PCA:

Dimensionality Reduction Technique.

Eg. Taking picture. Which angle to take picture. Taking the picture of the data still preserving most of the information in the data.

D => D’ , where D’ < D

Distance metrics cannot deal with high dimensional data : Curse of Dimensionality. We use PCA and t-SNE for high dimensional data.

PCA extracts linear feature. Creates new independent features. Maximizes the variance of data in low dimensional space. PCA finds the principal components that capture the maximal variance in the dataset.

PCA does not work well for data having non-linear relationships. Also, we cannot see dissimilar points well separated by PCA. Unlike PCA, t-SNE is a probabilistic approach to dimension reduction. First a pair-wise data similarity matrix of the input features is drawn based on the t-distribution. Also, a similar probability distribution is constructed over low dimension map. The embedding then tries to minimize the divergence between the two distributions such that there is as less input dimension as possible.

Steps followed in Sklearn for TSNE:

* Load Data
* Apply dimensionality reduction technique eg PCA.
* Visualize the two dimensions using Scatterplot.