Report

Asish Kakumanu (50288695) | Swapnika Pemmasani (50289464) | Vennela Veerisetty (50286879)

Objective

The whole purpose of the project is to perform **dimensionality reduction** on the given dataset using three different methodologies viz., PCA (Principle Component Analysis), SVD (Singular Value Decomposition) and t- Distributed Stochastic Neighbor Embedding.

Why?

As most of the datasets have more than 2 features. It becomes too hard to understand the data by plotting it in on a graph. So, we reduce the number of features to as less number as 2 or 3. So that we can project the data points on a 2D graph or a 3D graph. This helps us to better understand the data. To achieve this, we use one of the dimensionality reduction algorithms.

How?

Area of variance in data are where objects can be best discriminated. If two dimensions are highly correlated or telling us about the same underlying variance in the data, combining them to form a single measure is reasonable. Therefore, the new dimensions are linear combinations of the original ones and are uncorrelated with one another i.e they are orthogonal in dimensional space. These are principle components. First principle component is the direction of the greatest variability and the following components are next orthogonal direction of greatest variability.

Algorithm

- Read the dataset.
- Extract the data and labels from the dataset.
- Normalize the data by calculating the mean.
- Compute the covariance matrix from the normalized data.

$$S = \frac{1}{n-1} X X^T$$

- Find the Eigen Vectors, Values of the matrix.

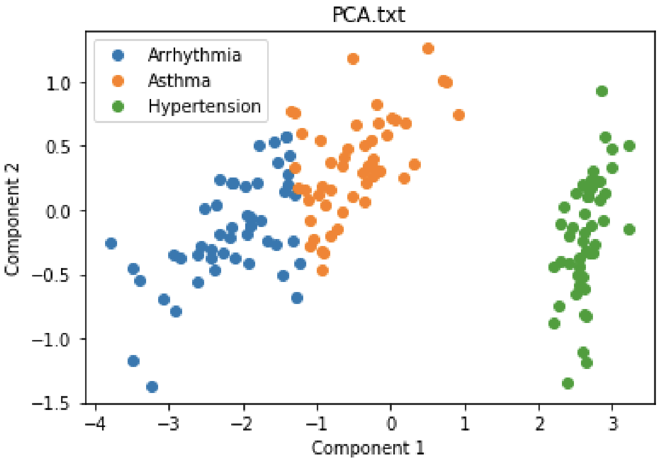
$$Sa = \lambda a$$

- Sort the Eigen Vectors, Values and plot the data points.

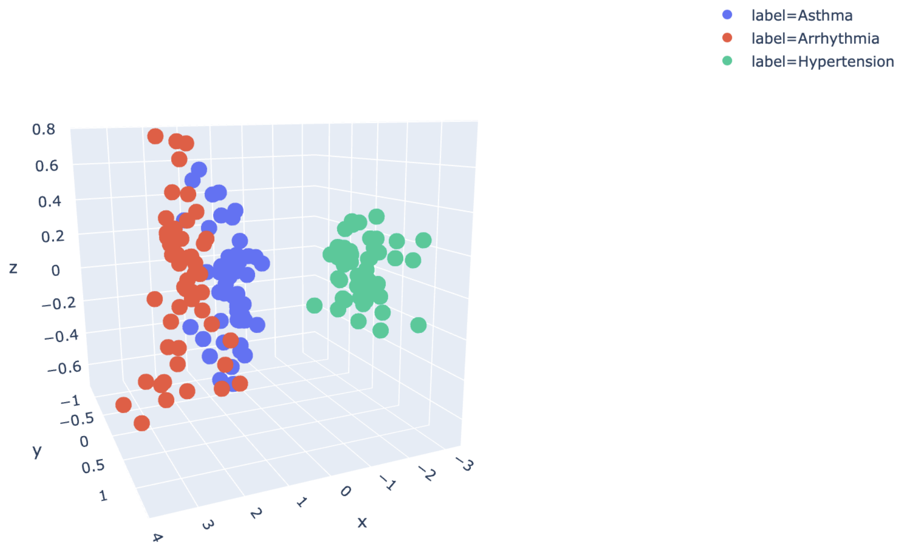
Screenshots

Dataset : pca_a.txt

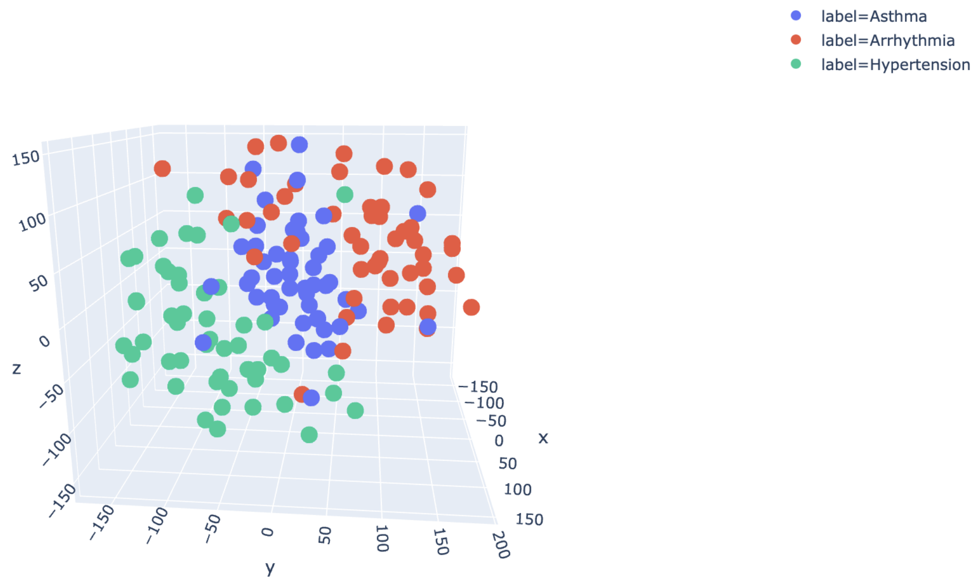
PCA



SVD

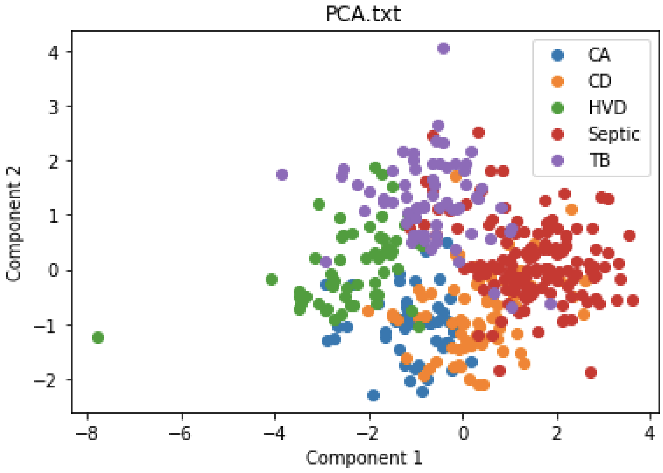


T-SNE

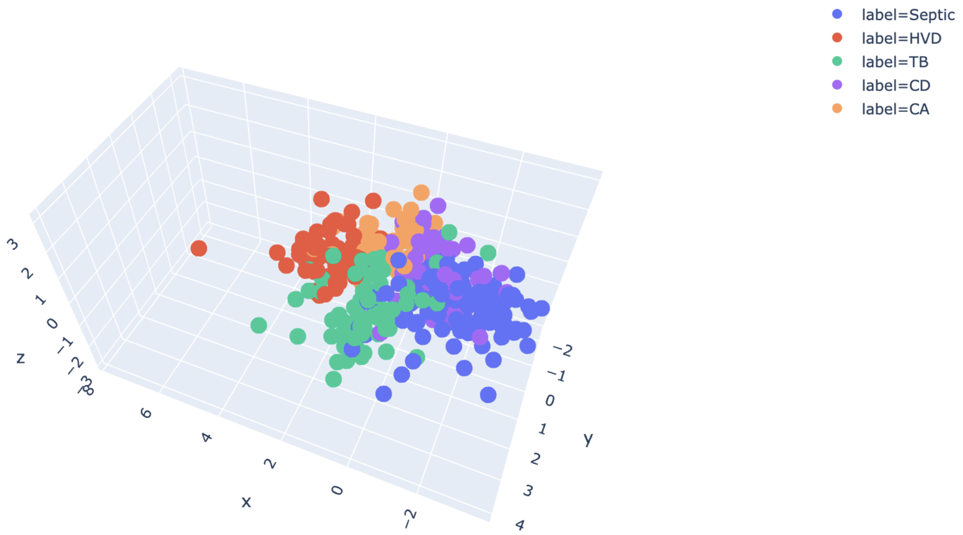


Dataset : pca_b.txt

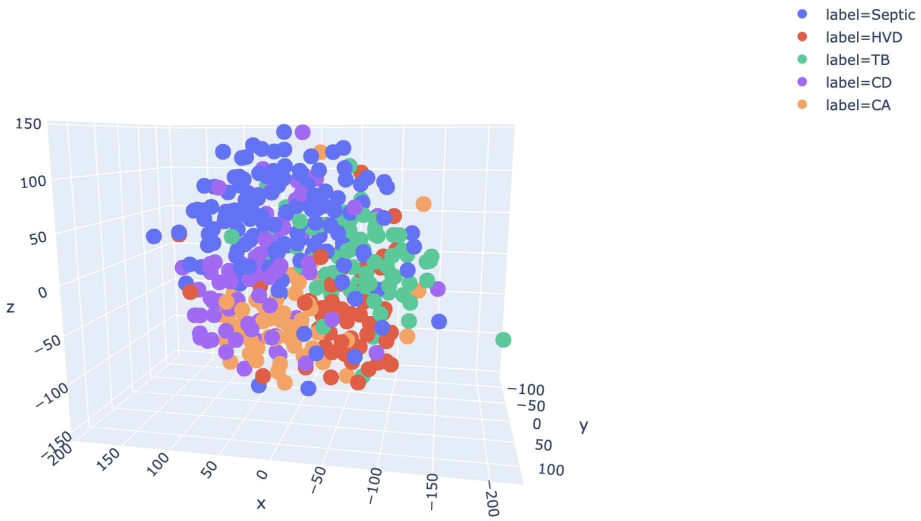
PCA


PCA.txt
CA
CD
HVD
Septic
TB
Component 2
4
3
2
1
0
−1
−2
Component 1
−8
−6
−4
−2
0
2
4

SVD

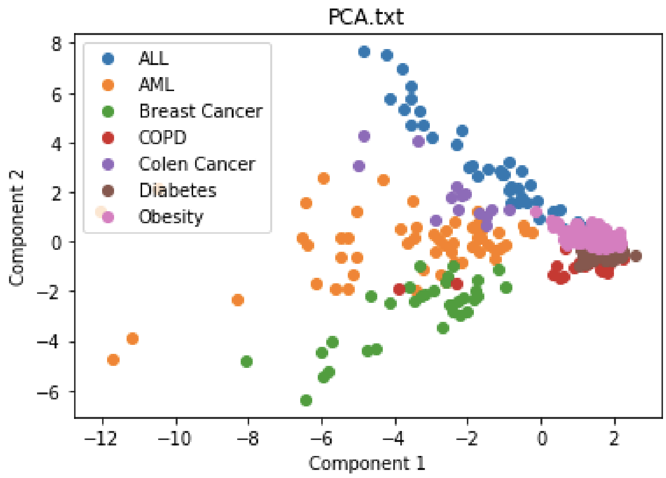

label=Septic
label=HVD
label=TB
label=CD
label=CA
z
3
2
1
0
−1
−2
−3
x
6
4
2
0
−2
y
−2
−1
0
1
2
3
4

T-SNE

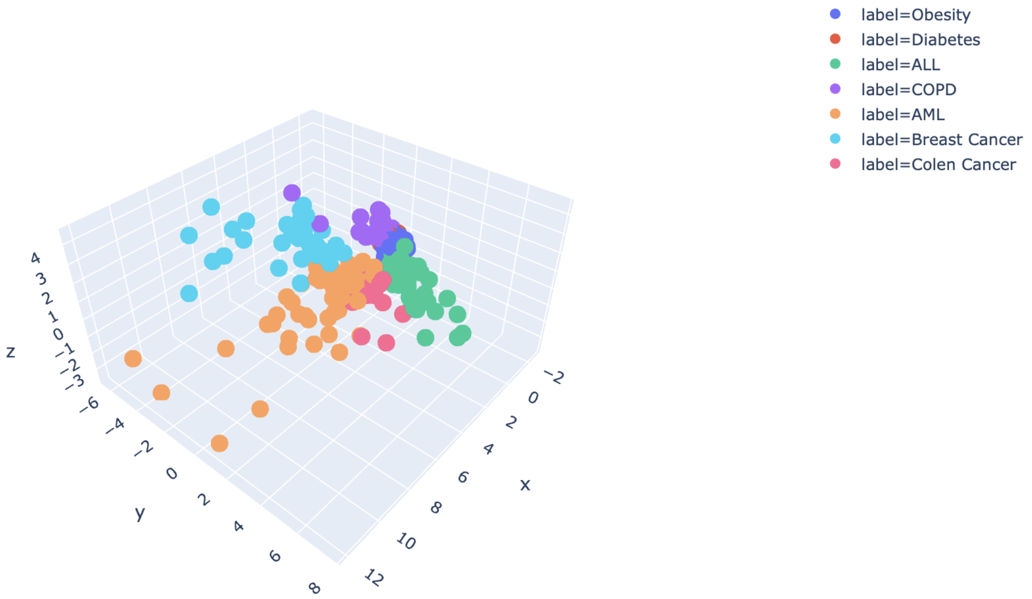


Dataset : pca_c.txt

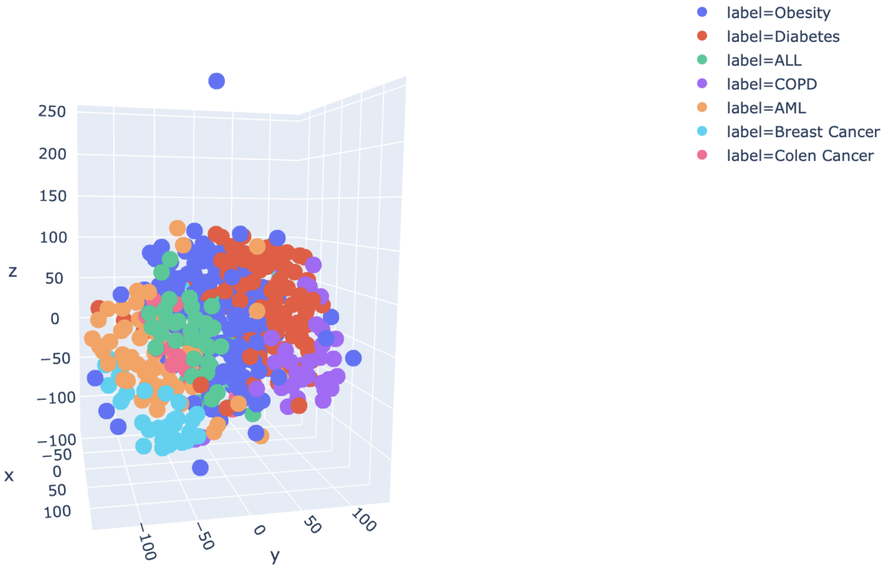
PCA



SVD


label=Obesity
label=Diabetes
label=ALL
label=COPD
label=AML
label=Breast Cancer
label=Colen Cancer
z
4
3
2
1
0
−1
−2
−3
y
−6
−4
−2
0
2
4
6
8
x
−2
0
2
4
6
8
10
12

T-SNE


label=Obesity
label=Diabetes
label=ALL
label=COPD
label=AML
label=Breast Cancer
label=Colen Cancer
z
250
200
150
100
50
0
−50
−100
x
−100
−50
0
50
100
y
−100
−50
0
50
100