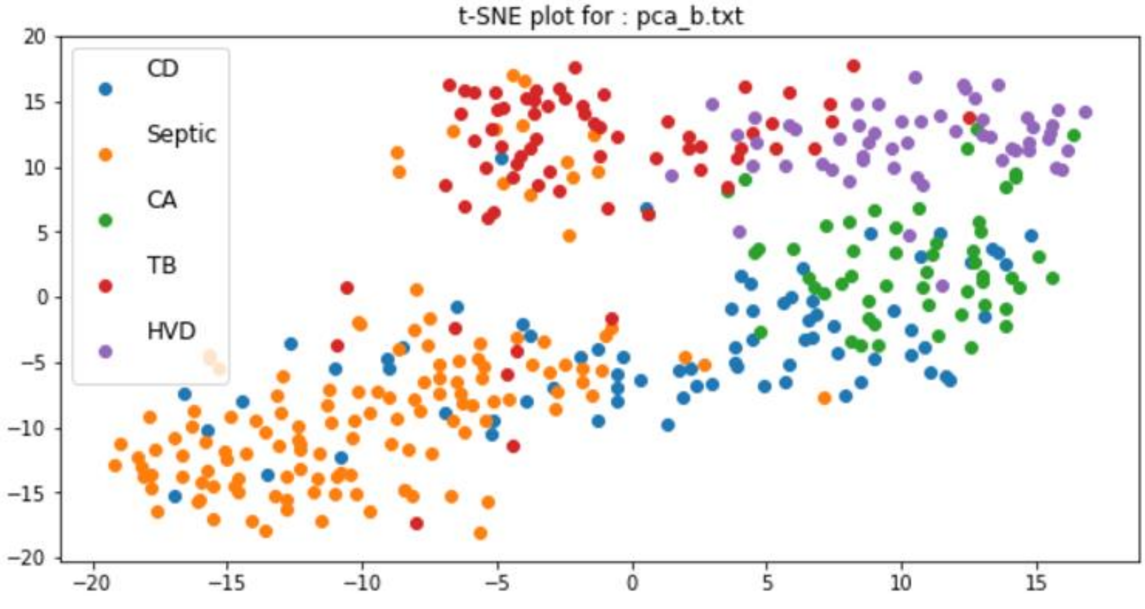

Figure 6: t-SNE plot pca_b.txt

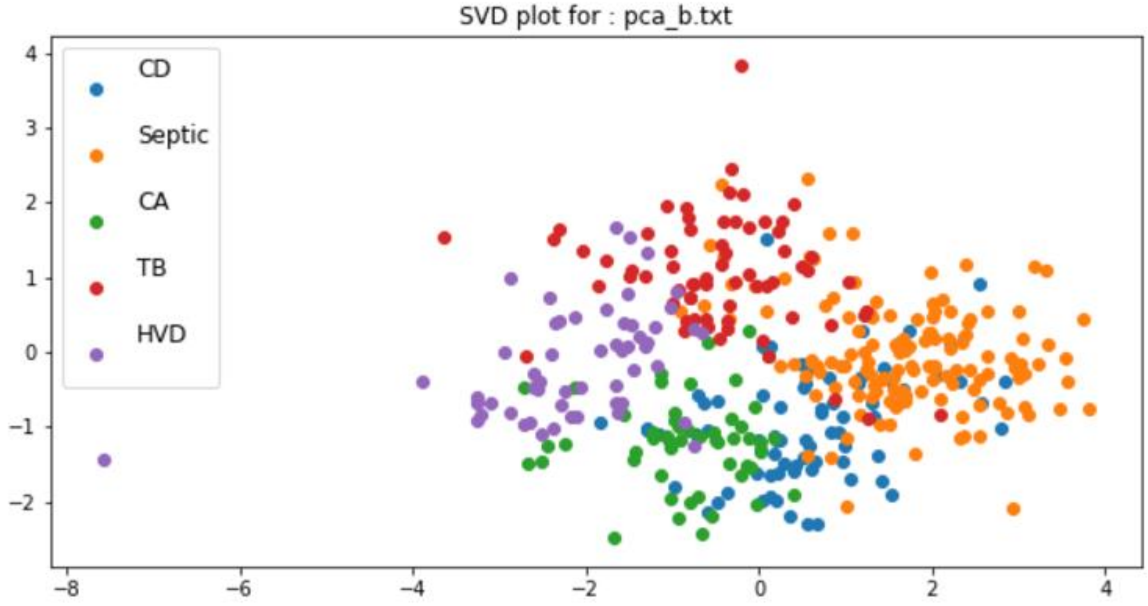

Figure 7:SVD plot for pca_b.txt

Plots for pca_c.txt file

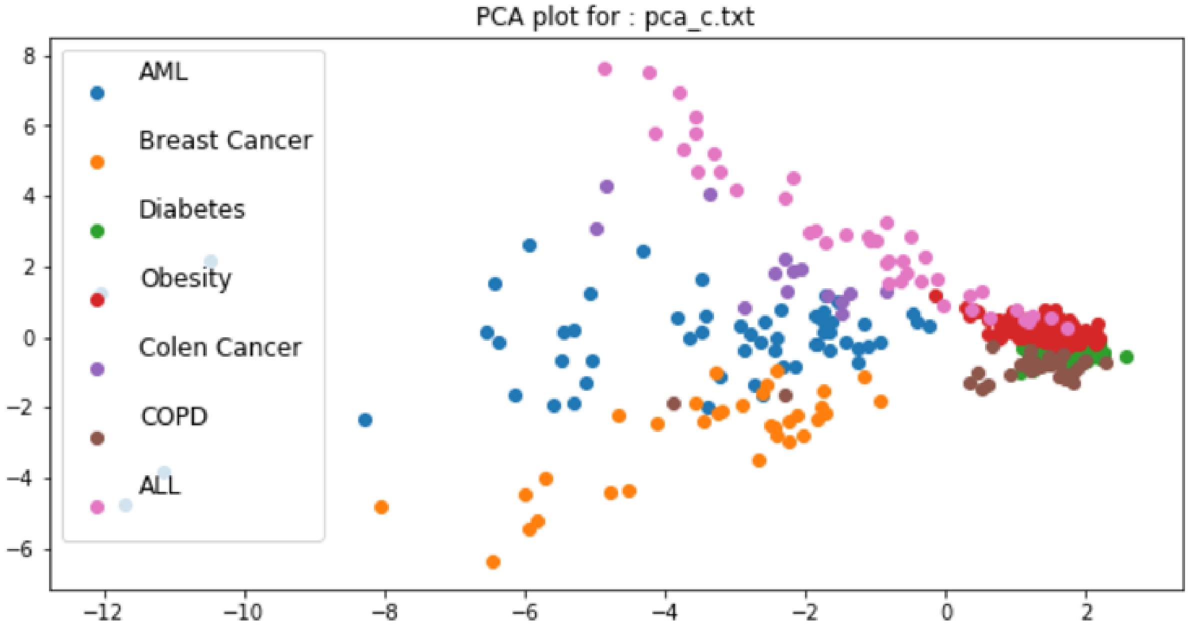

Figure 8: PCA plot for pca_c.txt

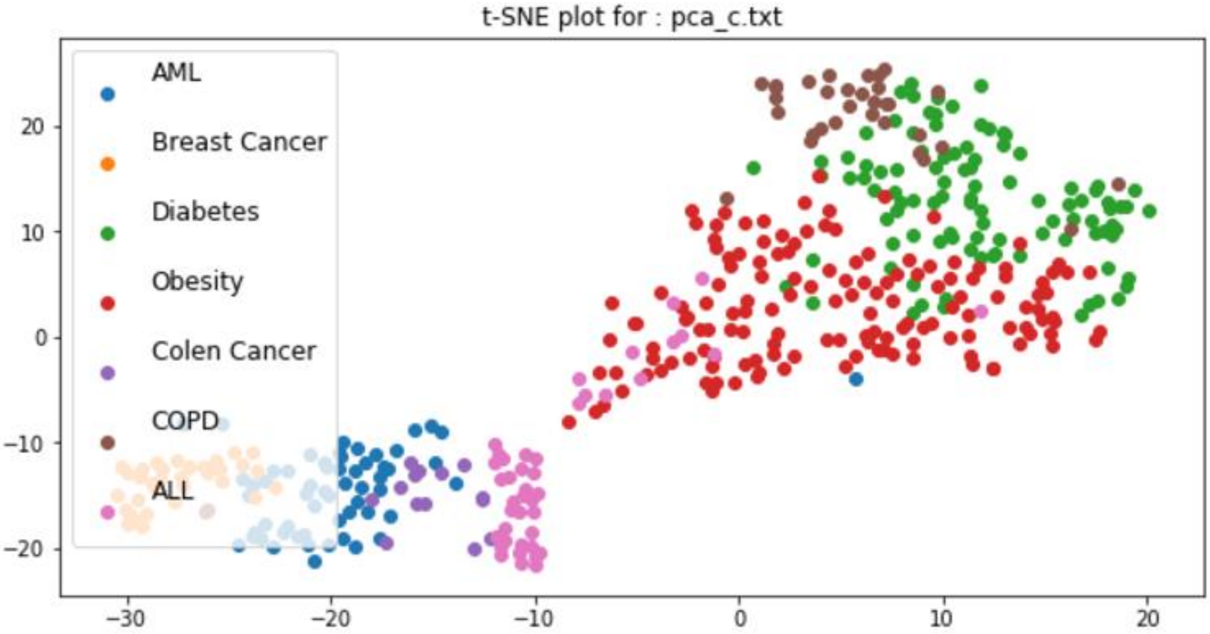

Figure 9: t-SNE plot for pca_c.txt

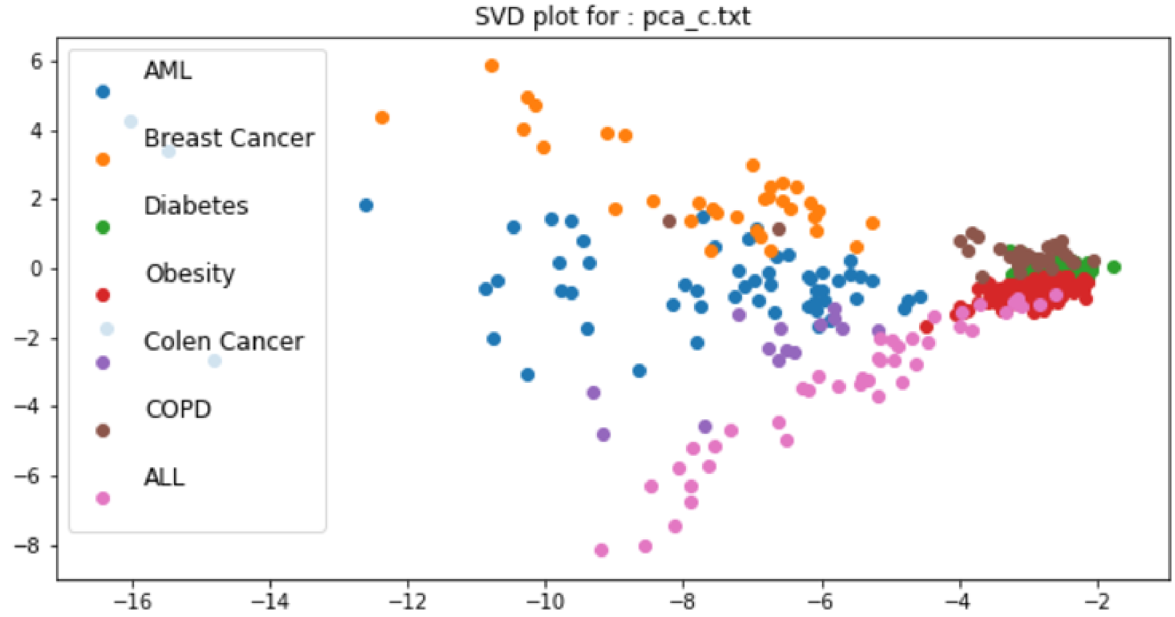

Figure 10: SVD plot for pca_c.txt

FLOW OF PCA IMPLEMENTATION:

Given below are the steps in which the PCA code has been implemented:

1. All the content in the given dataset is read line by line and stored in an array.
2. Exclude the disease label column from this array to get the input array.
3. Calculate the mean values of all the features in the input array and normalize the input array.
4. Compute the covariance of the resultant values obtained in step 3.
5. Compute the Eigenvectors and Eigenvalues of the covariance matrix.
6. Sort the eigenvectors based on the eigenvalues to find the eigenvectors corresponding to the highest eigenvalues. Select the top two eigenvectors which will serve as the principal components.
7. Compute the dot product of the input features and the principal components to plot the multidimensional input features along the two principal axes.
8. Plot the points in the resultant array and group them according to the diseases they belong to. Thus, we have scatter plot with eigenvalues along x-axis and y-axis which makes it evident that the multi-dimensional data has been reduced to two-dimensions.

ANALYSIS OF THE RESULTS OBTAINED USING DIFFERENT ALGORITHMS:

By analyzing the various plots for PCA and SVD we could see marginal difference between the plots. PCA and SVD both are linear dimensionality reduction techniques. Therefore, although the algorithms by both of them are different, the plots obtained are similar. We could also see that dissimilar data points are placed far apart in lower dimension.

On the other hand, t-SNE is a non-linear dimensionality reduction technique. Here similar data points are represented close to each other, unlike the case of PCA and SVD where these points would be placed far apart. Since t-SNE uses local approach, therefore the nearby points on the manifold are mapped to nearby points in a lower-dimensional representation.