CSE 601 - Data Mining and Bioinformatics

Project 1.1 - Dimensionality Reduction

Goal: The goal of this part in the project is to conduct dimensionality reduction using algorithms such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD) and t-Distributed Stochastic Neighbor Embedding (t-SNE) and produce the 2D datapoints as a scatter plot.

Dataset: Biomedical datasets

Language: Python

Libraries: Numpy, Pandas, Matplotlib

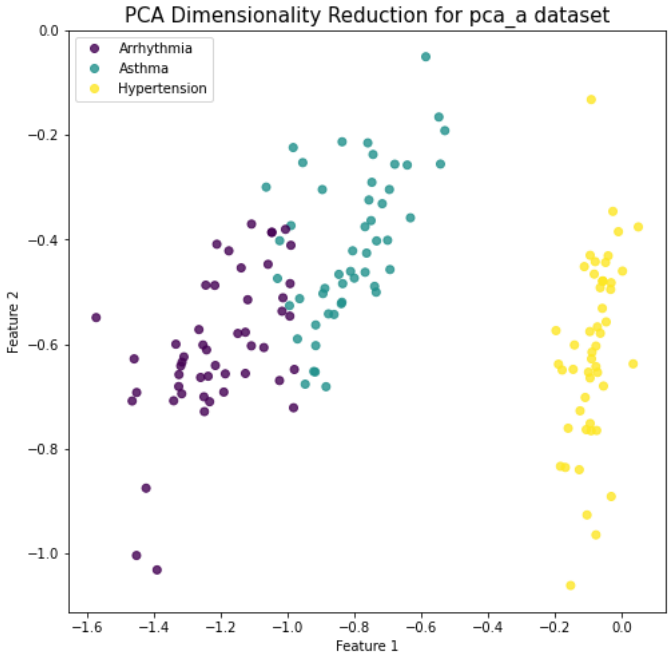
PCA:

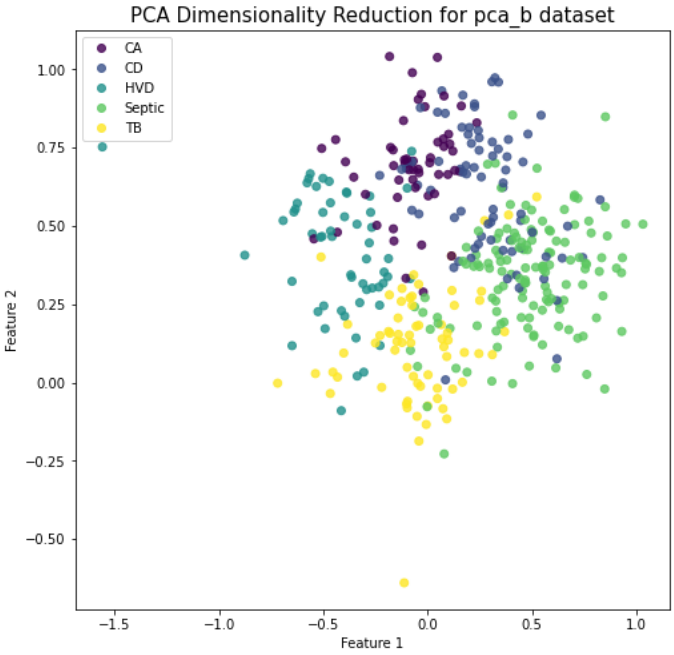
Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables. PCA is a most widely used tool in exploratory data analysis and in machine learning for predictive models. Moreover, PCA is an unsupervised statistical technique used to examine the interrelations among a set of variables. It is also known as a general factor analysis where regression determines a line of best fit.

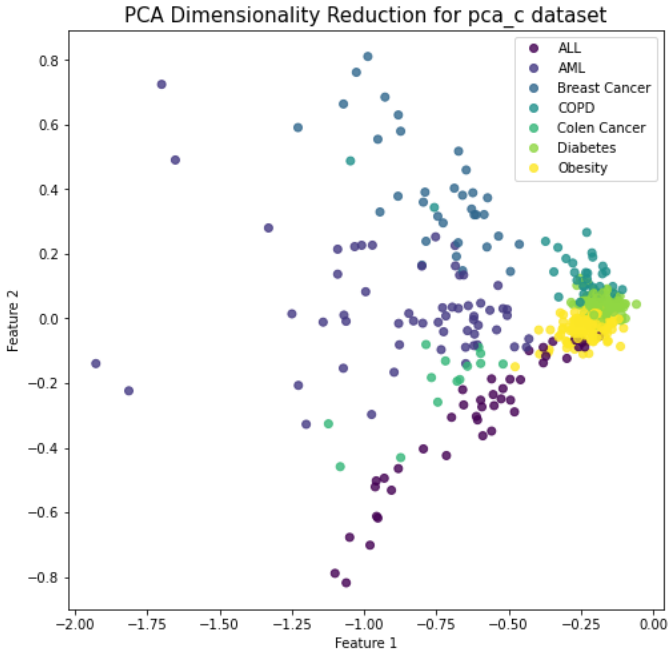
Implementation of PCA algorithm:

- The datasets are read using pandas and stored as a data frame.
- Data preprocessing is done where the attributes are separated from its target values.
- Normalization is done on the dataset using numpy.
- As first step of PCA, Mean is calculated for the entire data and their Centers from the data are calculated.
- Next, the Co-Variance matrix is calculated by taking the centers obtained from the previous step.
- Finally, the eigen values and their corresponding eigen vectors are obtained from the covariance matrix using numpy functions.
- At last, the first 2 columns from the resultant eigen vector matrix are chosen as Principal Components and they are plotted as feature 1 and feature 2 using scatter plots in matplotlib.

Results:


PCA Dimensionality Reduction for pca_a dataset
Arrhythmia
Asthma
Hypertension
Feature 1
Feature 2


PCA Dimensionality Reduction for pca_b dataset
CA
CD
HVD
Septic
TB
Feature 1
Feature 2



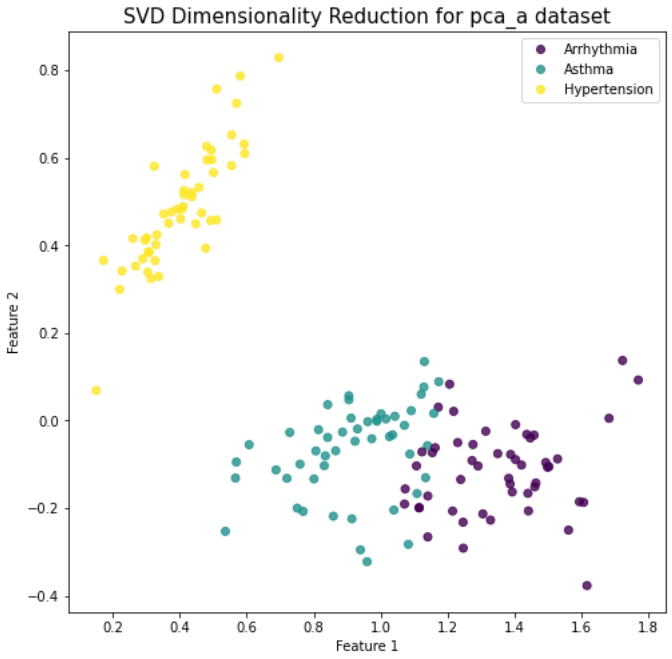
SVD:

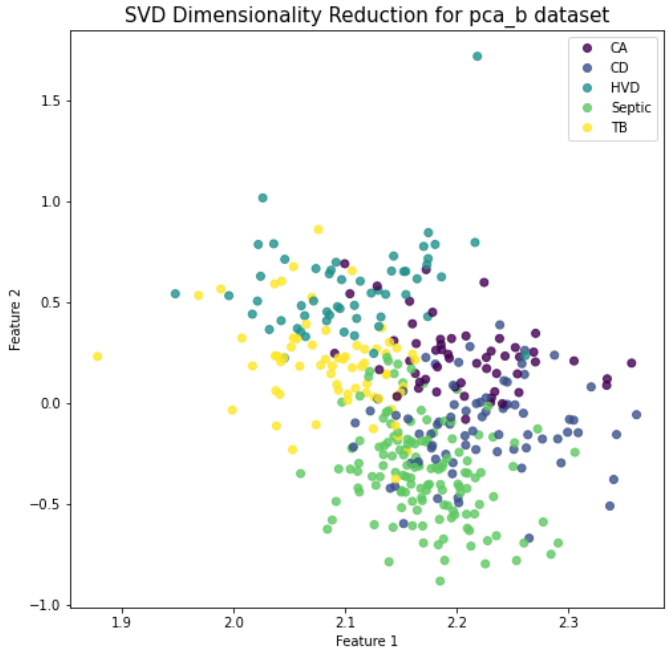
Singular Value Decomposition(SVD) is one of the most widely used Unsupervised learning algorithms and matrix decomposition methods that is at the center of many recommendation and Dimensionality reduction systems. A matrix decomposition method is used for reducing a matrix to its constituent parts to make certain subsequent matrix calculations simpler.

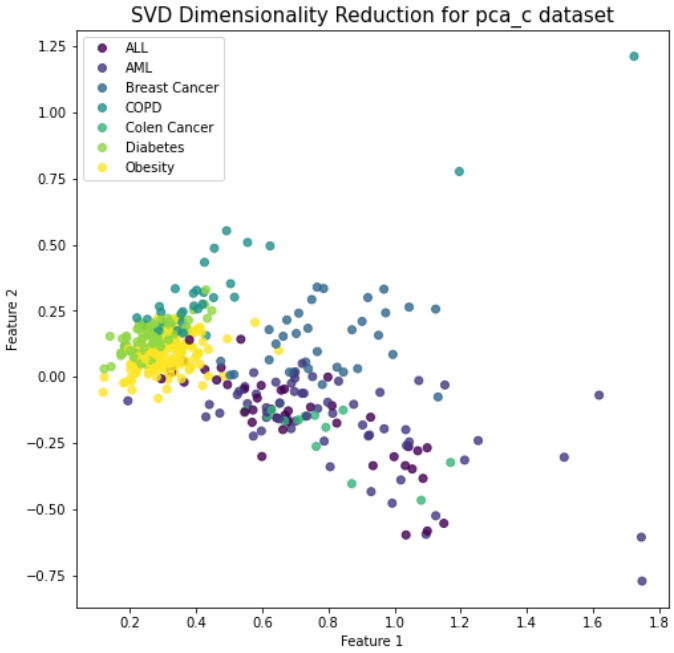
Implementation of SVD algorithm:

- The datasets are read using pandas and stored as a data frame.
- Data preprocessing is done where the attributes are separated from its target values.
- Normalization is done on the dataset using numpy.
- SVD is performed on the dataset using the scikit-learn libraries.
- Finally, the results are plotted as a scatter plot using matplotlib.

Results:


SVD Dimensionality Reduction for pca_a dataset
Arrhythmia
Asthma
Hypertension
Feature 1
Feature 2


SVD Dimensionality Reduction for pca_b dataset
CA
CD
HVD
Septic
TB
Feature 1
Feature 2



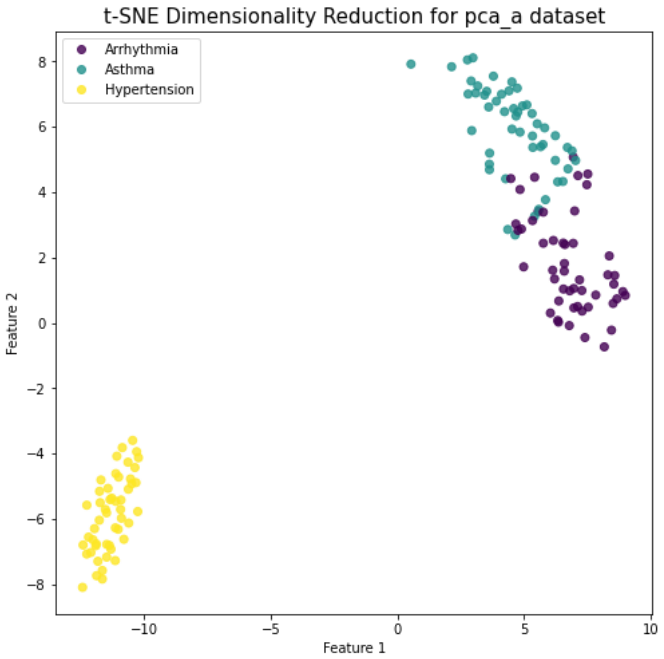
t-SNE:

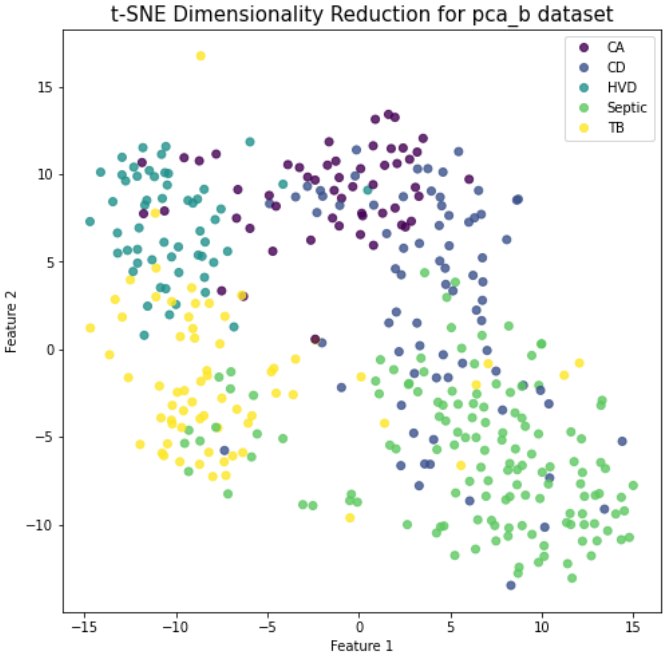
t-distributed stochastic neighbor embedding (t-SNE) is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

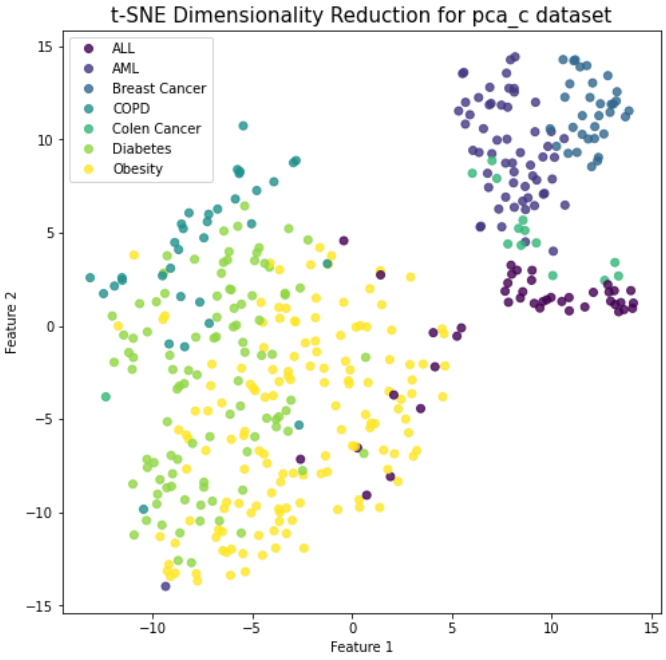
Implementation of t-SNE algorithm:

- The datasets are read using pandas and stored as a data frame.
- Data preprocessing is done where the attributes are separated from its target values.
- Normalization is done on the dataset using numpy.
- t-SNE is performed on the dataset using the scikit-learn libraries.
- Finally, the results are plotted as a scatter plot using matplotlib.

Results:


t-SNE Dimensionality Reduction for pca_a dataset
Arrhythmia
Asthma
Hypertension
Feature 1
Feature 2


t-SNE Dimensionality Reduction for pca_b dataset
CA
CD
HVD
Septic
TB
Feature 1
Feature 2



The results show that PCA and SVD visualizations are very similar to each other whereas t-SNE is unlike PCA, a non-linear data visualizer and a probabilistic technique. It means, it does not form a linear line to separate the classes or to calculate the variance and it does not use any norm or distance metric to calculate the distance between points. It tries to preserve the local structure(cluster) of data whereas PCA preserves the global structure of data.

Team Members:

Team Members	UB IT Name	UB Person No	UB Email
Charan Reddy Bodennagari	charanre	50338186	charanre@buffalo.edu
Sri Charan Chintapenta	schintap	50313858	schintap@buffalo.edu
Varsha Ravichandiran	varshara	50315099	varshara@buffalo.edu