**Dimensionality Reduction - Principal Component Analysis**

When training different machine learning models, we come across many high dimensional datasets. In order to train models efficiently with such types of datasets, the dimensions should be reduced. Principal Component Analysis is one of the most widely used technique for this purpose.

One main constraint to dimensionality reduction is that, while reducing the dimensions – the variability or the statistical information contained by the dataset should be preserved. While reducing the dimensions, sometimes a small trade-off is done with accuracy in order to maintain simplicity.

There are some practices that should be followed which performing PCA:

* Data Standardization: Data Standardization plays a crucial role before doing PCA. If it’s not done, PCA would consider the larger values to be dominant over the small values. While doing dimensionality reduction, one has to make sure that the contribution of the variables is not compromised. Standardizing data prevents the result to be biased.
* The covariance matrix is computed in order to remove redundant data as sometimes the variables are strongly correlated.
* Computation of Eigenvectors and Eigenvalues: Eigenvectors represent the direction of axes where most of the information or variance lies. Each eigenvector has an eigenvalue which is the amount of variance.

Now, we have the corresponding Eigenvectors and their eigenvalues, principal components are ready to be computed.

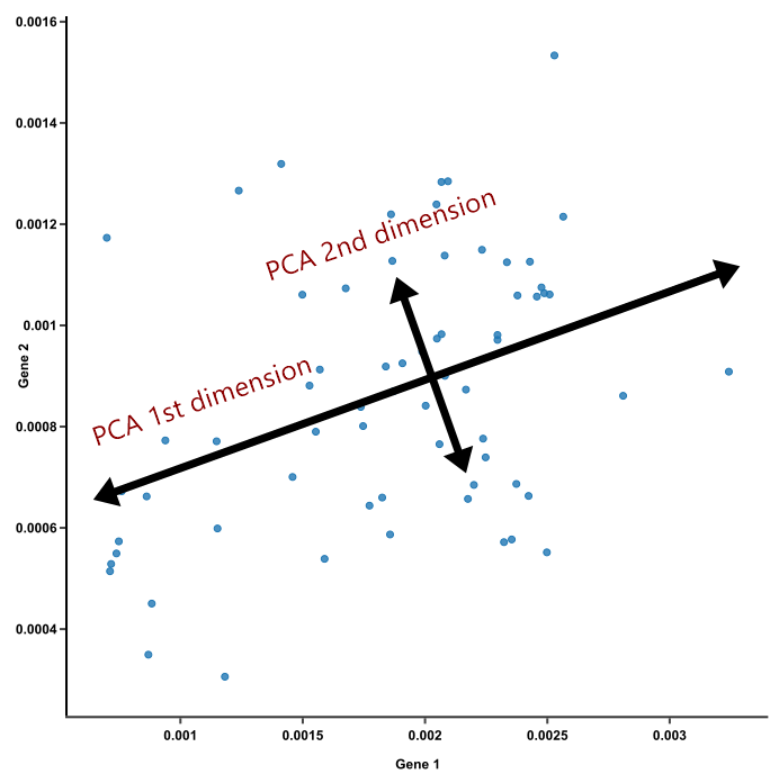
Principal Component one is computed such that the variance contained by it is maximum in the dataset. Similarly, the second principal component is computed in a way that it has the least corelation with PC1 and the second highest variance.

* Creation of the Feature Vector: This is one the main steps in the process of principal component analysis as after this we have the components which we decide to keep. i.e., a subset of components is chosen and the ones with less significance are removed.
* Finally, the data is reoriented using this feature vector and now, the axes are redefined using principal components.

Principal Component analysis reduces the chances of model-overfitting, also increases the algorithm performance and data visualization is made easier since the dataset is low dimensional.

Some disadvantages that come along with PCA are: it is important to perform data standardization, the data becomes less interpretable since, the data is manipulated in PCA.

The below diagram shows a brief depiction of Principal Component Analysis:



**Source: blog.bioturing.com**

**Accuracy on datasets after applying Principal Component Analysis is as follows:**

**Voice Gender recognition:**

**BVC Gender and Age classification:**