Summary:

Assume you want to predict India’s GDP of next year and you collected that there are lots of factors that affect this.

Some of them may affect significantly and some of them don’t affect much.

Now you can still consider all the data but it will take much time to make an ml model on it and it may lead to over fitting of data.

So now you randomly choose few of the features but it will compromise the accuracy of the model as you don’t know which of the features are more significant.

Dimensionality Reduction is the process of reducing the dimensions of the feature dataset which have fewer but significant information about the actual dataset

These are the two ways to it, first is we can do feature selection in which we selectively choose the features manually

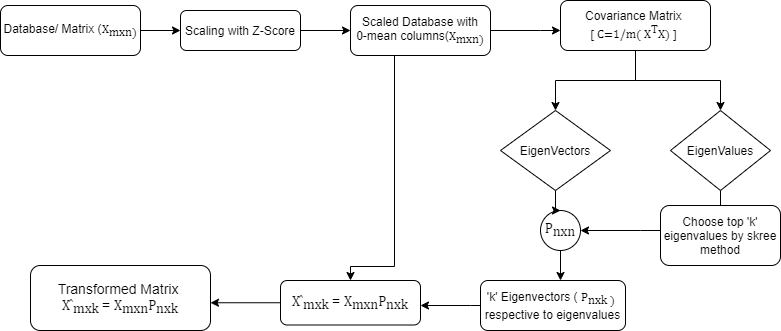
Second way is to do the feature extraction on the dataset which gives more informative data with less dimensions.

Data Standardization is used to convert all the feature variables to similar scale or common format and centralizes the data.

It should be done to remove anomalies due to large and small scaled variables

The common method to do Data Standardization is Z score

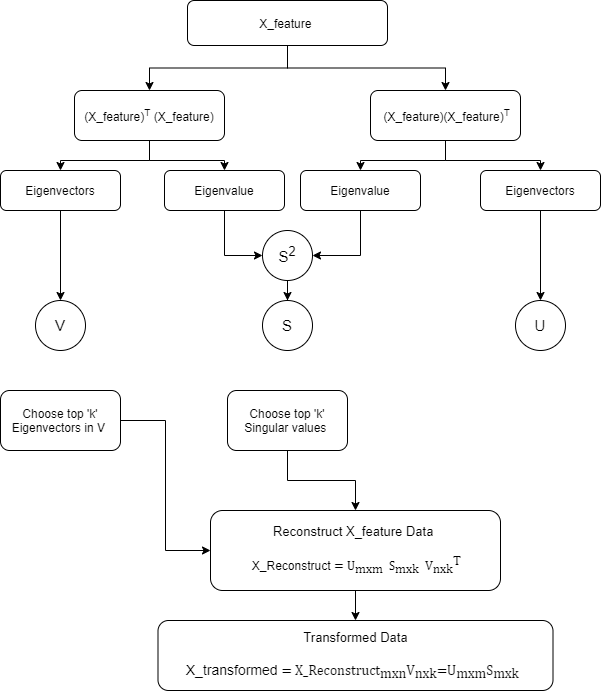
There is a Principle Component Analysis which is a dimensionality reduction technique. The complete mathematical process is shown is the block diagram below. However, to learn more, you can check the literature review.

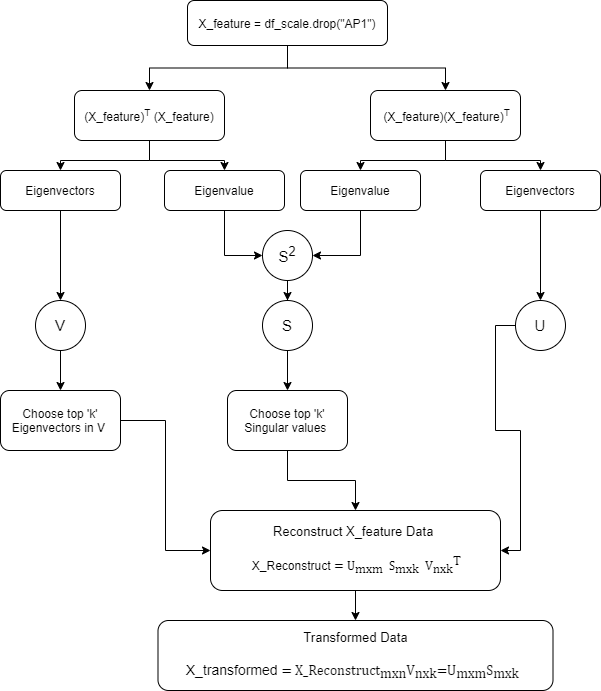


There is a PCA library in the decomposition python module which gives the same results for the mathematical process applied on the database.

Singular Value Decomposition is similar to EVD but it used for non-square matrices. The problem for rectangular matrices is that we can’t do EVD directly.

The mathematical process is shown below.





There is Truncated SVD function in decomposition python module which gives the same results as mathematical.

The PCA and SVD are widely used in image data analysis in which we reduce the image size to save lots of data.